

# **Profiling and Grouping Space-time Activity Patterns of Urban Individuals**



Jianan Shen

Department of Civil, Environmental and Geomatic Engineering

University College London

Supervisors:

Prof. Tao Cheng

Dr. Edward John Manley

A thesis submitted for the degree of

*Doctor of Philosophy in Civil, Environmental and Geomatic Engineering*

November 2017

### **STATEMENT OF ORIGINALITY**

I, Jianan Shen 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

With the widespread use of new geospatial and information technologies, huge amounts of human dynamics data have been collected with high spatio-temporal resolution, especially in urban areas. Applying data mining techniques to these datasets can reveal activity patterns of individuals in greater detail and finer scale. Urban areas are home to millions of individuals' complex and dynamic activities and interactions. Given the exponential population growth and expansion of cities nowadays, understanding activity patterns of massive groups of people moving in the urban environment is therefore playing a more and more important role in the building of smarter cities.

The study of human activity patterns seeks to determine how to describe how people keep different routines, and how people play out different roles and possess different preferences and inclinations to behave in certain ways. In this thesis, we introduce the concept of **“the place you go, when you go and how long you stay is who you are”**, expanding the focus from traditional physical locations to places by integrating the temporal attribute and semantic meaning of places into the analysis. This evolution is achieved by a new methodological framework that enables us to more realistically analyse large-scale mobility data with the awareness of the network representation of urban space and the dynamism of human activities. In this framework, advances of methodologies are made to improve the representation of individual activity profiles and a novel spatio-temporal clustering algorithm is designed to detect regions of interests. Following this stage, we also carry out the semantic enrichment of places and activity patterns based on state-of-the-art text mining techniques. The final phase of the framework brings in network-based transformations of the proposed methods in the framework to further enhance their space-time accuracy and applicability for activities occurring in urban areas.

The framework integrating the spatial, temporal and semantic considerations is then implemented within a large-scale police movement dataset in Inner London, United Kingdom. The case study shows how the proposed framework enables a marked improvement in the aggregative analysis of semantic activity patterns from that offered by conventional approaches.

## IMPACT STATEMENT

The innovations in this thesis can be of beneficial use both inside and outside academia. The overall advantage of the framework developed in this thesis is that it provides a useful toolkit to automatically extract activity patterns from GPS data and make sense out of these GPS-based activity logs by analysing the staying behaviours in a semantic environment.

Within the academia, the designed methodological framework firstly integrated spatial, temporal and semantic dimensions of activities into the analysis, which can provide more complete and realistic personal profiles for human behaviour study. Second, the ST-Net-DBSCAN algorithm developed in the research was the first attempt to fit spatio-temporal density-based clustering algorithm into urban street networks, which enabled the space-time hotspot detection with the awareness of the topology of streets. This improvement increased the spatial accuracy of space-time hotspot detection in urban context and enabled more detailed studies on urban dweller behaviours. Third, the semantic enrichment methods introduced in this thesis transformed the traditional spatial analysis approach of areas into a place-based approach. Apart from considering where the place locates, questions of what the place is and how the place semantically relates to different visitors was also taken into the analysis of human movement behaviours. Unlike traditional semantic enrichment methods that view the function of places as a constant attribute, this semantic enrichment module took temporal information such as opening hours of different facilities into account, which enabled the automatic detection of functional changes of places throughout the day and therefore better depict the highly dynamic nature of modern cities.

In non-academical domain, the innovations in this thesis can contribute to professional practice, public policy design, and commercial applications. During the experimental stage of this research, we have cooperated with the London Metropolitan police. Using their officer movement data, the framework can detect and profile each officer's preference in patrol movements and find outlier and abnormal behaviours in the behaviour profile clustering process, which provided the police with a quantified method for officer performance evaluation. With the wide-spread use of intelligent policing equipments, movement data of police officers have become more accessible worldwide. After proper adaptations, this framework can also be used to improve the professional practices of the police officers in other cities around the world. Moreover, this toolkit can be used to depict the time-varying semantic meaning of places for human activities so that urban planning authorities can be provided with well-rounded and high-temporal-resolution information of the changing hotspots in the city, which helps policy makers make faster and better informed decisions. It can also be applied to the emerging location-based social networks, bridging the gap between users' online social features and their real-world activities to provide a more accurate and complete profile for the users. The implementation of the framework can aggregate users sharing similar semantic activity profiles and space-time preferences in activities, and facilitate smarter friend recommendations in the next-generation social network applications.

Impact of this research will be brought about through the publication of multiple journal articles and commercialised software products.



## ACKNOWLEDGEMENTS

Many people have inspired, guided and supported me in the journey of my PhD study. Without the guidance and encouragement of my primary supervisor Prof Tao Cheng and her support since 2013, this work would not have been possible. Since I started my PhD directly after my undergraduate study, Tao had guided me step by step to build up my research experience and all abilities necessary for a full-fledged PhD candidate from scratch. The endless discussions, project works, and case study attempts under her supervision have been the corner stone of progression in my research career, and therefore, I would like to sincerely thank her for her continuous encouragement and wise advice both at work and in life.

My second supervisor, Dr. Ed Manley from UCL, has also been a great mentor and friend of mine. He guided me through the academic journey with his critical comments, patience during our consultative meetings and his limitless effort in helping me improve my academic writing skills.

I am grateful to Trevor Adams at Metropolitan Police, who have organised our first-hand practices with the police at work and provided crucial field knowledge on policing operations and explanation of the officer's behaviours. I also thank China Scholarship Council and UK Engineering and Physical Sciences Research Council (EP/J004197/1) for their financial support, and the Metropolitan for providing the valuable data used in this thesis.

I owe a lot respect and gratitude to the members of the CPC project, especially Dr. Gabriel Rosser, Dr. Toby Davies, Huanfa Chen who helped me improve my programming skills and provide insightful suggestions and challenges to my ideas. I also owe a debt of gratitude to my warmhearted roommates in Warwick Avenue, who have created the coziest and most harmonious place for me to live in. My close friends have also been my torch in the darkest of times. The long list starts with, and is not limited to, Dr. James Haworth, Dr. Garavig Tanaksaranond, Juntao Lai, my colleagues/friends in the SpaceTimeLab, administrative personnel in the Department of Civil, Environmental & Geomatic Engineering at UCL and other close friends that made my life in London some of the most enjoyable years of my life.

Above all, I would like to deeply thank my parents, Yuejin Shen and Xingmiao Peng, for always trusting and supporting me in my pursuit of my career passions, despite the long physical distance. They have taught me by example and benevolence and made me who I am today.

## CONTENTS

Abstract .....	3
List of Figures .....	11
List of Tables .....	15
Nomenclature .....	16
1 Introduction .....	18
1.1 Progress in activity pattern and human dynamics studies .....	19
1.1.1 Movement logging .....	19
1.1.2 Extracting region of interest and activities from movements .....	20
1.1.3 Adding semantic meaning to places .....	21
1.1.4 Aggregative analysis of activity patterns .....	22
1.2 Limitations and challenges .....	22
1.3 Research aim and objectives .....	23
1.4 General thesis organisation .....	24
2 Literature Review .....	29
2.1 General theories and frameworks in activity pattern studies .....	30
2.1.1 Geospatial process .....	32
2.1.2 Semantic process .....	34
2.1.3 Knowledge discovery .....	35
2.2 Detecting regions of interest .....	36
2.2.1 Trip segmentation and stop identification .....	36
2.2.2 Detecting regions of interest in Cartesian space .....	39
2.2.3 Detecting regions of interest in space and time .....	42
2.2.4 Detecting regions of interest in spatial networks .....	44
2.2.5 Summary .....	46
2.3 Semantic enrichment of places .....	48
2.3.1 Space-time boundaries of places .....	49
2.3.2 Semantic enrichment based on contextual data .....	49
2.4 Similarity and aggregation of activity profiles .....	52
2.4.1 Similarity metrics based on physical features .....	53
2.4.2 Similarity metrics based on travelling sequences .....	54
2.4.3 Similarity metrics based on semantic location histories .....	57
2.4.4 Profiling and aggregative analysis .....	59
2.4.5 Summary .....	60

2.5 Spatial analysis toolkit.....	60
2.5.1 Map matching .....	61
2.5.2 Spatial query and spatial network query.....	62
2.5.3 Visualisation analysis of activities .....	62
2.6 Chapter Summary.....	65
3 Methodological Framework.....	68
3.1 Framework description .....	69
3.2 The two paradigms .....	71
3.3 The Euclidean paradigm.....	72
3.3.1 Module I: Data Pre-processing .....	72
3.3.2 Module II: ST-ROI detection in time and Cartesian space.....	74
3.3.3 Module III: Semantic enrichment of ST-ROIs.....	74
3.3.4 Module IV: Aggregative analysis of the semantic profiles.....	75
3.4 The network paradigm .....	75
3.4.1 Module I: Pre-processing for network-based movement analysis.....	76
3.4.2 Module II: ST-LOI detection .....	76
3.4.3 Module III: Semantic enrichment of ST-LOIs.....	77
3.4.4 Module IV: Aggregative analysis of the semantic profiles .....	77
3.5 Addressing research aim and objectives .....	77
4 Knowing about the Data.....	80
4.1 Case study area.....	80
4.2 Datasets.....	81
4.2.1 Movement data .....	81
4.2.2 POI data.....	84
4.2.3 Street network data.....	88
4.3 Exploring basic patterns of the data .....	88
4.3.1 Spatial ROIs.....	88
4.3.2 Statistical patterns of people's visits to ROIs .....	91
5 The Euclidean paradigm .....	95
5.1 Introduction .....	95
5.2 Module I: pre-processing.....	96
5.2.1 Trip segregation .....	96
5.2.2 Stay point identification.....	97
5.3 Module II: ST-ROI detection.....	100

5.3.1 ST-DBSCAN.....	101
5.3.2 The simplified representation of trajectories .....	103
5.3.3 Space-time profile and its similarity .....	105
5.3.4 Hierarchical clustering of space-time profiles .....	107
5.4 Module III: Semantic enrichment of ST-ROIs.....	108
5.4.1 Space-time boundaries of ST-ROIs .....	109
5.4.2 POIs in the space-time boundaries .....	110
5.4.3 Reweighting POIs for semantic annotation of ST-ROIs.....	111
5.5 Module IV: Semantic profiling and hierarchical clustering.....	113
5.6 A single-borough case study .....	114
5.6.1 Module I: Pre-processing.....	114
5.6.2 Module II: ST-ROIs in foot patrol activity .....	115
5.6.3 Intermediate outcomes: Space-time profiles .....	116
5.6.4 Explanation of the outcomes.....	118
5.6.5 Discussion: The need for semantic enrichment .....	122
5.6.6 Module III: Semantic enrichment of ST-ROIs.....	123
5.6.7 Module IV: Aggregative analysis of semantic profiles .....	125
5.7 An extended case study in multiple boroughs .....	128
5.7.1 Semantic ST-ROIs .....	129
5.7.2 Semantic profiles and profile aggregation .....	130
5.8 Chapter summary .....	131
6 The Network Paradigm.....	134
6.1 Introduction .....	134
6.1.1 From trip trajectories to trip routes .....	134
6.1.2 From ST-ROI to ST-LOI .....	136
6.1.3 From 3D scatter map visualisation to 3D wall map visualisation .....	136
6.1.4 Work flow of the network paradigm .....	136
6.2 Basic toolkit for spatial network analysis .....	137
6.2.1 Shortest path finding tool .....	138
6.2.2 Range searching and nearest neighbour searching tool.....	138
6.3 Module I: Pre-processing of movement data in urban networks .....	139
6.3.1 Trip segmentation .....	139
6.3.2 Map-matching.....	139
6.3.3 Network-based stay point identification.....	142

6.3.4 Complexity .....	144
6.4 Module II: ST-LOI detection.....	144
6.4.1 Definitions .....	147
6.4.2 Describing the algorithm.....	148
6.4.3 Space-time neighbour retrieving strategy.....	150
6.4.4 Visualisation of ST-LOIs.....	153
6.4.5 Complexity .....	155
6.5 Module III: Space-time semantic enrichment in road networks.....	155
6.5.1 Network POIs .....	155
6.5.1 Visualisation.....	157
6.6 Module IV: Profile aggregation.....	157
6.7 Case study .....	157
6.7.1 Map-matched observations.....	158
6.7.2 ST-LOIs and their visualisation .....	159
6.7.3 Semantic enriched ST-LOIs .....	162
6.7.4 Semantic profiles and profile aggregation .....	164
6.8 Chapter summary.....	167
7 Improvements in Semantic Enrichment Module .....	170
7.1 Method description .....	170
7.2 Case study .....	172
7.2.1 The multiple-borough case.....	172
7.2.2 Model evaluation.....	174
7.3 Chapter summary .....	177
8 Further Model Comparison and Validation.....	179
8.1 Module I .....	179
8.1.1 Artificial route and trajectory generator .....	180
8.1.2 Accuracy of stay point identification methods.....	180
8.1.3 Map-matching accuracy .....	182
8.2 Module II .....	183
8.2.1 Spatial cohesion of points in ROIs.....	183
8.2.2 Temporal cohesion.....	185
8.2.3 Optimisation of ST-network-DBSCAN .....	186
8.3 Module III .....	187
8.3.1 Closeness of ROIs to POIs.....	187

8.4 Chapter summary.....	189
9 Conclusion and Future Work.....	192
9.1 Review of findings in research.....	192
9.2 Theoretical contributions and Technical innovations.....	196
9.3 Applications and policy implications .....	198
9.3.1 Application perspectives .....	198
9.3.2 Policy implications.....	199
9.4 Critique of limitations.....	200
9.4.1 Data limitations.....	201
9.4.2 Limitation in the hybrid spatial representation of places.....	202
9.4.3 Information losses in ROI detection.....	202
9.5 Future work .....	202
9.5.1 Data pre-processing.....	203
9.5.2 ROI detection.....	204
9.5.3 Semantic enrichment.....	205
9.5.4 Aggregative analysis of activity profiles.....	205
9.6 Final conclusion .....	205
References .....	207
Appendices.....	220
Appendix A: Data sample of APLS.....	220
Appendix B: Merged POI dataset .....	221
Appendix C: POI Classification of Google Places.....	222

## LIST OF FIGURES

Figure 2.1 Individual and environmental factors influencing activity patterns by level ..	31
Figure 2.2 The three-step extraction of activity patterns .....	32
Figure 2.3 Representation of trip trajectories in a space-time cube .....	33
Figure 2.4 Stay points in a trip trajectory.....	37
Figure 2.5 Special cases in which conventional stop identification methods may make mistakes.....	38
Figure 2.6 A region of interest that attracts multiple persons' visits .....	40
Figure 2.7 An example of DBSCAN clustering result when $MinPts = 3$ .....	41
Figure 3.1 The three aspects of a complete activity profile .....	68
Figure 3.2 Work flow of the Euclidean paradigm.....	72
Figure 3.3 Work flow of the network paradigm.....	76
Figure 4.1 The study area of the thesis .....	81
Figure 4.2 APLS's EPE (estimated positional error) distribution (London, UK, August 2015) .....	83
Figure 4.3 The POI number of each major category in the study area.....	87
Figure 4.4 An example of ITN network structure .....	88
Figure 4.5 Average Euclidean speed in move episodes of APLS.....	89
Figure 4.6 reachability plot showing the OPTICS clustering results.....	89
Figure 4.7 ROIs detected by conventional P-DBSCAN in Camden APLS, February 2012	90
Figure 4.8 Visualisation of two officers' (102PO and 619PO) raw movement trajectories (as polylines) in a space-time cube .....	91
Figure 4.9 Heat map of hourly intensity of activities in Camden police patrol.....	92
Figure 4.10 Heat map showing individual active intensity of two officers.....	92
Figure 4.11 Heat map of dwelling time of officers "102PO" and "619PO" .....	93
in different ROIs in February 2012 .....	93
Figure 5.1 Trip segmentation example of one person's sequential location updates.....	97
Figure 5.2 $p_1$ and $p$ are in the temporal scanning window while $p_2$ is not .....	98
Figure 5.3 A kernel-based temporal window scanning through a trajectory.....	100
Figure 5.4 An example of ST-DBSCAN .....	102

Figure 5.5 The simplified representation of two example users' movements.....	104
Figure 5.6 Histogram showing the percentage of the time two users allocate to ST-ROIs .....	106
Figure 5.7 An ST-ROI and its space-time boundary .....	109
Figure 5.8 The space-time relationship between POIs and ST-ROIs .....	110
Figure 5.9 One typical working day of officers is separated into 3 shifts. 28 ST-ROIs of foot patrol officers are detected by ST-DBSCAN and are labelled with different colours (Shen and Cheng, 2016).....	116
Figure 5.10 Dendrogram showing the clustering results of officers with different patrol patterns (Shen and Cheng, 2016) .....	117
Figure 5.11 Evaluation of hierarchical clustering results based on two different similarity metrics (Shen and Cheng, 2016) .....	117
Figure 5.12 Histograms of average space-time profiles of different officer subgroups ...	119
Figure 5.13 The stay points of 4 chosen generated officer subgroups.....	120
Figure 5.14 They stay points of 4 chosen generated officer subgroups .....	122
Figure 5.15 Spatial boundary of ST-ROI No.23 in the 28 ST-ROIs detected by Module II .....	123
Figure 5.16 The Ordnance Survey POIs in the space-time boundary of ST-ROI No.23..	124
Figure 5.17 The space-time profiles of three police officers.....	126
Figure 5.18 The summarised semantic profiles of three police officers .....	127
Figure 5.19 Dendrogram showing the clustering results based on semantic profiles ....	128
Figure 5.20 The 54 ST-ROIs in three BOCU areas chosen for demonstration .....	129
Figure 5.21 The average semantic profiles of five officer groups generated by the aggregative analysis .....	131
Figure 6.1. The Euclidean distance between two points is 78m, however, people need to move 1240m from one point to the other in the network.....	135
Figure 6.2 Work threads and work flow of the network paradigm for large study area .	137
Figure 6.3 The task of map-matching.....	140
Figure 6.4 Candidate transition graph $G_{Tr}$ of trip $Tr$ (Lou et al., 2009) .....	141



Figure 6.5 The kernel-based temporal scanning window based on network route length .....	143
Figure 6.6 Comparison between the stay point identification in Euclidean paradigm and network paradigm in space .....	144
Figure 6.7 Pseudocode of ST-Net-DBSCAN.....	149
Figure 6.8 The Euclidean coverage of map-matched stay point A and its reachable points in the network .....	150
Figure 6.9 The Euclidean filter area of $s'$ and the reachable street segments of $s'$ when Euclidian Spatial_Eps = Network_Eps .....	151
Figure 6.10 Detected ST-LOIs visualised in a space-time cube and their projections on the streets .....	153
Figure 6.11 The 3D wall map visualisation of one ST-LOI in a space-time cube .....	154
Figure 6.12 The relationship between ST-LOIs and the POIs along the streets.....	156
Figure 6.13 (a) Raw observations; (b) Map-matched observations.....	159
Figure 6.14 Space-time cube visualisation of the 67 Net-ST-ROIs detected.....	160
Figure 6.15 (a) The spatial boundary of an ST-ROI in Oxford Circus; (b) The network spatial boundary of an ST-LOI in Oxford Circus .....	160
Figure 6.16 3D wall map visualisation of the 67 ST-LOIs .....	162
Figure 6.17 Semantic contributions of POI categories in ST-LOIs.....	163
Figure 6.18 3D wall map visualisation of semantic ST-LOIs .....	164
Figure 6.19 The 620 active officers separated into 7 groups .....	165
Figure 6.20 Average semantic profiles of officer subgroups.....	166
Figure 6.21 Work type composition within the officer subgroups .....	167
Figure 7.1 3D visualisation of the outcomes of LDA semantic enrichment .....	174
Figure 7.2 Loglikelihood values of LDA outcomes based on different scales of inputs..	176
Figure 7.3 Alluvial diagram showing the outcome comparison of TF-IDF and LDA approaches in Camden, Islington and Westminster .....	177

Figure 8.1 Three examples of k-NN queries of chosen points (a) the top 3 nearest neighbour points of point A; (b) the top 4 nearest neighbour points of point A; the top 4 nearest neighbour points of point B.....	184
Figure 8.2 Average k-NN distance of the ROIs generated by different approaches.....	185
Figure 8.3 The different distances from a stay point to its nearby POIs .....	188
Figure 8.4 The relative location of stay points to POIs: .....	188
(a) stay points in an ST-ROI; (b) map-matched stay points in an ST-LOI.....	188
Figure 8.5 Average closeness of ROIs to their k-th nearest POIs .....	189

## LIST OF TABLES

Table 2.1 Brief summary of classical ROI detection algorithms .....	46
Table 3.1 Input, output and method options for constructing the modules in the methodological framework.....	71
Table 4.1 Non-spatio-temporal information contained in the APLS dataset .....	84
Table 4.2 The reclassified POI categories based on the Ordnance Survey POI classification scheme.....	85
Table 5.2 The semantic contribution of different categories of POIs in ST-ROI No.23 calculated with Equation 2.2 and Equation 5.9.....	124
Table 5.3 The TF-IDF weighted semantic contribution of different categories of POIs in each ST-ROI in Camden.....	125
Table 5.4 The TF-IDF semantic meanings and names of five chosen ST-ROIs in the extended case study.....	130
Table 7.1 Analogy from textual topics to semantic analysis of places .....	171
Table 7.2 Semantic categories summarised by LDA though the combination of popular POI subcategories .....	173
Table 8.1 Accuracy comparison of stay point identification methods .....	182
Table 8.2 Accuracy comparison of map-matching methods.....	182
Table 8.3 Average temporal density of ROI detect by different approaches .....	185
Table 8.4 The basic network information and the percentage of avoided distance computations.....	186

## **NOMENCLATURE**

API: Application Programming Interface

APLS: Automatic Personnel Location System

POI: Place of Interest

GPS: Global Positioning System

GIS: Geographic Information System

UCL: University College London

UK: United Kingdom of Great Britain and Northern Ireland

Inner London: City of London, and the London boroughs of Camden, Hackney, Hammersmith and Fulham, Islington, Kensington and Chelsea, Lambeth, Lewisham, Greenwich, Southwark, Tower Hamlets, Wandsworth, and Westminster

GPS Point Update/Record: A positioning capture by a GPS device containing the information of Northing and Easting coordinates of a moving object at a given point in time.

GPS Sampling Rate: A specific instant in time. GPS carrier phase measurements are made at a given frequency (e.g. every 30 seconds)

## **Chapter 1**

# **Introduction**

## 1 INTRODUCTION

Things that people do in space and time have long been a research topic in behavioural and socio-economic studies, with particular focus on the highly dynamic urban environment (Chapin, 1974; Cullen, 1972). The term "activity pattern" in this research is used to describe patterned ways in which groups of people carry out their daily activities. These activities are naturally linked to the places where they are undertaken and the times (e.g. time of day, day of week or year) at which they take place. By segregating communities or aggregating individuals into groups of people sharing similar activity patterns, many socio-economic and socio-demographic problems and their ties with individual behaviour preferences can be revealed (Chapin, 1974). Research into these patterns attempts to answer questions about the life styles, behaviours, routines and preferences of different groups of people.

Early studies of human activity patterns were confined to traditional statistical and survey studies because of a lack of large scale activity data and the tools/measures to enable the tracking, logging, management and analysis of detailed lifecycles of individuals. Nowadays, thanks to the ubiquity of telecommunication and sensor technologies, such data are now available at decreasing cost in the form of GPS trajectories and mobile phone user data. Movements are continually recorded as trajectories, which are sequences of geo-located and time-stamped points, often with associated information (Kuijpers & Vaisman, 2007). GPS, mobile phone service and location-based app data are typical examples of these new datasets. They are often large and possess high spatial and temporal resolutions, which enable researchers to explore movement patterns in greater detail than before.

Most current research trying to make use of this kind of data for behavioural analysis focuses on the spatial, temporal or semantic aspects in isolation (Andrienko et al., 2011; Kwan, 2004; Li, 2011; Timmermans et al., 2002). For instance, Li et al. (2008) uses space and place as a depiction of human activity patterns, while Wilson (2001; 2007) analyses human activities in time based on duration and time sequences. However, these studies neglect the fact that space, time and semantic contents play equally significant roles in the description of people's activities and therefore do not provide a complete indication of people's activity patterns. In reality, people carry out different activities at different places at different times of the day. The activity they are doing is not only indicated by where they are, but also how long they spend in the place and when they do it. This is also because time is a resource; how people allocate the length of their time resource on particular activities also varies (Szalai, 1966). Goodchild (2015) also argued that "Platial views offer new insights beyond traditional spatial perspectives as human activity is more aligned with place rather than geometric space", which indicated the importance

of “what the place/activity is about” over the pure physical location in the activity and place related studies. In the light of this, the thesis aims to build on previous work that views the spatial and temporal domains in isolation, and establish a universal framework that enables a comprehensive analysis of space, time and semantic meaning in order to group people with similar behaviour patterns based upon trajectory data. The framework segregates individuals into subgroups based upon where (place), what (semantics), when and how long (duration) the activities are conducted for each individual.

This chapter describes the general progress and limitations of activity pattern profiling and grouping in human dynamics research. We first discuss as a background the progress of related research and the motivation that inspires the carrying out of this work. Then, we summarise the aim and objectives according to the existing limitations and problems, which is followed by a description of the structure of the thesis describing the progressive process of the research idea and the logical relationship between different paradigms of methodological frameworks and the chapters of the thesis.

## **1.1 PROGRESS IN ACTIVITY PATTERN AND HUMAN DYNAMICS STUDIES**

The term human dynamics is a concept derived from the realm of physics, which analyses the movements and flows of objects with the help of mathematic models, computation tools and physical laws. Physical dynamics focuses on both micro (individual) and macro (group) perspectives and specifically investigates the spatio-temporal relationships of the observations and the patterns of changes of the observed movements. Human dynamics, on the other hand, takes the movements and activities of humans as research objects. It not only investigates the patterned features in activities, but also examines the semantic causes that give rise to the patterns. Many applications such as tourist activity study (Edwards et al., 2009), location-based social networks (Zheng, 2011) and urban planning rely on the analysing tools of human dynamics, and progress in these studies also in return facilitates the development of techniques and theories of activity patterns and human dynamics’ studies. Below is a general summary of the progress in these studies that can facilitate the construction of the proposed theories and framework of this thesis.

### **1.1.1 Movement logging**

To collect the data of movements and activities, travel logs, recording where people travel to and what they do, are proposed and are used as one of the most crucial measures to obtain the critical information needed for pattern analysis. Travel logs

contain information about the space, time and demographics, as well as socioeconomic characteristics of individuals, so that the scheduling and purpose of daily activities can be reflected. Although traditional manual travel logging is labour intensive, it enables us to link the movements and stops of people with the physical environments they are in and look into the higher behaviour characters from the activities in space and time.

Thanks to the increasingly ubiquitous application of mobile location-aware sensors and the progress of information and communication technology (ICT) during the past two decades, large-scale data collection of the ever-changing position of moving individuals, such as Global Positioning System (GPS) data, has become technically feasible and economically affordable. These changes have replaced the previous challenges of data scarcity and a lack of computational power with the unprecedentedly large movement of data that contain much more detail and higher spatial and temporal granularity than ever before via various sources, such as smartphone apps, e-commerce and public information infrastructures. More importantly, they transformed the focus of Geographic Information Science towards spatio-temporal and dynamic relationships of human behaviours and the environment.

Individual movement data collected by mobile devices are commonly seen as simple spatio-temporal points (Spaccapietra et al., 2008). In other words, they are a set of points with geographic locations, times and sometimes a few other relevant attributes that are associated in the form  $(x, y, t)$  or  $(x, y, z, t)$ , where  $x$ ,  $y$  and  $z$  are spatial dimensions and  $t$  is the temporal dimension (Kuijpers, 2007). A trajectory is naturally formed when these points of an individual moving object are linked in chronological order. Trajectories can be generated by service providers (e.g. telecom companies, taxi companies, airline companies, etc.), social media services (e.g. Twitter, Flickr, Instagram, etc.), life-logging applications (e.g. Nike+ and Mytracks), government and nongovernmental organisations (e.g. maritime traffic management, aviation management, police force activity, etc.). Space and time behaviours in these datasets are originally logged for typical operational and management purposes, rather than knowledge extraction. Abundant information embedded in these datasets remains untouched and can be of great value for human dynamics research if handled properly with the emerging data mining approaches.

### **1.1.2 Extracting region of interest and activities from movements**

Region of interest (ROI) has many synonymous names in activities studies, such as hotspot, interesting place and interesting region. This concept is widely used in travel pattern, crime study and epidemiology, in which the occurrence of events are



represented by point records in space and hotspots are significant aggregations of the point records. In human dynamics studies, the ROIs are the places attracting high volumes of people visits. With more and more location point data generated by modern sensors, such as GPS devices and mobile phone networks, ROIs where people gather and interact have been a hot research topic in geographical-related human behaviour research nowadays. They are usually detected by finding dense aggregations of stopping behaviours, information posted via telecommunication devices or check-ins with LBS applications. Traditional ROI detection methods only look for aggregations in planar space and thus generate Spatial ROIs that distribute on 2D Cartesian surfaces. The latest progress of the ROI detection method can be divided into two directions. One of them is to add temporal and non-spatial dimensions into the detection analysis via novel clustering or density calculation methods. This improvement allows researchers to not only detect ROIs in space but also find temporal aggregations and patterns in other non-spatial attributes of events. The other direction is to develop the detection method into network-based versions. In these methods, the traditional Euclidean and planar representations of urban space and locations of places are replaced by the topology of interconnected road links and positions on the road segments. The ROIs can be detected in urban areas and realistically aligned with the street geometry through the network-based ROI detection methods.

### **1.1.3 Adding semantic meaning to places**

Activities and interactions taking place in virtual space are not independent from the activities and interactions in physical geographic space (Shaw & Yu, 2009). In fact, they usually influence and interact with each other. Although there has been a long tradition in activity studies to link the relational spaces or the physical environment with activities of people, it is the recent advancements in computing technology, collection of large-scale mobility data, and development of theories on social-spatial processes that has revolutionised the way in which we investigate social-spatial events. The corpus of human dynamics has recently been greatly expanded by automatically combining the raw movement trajectories and ROIs with data mining operations on the related contextual knowledges. This process is known as semantic enrichment. It can provide applications with meaningful knowledge about movement and introduce the semantic aspect into the traditional space-time analysis of human activities.

#### **1.1.4 Aggregative analysis of activity patterns**

Aggregating (grouping) research objects to discover information in further detail, instead of regarding the objects as single entities, has long been a first step for many types of study, especially human behaviour studies. Aggregation study is important for a better understanding of human dynamics because people's different movement patterns characterise their respective motivations and behavioural preferences.

One of the earliest attempts to group people's urban activities is Chapin's (1974) aggregative model for the patterned forms of aggregates of individuals. Spatial aspect is a key factor of the activity in this study because space provides availability of physical access to functional facilities and services in the city and contributes to people's motivation to act. Chapin simplified the model for explaining activity pattern by the "motivation generated by people and place → choice of activity → Action" framework sequence. However, being limited by the information gathering technology and the cross-discipline ability at Chapin's time, the scale and approaches of his research is still based on statistics of predefined groups and manual surveys on committed volunteers.

With the increasing usage of smart phones and LBS technology, a new trend emerged that used GPS and mobile devices to automatically log people movement history in greater detail. Zhong (2015) contends that users' intrinsic visiting preferences of semantic places reflect the personal "profile" of the online check-in software user. Similarly, Zhong proposed "you are where you go". The idea is to use user's check-ins of POIs to infer his/her demographic features and design a "location to profile" (L2P) framework to depict users' activity patterns by bridging the gap between online and the physical world. In location-based social networks, Xiao (2014) depicts a person by his/her semantic location history (SLH). Clustering analysis based on the semantic similarity of the low-level activity and movement history of individuals can aggregate people sharing similar patterns.

## **1.2 LIMITATIONS AND CHALLENGES**

As a result of the ongoing trend mentioned above, several research studies have emerged over the past decade aiming to generate activity patterns from space-time mobility information. Although different solutions have been proposed for problems of human dynamic analysis and aggregation of profiles, many of these studies still suffer from several limitations. These limitations are mainly related to the deficiency in harnessing the temporal aspect of data and the insufficient awareness of urban network structures during their method development.

Deficiencies in the usage of temporal data are twofold: namely, the lack of temporal consideration in activity patterns and the ignorance of time in semantic enrichment of places. Limitations of activity pattern analysis for the temporal aspect can be further broken down into three issues: namely they are failing to incorporate time and sequence of people's place visiting behaviours, they are ignoring the time span of the detected interesting regions, and they do not account for the impact of the duration of stays on individual profiles. Limitations of time-sensitive semantic enrichment indicate that the semantic meaning of a place remains unchanged during the entire day in existing semantic enrichment studies. This static view of places fails to take the ubiquitous changes of activities and semantic meaning of places into account and is insufficient for the all-round description of highly dynamic urban places and activities.

On the other hand, limitations related to network analysis include three issues. The first issue is using straight Euclidean distance in stop identification and the pre-processing stage. This often causes errors because the actual routes taken by the moving individuals are ignored and the distance is misrepresented. The second issue is that the ROI detection approaches in most existing researches are Cartesian-based and cannot accurately pinpoint the covered area. The ROI detection methods in these studies are based on spatial or spatio-temporal clustering techniques that are not suitable for the urban environment. The last issue is that there is no work that has achieved the space-time semantic analysis and visualisation on a street segment level.

To sum up, very few of existing studies perform the clustering in the spatio-temporal domain and there is an immense lack of using urban street networks for activity profiling in finer scales. Therefore, in this thesis we aim to address these limitations in order to enhance and standardise the urban ROI detection and activity profiling from GPS data. The next section describes this aim in more detail, setting out the list of objectives that would help achieve this aim.

### **1.3 RESEARCH AIM AND OBJECTIVES**

The main aim of this research is to develop a methodological framework to automatically profile and aggregate people's space-time activity patterns based on space-time human mobility data in an urban semantic environment. The challenge is to produce a paradigm that overcomes limitations listed in the previous section and bridge the gap between time, network space and semantic meaning information of people activities. The developed method must also be robust enough to work with no information but with large-scale raw GPS trajectories and public POI data that were

originally collected for operational and navigation purposes. In order to achieve this aim, a number of objectives must be met, defined as follows:

- Objective 1: Review existing approaches and methodologies for activity pattern analysis based on mobility data from both geography and other disciplines. Summarise and critically assess the fundamental assumptions and experiment settings in these approaches.
- Objective 2: Develop an ST2P (Space-Time to Profile) framework of multiple modules. Each module is designed to address part of the limitations listed in section 1.2.
- Objective 3: Design two paradigms, according to the structure of the framework in Objective 2, to incorporate the spatial, temporal and semantic information for activity profiling and aggregation in Cartesian space and urban networks.
- Objective 4: Compare the performance of the two paradigms proposed in Objective 3 and the existing mainstream approaches. Examine and demonstrate our framework's suitability on real people's large-scale movement trajectories in urban road networks from various aspects.

## **1.4 GENERAL THESIS ORGANISATION**

The objectives of the thesis are fulfilled across nine chapters. As highlighted in this chapter, the thesis is dedicated to aggregate activity patterns from GPS-based human mobility data in cities. The method-related challenges and limitations described in subsection 1.2 can be further divided into five problems to be addressed across the development of the framework. The train of thought in the report is chronological. It records related studies that forge the ideas of research and proposes a path towards our newly proposed methodological framework before testing it via case studies. A summary of each chapter is provided accordingly in this section. Each of these chapters counts as a step in the train of thought, and possesses several subsections that describe a method or address the detailed research questions in their own respect.

After the Introduction, Chapter 2 reviews the literature, in particular the current state-of-the-art approaches and models for trajectory analysis, regions of interests (ROI) detection, semantic enrichment of places and the similarity definition and profiling of activity patterns. Chapter 2 describes these approaches along with a critical argument of benefits and defects of each listed method. Particularly, the gaps that are not yet filled by existing studies in each step are summarised in this part as well.

On the basis of the gaps summarised in Chapter 2, Chapter 3 lays out a hypothesis: “where (what place), when and how long you stay is who you are”, and presents our methodological framework to integrate the spatial, temporal and semantic aspects for profiling people’s urban activities. Lists of conceived methods to achieve the phasal goal of the framework modules are organised in 4 chronological modules: pre-processing, ST-ROI detection, semantic enrichment, profiling and aggregative analysis. The rationale for using the methods in each module of the framework is briefly explained, and the process of how certain methods bridge the gaps are depicted. The end of the chapter presents two paradigms, each of which are adapted to operate the general framework in Cartesian space and urban street networks respectively. The Euclidean paradigm achieves part of the research objectives and the network paradigm inspired by its predecessor fulfils all objectives.

Chapter 4 describes the data-related issue of this research. It is divided into 3 sections. Section 4.1 is the general definition of the study area and section 4.2 describes the datasets that participate in the analysis as inputs to the proposed framework. Within section 4.2, section 4.2.1 provides the basic knowledge about the human mobility data in our case studies, while section 4.2.2 presents the environmental data including the POI information and the street networks describing the space for the movements and activities to take place. The basic characteristics of these datasets are briefly explored in section 4.2.3 before the general framework of methodology is introduced. Through this better understanding of data, it is hoped that an improved model of existing methods may be derived, designing a framework to evolve existing challenges with novel approaches will contribute to the improved understanding of activity patterns in city-wide human dynamics.

Chapter 5 describes the Euclidean paradigm, in which the adopted methods are organised to follow the flow of the general framework. The Euclidean paradigm is tested with the police patrol movement data in 3 central London boroughs and demonstrates its advantages over existing approaches. The first module of the paradigm is pre-processing. A kernel-based scanning window slides through the time dimension of each trip to identify stop episodes. The ST-DBSCAN algorithm with parameterisation is used in the second module for dense aggregation detection of stay points in space and time. In the third module, the detected aggregations are defined as ST-ROIs, and their semantic meanings are explored by running a term frequency-inverse document frequency (TF-IDF) method on data of adjacent POIs and functional buildings. Specifically, we add opening time information of POIs to the semantic analysis of ST-ROIs to make this process more accurate. This step is motivated by the following observations. Sometimes, only a part of the POIs are open in the time period of intensive

visits; therefore, the semantic meaning of the street segments is only contributed to by the opened POIs. On the other hand, as different POIs open at different times, the same set of streets can serve different purposes at different times of the day. For example, a shopping street can turn into a bar street when most stores are closed and the roadside bars start working at night. By annotating all the ST-ROIs with semantic information, an individual space-time profile can also be turned into a semantic profile to describe his/her time allocation with semantic knowledge. Lastly, the semantic profiles depicting the individuals' time allocations on semantic ST-ROIs are aggregated by a hierarchical clustering method in the last method. A Euclidean paradigm enables the synergetic clustering analysis of space and time and brings a temporal dimension into the semantic profiling of people's pattern of activity. The ubiquitous street structure in the city and the influence of buildings' opening hours on semantics, however, are completely ignored by this paradigm.

Aiming at the unfilled gaps of the Euclidean paradigm, Chapter 6 moves on for a more advanced network paradigm to analyse human mobility in finer scale. The central London boroughs with dense streets are chosen as a case study. We modified the ST-DBSCAN algorithm and combined it with map-matching techniques to detect ST-ROIs based on human movements within the street networks. We proposed spatio-temporal line of interest (ST-LOI) as the novel definition of ST-ROIs under the network representation of urban space. This improvement enables us to better locate the stay points and interesting regions in street segments for detecting ST-LOIs (Spatio-Temporal Lines of Interests) instead of the approximate areas generated by the Euclidean paradigm. Furthermore, we proposed a 3D wall map for better visualisation of the ST-LOIs. The space-time boundaries of ST-LOIs are also confined in street networks to more accurately associate the POIs with ST-LOIs spatially in the semantic enrichment process.

In Chapter 7, the semantic enrichment module of the network paradigm adopts an LDA (Latent Dirichlet Allocation) topic modelling algorithm to further improve the adaptability and accuracy performance on large scale human dynamics datasets. The semantic ST-LOIs generated by TF-IDF and LDA are compared to show LDA's ability to detect less significantly meaningful places and provide a detailed composition of the semantic meaning in each place without a priori hierarchical definition of POI/building categories. The experiment also shows that LDA's stability rises as the sizes of the data and study area increase, indicating its better suitability to work with scaled up datasets.

In Chapter 8, the proposed methods in the framework are validated and evaluated and the Euclidean paradigm and the network paradigm of the framework are compared. We

evaluate the methods in every module by their accuracy from spatial, temporal and semantic aspects to evaluate the performance of the two paradigms against conventional approaches on urban movement data.

Finally, Chapter 9 summarises the significance of the major findings and the completed objectives. Current deficiencies are critically reviewed. Improvements to be performed in further studies are outlined to overcome the deficiencies and expand the finding to wider applications. In the end, the summaries in the chapter lead to the policy implications and the final conclusion of the work conducted in this thesis.

## **Chapter 2**

# **Literature Review**



## **2 LITERATURE REVIEW**

Urban activity pattern analysis has been a popular research topic since the 1970s. Chapter 1 has demonstrated the importance of this research for many applications such as planning for facilities and linking virtual and material services. The first and foremost step in undertaking research in this area is to understand previous related works and theories, which can be of significant value as references and inspiration for the research. In this chapter, a comprehensive review of existing theories and conventional approaches to profiling and aggregating the activity patterns of urban individuals will be presented.

In the first section, theories and general concepts of trajectory-based activity pattern studies are introduced. The review focuses on the modelling of movements and stops and several aspects of activity patterns.

The second section reviews existing approaches for the detection of regions of interest (ROIs) (i.e. hotspots). This section will examine how algorithms have used different features of movement trajectories to detect ROIs, mostly in Cartesian space. Particular focus will be placed on the stop identification methods and the representation of urban space.

The third section is the review of methods for the semantic enrichment of places. This section will detail how environmental information is used to annotate and explain the meaning of interest regions for human activities and how these approaches play a fundamental role in the complete description of activities.

The fourth section focuses on the methods used for profiling and describing patterns in activities and trips. The similarity metrics used in preceding works for aggregative and grouping analysis are summarised and categorised. Advantages and deficiencies of various similarity metrics are critically reviewed to provide insight for the novel profiling method proposed in the following chapter.

The fifth section outlines the alternative network analysis tools to evolve the research in Cartesian space into a network-based version of the work. One of these tools, map matching, improves the accuracy of raw trajectories in urban streets. Spatial querying algorithms are another type of tools to ensure the efficiency of neighbour searches and massive distance calculations in networks. These tools are ubiquitously used across various network analysis methods.

The sixth section discusses the visualisation techniques for the presentation of research results. The trade-offs and advantages of traditional and state-of-the-art space-time visualisation methods are reviewed in this section.

The chapter will conclude with an assessment of the challenges and limitations of conventional approaches that this thesis is aimed to resolve. The implications of previous theories and models are summarised to provide insight and a foundation on which the proposed methodological framework will be built.

## **2.1 GENERAL THEORIES AND FRAMEWORKS IN ACTIVITY PATTERN STUDIES**

Before Internet and mobile devices became part of everyday life, early studies of human activity patterns were confined to traditional statistical and survey studies involving tracking, logging, managing, and analysing of the massive and detailed life cycles of individuals. In early time-use and activity studies (Chapin, 1974; Cullen et al., 1972; Szalai, 1966), it was of great difficulty to simultaneously analyse complex trip trajectories generated by human trips and activities in space and time. This is because these trajectories possess multiple interacting features including location, timing, duration, sequences, speed, and semantic type of activities. Modern ubiquitous telecommunication and sensor technologies make this simultaneous analysis possible. Large scale data collection of the movement trajectories of massive users, such as through GPS data, smart card data, and mobile phone user data, has become technically feasible and economically affordable. In particular, location-based services (LBS) has been a popular industry with the wide spread of the above-mentioned technologies in recent years. Some applications of LBS, such as Foursquare and Twitter geotagging, have penetrated into all aspects of daily life and provided a huge amount of data recording the “check-in” and place-visiting behaviours of millions of users. These data provide continuously updated “4W” information (Shaw et al., 2016): “who the person is, when s/he visited/stopped at a place, where the place is, and what the place/activity is about”. Thus, they capture the spatio-temporal and semantically meaningful snapshots of personal activity patterns and indicate “what the person is like”. The above features make this new type of dataset particularly suitable for the research of human dynamics. Harnessing the advantages, however, requires systematic and multi-disciplinary insights and efforts.

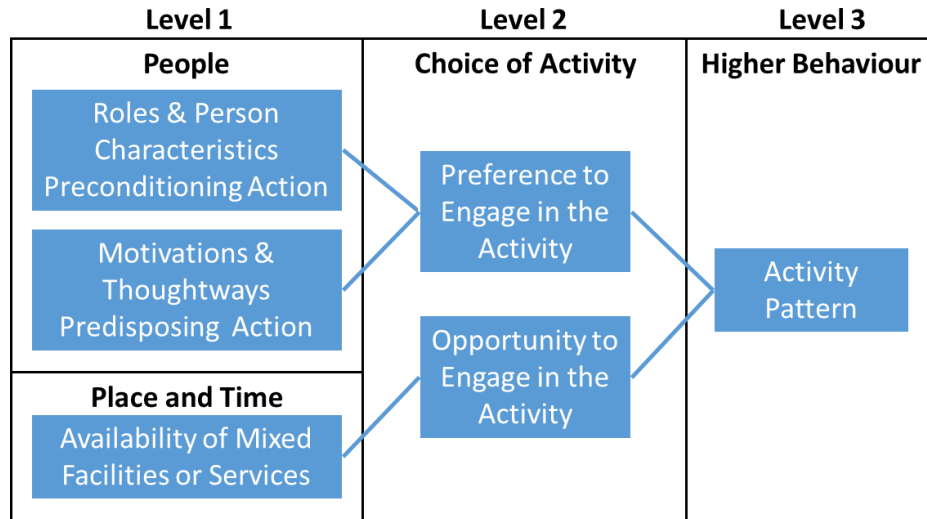


Figure 2.1 Individual and environmental factors influencing activity patterns by level

By generalising the reviewed works (Ashish & Sheth, 2011; Yan & Chakraborty, 2014), the factor composition of individual activity patterns can be broken into three levels. Level 1 is the basic physical level of activities. The personal characteristics and motivation features in this level determine the activities people are able or willing to undertake, while the spatial location and temporal availability of the activity-related place determine the opportunity to engage the activity. Combining the influences of individual preference and external/environmental opportunity in Level 2, the higher-level semantically patterned activities are developed.

Corresponding to the three-level composition of patterned activities, most data-driven approaches (e.g. Parent, 2013; Renso, 2013; Zhang, 2016) commonly indicate that meaningful activity patterns can be extracted and understood from raw trajectory data through three general steps: **geospatial process**, **semantic process**, and **knowledge discovery**. As Figure 2.2 shows, the geospatial process detects activities in space and time and models the trajectories as place-visit histories, the semantic process explores the meaning of the places where the activities take place, and knowledge discovery finds patterns in the semantic trajectories or semantic location histories. General definitions and theories of these three steps are summarised in the following sections.

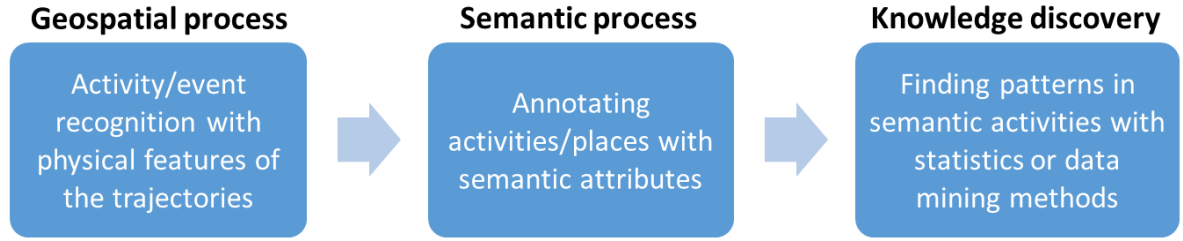


Figure 2.2 The three-step extraction of activity patterns

### 2.1.1 Geospatial process

In Level 1 of the activity factors presented in Figure 2.1, the individual person intends to fulfil meaningful goals by staying at a certain place for a certain period and travelling from place to place for multiple stays. This process can be visualised as continuous lines in a space-time cube (Andrienko et al., 2010) as demonstrated in Figure 2.3. When these continuous movements are recorded by on-body sensors, they are actually discretely recorded with certain sampling rates and positioning errors. Therefore, in data-driven approaches, each trip trajectory is discretely represented by a sequence of time-stamped location points,  $\{(x_0, y_0, t_0), (x_1, y_1, t_1), (x_2, y_2, t_2), \dots, (x_N, y_N, t_N)\}$ , where  $x_i, y_i, t_i \in R$ ,  $i = 0, 1, 2, \dots, N$  and  $t_0 < t_1 < t_2 < \dots < t_N$ . In the trajectory,  $x_i, y_i$  are the spatial coordinates of the moving object at the instant  $t_i$ ,  $t_0$  marks the start of the trip, and  $t_N$  is the instant when the trip terminates. In accordance with Level 1, the essence of the geospatial process is the spatial and spatio-temporal analysis of personal trips for the extraction of non-semantic spatial knowledges from the raw trajectories.

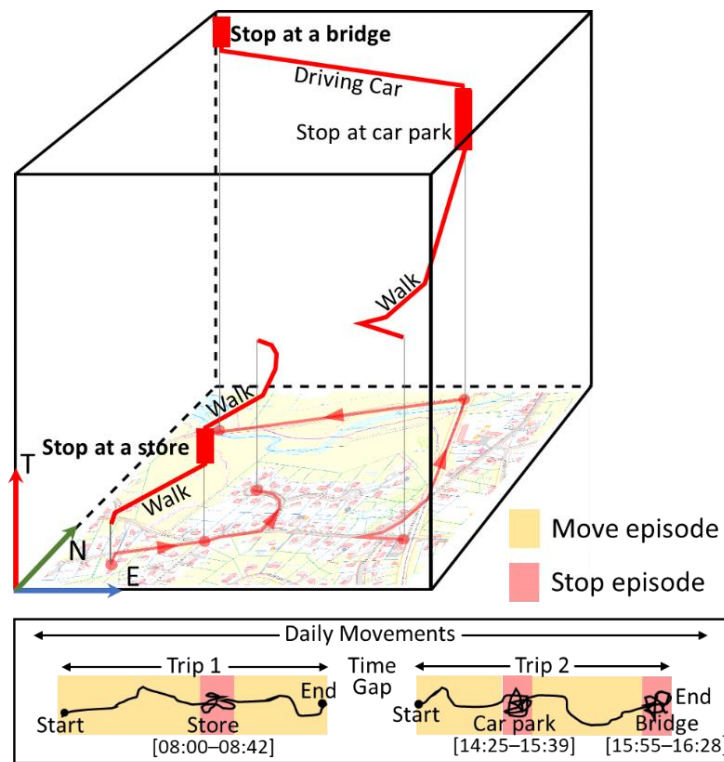


Figure 2.3 Representation of trip trajectories in a space-time cube

Both moving and stopping actions during the trip take finite amounts of time in order to achieve the trip purpose. Palma et al. (2009) separate trips into stop episodes, in which a person stops and visits a place, and move episodes, in which a person travels between places. The common assumption in activities pattern study is that patterned stops during movements in places indicate that a person is undertaking an activity (Kwan et al., 2004; Palma et al., 2008; Thierry et al., 2013; Zheng et al., 2009). Since these studies focus on patterns in activities instead of travelling patterns, the stop episodes in trip trajectories are of greater interest than the move episodes. Thus, many relevant works transform the trip trajectories into place-visit histories and the detection of regions of interest (Li et al., 2008; Xiao et al., 2010), and the occurrence of events and activities is an important step in these studies.

Spatial aspect is not the only concern in the geospatial process; time plays an equally significant role because “the trajectory is by definition a spatio-temporal concept” (Spaccapietra, 2008). For this reason, concepts in time geography are often introduced into the activity study to explain people’s constraints and trade-offs when they have only a limited amount of time to spend on multiple activities in different places (Miller, 2005). The development of portable location sensors and digital communication technologies has further improved the volume and resolution of the temporal dimension of mobility data to a degree beyond the reach of traditional activity survey studies. These advances

present opportunities to discover temporal and spatio-temporal patterns of activities in addition to the purely spatial aspect of trips; however, they have also multiplied the complexity of data mining tasks (Zheng & Zhou, 2011). Approaches for detecting ROIs of human activity and events are reviewed in Section 2.2.

As depicted in Figure 2.1, movement trajectories follow the geometry of the streets. Urban networks are another constraint of movements and activities in the city, since urban streets provide the space people navigate to access the places they intend to visit. A network-based spatial representation is necessary because most activities that involve public interests are located somewhere in the network. Transforming the raw trajectories into network-based trip routes (i.e. street segments passed by a moving individual in sequence) and switching from the Cartesian spatial representation to the network representation of space reveals different space-time patterns, which can carry different meaning from the patterns discovered based on unconstrained raw trajectories (Zheng & Zhou, 2011). People and objects moving in the city can be modelled as network time – geographic entities (Chen et al., 2016) by combining time geography and network analysis. The review of network analysis methods is detailed in Section 2.5

### 2.1.2 Semantic process

The semantic activity factors in Level 2 of Figure 2.1 show how the occurrence of an activity is triggered by the lower-level factors. The semantic process transforms the spatial and temporal factors in Level 1 into semantic activity attributes in Level 2. The term “semantics” means the “philosophical study of meaning”. Semantic study mainly investigates the relationship among words, phrases, signs, and symbols as well as their denotations. Extending this concept into geographic information science, the aim of the semantic process is to extract more meaningful information from the spatio-temporal data collected by the location sensor or from environment data depicting the external context in which the activity takes place. For example, after the semantic enrichment process with map data, a location with coordinates can be annotated as a place of tourist attraction or a commercial street. These annotations are also called “platial (i.e. belonging to a place) signatures” or “semantic meaning of places” in different works. The enrichment process turns the traditional ‘space-centred’ mobility study into the “people-centred” activity study (Yan & Chakraborty, 2014) and provides the “platial” thinking beyond traditional spatial perspectives of human activity (Goodchild, 2015). Corresponding to the conceptual differences between a “location” and “place”, the geospatial analysis process and semantic analysis process are often used separately for

different research objectives. Nevertheless, there are strong conceptual overlaps between the methods used in the two processes (Luo & MacEachren, 2014). Other definitions of trajectories that combine physical and semantic features include the notion of lifelines or periods of life, which are associated with people's stop locations at regular or irregular intervals (Yuan et al., 2004). Similarly, Thériault et al. (2002) use a spatio-temporal database model to handle multi-dimensional lifelines with temporal GIS. Their core contribution is to find spatial clusters of activity locations and use them to determine patterns in certain aspects of people's lives, such as professional activities. The semantic patterns support the statistical and data mining analyses in the afterwards knowledge process. Related methods for semantic enrichment of places with external context data are outlined in Section 2.3.

### 2.1.3 Knowledge discovery

The knowledge-discovery process extracts high-level behavioural information from the semantic trajectories or place-visit histories. Statistical and data mining approaches can be used to find patterns in spatial, temporal, or semantic senses after the trajectories are described with corresponding features in previous processes. Early studies such as Axhausen and Gärling (1992) and Chapin (1974) could only use statistical methods to summarise the activity patterns of pre-defined social groups or communities. For the activity study in policing applications, some papers have looked at police presence in general and tested the influences on an overall statistical scale (e.g. Sherman, 1995; Ratcliffe & Taniguchi, 2011). Critically, however, the detailed process of foot patrol was not empirically studied with data that contain more detailed or contextual information.

With the progress of data mining, clustering/unsupervised machine learning techniques are introduced into human dynamics studies. Some researchers have used data mining to extract non-semantic activity patterns or travel patterns purely based on the physical features of trajectories, such as spatial coordinates and speed (Ashbrook, 2003; Hariharan & Toyama, 2004; Liao, 2005). Other data mining studies of activity patterns are based on semantic location histories.

Semantic location history (SLH) comprises a series of ROIs of various functional categories visited by a person in a trip in order, e.g. shopping malls → restaurants → cinemas. The semantic similarities of moving individuals can be measured by comparing their semantic location histories. For example, Zheng et al. (2009) extracted stay points from user trajectories and applied tree-based hierarchical graph (TBHG) clustering to

model the location histories of multiple users. Exploring patterns among trips and activities can also be based on other similarity indices regarding spatial and/or temporal dimensions, such as place-visit sequences. These similarity metrics for aggregative knowledge discovery are reviewed in Section 2.4.

## **2.2 DETECTING REGIONS OF INTEREST**

Lew and McKercher (2006) suggested that uneven distribution of human dynamics in the space and time of cities is ubiquitous; most activities take place in more “popular” regions of a city, making human flow to some areas much more crowded and denser than to other places. Due to this tendency, three steps are normally executed for ROI detection: stop detection and determination (Andrienko & Andrienko, 2011), density-based aggregation (Zhou, 2004) and validation. Since validation requires external information and varies across cases, this section reviews the first two steps, stop detection and density-based aggregation. Points where people stay for a specified length of time are defined as stay points, and dense aggregations of these stay points as ROIs. In other words, an ROI is a place with a high-density gathering of stops or user visits. Most studies discuss spatial ROIs, while a few extend the concept into spatio-temporal ROIs by taking the time dimension into account. In this section, some of the existing techniques for identifying ROIs from trajectories collected with relatively high and regular sampling rates, especially GPS data, are reviewed.

### **2.2.1 Trip segmentation and stop identification**

A generic and well-known paradigm for pre-processing raw GPS trajectories is to extract the location and time of the activities and events. For this purpose, stay points within trips need to be detected from trajectories, and they should be distinguished from the ends of trips. Therefore, stop identification often accompanies trip segmentation, which has been standardised for use separating different fragments of personal daily movements in mobility studies. As demonstrated in Figure 2.3, the trajectory of a sequence of consecutive GPS points is called a trip, and the moving individuals’ movement records have been segmented into two trips. Time gaps longer than a defined time threshold divide the time-stamped points into different trips, and a person can launch multiple trips in a day.



Biljecki (2010) argues that most trips should be separated from each other because long stops such as staying at home at night or eight-hour working periods often reduce the continuity of location updates of location sensors. Hence, Biljecki (2010) applies an easy rule-based algorithm to segment different trips. Stopping time is the rule, and extremely long stops are considered indicative of the ends of trips. In his study, the stopping time threshold can be altered for different sampling rates or moving objects. Similar rule-based trip segmentation methods are also used by Bolbol et al. (2012) and Zheng and Zhou (2011).

As for the stop identification within the segmented trips, Figure 2.3 shows that every trip can be divided into “move episodes” and ‘stop episodes” with stop identification methods (Alvares et al., 2007, Spaccapietra et al., 2008), and semantic meaning or contextual information can be added to each episode. The “move episodes” are the parts of a trip in which the moving object continuously changes location, whereas the ‘stop episodes” are the parts of a trip in which the object stops moving for a while or slowly “wanders around” within a small and confined area as illustrated in Figure 2.4 (Palma, 2008). ‘stay points” are the location updates recorded in the stop episodes, and most of the semantic analysis focuses on these (Parent et al., 2013; Ying et al., 2013; Zheng et al., 2009).

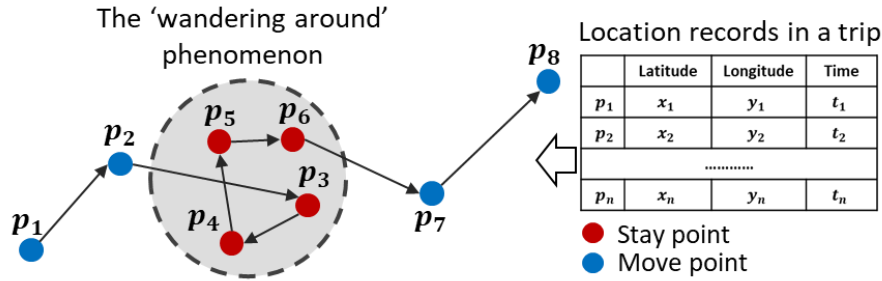


Figure 2.4 Stay points in a trip trajectory

There are two main types of methods for stop identification: density-based and threshold-based. The former is based on calculating the density and number of point records within a confined space. The latter uses speed or distance threshold as the condition to differentiate stops from movements. Both types have flaws.

In density-based methods, a stop can be detected by searching for enough nearby points only in the spatial domain, as exemplified by Ashbrook and Starner (2003). Thierry et al. (2013) proposed a spatial kernel density estimation (KDE) to create a smooth density value surface for each individual trip trajectory so that the identification process can be less sensitive to the large positioning errors in some records. However, this type of

method completely ignores time information and may misidentify a cluster of stay points at the same location but different times as a stop. Trip 1 in Figure 2.5 is a typical example showing a moving person who has visited a place already and then visits it again during his/her return trip. Spatial density-based methods will interpret the two stops at different times as a single stop by mistake. This method is, however, robust under noisy conditions; a few stay points with large location errors (e.g. Trip 2 in Figure 2.5) will not undermine its accuracy. Siła-Nowicka et al. (2016) avoided the misidentification of stops in returning trips in trip 1 of Figure 2.5 by adding a temporal sliding window to the KDE approach. However, their method only worked for trajectories with constant sampling rates because it simply counted the number of points in the time window to decide whether the points are stay points and ignored the duration of stays.

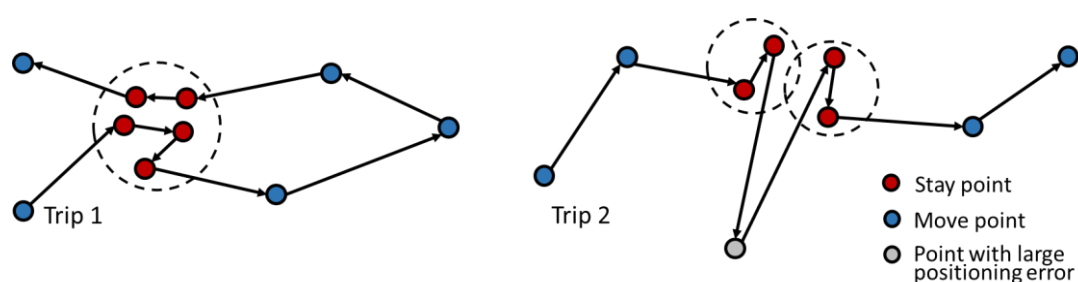


Figure 2.5 Special cases in which conventional stop identification methods may make mistakes

Another very common method is to find the point where a person stops moving, moves slower than the pre-defined speed threshold (e.g. less than 1 km per hour for pedestrians), or moves a very short distance from the previous record. These methods are called threshold-based stop identification. They look at the temporal sequence of recorded locations and use a set of decision rules based on distance and time to identify stay points. This type of method iteratively tests observations to determine whether they remain within a given “wandering around” distance of previous observations and checks whether the time between a point and the previous point exceeds a predefined stop duration.

For movement datasets that possess relatively high spatial and temporal accuracy, such as GPS tracking data, one person’s movement speed can be estimated by dividing the spatial displacement between two consecutive updates by their time difference. This estimated speed is called median speed and can be associated with the point of the latter update (Bolbol et al., 2012; Lou et al., 2009). The calculation of median speed can be expressed as Equation 2.1.

$$V_1 = \frac{\text{Distance}(p_1, p_0)}{t_1 - t_0} \quad \text{Equation 2.1}$$

where  $\text{location}(t_1)$  is the  $\text{Distance}(p_1, p_0)$  is the Euclidean distance between two consecutive updates,  $p_0$  and  $p_1$ .

The threshold-based methods are aware of the time differences of stops, so the stops in Trip 1 in Figure 2.5 can be accurately differentiated. Nevertheless, when this type of methods meets an erroneous stay point that is observed far from its actual location, the stay point will not be identified as a stay point because the estimated median speed will be large according to Equation 2.1. Some stay episodes will be identified as two stay episodes with the wrong staying time.

Apart from the two types of mainstream stop identification methods, some researchers achieved stop identification with supervised machine learning methods. For example, Yang et al. (2014) used a support vector machine (SVM) to achieve high accuracy stop identification of vehicles in an urban scenario. This type of method takes advantage of state-of-the-art machine learning techniques; however, it is not practical because it requires carefully arranged model trainings of the SVM or other classifiers. In the training process, ground-truth data are required, and the training samples and manual truth data collection is expensive and laborious work. Furthermore, the well-trained classifier can only be used to identify the stops of the same group of vehicles in the same case study. Identifying stops of other vehicles or moving objects must be based on yet another ground-truth collection for new data. Since most mobility datasets do not include ground-truth information when collected (Biljecki, 2010), the supervised-learning-based stop identification methods are not commonly used.

### 2.2.2 Detecting regions of interest in Cartesian space

Spatial ROIs are the ROIs generated when only spatial variables, such as longitude and latitude, are used to measure the distance and sparseness between stay points. Some researchers are simply looking at all stay points as ROIs (Zhong, 2014), while most studies choose density, quantity, or spatial closeness of points as the major indicator of ROIs (e.g. Ashbrook & Starner, 2003; Cao, 2010; Ester et al., 1996). Conventionally, as illustrated in Figure 2.6, ROIs are determined as the regions where there is a high-density aggregation of stay points by multiple moving objects in space (Parent, 2013; Ying, 2013). Li et al. (2008), Palma et al. (2008), Cao et al. (2010), and Lee et al. (2013) all defined the place where multiple users stay as their common ROI. Zhao et al. (2011) used

minimum bounding boxes (MBB) to define the highly active region of moving objects' trajectories for detecting ROIs. Yan et al. (2013) identified them using spatial bounding rectangles of the stop episodes or the centre of a group of aggregated stay points. Downs (2010) used kernel density estimation (KDE) to enable visual analysis of a home range or interesting places in animal group activities (Downs et al., 2011). Various clustering methods are also applied to extract the places where users frequently visit and stay over a certain period from trajectory data. Jain (2008) used a variation of k-means clustering.

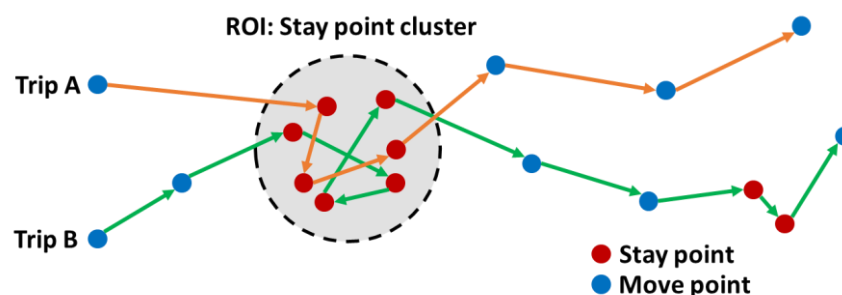


Figure 2.6 A region of interest that attracts multiple persons' visits

DBSCAN (Density Based Spatial Clustering of Applications with Noise) and its variants are the most common methods for ROI detection (Giannotti et al., 2008; Güting et al. 2006; Karlis, 2009; Li et al., 2010; Palma et al., 2008; Parent, 2013). As exemplified in Figure 2.7, DBSCAN functions as follows:

The inputs of basic DBSCAN include  $Eps$  and  $MinPts$ , where  $Eps$  is the neighbourhood searching radius of a selected point  $p$ ,  $MinPts$  is the minimum number of points within the neighbourhood to make  $p$  a core point. If  $p$  is a core point, and point  $q$  is within the  $Eps$  radius of  $p$ , then  $q$  is defined as directly reachable from  $p$ .  $q$  is reachable from  $p$  if there is a path  $p_1, \dots, p_n$  with  $p_1 = p$  and  $p_n = q$ , where each  $p_{i+1}$  is directly reachable from  $p_i$ .

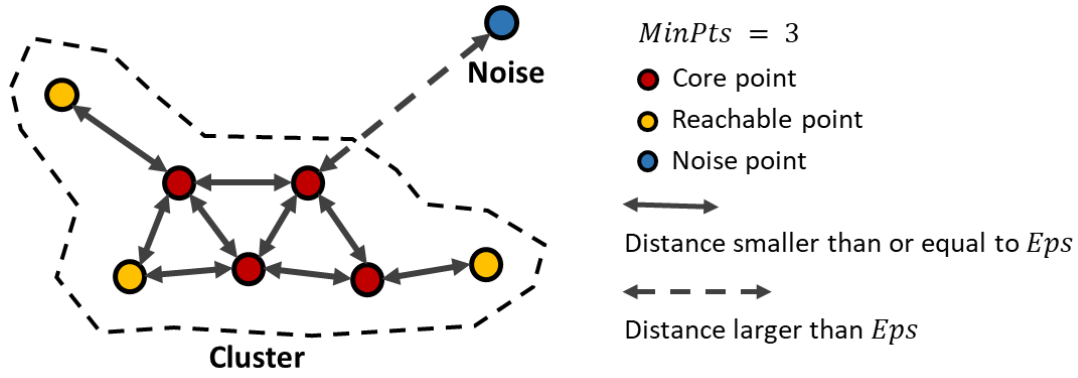


Figure 2.7 An example of DBSCAN clustering result when  $MinPts = 3$

- Step 1: An unlabelled point is picked. Count the number of points within  $MinPts$  of this selected point.
- Step 2: If the selected point is a core point, find all points reachable from the selected point and label these as a new cluster.
- Step 3: If the selected point is not a core point, find another unlabelled point and start from step 1 again.
- Step 4: Repeat the processes above until all points are label as points in clusters or noises.

Li et al. (2008) introduced OPTICS (Ordering Points To Identify the Clustering Structure), a variation of DBSCAN to take advantage of both hierarchical clustering and density-based clustering to exempt the parameterisation process of conventional DBSCAN. The wide adoption of density-based clustering methods in ROI detection is due to its working mechanism enabling them to detect clusters of arbitrary shapes such as linear, concave, oval, etc. Furthermore, in contrast to other conventional clustering algorithms, such as K-means and hierarchical algorithms, density-based clustering methods can work without specifying the number of clusters a priori and avoid the “initiating problem” (Kriegel et al., 2001). DBSCAN also has the ability to process large databases with the help of spatial query trees (Ester, 1996; Ester, 1998; Zhou, 2000). For example, DCPGS-G (Shi et al., 2014) is a direct combination of basic DBSCAN and R-tree. R-tree speeds up the range query centred at  $p$  with  $Eps$  radius to determine the number of points reachable from  $p$ .

### 2.2.3 Detecting regions of interest in space and time

Spatial location alone is not sufficient to depict the reality of interesting places because the meaning of the places and activities that occurred in the places are dependent not only on location, but also on time. For semantic and activity-related analysis, temporal information is of equal importance to location, because interesting places are not always active all day and their meanings vary over time. Therefore, some researchers (Chen et al., 2015; Kwan et al., 2004; Zheng & Zhou, 2011) argue that places and activities should be described with spatial-temporal ontology, and time-geography should be introduced into the traditional ROI studies. Ren and Kwan (2007) proposed the use of information cubes in spatio-temporal activity studies. The information cubes are attached to every stop in trips to contain non-distance-based information such as what activities an individual engaged in during the stops and where and when the person conducted these activities throughout a day. The spatial boundary of an information cube is, as its name suggests, rectangular (see Figure 2.8 (a)). Chen et al. (2015) used space-time path bundling methods to create cylinder-shaped space-time bundles of multiple trajectories to visually represent the common areas and time periods in which multiple meets and interacts in activities. The spatial boundary of a space-time path bundle is circular (see Figure 2.8 (b)).

Apart from the information cubes and space-time path bundles, Shen and Cheng (2016) extended the idea of traditional spatial ROIs, taking the closeness in space and time dimensions into joint consideration. They introduced a similar concept called spatio-temporal ROIs (ST-ROIs), i.e. ROIs that are defined with spatial locations and boundaries as well as start and perish times. In other words, an ST-ROI is a region of high-density clustering of stay points in space and time. The spatial boundary of an ST-ROI is a polygon encircling the stay points in the ST-ROI. Demsar et al. (2015) also used space time kernel density to create the utilisation distribution (UD) surface from animal movement data and then use the peak areas of the UD surface as the so-called home range areas (i.e. the most important regions of interests) of the animals.

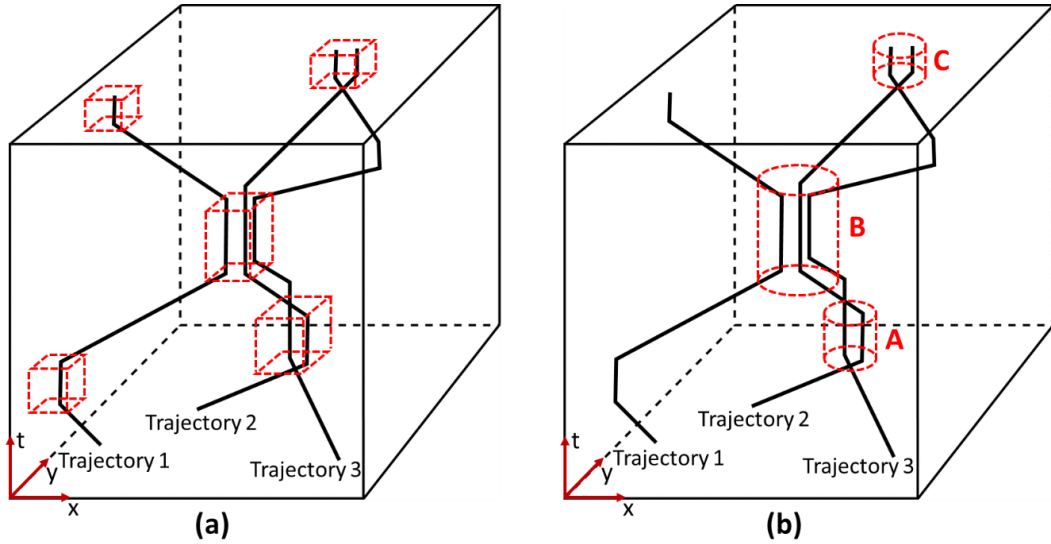


Figure 2.8 Representation of places and activities in a space-time cube: (a) information cubes; (b) space-time path bundles

For the detection of the ST-ROIs, space-time path bundles, and entities with similar concepts, researchers select methods according to their research purposes. Most of these methods are improved and modified based on the methods used for spatial ROI detection by adding temporal operations to their algorithms. Loecher et al. (2009) used scan statistics to look for locations in time and space which are the most likely circular/elliptical/rectangular regions for users to visit subsequently. Webb (2008) applied a similar approach to identify the killing sites of wild wolves tracked by GPS. In addition to discovering areas with high point densities, scan statistics can provide the statistical significance of the generated places. The space-time extension of the conventional KDE method is also used for the same purpose. Demsar et al. (2015) propose a stacked space-time KDE, which adds time as a new dimension, and visualised the meaningful places of multiple types of moving objects in a space-time cube. The space-time trajectory data test by Demsar included pedestrians (McArdle et al., 2013; McArdle et al., 2014), wild animals (Demsar & van Loon, 2013), ships (Demsar et al., 2010), and the focuses of eyes on a screen (Demsar et al., 2015).

Most DBSCAN variations are designed to aggregate point objects only in space, and they ignore the influence of time on the semantic meaning of ROIs. Some researchers also add space-time adaptations to the traditional density-based method to cater to their space-time activities studies. Palma et al. (2008) proposed the CB-SMoT algorithm by replacing the parameter *MinPts* in conventional DBSCAN with *MinTime*, which is a staying time threshold. The difference of the maximum and minimum time value  $t$  of the reachable points from point  $p$  in space must be larger than *MinTime* to make this group of points a valid space-time cluster (i.e. ST-ROI). This algorithm used stay time

as an extra standard to detect ST-ROIs but ignored the number of stay points that indicates the visit intensity of places. It also ignored short term intensity visits and activities in places, so if an ST-ROI attracted a large number of visits in a period shorter than *MinTime*, it was ignored.

Unlike Palma et al. (2008), Birant and Kut (2007) directly introduced time as an independent dimension in their ST-DBSCAN algorithm. ST-DBSCAN is a variation specially developed to handle comprehensively point density in space and time (or other non-spatial dimensions). By counting reachable points within both *spatial Eps* and *temporal Eps*, the ST-DBSCAN can detect point aggregations according to not only the spatial distance, but also the closeness in time. An object point existing in space and time must be simultaneously reachable according to the spatial maximum reachable distance and the temporal maximum reachable interval to be included in a spatio-temporal cluster. In addition to the advantages of ST-DBSCAN inherited from DBSCAN, ST-DBSCAN has features of its own that make it even more effective for detecting ST-ROIs. Shen and Cheng (2015) were the first to applied ST-DBSCAN on the stay points in human movement trajectories to detect spatio-temporal regions of interest (ST-ROIs). The generated ST-ROIs contain information about their spatial boundaries as well as their lifespans, revealing where ST-ROIs are, when they emerge, and when they perish. Zimmermann et al. (2009) designed a time-based OPTICS algorithm to cluster stay points, considering both spatial and temporal properties of a trajectory, and achieved similar results to ST-DBSCAN's without a parameterisation process.

#### 2.2.4 Detecting regions of interest in spatial networks

As contextual data for activity description can now be geocoded to street addresses, new analytical methods that can handle the analysis of activities and movements in network space are needed. On the other hand, most spatial clustering methods use Euclidean distance and ignore the urban context and road network within which most people move. These conventional clustering algorithms cannot correctly identify the true shape of spatial ROIs. Although DBSCAN is purported to be better than other clustering methods in detecting ROIs with arbitrary shapes, the ROIs detected by DBSCAN are still somewhat round-shaped and cannot fully represent the true coverage area in cities with dense streets because the GPS observations themselves have positioning errors and are distributed around the true location in nearly circular areas. Yan et al. (2011) argue that points, regions (or areas), and lines are all standard spatial data types in GIS and that activity study and semantic annotation of interesting places based on movements in



urban networks should be integrated with road structures. In their works, they have proposed the concept of Line of Interest (LOI) so that semantic meanings of places were studied on the street-segment level instead of the approximate areas.

Yiu and Mamoulis (2004) are the first to attempt the adaptation of traditional spatial clustering algorithms to network analysis. In their work, they revised and tested three major types of algorithm (portioning based, density-based, and hierarchical clustering) to generate spatial clusters according to the network distances between points on a road network. To mitigate the increase in computation burdens brought by the network distance calculations, the spatial query algorithms are optimised accordingly. Through comparison, they concluded that the Network Eps-Link, a variation of a density-based clustering algorithm, is the most appropriate for clustering objects on a spatial network in terms of complexity and robustness with noises. However, their work focused solely on innovations of algorithms per se and was not used for ROI detection based on movement data or any other application.

Zhang et al. (2016) also argues that links and segments in the road network enable movement analysis and interesting region detection on a finer scale than Euclidean approaches in highly dynamic cities. They took taxi trajectories in a Chinese city as a case study and used the map-matching technique to accurately associate the trajectories with the streets and conduct all subsequent movement analyses in street networks. The spatial representation of location records is also transformed from  $(x, y, t)$  to  $(Segment\_ID, P\_in\_segment)$ , where *Segment\_ID* is the street the point was snapped to and *P\_in\_segment* described the point's location using the distance from the starting node of the segment to the position of the map-matched point on the segment. They applied a Latent Dirichlet Allocation (LDA) algorithm to calculate the significances of streets to the taxis' pick-up and drop-off activities in their trips. Streets above a defined significance level were labelled as interesting places in the network and semantically analysed. This method enabled the LOI detection in streets but only focused on the starts and ends of trips. Stop episodes within the trips were all ignored.

Buchin et al. (2009), Shine (2007), Oliver et al. (2010), Okabe et al. (2006), and Shi et al. (2014) also proposed network-based spatial clustering algorithms to detect hotspots and point-aggregation patterns in streets. These approaches can take into account graph properties such as edges, connectivity, and directionality while finding hotspots or ROIs, but none have made use of the temporal information or tested with large-scale movement data.

### 2.2.5 Summary

In this section, some of the existing techniques to identify ROIs within a collection of space–time trajectories are reviewed and critiqued. The existing stay point detection methods are first summarised into density-based methods, threshold-based methods, and supervised machine learning methods as a general pre-processing step before ROI detection. Of these methods, supervised machine learning methods suffer from their poor versatility for general use. Density-based methods cannot distinguish stop episodes at different times in returning trips, while threshold-based methods are not robust with regard to large sporadic positioning errors. Strengths of these methods should be combined to overcome their weaknesses.

As for the detection of ROIs, some visual studies applied KDE and its variations, while studies that required statistical significance or prediction capabilities applied scan statistics. Nevertheless, most researchers choose to use DBSCAN or develop their own ROI detection algorithms based on DBSCAN because of their advantages in detecting clusters of arbitrary shapes without knowing the cluster number, robustness to noises and outliers, and intuitively adjustable parameters.

These conventional methods for detecting purely spatial ROIs are evolved into space–time clustering methods to detect ST-ROIs, which takes the advantage of the theoretical advancement in time geography and space–time analytics. As there is no universal model to combine spatial and temporal expressions of ROIs, researchers from different backgrounds have proposed various methods to detect ST-ROIs to solve the different problems they have encountered in case studies. ST-DBSCAN and CB-SMoT, popular clustering approaches for ST-ROI detection, both inherited the advantages of DBSCAN and the ability to deal with large space–time datasets. For detecting LOIs in urban spatial networks, the distance metrics of conventional ROI detection methods are replaced with network distance. Some researchers have even added map matching as a pre-processing measure before LOI detection to increase spatial accuracy for urban environments. The characters of some classical ROI detection algorithms are summarised in Table 2.1.

Table 2.1 Brief summary of classical ROI detection algorithms

	<b>DBSCAN</b>	<b>DCPGS-G</b>	<b>CB-SMoT</b>	<b>ST-DBSCAN</b>	<b>Network Eps-Link</b>
Overview	Uses point density in space for clustering	Grid-based spatial point density clustering	Extension of DBSCAN with a stay	Extension of DBSCAN to use both spatial and temporal	Using network distance in spatial clustering

			time condition	attributes in clustering	
Running time	$O(n^2)$	$O(n^2)$	$O(n^2)$	$O(n^2)$	$O(n^2)$
Spatio-temporal	No	No	Yes	Yes	No
Network awareness	No	No	No	No	Yes
Problems	Noise and information loss	Yes	Yes	Yes	Yes
	Misidentifying ROIs in returning trips	Yes	Yes	Yes	No
					Yes

To sum up, the existing ROI detection approaches have the following limitations:

1. The partitioning-based and hierarchical clustering methods adapted for network analysis in many previous works (Yiu & Mamoulis, 2004) have been proven slow, ineffective, and over-sensitive to noise.
2. Existing network-based clustering methods are purely spatial and are not designed for ST-LOI detection. The temporal dimension is totally ignored in those methods.
3. Existing spatio-temporal clustering methods use Euclidean distance in space to determine reachable points and are not designed for LOI detection. Urban networks and true movement trajectories on streets are totally ignored in those methods.
4. Existing network-based spatial clustering methods are applied and tested on synthesised point data (Yiu & Mamoulis, 2004), points representing independent issues such as crimes (Buchin et al., 2009), and starting and ending points of movements (Zhang et al., 2016). None of them are designed for the analysis of point data generated by consecutive movements. Among these methods, only a few (Zhang et al., 2016) have used map-matching techniques as a pre-processing step to guarantee that the points themselves are precisely located and the speed correctly calculated. None of the existing works have used map-matching and space-time clustering techniques in combination to generate highly accurate ST-LOIs.

Further works may include the development of a spatio-temporal stop identification method to overcome existing limitations and a clustering algorithm that can detect dense aggregation of stay points in spatial networks and temporal domains. The space-time clustering method should also work with map-matching algorithms for better accuracy.

## 2.3 SEMANTIC ENRICHMENT OF PLACES

Identifying “hotspots” in raw trajectories is known as ROI detection, while inferring the meaning of a place or identified ROI for people’s activities with contextual knowledge is called semantic enrichment. Emerging semantic enrichment theories and methodologies, along with new data sources, have enabled the extraction of useful geospatial and semantic information from high-dimensional and heterogeneous datasets. Semantic enrichment provides semantic and “patial” (Jenkins et al., 2016) views beyond traditional spatial perspectives as human activities are more aligned with social places than purely geographic locations.

Semantic enrichment is achieved by associating raw movement trajectories and ROIs with related contextual data describing nearby geographic objects and backgrounds, such as POIs, as suggested in Figure 2.9. Cao et al. (2010) and Li et al. (2008) have introduced models for semantic GPS movement trajectories. They define the trajectories as sequences of stops and movements from place to place with temporal and semantic labels in addition to geographic background. The boundary definition of semantic analysis and algorithms for semantic enrichment are reviewed in this section.

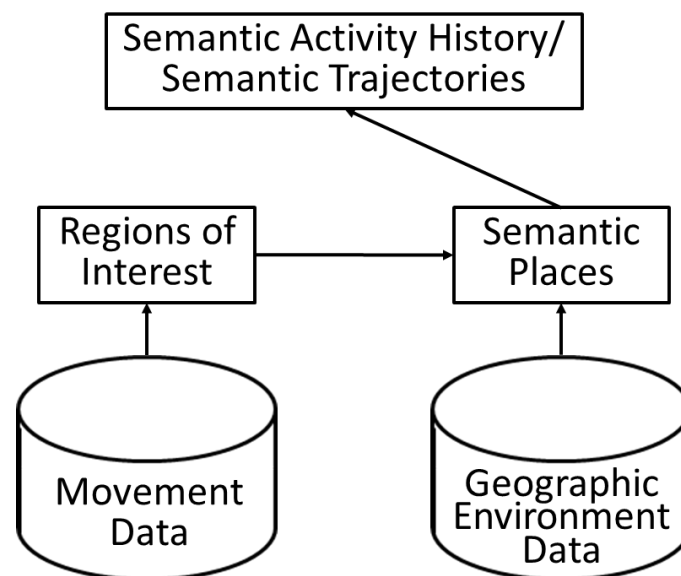


Figure 2.9 The semantic enrichment of places

### 2.3.1 Space-time boundaries of places

Since semantic enrichment methods use information about the geographic environment adjacent to places, “adjacency” must be defined. The social meaning assignment to places derived from social data can be problematic due to vague boundaries (Goodchild & Li, 2012). Thus, the boundary of a place or a ROI needs to be clarified so that the semantic enrichment process can be performed based on the geographic environment within a well-defined area. Many boundary definitions of places are proposed in semantic analysis. Yuan et al. (2012) cut and split their urban study area into multiple blocks with road links of the city and analysed the function of each divided block. They enriched the semantic meaning of these blocks with POI data. The spatial boundaries of the blocks are lines of roads which are actually existing spatial entities. Lew and McKercher (2006) used grids as basic spatial units. Yan et al. (2011) used bounding rectangles in semantic enrichment. The spatial boundaries of Ren and Kwan’s (2007) information cubes are rectangular areas, while Chen et al.’s (2015) space-time path bundles are circular. Jahnke et al. (2010) limited semantic enrichment inside the circular buffer zones of ending locations of trips. Monajemi (2013), Polisciuc et al. (2015), and Shen and Cheng (2017) used bounding convex hulls of stay point clusters to define the polygon areas of ROIs.

Time geography has added time constraints to human activities, evolving hotspots from spatial entities to spatio-temporal entities (Shoval & Isaacson, 2009). Time span is added as the temporal constraint to a place because activities cannot last forever in the place, and different functions and meanings can be found in the same place at different times. This representation of semantic places embraces the spatio-temporal ontology of places and activities. For example, Both Chen et al.’s (2015) space-time path bundle and Shen and Cheng’s (2016) ST-ROI have temporal boundaries of activity places that can last several hours.

### 2.3.2 Semantic enrichment based on contextual data

Luo and MacEachren (2014) argue that the First Law of Geography can be extended from space to social and semantic concepts. By taking contextual environment as part of human dynamics, we can model the interactions between social and physical spaces in space and time with various granularity to discover the meaning of activities, thus bridging the gap between spatial movement analysis and social activities studies. A typical example of this approach is the idea of “How they move reveals what is happening”

proposed by Mazimpaka and Timpf (2017). To extract high-level activity information from raw trajectories with only geographic features, the semantic features of places along those trajectories should be enriched to explain the behaviours of the trip. After detecting ROIs or hotspots and determining their boundaries, the next step for activity study is to add semantic information to them. Semantic enrichment is a challenging task because of the irregularity of various contextual data sources. Generally speaking, contextual environmental data have two types: officially collected datasets and crowdsourced datasets. Officially collected contextual data include POI, land use, and building function registry data that are strictly collected by administrative or business organisations according to industry standards (Krüger et al., 2015). This type of data was originally used in user-centred navigation applications to answer questions such as “What is this place?” Crowdsourcing data include geo-tagged tweets, geo-tagged photographs, open-source map applications, and e-business transactions generated in location-based services. Some of the better-organised crowdsourcing data that are contributed, managed, and freely available are called volunteered geographic information (VGI) data (Haklay, 2010). Andrienko et al. (2013) semantically annotated places based on GPS, GSM (cell phone), and Twitter data. Krueger et al. (2015) associated Foursquare data with places to automatically extract activity information about the latter. Zhang et al. (2012) present another example of semantic enrichment in which information of nearby building functions was used. Yan et al. (2013) used well-defined land use data to annotate the meanings of places and road segments. Tardy et al. (2016), Jenkins et al. (2016), and Cai et al. (2016) analysed semantic places with VGI data, such as geo-tags of photos uploaded by social media users. POI data were the most widely used contextual data source for this purpose. Jahnke et al. (2010), Braun et al. (2010), Yuan et al. (2004), Niu et al. (2017), Siła-Nowicka et al. (2016) and Shen and Cheng (2016) used urban POI data to extract meaningful information about activities from ROIs. Krueger et al. (2013) also used POI for the same purpose, making a simple assumption that the meaning of a place is determined by the dominant type of POIs in the region. To transform the physical location records of individual trips into semantic activity histories meaningful to people and society, researchers (Palma, 2008) used the nascent concept of semantic trajectories, in which background geographic information is integrated with points in trajectory. In this new concept, a trajectory is observed as a sequence of visiting behaviours to various ROIs whose semantic meanings were semantically enriched.

As to the annotation methods, Jahnke et al. (2010) manually annotated the semantic meaning of places where POIs are sparsely distributed. Jenkins et al. (2016) uses statistical significance to quantify semantics in Manhattan. Nishida (2014) uses supervised machine learning approaches to use labelled semantic places to predict the

unlabelled functional buildings a person was in. These methods add the meaning of buildings into movement trajectories and can be used to enrich semantic meanings of location histories. In accordance with Waldo Tobler’s (1970) First Law of Geography, which proclaims that “everything is related to everything else, but near things are more related than distant things”, POIs and buildings of identical functions are likely to aggregate in the same neighbourhood. Based on this phenomenon, Krüger et al. (2013) and Polisciuc et al. (2015) used the quantity of POIs in a place to explain its major semantic meanings in a simple manner. The semantic significance of one type of POI to the place it falls in can be expressed as Equation 2.2.

$$SC_{I,j} = f_{I,j} = \frac{count_{I,j}}{\sum_k count_{k,j}} \quad \text{Equation 2.2}$$

where  $count_{I,j}$  is the number of category  $I$  POIs in place  $j$  and  $\sum_k count_{k,j}$  is the sum of all categories of POIs.  $f_{I,j}$  is called the emergence frequency of a type of POI.

Furthermore, Krüger et al. (2013) added the influence of distance between the POIs and the location of people’s activities into consideration as an improvement to his semantic enrichment approach and compared the suitability of the POI data provided by Foursquare, Facebook, and Google for semantic analysis. In their work, POIs closer to the individual’s locations are considered to influence more significantly the person’s activity than other POIs in the area. Damiani et al. (2011) also weighted POIs differently for protection of users’ privacy in sensitive stops. Since POIs near users’ locations may have different degrees of sensitivity according to their semantic meanings, lowering the significance of certain types of sensitive POIs can keep users’ sensitive behaviours from being compromised. For instance, stopping at normal restaurants is considered less sensitive than being in a hospital, so restaurants and hospitals should be given different weights. Notably, all land use and POI data used for semantic analysis in the works above are hierarchically classified into multiple categories and subcategories so that the semantic meaning can be summarised in different levels of details.

Nevertheless, quantity and distance are not the only indicators of semantics, and the numbers of different categories of POIs are heavily biased. For example, convenience stores are far more densely distributed than airports in a city, and relying solely on absolute POI quantity will cause mistakes. This problem is similar to understanding the meaning of an article in which auxiliary words like “the” far outnumber the proper nouns that contribute significant meaning. Therefore, a few topic-modelling algorithms of text mining have been used in geographic semantic studies. Topic-modelling algorithms are adapted for semantic enrichment of places to find hidden information from the biased

quantities of different POIs. Topic-modelling algorithms can overcome the drawbacks of traditional statistical information retrieval methods and can annotate places automatically with massive POI datasets. Yuan et al. (2012) and Shen and Cheng (2017) have used term frequency-inverse document frequency (TF-IDF) algorithms to mitigate the semantic significance bias caused by the POI quantity bias. Zhang et al. (2016), Sizov (2010), Chon et al. (2012), and Yuan et al. (2012) used the more advanced Latent Dirichlet allocation (LDA) on POI data and crowdsourced data. Yuan et al. (2012) compared the performance of LDA and TF-IDF in their case study of Beijing and concluded that LDA is a more suitable method for semantic enrichment.

Because of the highly dynamic nature of urban activities, the semantic meaning and the human perception of a place are constantly changing as places are refilled with new activities and different people over time (Batty et al., 1999). Therefore, temporal information is also used, in addition to spatial context, for semantic enrichment since most trajectories contain temporal records. Liao et al. (2006) proposed that different activities have different temporal durations and temporal patterns, which can be used to distinguish activities near multiple POIs. Andrienko et al. (2013) have suggested interpreting semantic meanings of places based on cyclic temporal patterns of visiting times. Reumers et al. (2013) designed a classification tree to identify semantic places which relies purely on temporal stop durations.

## **2.4 SIMILARITY AND AGGREGATION OF ACTIVITY PROFILES**

As suggested by Chapin (1974), it is necessary to simplify and combine similar activity patterns into more general categories to explore activities and population aggregations. For knowledge-discovery purposes, many grouping and aggregative clustering methods are used to discover patterns in activities and movements. This section examines the literature that explains the various definitions of activity similarities and the methodology for aggregative clustering and grouping analysis. Exploring similarities among different individuals' trips, where trip similarity can be defined regarding spatial, temporal, and/or semantic dimensions, and applying data mining methods to these similarities have potential to reveal patterned characteristics in activities and identify who the people are (Tsou, 2015).

Before people are grouped, clustered, or classified with different activity patterns, the metrics of similarity or dissimilarity between them should be defined. The similarities of moving objects or trajectories represent a fairly new topic in spatio-temporal data mining. The metrics used to define similarity vary for different applications and research



purposes. Therefore, the features used for calculating similarity can range from physical ones, such as the speed and geometric shape of routes, to semantic ones, such as information regarding places visited or activities undertaken. Three major categories of activity similarity metrics are discussed in the rest of the section.

#### 2.4.1 Similarity metrics based on physical features

Most of the traditional methods proposed seek similarities in the geometric shapes of trajectories based on a defined distance function. Generally, p-norm distances are used as a spatial similarity measure. The Euclidean distance between two location sequences is a special case of a p-norm when  $p=2$ . Many studies have extended and applied this metric (Agrawal, 1993; Goldin, 1995; Johnson, 1996; Kahveci, 2001; Keogh, 2001). In these studies, the dissimilarity of two sequences,  $\langle w_1, w_2, \dots, w_n \rangle$  and  $\langle w'_1, w'_2, \dots, w'_n \rangle$  is defined as

$$D(c, c') = \sqrt{(w_1 - w'_1)^2 + \dots + (w_n - w'_n)^2} \quad \text{Equation 2.3}$$

where  $w$  is the coordinates of sequential updates, which can include longitude, latitude, time, and other user-defined variables.

However, trajectories have not only time series, but also other information such as directional variables. For this reason, Yanagisawa (2003) modified the method to perform the similarity query based on the time interval between updates. Lin (2008) defined the one-way distance (OWD) to improve the precision and efficiency of shape-based trajectory matching to find trajectories of similar shapes. Lee (2000) extended the Euclidean distance to define a distance metric between two series of minimum bounding rectangles (MBRs) encompassing two spatio-temporal sequences. This method achieved a very high efficiency and successfully dealt with sequences of different lengths. However, it is not reliable against noise and time shifting, which usually exist in real movement data. Cai (2004) used a similar dissimilarity definition as Lee (2000) and inherited similar defects. Ranacher and Tzavella (2014) and Demsar et al. (2015) reviewed multiple classical definitions of spatial and temporal distances to measure the similarity between trajectories including Euclidean distance, Minkowski distance, Hausdorff distance, earth mover's distance, relative direction and qualitative trajectory calculus. To summarise, shape-based approaches are sensitive to outliers and require some ideal

prerequisites (such as equal lengths of tracks and intervals between nodes) that are uncommon in real data.

Studies have also focused on the movements of different objects by depicting their moving behaviours using other dynamic physical features including speed, acceleration, duration of movement, sinuosity, travelled path, displacement, direction, etc. (Qu, 1998). These techniques often aim to predict transportation modes and to distinguish objects with different moving features based on the previously learned sample features. Schüssler and Axhausen (2009) introduced a fuzzy-logic method based on speed and acceleration to distinguish five transportation modes: walking, cycling, car, bus, and rail. However, these methods do not satisfactorily cope with journeys with similar speed ranges. Dodge et al. (2009) extracted multiple movement parameters to describe movements, including speed and angle, to generate description profiles and analyse them comprehensively. In studies by Dodge et al. (2008), Giannotti and Pedreschi (2007) and Laube, et al. (2007), researchers introduced the parameters of a trajectory generated by a moving object, such as speed, acceleration, duration of movement, sinuosity, travelled path, displacement, and direction. These descriptors form fundamental building blocks for characterising the movement of an object, and they can be defined in an absolute sense (i.e. with respect to the external reference system) or in a relative sense. Bolbol et al. (2012) chose speed and acceleration as the descriptive features and tested their suitability with analysis of variance (ANOVA).

The main idea of these methodologies is to use an algorithm that calculates the track similarity based on the profiles generated from the physical variables of movements to distinguish types of moving objects and aggregate similar movements. Almost all the methods proposed primarily considered motion details, and they do not promise the ability to cope with the behaviours and patterns of higher levels. Some of the above methods cannot perform well on data with close speed or angle ranges. Therefore, there is still a need for approaches that incorporate locational information and semantic meanings in the environment.

#### 2.4.2 Similarity metrics based on travelling sequences

In terms of similarity based on the sequences of places visited in personal trips, it is assumed that users usually stop at places for specific objectives. Different social groups may have different preferences and habits that may lead to dissimilarities in their movement patterns and reactions to certain events (Chapin, 1974). Because trajectories

are generated by people moving from time to time, many methods that have been used for the analysis of time-series data with sequential attributes have also been used on movement trajectories in human activity studies.

The most commonly used similarity metric is that which considers the sequences of the visited ROIs. Suppose Trajectories 1, 2, and 3 in Figure 2.8 (b) are generated by trips of three different persons and that the individuals in both Trajectory 2 and 3 included three ROIs in the sequence of ROI (A)→ROI (B)→ROI (C), while that of Trajectory 1 visits only visited ROI (B). As a result, users 2 and 3 are considered to be more alike than user 1 in accordance with their sequential similarity. Based on this general idea, different methods, such as Longest Common Subsequence (LCS) (Dodge et al., 2009), Multiple Sequence Alignment (Kwan et al., 2014), Edit Distance (Chen et al., 2005), and trajectory clustering (Nara et al., 2011) have been used for measuring similarity in terms of sequence relationships.

Little and Gu (2001) used path and speed curve changing in time as profiles to measure the dissimilarity between two trajectories using dynamic time warping (DTW), which is often used in matching wave forms. Vlachos (2004) used DTW on a rotation invariant to compare sequences of angle and arc-length pairs. However, DTW cannot match geometrically similar tracks with gaps during the trips because this method requires continuity along the warping path. Gaps, the sub-trajectories between similar sections of two trajectories, weaken DTW's robustness to noises. Unlike weather, customer flows, or other ordinary spatial or temporal data, trajectory data are collected by devices with positioning errors and a higher probability of functional failures. This makes noises and outlier records especially common obstacles in trajectory analysis. Longest common subsequences (LCSs), which are robust to noise, were presented by Vlachos (Dodge, 2009; Vlachos, 2002) to handle the defects of the above-mentioned approaches. Yet this method cannot accurately deal with sub-sequences of similar shapes with dissimilar gaps and different trip lengths. Bozkaya (1997) modified LCSs into an extended metric that measures similarity that can be used for clustering analysis. Edit distance on real sequence (EDR) (Chen et al., 2005) is a novel dissimilarity metric introduced to remedy the defects of LCSs. The trajectories are first normalised to remove the spatial shiftings of trajectories in different places. Edit distance (Ukkonen, 1983) was originally used in string and text matching in accordance with the minimum number of edit operations required to change one string to the other. This enables the EDR to tolerate the negative effects of time shifting gaps within sub-trajectories. Noises are also eliminated by quantifying the distance between each pair of elements in the two sequences being compared.

Many studies have used multiple sequence alignment methods, which are traditionally used in gene sequence analysis in bioinformatics, to match the sequences of GPS, Wi-Fi, and Bluetooth log-in movements (Delafontaine, 2012; Shoval, 2007; Shoval, 2009). Biochemists used to grapple with the problem of analysing protein and DNA sequences. Social scientists faced a similar challenge when studying sequences of events. Sankoff and Kruskal (1983) published a groundbreaking work in which they set forth a series of basic algorithms capable of efficiently analysing complete sequences. Their work, which formed the basis of most subsequent sequence analysis algorithms, was instrumental in the eventual breakthroughs made in DNA and protein analysis. It took about ten years before the sequence analysis methods used in the natural sciences were ported to the social sciences. Abbott (1995) used sequencing algorithms to analyse socioeconomic data in investigations of the progress of musicians' career histories. Later, in the field of travel research, Bargeman et al. (2002) aligned information describing vacation patterns. Transport and time allocation research have benefited from work performed by Wilson (2001; 2006), Joh et al. (2002), and Joh et al. (2001; 2005) that aligned patterns of activity.

There are generally two types of activity sequence analysis. The more commonly used product is utilised to generate groups based on their overall activity patterns. Programs using this kind of sequence analysis produce "trees", which divide sequences taxonomically. The second type of sequence analysis, less frequently employed, is used to match and detect patterns of behaviour in some or all of the sequences scrutinised (Wilson, 1999). The first type of utilisation is more relevant to our research. Early studies (Wilson, 1999) of sequence alignments, including "Activity Settings, Sequencing, and Measurement of Time Allocation Patterns", were based on software called ClustalG. ClustalG is a general version of the original biochemistry-oriented ClustalX program with an extended alphabet that enables it to cope with a wider range of more complex human activities using adjustable setting and parameters. Recently, with the increasing utilisation of sequence alignments in geoinformatics, ClustalXY (Shoval, 2009) was developed based on the Clustal software series to deal explicitly with the alignment of spatial data. This method allows for the alignment of activity regions for each participant in the research, as demonstrated in Figure 2.10, so that similar sequences can be found and grouped. For example, Person 1 and Person 4 in Figure 2.10 shares the same sequence and timing of movements and have the highest similarity between each other. In contrast, Person 3 never visits place B and C. Instead, he/she visited E twice while Person 1 only visited E once before D. Therefore, Person 3 and Person 1 are considered relatively different.

Person 1	C	D	A	A	E	D	B	B	A
Person 2	A	D	A	A	E	B	B	A	A
Person 3	A	D	A	A	E	D	E	E	A
Person 4	C	D	A	A	E	D	B	B	A
Person 5	C	D	A	A	E	D	D	C	C


Aligned ROI visiting sequence  


Figure 2.10 Similarities and differences in ROI visiting sequences of people

#### 2.4.3 Similarity metrics based on semantic location histories

All movements of people and animals are undertaken in geographic contexts which directly or indirectly influence these movements. The analysis reviewed in the last two sections for trajectory similarities have so far ignored the context, which severely limits their applicability. Similarity metrics based on semantic location histories are generated according to patterns of visits to semantic ROIs semantically enriched with contextual information that already exists in given databases. From a spatial point of view, the concept of “where you stop is who you are”, proposed by Spinsanti et al. (2010), posits that individuals’ activities are associated with semantic places. Zhong et al. (2015) proposed “you are where you go” with the similar idea. Progress has been made in defining movement patterns with series of semantic locations according to users’ travel sequences, which can be used to group users’ activity profiles (Li et al., 2008; Mckenzie, 2014; Xiao et al., 2010). Xiang (2011) used the statistical summary of the port-visiting history of ships as a similarity metric to determine the similarity paths and association rules of ships. Buchin et al. (2012) proposed that context should be integrated into the similarity analysis of hurricane movement data. By taking into contextual information, they were able to distinguish hurricanes that were spatially close but influenced the surroundings differently. Li et al. (2008) presented a moving behaviour-modelling framework called a hierarchical-graph-based similarity measurement (HGSM), which takes both sequential and hierarchical properties into account. In this multi-layered framework with various scales in each layer, the similarity of two moving persons in a single layer is first formulated as Equation 2.4.

$$S_l = \frac{1}{N_1 * N_2} \sum_{i=1}^n 2^{m-1} \sum_{i=1}^m (k_i, k_i') \quad \text{Equation 2.4}$$

where  $N_1$  and  $N_2$  denote the number of different regions visited by the two individuals, respectively, and  $m$  is the total number of places visited in each person's trip.

Next, this simple equation is extended to define the similarity across multiple layers:

$$S_{overall} = \sum_{l=1}^H \beta_l S_l \quad \text{Equation 2.5}$$

where  $H$  is the total number of layers in the model,  $\beta_l$  is equal to  $2^{l-1}$  and represents the weight attached to each layer (the lower the layer, the higher the resolution it has and the higher the weight is) (Doherty, 2006).

Based on the pattern in which individuals stop at a series of semantically enriched places, various similarity metrics have been proposed that emphasise different features of the movements. Similarity of movement patterns was defined in earlier studies by commonly visited places. A typical expression of place-based similarity between “GeoLife2.0” Users 1 and 2, as proposed by Zheng et al. (2009), is calculated as follows:

$$SIM_{user}(1,2) = \frac{\sum_{p \in ROIS_{1,2}} \frac{1}{F_p}}{\sqrt{\left(\sum_{p \in ROIS_1} \frac{1}{F_p}\right) * \left(\sum_{p \in ROIS_2} \frac{1}{F_p}\right)}} \quad \text{Equation 2.6}$$

Here,  $ROIS_{1,2}$  is the set of places visited by both Users 1 and 2, while  $ROIS_1$  and  $ROIS_2$  represent the sets of places visited by Users 1 and 2, respectively.  $F_p$  is the popularity index of these places and is calculated according to the number of people that have been there. The popularity indices are used as the denominator in the weight attached to different places. Weighting places according to popularity decreased the similarity generated by the case of two users going to a common place that is visited by many other users. In such a case, the impact on the behavioural similarity of this place should be smaller than that of places that have been visited by Users 1 and 2, but which are not usually visited by other people.

Other research also added temporal information into semantic profiling of activities. The concept of “what you are is when you are”, proposed by Ye et al. (2011), uses temporal activeness profiles to define the similarity between check-in activities in location-based social networks. Such temporal profiles have also been applied to quantify the description of human mobility and for behaviour similarity analysis (Andrienko et al., 2015; Jankowski et al., 2010; Vazquez-Prokopec et al., 2013). The pioneering work of Chapin (1974) introduced the description of patterned activities by allocation of “time budget” and surveyed how people from different socio-economic backgrounds spend

their time in different places carrying out activities. The assumption is that the more time is spent in a place, the more important the place is for the moving person. A similar assumption is used in the studies of Zheng et al. (2011) and Shen and Cheng (2016). Following Chapin's idea, Sideridis et al. (2015) used people's dwelling time allocation on various types of semantic places to depict individual activity differences in a quantitative manner. Shen and Cheng (2016) defined these time allocations as individual space-time profiles, applying Jensen-Shannon Divergence (JSD) as the similarity metric to cluster people sharing similar profiles. Sizov (2010) also applied JSD to define similarity. Yan (2013) compared the stopping and moving time distribution of vehicle trajectories in different types of semantic places to show the behavioural differences of cars, buses, taxis, and trucks.

#### 2.4.4 Profiling and aggregative analysis

Clustering methods are the most popular machine learning approaches for grouping similar objects. For real-world patterned activity of people, relatively few clustering methods have been used to obtain knowledge from moving trajectories, as these datasets are still quite new to researchers. However, the applications of clustering methods on Internet page browsing behaviours and other time-sequence data are well developed. This provides us inspiration for applying them to movement behaviours in the real world.

Clustering algorithms with different rationales have been used to group time-sequence data. Nasraoui (2000) used fuzzy clustering to extract web users' group profiles based on a series of online behaviours recorded in web log data. Ozer (2011) also used fuzzy clustering to identify homogenous groups of potential users possessing different attitudes, interests, and opinions about the service and computers so that companies can customize strategies for each group. Xie (2001) proposed a belief function for web user clustering, while Xu (2005) used k-means clustering to achieve a similar objective. Nanni (2006) used OPTICS clustering on the trajectories of moving objects, achieving superior stability and better segmentation of results than k-means and hierarchical methods. Sims (2009) used Jensen-Shannon divergence and hierarchical clustering (Ward et al., 1963) to match the gene sequences of animals and replace the traditionally used sequence alignment methods to achieve promising results. Shen and Cheng (2016) extended this idea to the clustering of space-time activity profiles. Because hierarchical clustering uses similarity matrices as inputs, researchers can define their own distance

or similarity metrics to generate a similarity matrix for hierarchical clustering according to their research purposes.

#### 2.4.5 Summary

The selection and definition of similarity serves the demand of a given research purpose. Studies seeking to identify similar routes usually use geometric shapes and geographical proximity as their similarity metrics, while studies trying to find sequential patterns in individual trips often use sequence or temporal closeness to calculate similarity. Studies focusing on user preferences in multiple trips, instead of one particular trip, consider dwelling time within commonly visited places and the semantic meaning of the places indicative of similarity of activity patterns.

Review of various similarity metrics reveals that similarity based on physical features cannot be used to discover high-level activity and behavioural information. In sequence-based activity similarity analysis, the time information of when events happen is lost. Moreover, because places need to be represented by characters in the alignment process, the time and spatial scales must be discretised, and the resolution of the discretisation becomes an inevitable problem in sequence alignments.

As a given activity or visit to an activity location is theoretically linked to the exposure or accessibility and space-time budget at the previous or subsequent activity location as well as the meaning of the location (Thierry et al., 2013), defining similarity based on semantic location histories is the most appropriate method for aggregative analysis of activities.

### 2.5 SPATIAL ANALYSIS TOOLKIT

This section reviews the universal analytical tools used in most spatial and spatio-temporal operations. They include map-matching, spatial-query, and geo-visualisation techniques.



### 2.5.1 Map matching

Map matching is a fundamental pre-process solution for many trajectory-based applications, such as moving-object management, traffic-flow analysis, and driving directions. Map matching is the process of assigning every location observation point to its corresponding network segment in a given network on digital maps. Methods that use map matching apply it for smoothing GPS data, decreasing positional error and thus reducing distance and speed errors. Therefore, map matching is also advantageous as a for increasing the accuracy of modal classification and significant stop identification.

There are basically two classes of map matching: local and global. Local methods usually measure orientation similarity or distance similarity to snap each movement point onto the most likely street edges segment by segment (Chawathe, 2007; Greenfeld, 2002), while global methods match the entire trajectory with the road network according to the proposed probability to match all the observations in one trajectory with all street segment candidates (Alt et al., 2003; Yin & Wolfson, 2004). However, most methods cannot avoid degradation of accuracy caused by low-sampling-rate GPS data, nor do they consider time an important indicator in identifying routes. Lou et al. (2009) devised the ST-matching algorithm that incorporates the advantages of global matching methods and can easily be localised to achieve high efficiency. Time and speed are also incorporated to account for the likelihood calculation of ST-matching, which improves the method's performance on low-sampling-rate GPS observations.

Most ROI detection and semantic enrichment methods are based on human movement in cities, where huge amounts of movement trajectories and POI/building function data are generated and collected. However, in the problem setting of these studies, no researchers have taken into account the influence of road network structures on the movement of humans, nor do they consider that the semantic meaning of a place can change with the time of day. In reality, urban canyon effects (Misra & Enge, 2006) usually cause larger position errors (i.e. displacements) in location data sets such as GPS data, and pedestrians and drivers navigate along streets rather than moving in a straight line to their destinations. Moreover, accessibility of most spatial objects in cities is constrained by spatial networks. Knowing the precise route that people have taken can provide more credible movement details than methods based purely on discrete raw location data. It is therefore realistic to define the distance between objects by their network distance rather than Euclidean distance and apply map matching before spatial operations in urban activity analysis.

### 2.5.2 Spatial query and spatial network query

Speedy searches for points in space must be based on appropriate structural organisation and indexing methods of the stored point data. Massive neighbourhood and range searches in the space-time clustering process of our framework are in particular need of suitable indexing methods. Spatial query trees such as Quadtree (Samet, 1990), R-tree (Guttman, 1984) and K-d tree (Ooi, 1987) are the most common indexing techniques. Of these, the K-d tree is relatively easy to implement in memory. The K-d tree is a binary tree in which nodes are k-dimensional points and represent separating planes. The hyperplane binarily splits the data space. Points on the left side of the hyperplane are represented by the left branch from the current node and those on the right by the right branch. In our study, we apply the K-d tree indexing method for the spatial range query in our ST-network DBSCAN clustering.

The network distance between two objects is defined by the distance of the shortest path from one object to the other over the network. Such distance acts as the spatial-closeness indicator in the map-matching process of network-based spatial clustering algorithms. Nevertheless, replacing the commonly used Euclidean distance with network distance in clustering methods involves shortest-route computations, resulting in a much higher complexity and inconstant cost for computation. To alleviate this drawback, researchers have focused extensively on algorithms for fast network-distance calculation and spatial network query over decades. Dijkstra (Cormen et al., 2001) is the basic technique for shortest-path computation in a graph or network structure. Dijkstra starts the search from a source node and lists its adjacent nodes in a priority queue according to their distance from the source node. The rationale of this method inspired many subsequent heuristic shortest-path algorithms. A more efficient option for short-path searches is the A\* algorithm (Hart et al., 1968), which guides the query towards the destination with a heuristic function and demonstrates a 40–60% saving of computational cost in a medium-scale network (Fu et al., 2006) over the Dijkstra algorithm. Other classical search strategies in practical application include hierarchical search (Jing et al., 1996), search decomposition (Dillenburg & Nelson, 1995), etc.

### 2.5.3 Visualisation analysis of activities

Visualisation is an exploratory process of creating graphical representations of data to improve human understanding. Geo-visualisation, on the other hand, is a powerful tool for analysing large and complex activity and movement data with human visual abilities.

In early studies, 2D geo-visualisation methods were used to portray patterned human activities (Chapin, 1974). These methods provided purely spatial analytical environments and had difficulty expressing non-distance-based attributes of activities such as time, sequence, semantic meaning, and population. For example, location point updates recorded at different times will overlap in a 2D scatter map (Figure 2.11[a]), causing ambiguity and mass loss of information. With the advancement of time-geography came the concept of space-time cubes or space-time aquariums. These developments indicate that GIS-based geo-visualisation has considerable potential for presenting the space-time analytical results for better and well-rounded understanding of activities without losing non-spatial information.

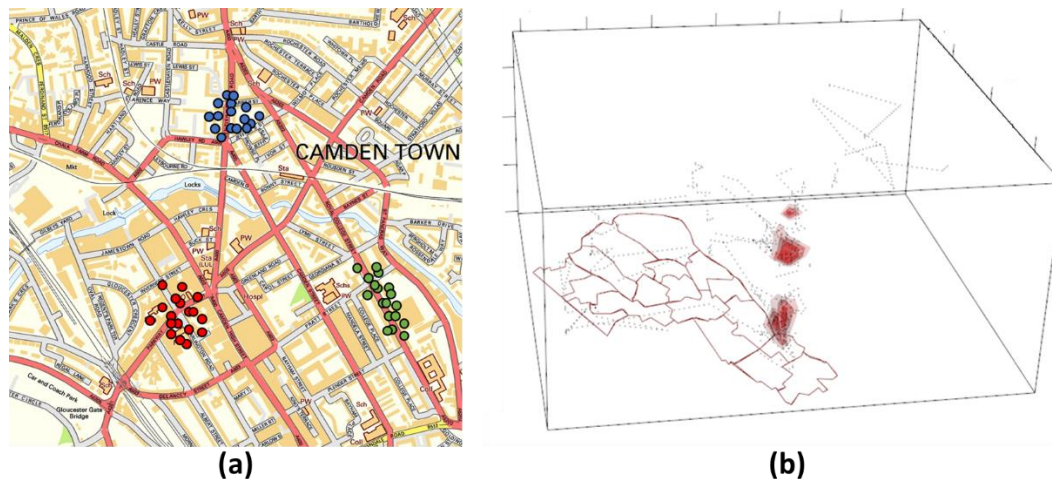


Figure 2.11 (a) 2D scatter plot of regions of interest; (b) Space-time hotspots in space-time cube

Unlike methods which attempt to present data by reducing their dimensionality, 3D geo-visualisation can preserve the original data's complexity to the extent that humans can still visually comprehend it. Kwan (2004) implemented 3D visualisation of space-time paths in space-time cubes (Hägerstrand, 1970). Additionally, Ren and Kwan (2007) represented activities by adding information cubes (Figure 2.8 [a]) onto the space-time paths. Demsar and van Loon (2013) rendered KDE values as volumes in a space-time cube to depict home range and in-home duration of wild animals. Lukasczyk et al. (2015) visualised point-based hotspots in both spatial and temporal dimensions to keep track of the relationship between hotspots over time, as shown in Figure 2.11 (b). These methods enabled the visualisation of time-varying attributes; however, they lacked the ability to precisely present the data within the true spatial structure of activities happening in urban streets.

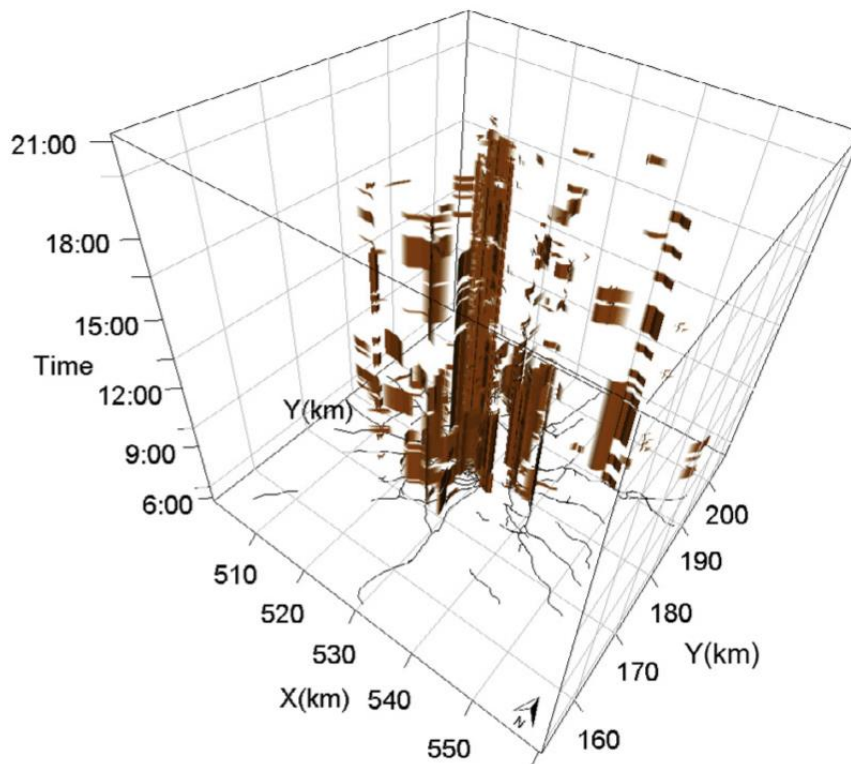


Figure 2.12 3D wall map visualising congested road links in space and time (Cheng et al., 2013)

The combination of space – time visualisation and urban networks is achieved by a method called the 3D wall map. The 3D wall map is made by adding a time dimension into a 2D network link map (Becker et al., 1995). This method was first employed to demonstrate time-varying vehicle counts (ITO World Blog, 2009) on highway networks and travel time (Cheng et al., 2010) for vehicles to pass through multiple road links. The latest progress is to apply a 3D wall map for visualising kernel density interpolation values to show hotspots of traffic congestion on road links with congested time spans (Cheng et al., 2013). As described in Figure 2.12, the 3D wall map shows exactly which road links were congested, when the congestions happened, and how long they lasted. Tominski et al. (2012) were the first to display the physical attributes of movement trajectories with 3D wall maps. These maps generate far more complex and realistic representations of the urban data than conventional 3D visualisations with Cartesian spatial representations.

## 2.6 CHAPTER SUMMARY

This chapter has sought to outline the origin of activity pattern studies in the urban realm and state-of-the-art research methods based on advancing GIS technology that serve the same purpose. It has focused on the techniques and theories used in the geospatial process, semantic process, and knowledge-discovery process of urban activity patterns analysis.

The literature review was undertaken in five sections. The first section reviewed and summarised the overall framework of location-based activity pattern analysis. Theories proposed by previous researchers for the geospatial, semantic, and knowledge-discovery processes of the general framework were critically reviewed and summarised to provide insight for a new analytical methodological framework. The following three sections reviewed methods and approaches for each of the process in the framework. The second section examined the progress of ROI discovery methods in the geospatial analysis process in view of previous research. The review started from the common definition of activity stay points and extended the discussion from purely spatial ROI discovery to spatio-temporal and network-based ROI step by step. The third section described existing approaches of semantic processes and introduced topic-modelling algorithms into the semantic enrichment of places. The knowledge-discovery process is reviewed in Section 2.4 by summarising and comparing different similarity metrics of the features in trips and trajectories for profiling and aggregative analysis of activity patterns. Section 2.5 reviewed the basic analytical toolkits necessary for the methods in all previous sections.

The literature review has shown the variety of approaches proposed as the solutions for the three processes in the framework. Yet, in spite of the wide range of research of activity patterns based on location data, limitations and defects can still be found unresolved in existing studies. These defects can be summarised into the following five problems.

- Problem 1. Spatio-Temporal stop identification: As demonstrated in Figure 2.5, conventional stop identification methods used spatial or temporal information only and cannot correctly identify stops in some very common cases.
- Problem 2. Spatio-Temporal ROI detection in urban networks: There is no method to simultaneously discover ROI in time and spatial networks. Existing ROI detection algorithms are either network-based clustering methods that ignore time or space-time ROI detection methods that ignore the structure of street networks.

- Problem 3. Semantic enrichment of places: No existing semantic enrichment method regards the semantic meanings of places as time-varying features, which fails to reflect the highly dynamic nature of cities and urban activities.
- Problem 4. Profiling of activity patterns: Through the use of multivariate clustering methods, complex activity patterns can be represented by the chosen similarity metric and organized into a relatively small number of homogenous groups. However, no existing approach has incorporated activity-time budget allocation, semantic meaning, and sequential factors in a single similarity metric for aggregative analysis.
- Problem 5. Adaptation to urban networks: The entire framework needs to be practiced in an urban network environment to truly reflect the space where trips and activities take place. No previous research has focused on activities at this scale in its entire analytical processes. There is also no previous visualisation technique to present semantically ROI and results of activity studies in spatial network and time.

The next chapter will introduce a methodological framework to incorporate the newly proposed methods and make improvements based on the current state of the theories and approaches in the literature reviewed. This process will describe the framework as well as the methods for solving the problems summarised in the literature review.

**Chapter 3**

# **Methodological Framework**

### 3 METHODOLOGICAL FRAMEWORK

The literature review demonstrated that there exists a need for a reassessment of the way in which the people's behaviour in the urban realm is represented within human dynamics models. In conducting this review, five areas of research were identified – namely stop identification, Spatial-Temporal ROI detection in urban networks, semantic enrichment of places, profiling of activity pattern, adaptation to urban network – to have been neglected within existing human dynamics pattern models, relative to the extent of reported research. These findings demonstrate a strong indication that current human dynamics models poorly capture the combined effect of space, time and semantics, as well as the true urban road network influencing people's movement and activities. The review furthermore showed that no existing framework or approach completely describes the full extent of people's behaviours in space and time, nor did they incorporate urban street networks in the semantic explanation of the activity patterns. Inspired by the ideas of related works, we model the relation between space, time, people and the activity patterns as shown in Figure 3.1, and use it as the theoretical basis of the semantic place analysis and activity profiling in our methodological framework.

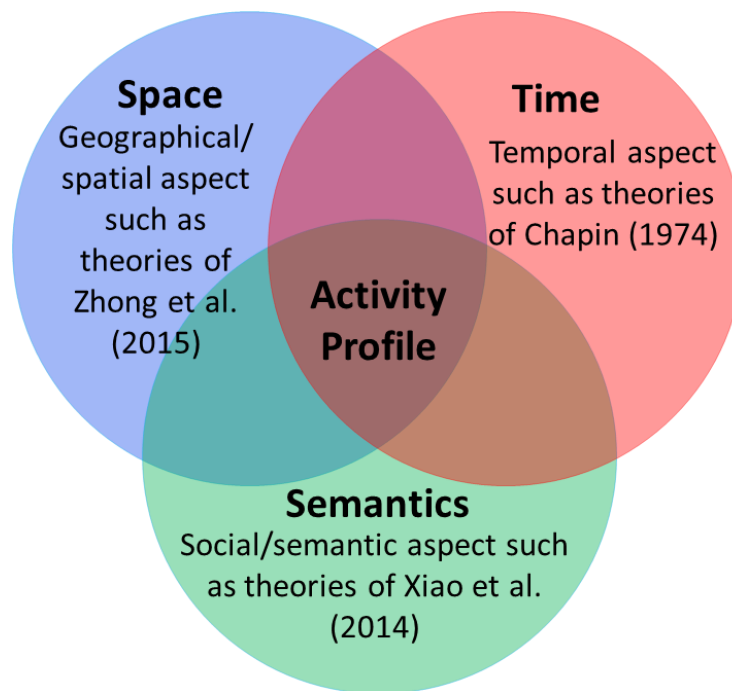


Figure 3.1 The three aspects of a complete activity profile

This chapter is composed of five sections. An overall description of the 4-step framework is provided in the first section. This part demonstrates how the methods in each step address certain limitations highlighted in the previous (literature review) chapter. The



second section describes how this framework achieves the aim and objectives, specified at the beginning of the introduction chapter of this thesis, with four functional modules. The third and the fourth sections demonstrate how the ideas of the framework are implemented in detail by the two paradigms in Cartesian space and environment of urban street networks respectively. The final section of this chapter discusses the comparison, validation and discussion of results.

### 3.1 FRAMEWORK DESCRIPTION

The methodological framework introduced here will describe only the integration of the different modules of work contributing to this aim. The complete dataset, methodology and detailed algorithms employed during each module of the work will be further discussed across the four following chapters (Chapters 4, 5, 6 and 7) of the thesis. Among them, Chapter 5 and Chapter 6 both go through the four framework modules in chronological order. Chapter 5 describes a Cartesian model in which all analyses in the four modules will be put to practice under the condition that all spatial distances are defined by the straight line Euclidean distance. Chapter 6 goes further by establishing a network paradigm and proposing a series of methods to calculate all spatial distances along the urban street networks or the actual persons' routes for all four modules. A more advanced text mining algorithm will be introduced in Chapter 7 to improve the framework's semantic enrichment ability.

For overcoming the five major problems highlighted in the literature review, the proposed framework should provide five corresponding solutions:

- **Solution 1:** Space-time stop identification: Comprehensively using the spatial and temporal information to identify stops in the movement trajectories.
- **Solution 2:** Spatio-temporal clustering: Proposing a clustering algorithm to detect region and time period of high density stay point aggregations.
- **Solution 3:** Space-time semantic enrichment: Associating the meaning and function of the visited places to people's activity patterns.
- **Solution 4:** Profiling and grouping activity patterns: Quantifying the pattern differences between people's movements and grouping similar patterns with respect to the semantic meaning of the activity.
- **Solution 5:** Urban-network-friendly improvements: Adding spatial network analysis approaches to every step of the above four solutions and enabling the framework to generate more precise results with network awareness in the city.

For **Solution 1** to **Solution 4**, we propose a four-module framework of methodology. Each of the four modules within the framework resolves an existing problem accordingly. These modules represent the core contributions of the thesis, contributing in ultimately advancing the state-of-the-art in activity pattern analysis. The flow path of the framework is as follows:

- **Module I (Pre-processing):** Pre-processing trajectory data is a mandatory step to remove errors for various research purposes. It is crucial in creating simple and correct low-level representations of movements. In this module, we identify stay points and segment trips with a kernel-based temporal scanning window. The stop identification process is robust for a location dataset with sporadic positioning errors. Conventional threshold-based decision methods and density-based methods can be used to play the function of this module. However, they cannot make full use of the temporal dimension of the movement data and tend to misidentify stops when there are positioning errors within the trips. Hence, we present a kernel-based temporal scanning window to jointly use spatial and temporal information in stop identification. In further improvements, we will also use map-matching to match the movement points to appropriate road networks so that the framework can work better with movements in dense city streets.
- **Module II (ST-ROI Detection and Space-Time Profiling):** Applying a point-based space-time clustering method on the stay points identified by Module I to detect Spatio-Temporal Region of Interest (ST-ROI). An ST-ROI is a region experiencing intensive visits of people with a time duration. Density-based clustering, kernel density estimation and their variations can be used to play the function of this module. Since a person's time budget allocation of activities in a day is the major indicator of his/her activity pattern, we simplify a person's daily activities as a space-time profile showing how much time he/she has spent on the detected ST-ROIs.
- **Module III (Space-Time Semantic Enrichment):** This module is designed to generate higher-level semantic abstractions from the low-level trajectory representation in the previous modules. Text mining methods are applied on POI data to relate the ST-ROIs detected in Module II with semantic/functional meanings that are of interest to the people. This process turns ST-ROIs into semantic ROIs. Statistic and topic modelling methods can be used to play the function of this module. By these methods, the semantic meaning of places can be attached to individuals' space-time profiles in Module II and turn the space-time profiles into semantic profiles for aggregation analysis in Module IV. This module also makes use of the opening hours in the semantic enrichment process, taking the temporal changes of semantic meaning of places into account.

- **Module IV (Semantic Profiling and Aggregation):** Summarising an individual's semantic movement pattern with a semantic profile represented by the time budget allocation on the semantic ROIs he/she visited. A similarity metric of profiles is defined and used in a hierarchical clustering method so that people with similar patterns can be aggregated. Several similarity metrics can be use as inputs to the clustering analysis method to play the function of this module.

Methods and algorithms adopted in each functional module can be replaced by their equivalents for performance comparison. Table 3.1 shows the flow chart of the framework as well as the methods that can act as potential options to serve the purpose of each functional module.

Table 3.1 Input, output and method options for constructing the modules in the methodological framework

	Module I	Module II	Module III	Module IV
Input	<ul style="list-style-type: none"> <li>• Raw GPS trajectories</li> </ul>	<ul style="list-style-type: none"> <li>• Stay Points</li> </ul>	<ul style="list-style-type: none"> <li>• POI data</li> <li>• Spatial Coverage and Time Span of ST-ROIs</li> <li>• Individual Space Time Profiles</li> </ul>	<ul style="list-style-type: none"> <li>• Semantic Profiles</li> </ul>
Output	<ul style="list-style-type: none"> <li>• Stay Points</li> </ul>	<ul style="list-style-type: none"> <li>• ST-ROIs/ST-LOIs</li> <li>• Individual Space Time Profiles</li> </ul>	<ul style="list-style-type: none"> <li>• Semantic Profiles</li> </ul>	<ul style="list-style-type: none"> <li>• Activity Groups</li> </ul>
Method Options	<ul style="list-style-type: none"> <li>• Threshold based</li> <li>• Density based</li> <li>• Kernel based Temporal Scanning Window</li> </ul>	<ul style="list-style-type: none"> <li>• DBSCAN</li> <li>• ST-DBSCAN</li> <li>• ST-Network-DBSCAN</li> </ul>	<ul style="list-style-type: none"> <li>• Statistical</li> <li>• TF-IDF</li> <li>• LDA</li> </ul>	<ul style="list-style-type: none"> <li>• Hierarchical Clustering</li> </ul>

**Solution 5** is achieved by realising the goals of **Solution 1** to **Solution 4** under the network representation of space. This requires fundamental transformations in all methods of the four modules and will be discussed further in the next section.

### 3.2 THE TWO PARADIGMS

To better illustrate the framework, we propose two paradigms to test the framework under different experimental settings. Both paradigms follow the 4-module workflow; however, they incarnate the framework with different combinations of approaches under different definitions of spatial distances.

The four modules will firstly be tested in a series of controlled experiment settings assuming all spatial distances are Euclidean distances. We call this first attempt the Euclidean paradigm because all location points in this process are in Cartesian space, just like the points in existing conventional methods. Keeping the spatial distance metrics same as in existing methods allows us to focus on the advantages of jointly

analysing time and space over separating the spatial and temporal aspects in conventional activity studies.

After that, a network paradigm is proposed to achieve **Solution 5** and provide a more realistic toolkit for trajectory analysis and activity pattern study in an urban context. Although the network paradigm is also composed of the four modules, it further adapts the framework to the real environment of the urban street networks. Unlike the methods used in the Euclidean paradigm, distance measurements in all modules of the network paradigm are based on the movement routes along the road segments and street network distances. This improvement requires fundamental changes to all distance-related operations and algorithms in all modules so that they can be adapted to the road network structure and optimised for the consequent computation cost. In the network paradigm, individual behaviours in the cities can be better represented than in its conventional counterparts.

In short, the Euclidean paradigm completes the contents from **Solution 1** to **Solution 4**, whereas the network paradigm makes further improvements to the Euclidean paradigm and resolves all listed problems and achieves **Solution 5**.

### 3.3 THE CARTESIAN PARADIGM

The workflow of modules and the combination of proposed methods of the Euclidean paradigm are shown in Figure 3.2.

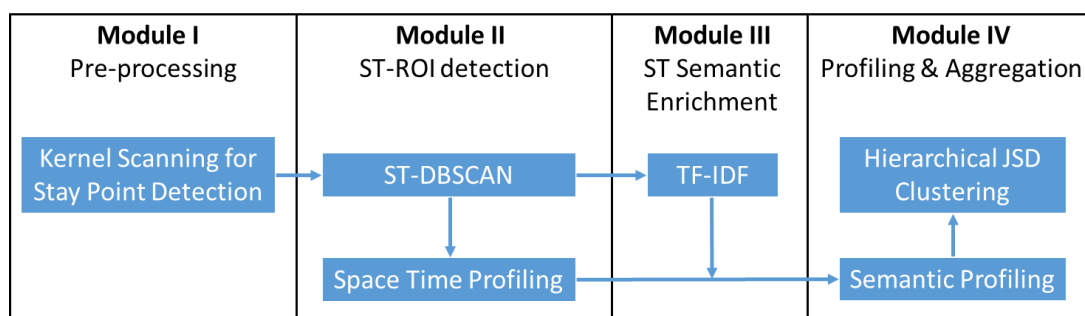


Figure 3.2 Work flow of the Euclidean paradigm

#### 3.3.1 Module I: Data Pre-processing

The pre-processing module has three major functions: trip segmentation, stay point identification, and data cleaning. Each of the functions will be achieved by a solution listed below.

##### Trip segmentation:

Trajectory/Trip segmentation is driven by data-dependent and application-related criteria. In our research, a trip refers to a continuous sequence of outdoor movements and stops by an individual person, starting from place A and ending in place B (A and B can be the same place). The purpose of trip segmentation is to break an individual person's sequence of GPS records in a day into trips by respectively identifying gaps between them. If a person reaches all purposes of the current trip and stops all activities for very long hours, the trip ends. Since a person can make multiple trips in a day, the purpose of trip segmentation is to differentiate one trip from another. From this definition, trip segmentation can be easily achieved by ignoring inactive time periods and marking time periods with continuous location updates as one trip. This process enables us to differentiate a short term stop episode within a trip from the long gap between two separate trips.

### **Stay point identification in Cartesian space:**

In the study of the movement of individual users, the basic assumption in many existing works is that people stop at a certain place to undertake various activities and then leave for the next place. Therefore, the stopping behaviour is of greater interest than the moving for the detection of interesting places and activities (Palma et al., 2009). Identifying stops in the trajectories is the first step in these researches (Alvares et al., 2007; Zheng et al., 2009). As we have discussed in the literature review, the two conventional types of method (i.e. density-based and threshold based) are not accurate enough in stay point identification, especially when they are dealing with urban movement data with positioning errors. Hence, we propose a Kernel-based temporal scanning window to confine the kernel-based stop identification process in a temporal window. This method can mitigate the misidentification problem on sporadic positioning error data, leading to the elimination of the wandering effect that GPS inaccuracies result in. It can overcome the limitation of conventional stop identifying methods mentioned in Chapter 2, and its performance improvement can be seen in Chapter 8.

### **Data cleaning:**

Constant and continuous misreport of locations and very short records are counted as GPS logging failures and are removed from the dataset. At the same time, sporadic positioning errors and temporary signal blockades are preserved and checked by the Kernel-based temporal scanning window as usual. A kernel with Euclidean spatial bandwidth will be used to interpolate and smooth the space between the points and calculate the density of points for identifying stops. A temporal scanning window will make sure that only temporally close points can participate in the kernel density calculation to avoid the limitation of methods that ignore temporal information. After

the interpolation, Points in the high kernel density area are labelled as stay points. Besides, if adjacent points logged before and after a noise/error point are identified as stay points, the error point will also be labelled as a part of the stop and counted in the dwelling/stopping time duration of the people. This improvement enables us to detect stops and stopping time more precisely to guarantee the quality of the input data of Module II.

### 3.3.2 Module II: ST-ROI detection in time and Cartesian space

As highlighted in Chapter 2, all existing ROI detection methods ignored the influence of time dimension on the ROIs. None of them see ROIs and the semantic meanings of ROIs as dynamic and ever-changing phenomena. Here we apply Birant and Kut's ST-DBSCAN algorithm to detect ROIs that attract high visit volumes within a certain time duration (i.e. ST-DBSCAN). This algorithm is chosen because of its unique ability to discover high-density clusters according to the spatial, non-spatial and temporal values of objects.

The temporal, spatial and cluster size parameters of the algorithm can affect its cluster researching process and results (the point density and area coverage of the detected ST-ROI). Therefore, a parameterisation stage is designed to make sure the algorithm serves the application of finding ST-ROIs. In the Euclidean paradigm, we primarily focus on Spatial-Temporal clustering and simplify the spatial distance measurement by using straight Euclidean distance and ignoring the urban streets. Each of the detected ST-ROIs has its spatial coverage and time span, which are important indicators of the ROI semantic enrichment process in Module III. In the Euclidean paradigm, the spatial coverage of an ST-ROI is defined by an extended bounding convex hull enclosing all stay points in the ST-ROI.

Based on the idea of “**where, when and how long you stay is who are**” summarised in the literature review chapter, a person's activity pattern can be described by the dwelling time allocation on the ST-ROIs he/she visits. This ST-ROI dwelling time allocation is defined as the space-time profile of a person and can be generated by summarising when the person arrived at and left the ST-ROIs.

### 3.3.3 Module III: Semantic enrichment of ST-ROIs

Here we expand the concept “**where, when and how long you stay is who are**” to “**the place you go, when you go and how long you stay is who you are**”, to emphasise the importance of ‘place’ in understanding human dynamics. We develop a semantic

enrichment method to integrate the semantic meanings of places into people's activities by a TF-IDF topic modelling algorithm on categorised Points of Interest (POIs) in the coverage area of ST-ROIs. Besides, we also introduce the influence of POI opening hours into the topic model. Only POIs open within the time span of a ST-ROI can contribute semantic meaning to the ST-ROI. This method allows us to find the changes of a place's semantic meaning at different times of the day. An individual's profile is then built as a summary of the person's dwelling time allocated in different semantic places. In the Euclidean paradigm, the distances between stay points and POIs are Euclidean.

### **3.3.4 Module IV: Aggregative analysis of the semantic profiles**

Individual space-time profiles can be transformed into semantic profiles after the semantic enrichment of ST-ROIs in Module III. After that, the similarity/dissimilarity between the individual semantic profiles needs to be defined for the aggregative analysis in this module. Here, discrete Jensen-Shannon Divergence (JSD) (Lin, 1991) is used to measure the dissimilarity of the semantic profiles of two users. The advantages of JSD in quantifying profile differences and aggregative clustering analysis will be demonstrated in detail in Chapter 5. We generate a pairwise distance matrix containing the JSD between any two individuals. This matrix is input into the hierarchical clustering algorithm to group people sharing similar activity patterns.

## **3.4 THE NETWORK PARADIGM**

The previous section described a Cartesian version of the methodological framework, in which all spatial distances are based on Euclidean measurements. This Euclidean paradigm is not sufficient for the completion of **Solution 5** in section 3.1 because the Cartesian space misrepresents the true topological structure of city streets and the actual route of people moving in the cities. Countering the noted weaknesses of the methods in the Euclidean paradigm, specific focus will rest upon improving the framework's adaptation to the analysis in urban networks. To this end, we propose a network paradigm with four network-friendly modules. The new work flow of the network paradigm is shown in Figure 3.3.

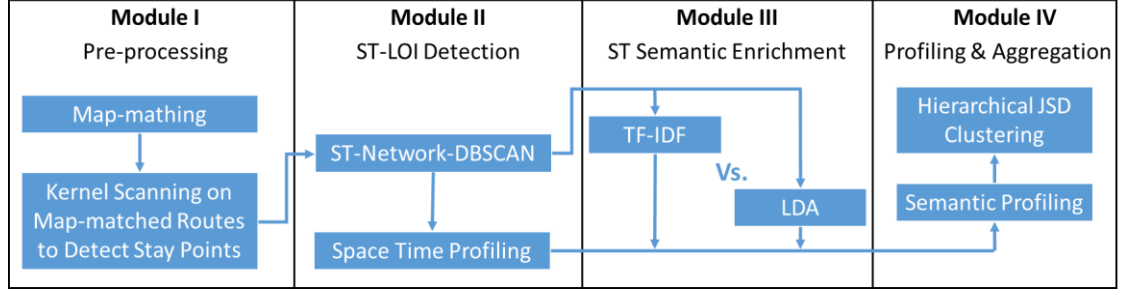


Figure 3.3 Work flow of the network paradigm

### 3.4.1 Module I: Pre-processing for network-based movement analysis

The pre-processing module in the network paradigm serves the purpose as Module I in the Euclidean paradigm. The trip segmentation and data cleaning process use the same simple rule-based methods as in Module I in the Euclidean paradigm. The difference is that a map-matching stage is added before all analysis to snap the sequences of points in individual trips onto the most probable street segments passed through by the person, so that the raw movement trajectories can be transformed into the routes of trips. To adapt the entire framework to the analysis of movement in the streets, all space and movement related methods and operations are based afterwards on those generated by the map-matching process in this module. We apply Lou et al.'s (2009) ST-matching algorithm as the map-matching algorithm because of its great performance on trajectories of low sampling rate.

The kernel-based temporal scanning window is also used in the module for stay point identification. The improvement we make in this module is that the bandwidth of the spatial kernel for stop identification is measured by the network distance alone, the routes of the individuals turning the method into a network-based kernel density calculation. Points on the road sections with high kernel densities are labelled as map-matched stay points. Like the Module I in the Euclidean paradigm, this stay point identification method is robust to sporadic errors and short-term signal losses.

### 3.4.2 Module II: ST-LOI detection

We ameliorate the ST-DBSCAN by replacing the Euclidean spatial metrics with network distance for the detection of ST-ROIs in the street networks. The ameliorated variation optimises the spatial query technique to mitigate the increase of computation burden brought by network distance calculation. We call this new variation ST-network-DBSCAN and apply it on the map-matched stay points. Unlike the Euclidean paradigm, dense space-time point aggregations detected by ST-network-DBSCAN are all located



on the road segments' line structure. We therefore call the ROIs detected in the network paradigm Spatial-Temporal Lines of Interests (ST-LOI) to distinguish them from ST-ROIs detected in the Euclidean paradigm. To better explain and showcase the results of ST-network-DBSCAN, we propose a 3D wall-map method to visualise network-based data in space and time in a joint effort.

### **3.4.3 Module III: Semantic enrichment of ST-LOIs**

To associate the POI information with the streets covered by ST-LOIs for semantic enrichment in street networks, each POI is snapped to the nearest segment or the segment sharing the POI's registered street address. Opening hours of POIs are also counted in to measure their semantic contribution to the ST-LOIs in TF-IDF.

In Chapter 7, the network paradigm is tested with a much bigger dataset in a larger study area and the TF-IDF is replaced by an LDA algorithm to achieve a better semantic enrichment outcome. The fundamental difference between these two methods is that TF-IDF is an algorithm based on term frequency, whereas LDA is based on probabilities. Moreover, TF-IDF generate semantic ROIs by reweighting the influence of existing semantic categories, whereas the LDA can generate a list of "top POIs" or "dominant POIs" of a group of interrelated POI types for the user to empirically summarise the meaning of the generated semantic category. The performances of TF-IDF and LDA are compared in Chapter 8.

### **3.4.4 Module IV: Aggregative analysis of the semantic profiles**

ST-LOIs are transformed into semantic LOIs after Module III. People's dwelling time allocation in ST-LOIs are also transformed into their semantic profiles. This module takes the semantic profiles as input and performs the same hierarchical clustering procedure as does Module IV of the Euclidean paradigm.

## **3.5 ADDRESSING RESEARCH AIM AND OBJECTIVES**

This chapter has outlined the methodological framework that will be elaborated upon during the next three chapters of this thesis, moving towards the overall aim of advancing the research agenda with respect to aggregate urban human activity patterns in a quantified manner.

In building towards this framework, it has become clear that although the review of literature provides some guidance with which to proceed, a great deal of work is required

for its completion. The limitations of conventional methods in the literature review chapter are summarised into five problems. For revolving these problems, our methodological framework has been divided into two paradigms, each of which consists of four modules. Modules I, II, III and IV in the Euclidean paradigm respectively achieve **Solutions 1 to 4** for the first four problems in a Cartesian representation of space, whereas the Network paradigm provides **Solution 5** for the last problem by bringing the scale of space-time analysing methods of the framework onto the streets.

In constructing the framework modules outlined in this chapter, the aims and objectives highlighted in Chapter 1 will be accomplished. At this moment, we have achieved Objective 1 (Chapter 2) and Objective 2 (Chapter 3). Chapters 5-7 will address Objective 3. The validation of methods and the comparison between two paradigms and conventional methods will be specifically summarised in Chapter 8 to address Objective 4. The final outcomes of this research will be reviewed in the conclusion chapter (Chapter 9).

## Chapter 4

# Knowing about the Data

## 4 KNOWING ABOUT THE DATA<sup>1</sup>

The success of a data driven approach for activity pattern study depends heavily on the characteristics of the datasets. This chapter describes all the datasets we used as inputs in the framework. As highlighted in Chapter 1, our goal is to profile and aggregate people's activity patterns by jointly considering the three aspects of people's urban activities (i.e. individual space-time mobility, urban spatial structure, and the semantic meaning of places) and their changes in time. Each of the aspects can be represented by a corresponding input dataset and used in the corresponding modules as demonstrated in Table 3.1. Three datasets therefore participate in our case studies and experiments. Among them, the GPS movement trajectories of London police officers are used as the space-time mobility dataset of people; the ITN (Integrated Transport Network) layer of London is used to represent the city's street network structure; and the Ordnance Survey (OS) POI dataset contains the semantic meaning of buildings in the city. These three datasets as well as the case study area in which they are collected are described in sections 4.1 and 4.2. Section 4.3 is a brief exploration of the patterns in the data. The data's characteristics and indications to the method options in the framework are summarised at the end of this chapter.

### 4.1 CASE STUDY AREA

This study takes place in London, UK. Greater London is made up of 32 boroughs (local authority districts), each of which is assigned a Borough Operational Command Unit (BOCU) of the Metropolitan Police. All BOCUs have police officers (regular and specials) who patrol and respond to emergencies. We specifically focus our study on the police activity of the 12 BOCUs corresponding to the 12 inner London boroughs (shown in Figure 4.1), statutorily defined by chapter 33 of the London Government Act 1963 (UK legislation, 1963), namely Camden, Greenwich, Hackney, Hammersmith and Fulham, Islington, Kensington and Chelsea, Lambeth, Lewisham, Southwark, Tower Hamlets, and Wandsworth and Westminster. Inner London is officially the wealthiest area in Europe according to the 2010 report of regional GDP per capita in the EU (European Commission, 2010). Its population density is more than double that of Outer London. The City of London, located in the centre of the map in Figure 4.1, is not included in the study area of this thesis because the law enforcement in this area is not administrated by the Metropolitan Police.

---

<sup>1</sup> Part of this chapter was presented in: Shen, J. and Cheng, T., 2014. *Group Behaviour Analysis of London Foot Patrol Police*. 23<sup>rd</sup> GIS Research UK, Leeds, UK.

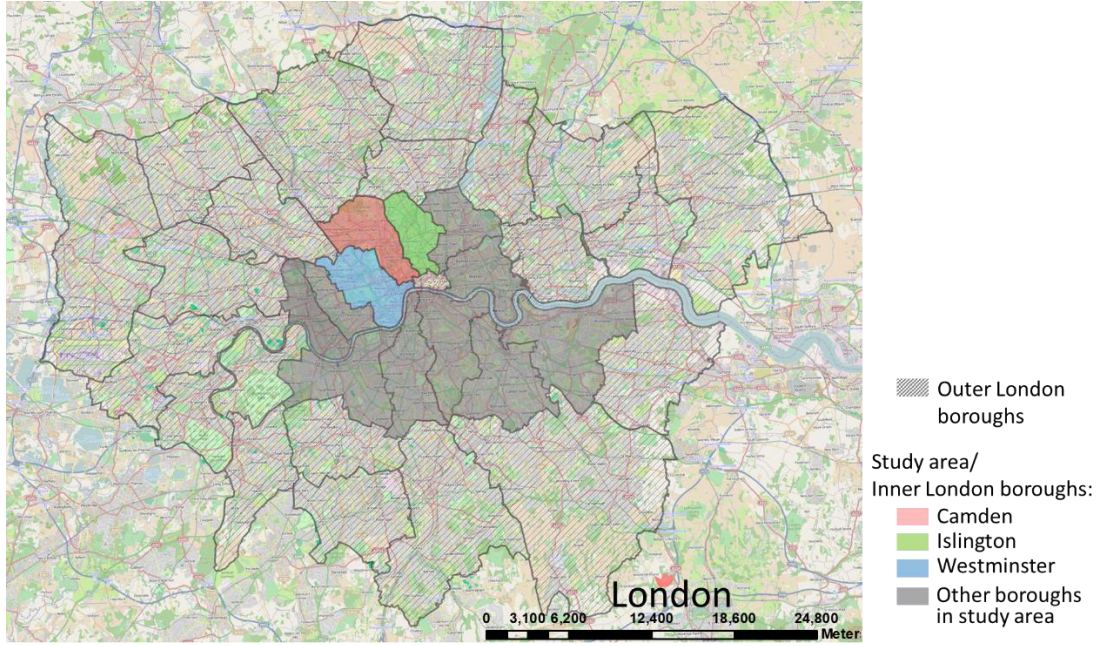


Figure 4.1 The study area of the thesis

Considering the convenience of demonstration and sensitivity of the police data, we will not display the results of all the 12 BOCUs. Instead, we mainly use a study area covering Camden, Islington and Westminster as an example to show the working mechanism and results of our stop identification, ST-ROI detection semantic enrichment and activity profile aggregation. These three BOCUs are chosen because they cover most of central London and they are the boroughs with the busiest urban activities. However, some of the results of the experiment in all of the 12 Inner London boroughs will be used in Chapters 7 and 8 for validation and comparison purposes.

## 4.2 DATASETS

### 4.2.1 Movement data

Human movement data is first and foremost a dataset to be processed by the framework. It contains the spatial locations as well as the time stamps of locations of the moving individual and represents one aspect of individuals' activities.

In our case study, we use the GPS data collected by the newly upgraded Automatic Personnel Location System (APLS) of the London Metropolitan Police. Two periods in the APLS dataset are specifically selected as our case studies. One of the periods is in February 2012 containing the movements in only one borough. It was used in the early stage study of the research in this thesis for initial testing of the developed algorithm. The other period of the APLS dataset is collected in August 2015. This dataset contains

police activities of all boroughs in London and was used for the full-fledged implementation and experiment of the Euclidean paradigm and the network paradigm. The terminals of the APLS are integrated into the portable radio sets of every police officer and stay connected as long as the officer is on duty outdoors, especially on missions and patrols and emergency responses. There were 17,983 officers working in the 32 BOCUs by the end of February 2015 according to the administrative report of the Metropolitan police service (Metropolitan Police, 2015). When working normally, the device generates GPS location records at a five-minute sampling rate. When the radio is powered off or blocked for some reason, the logging will be temporarily stopped. Under emergency situations, the location can be immediately updated when the officer pushes the emergency button on the radio set. Just like other equivalent location information systems, the APLS is originally designed for monitoring and dispatch applications in the policing operations. Operators and senior officers in an area control centre are able to see the latest updated location of every unit (including personnel and vehicles) and their working status on a base map. With this system, a dispatch operator is able to see who is the nearest officer to an incident or emergency call and whether he/she is available for mission dispatch. It should be noted that as the system was kept running for the past years, a huge amount of movement data was stored and became a valuable resource to analyse the activity pattern of officers and to evaluate their behaviours.

As recommended by Transport for London (TfL) (2010), London's public transportation authority, the preferred walking speed is 1.33 m/s in London's urban area. According to our calculation, the average length of street segments in inner London ranges from zero to 130m across different boroughs. A patrol officer walking continuously at this speed can move 400m in-between two updates of APLS. This distance is three times the average length of segments in London. Lou et al. (2009) experimented their method with data of vehicles travelling at 40km/h with a 2min GPS sampling rate (equivalent to 1333m between contiguous records). Bolbol et al. (2012) experimented their method with data of vehicles travelling at 18.65m/s with a 1min GPS sampling rate (equivalent to 818m between contiguous records). Comparison to data used by other researchers suggests that the APLS data is a typical low sampling rate but is an acceptable dataset for analysing movements that consist mainly of pedestrian trajectories. The other indicator of quality for movement of a dataset is the level of positioning errors. EPE (estimated positional error) is most commonly used indicator for positioning error. It is estimated (not measured) to predict the accuracy of GPS data by a confidence level according to the the navigation message of signal quality. The EPE formulas are the proprietary of the GPS receiver producers, and are not publically released. However, the confidence level are specified for each product. For example, a 5m EPE means that we can have a 95% confidence that the true location of the GPS device is within 5m of the coordinates

reported by the device in the current update. The 95%-confidence EPE information of APLS is contained in the movement dataset (see Appendix A). The EPE distribution of APLS complies to a slightly skewed normal distribution (Figure 4.2)  $N(20, \sigma^2)$ , where  $\sigma = 8$ .

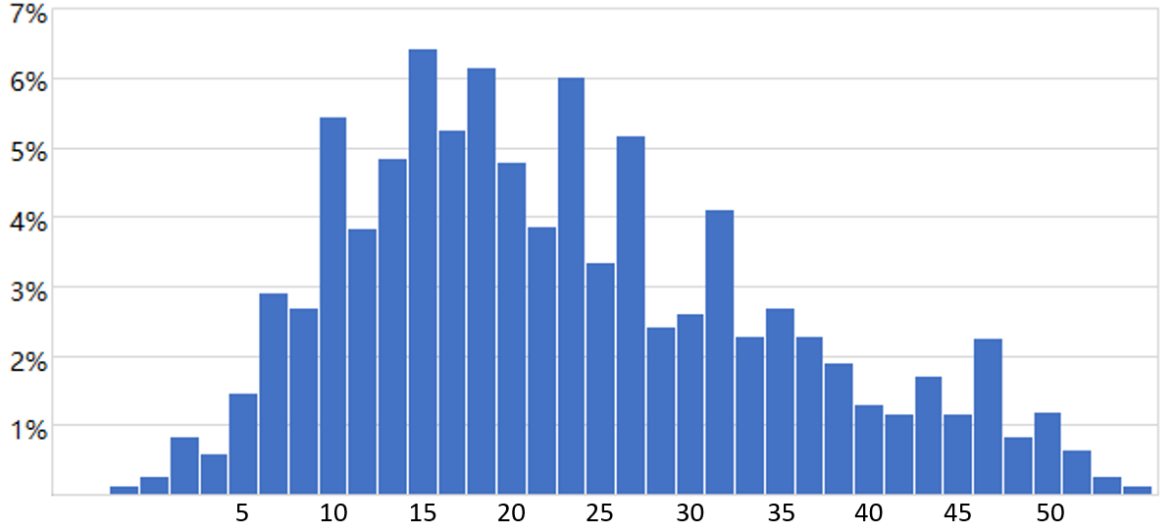


Figure 4.2 APLS's EPE (estimated positional error) distribution (London, UK, August 2015)

The movement dataset we used in section 4.3, as well as in Chapters 5 and 6, is collected only in Camden in February 2012. We choose this special period because there was a major security incident in February and the framework was able to automatically detect and generalise the reaction pattern of the police officers without a priori knowledge. The extended case study in the framework is tested and demonstrated based on its performance on the movement dataset of 12 inner London boroughs in August 2015. This allow us to show the framework's ability to deal with scaled-up datasets and guarantee better performance in Modules III and IV (i.e. semantic enrichment and profiling).

Besides location and time stamps, the dataset also contains multiple columns including call sign, status and work type. The explanation of these columns is shown in Table 4.1. A call sign is an identification number for each officer. A given call sign can be used only by one officer and it cannot be changed until he/she leaves his/her present unit. Thus, it can be assumed that one call sign uniquely represents one police officer. The status records the current mission and availability of dispatch of an officer and needs to be manually updated by the officers through their portable terminals. Work type is the role of an officer. There are eight common types of officer. Each type of officer has their own focus of work, although different officer types may sometimes share similar missions. An anonymised example of an officer's trip logged by the APLS can be found in Appendix

A. The homogeneity of professional background and diversity of assignment of responsibility provided us with a favourable dataset for behaviour study of people working in the same urban area. Methodologies developed from this movement dataset can be expected to work with similar data of individual movements of similar professional background.

Table 4.1 Non-spatio-temporal information contained in the APLS dataset

Column names	Call sign	Status	Work type
Explanation and Examples	A serial identification number of each officer	<ul style="list-style-type: none"> <li>• On Patrol</li> <li>• Committed</li> <li>• Assigned to a Vehicle</li> <li>• Off Duty</li> <li>• On Watch</li> <li>• Deployable</li> <li>• Limited Availability</li> <li>• Rest/Refreshments</li> </ul>	<ul style="list-style-type: none"> <li>• Foot Patrol Officer</li> <li>• Community Support Officer</li> <li>• Senior Officer</li> <li>• Special Constable</li> <li>• Foot Patrol – Plain Clothes</li> <li>• BOCU Management Officer</li> <li>• Detective Constable</li> <li>• Detective Sergeant</li> </ul>

#### 4.2.2 POI data

We acquire our POI data by combining the advantages of two POI data sources: Ordnance Survey POI dataset and Google Places dataset.

The major advantage of the Ordnance Survey POI dataset is that its function classification is very well-organised and includes many detailed sub-categories of POIs. The official Ordnance Survey POI classification scheme has a hierarchical structure of three levels, with nine major categories as the topmost level and 52 sub-categories that can be further broken down into more than 600 detailed classes. On the contrary, Google Places POIs can only be divided into 95 classes without any hierarchical summaries. Another advantage of the Ordnance Survey POIs is that Ordnance Survey customers can either directly adopt the official classification scheme defined by the Ordnance Survey or make changes according to their own research purpose. Customisation of researcher's own classification scheme by subsetting or merging POIs of different categories or sub-categories is encouraged (Ordnance Survey, 2016). We therefore made slight changes to the official classification scheme. By separating the



original “health and education” category into two independent categories and moving all “government and organisations” POIs out of “public infrastructure” to become a major category by themselves, a new 11-category classification scheme that fits our research purpose was generated (Table 4.2). Hence, we use the spatial information of Ordnance Survey POIs and a customised functional classification scheme to better preserve the semantic meanings of these POIs.

Table 4.2 The reclassified POI categories based on the Ordnance Survey POI classification scheme

Customised Classification Scheme	
<p><b>01 Accommodation, eating and drinking</b></p> <p>01 Accommodation</p> <p>02 Eating and drinking</p> <p><b>02 Commercial services</b></p> <p>03 Construction services</p> <p>04 Consultancies</p> <p>07 Contract services</p> <p>05 Employment and career agencies</p> <p>06 Engineering services</p> <p>60 Hire services</p> <p>08 IT, advertising, marketing and media services</p> <p>09 Legal and financial</p> <p>10 Personal, consumer and other services</p> <p>11 Property and development services</p> <p>12 Recycling services</p> <p>13 Repair and servicing</p> <p>14 Research and design</p> <p>15 Transport, storage and delivery</p> <p><b>03 Attractions</b></p> <p>58 Bodies of water</p> <p>16 Botanical and zoological</p> <p>17 Historical and cultural</p> <p>19 Landscape features</p> <p>18 Recreational</p> <p>20 Tourism</p>	<p><b>06 Public infrastructures</b></p> <p>34 Infrastructure and facilities</p> <p><b>07 Manufacturing and production</b></p> <p>37 Consumer products</p> <p>38 Extractive industries</p> <p>39 Farming</p> <p>40 Foodstuffs</p> <p>41 Industrial features</p> <p>42 Industrial products</p> <p><b>08 Retail</b></p> <p>46 Clothing and accessories</p> <p>47 Food, drink and multi-item retail</p> <p>48 Household, office, leisure and garden</p> <p>49 Motoring</p> <p><b>09 Transport</b></p> <p>53 Air</p> <p>59 Bus transport</p> <p>57 Public transport, stations and infrastructure</p> <p>54 Road and rail</p> <p>55 Walking</p> <p>56 Water</p> <p><b>10 Education</b></p> <p>27 Education support services</p> <p>31 Primary, secondary and tertiary education</p>

<b>04 Sport and entertainment</b> 22 Gambling 23 Outdoor pursuits 21 Sport and entertainment support services 24 Sports complex, gym 25 Venues, stage and screen  <b>05 Education and health</b> 26 Animal welfare 28 Health practitioners and establishments 29 Health support services	32 Recreational and vocational education  <b>11 Government and organisations</b> 33 Central and local Government
	Annotation:  <b>XX Major category code</b> xx Sub-category code

The number of different categories of POIs is counted to provide a general idea of the nature of the distribution of POIs in inner London. Figure 4.3 shows that POIs of commercial service, together with retail and infrastructure, accounted for a lion's share of total POIs, which is a common phenomenon in a metropolis like London. To use this unbalanced dataset for semantic enrichment, the weight of significance or semantic contribution of different POI categories will need to be quantified and reweighted.

On the other hand, as summarised in the literature review chapter, conventional semantic enrichment methods see the functional topic as a constant attribute of a place that does not change with time, which often leads to the misrepresentation of the place's semantic meaning. To avoid this misrepresentation, the opening/available hours of the POIs should be considered in the semantic enrichment module. Unfortunately, opening hours information is not collected into the Ordnance Survey POI dataset and my other conventional POI datasets. This disadvantage can be overcome by merging the Google Places dataset with the Ordnance Survey dataset, because POIs collected by Google contain very precise details of their opening hours.

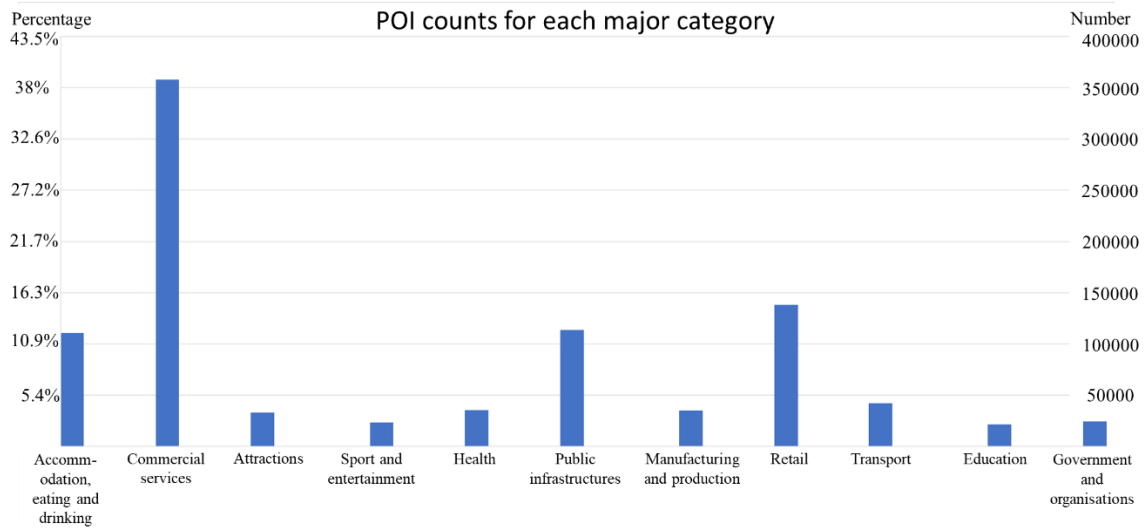


Figure 4.3 The POI number of each major category in the study area

The process of merging Ordnance Survey POIs and Google POIs is accomplished by a python-based web crawler programme that we developed. For every Ordnance Survey POI in the study area, the web crawler makes a spatial query through Google Places API to find the same POI in Google. If the Ordnance Survey POI's name matches with the same POI in the Google Places dataset, the opening hours in Google will be passed on to the Ordnance Survey POI. 11% OS POIs do not have any matches in the Google places dataset. Moreover, 16% of the matched POIs did not include information of their opening hours. For these two types of POIs with incomplete information, hypothetical opening hours are attached to them to complement the temporal information of POIs and mitigate the bias caused by the incomplete dataset. The hypothetical opening hours of each POI are determined according to the most popular opening and closing times of other POIs in its sub-category. For example, if a convenience store did not provide its opening hours in Google places and most of the other convenience stores open from 6:00 to 20:00, this convenience store with incomplete information will be estimated to be open from 6:00 to 20:00.

By merging information of OS POIs with temporal information in Google places, we can take advantage of both datasets to support our space-time semantic enrichment module. An example of merged POI data that include both spatial and temporal information can be seen in Appendix B. The category of each POI in this dataset is represented by a 4-digit serial code that shows the major category information with the first two digits and the sub-category information with the two last digits. For example, a primary school's category code is 1031 according to Table 4.2, and it falls in the major category of "Education".

### 4.2.3 Street network data

We use the Ordnance Survey Integrated Transport Network (ITN) urban theme layer dataset (Ordnance Survey, 2015) as the representation of London street networks. This dataset is a newly upgraded version of the conventional ITN originally designed for pedestrian and cyclist navigation. It extends the functionality of the normal ITN by joining up additional and more detailed local segments, including man-made footpaths, subways, steps and footbridges in a structured edge-and-node network in all urban areas. This enables us to match walking trajectories with minor streets in the map-matching process. The topological structure of the ITN network is composed of street segments (edges) and their intersections (nodes). A edge is a line or curve segment connecting two nodes at its ends. ITN data contains the ID and location information of each node and the length, connectivity and speed limit of each segment. A street contains a sequence of edges, and the edges are given the name and address of the street. The simple ITN network structure in Figure 4.4 shows the relationship between nodes and edges.

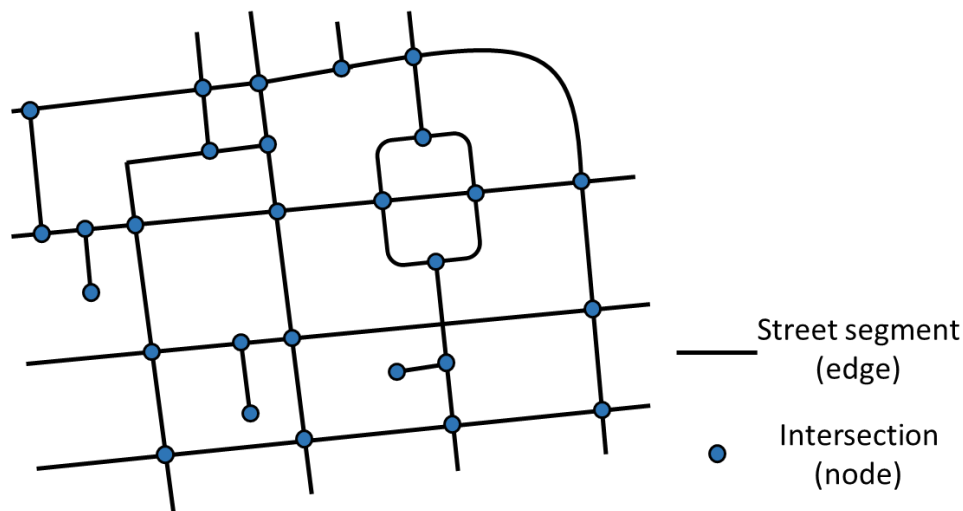


Figure 4.4 An example of ITN network structure

## 4.3 EXPLORING BASIC PATTERNS OF THE DATA

### 4.3.1 Spatial ROIs

To acquire some preliminary knowledge about the police movement in inner London, we apply conventional and statistical methods in this section to explore the basic space-time pattern in the data. We firstly explored the spatial ROIs for police activities by

applying a simple threshold-base stop identification proposed by Schönfelder et al. (2006) and Tsui and Shalaby (2006) on the police movement in Camden, February 2012. Euclidean distance is used as the spatial distance expression and points with speeds of less than 0.2 m/s for at least 10 minutes are identified as stay points.

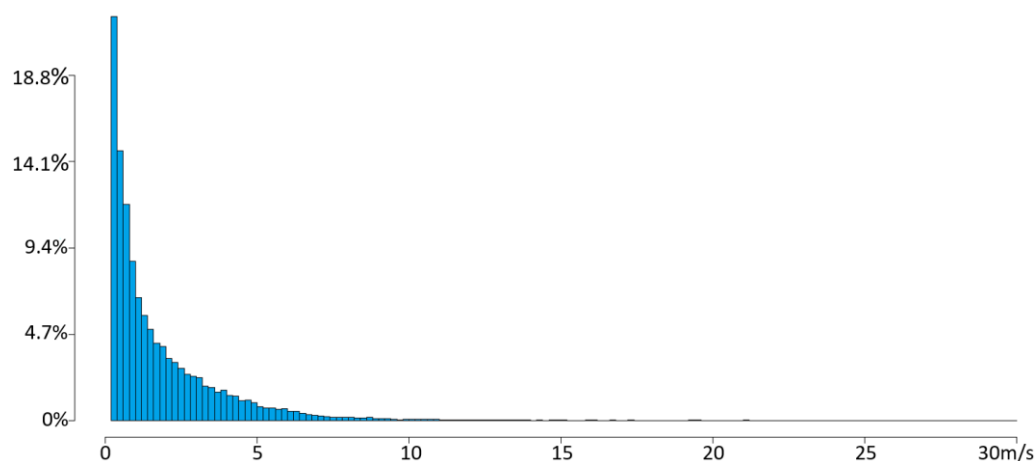


Figure 4.5 Average Euclidean speed in move episodes of APLS

According to this speed threshold, the police officers spend 50.9% of their time stopping at certain locations (i.e. stop episodes) in the city and 49.1% of their time moving around (i.e. move episodes). In the move episodes, the officers have an average moving speed of 1.874 m/s. Considering this speed is calculated with Euclidean distance and the actual distance covered by the trip from one place to another in the street network is always longer than the straight line Euclidean distance, the actual average moving speed of officers is slightly higher than 1.874 m/s. This figure is a little bit higher than a preferred walking speed suggested by TFL (2010), because in very rare cases officers catch a bus for travelling and watch the space from within the vehicle or go in a police car for emergency responses and travel as fast as 30 m/s during their patrols.

OPTICS, a variation of basic DBSCAN that does not require predefined parameters, is then used to cluster these stay points in space. Unlike DBSCAN, the OPTICS algorithm linearly orders points of the movement dataset so that points which are spatially closest to each other become neighbors in the ordering. The distance that represents the density needed to be accepted for a cluster in order to have two adjacent points grouped into the same cluster is stored for each point. This process is inspired by the hierarchical clustering method and its results can be represented as a reachability plot shown in Figure 4.6. This reachability plot (a special kind of dendrogram) shows the ordering of the points as processed by OPTICS on the x-axis and the reachability distance on the y-axis. Since points belonging to a dense cluster have a low reachability distance to their nearest neighbor, the clusters show up as valleys in the reachability plot. The deeper the

valley, the denser the cluster. In this way, the hierarchical structure of the clusters can be obtained easily. Visual inspection is needed to extract clusters from this plot, which is done manually by selecting a range on the x-axis after, by selecting a threshold on the y-axis. In Figure 4.6 the points identified as noises are coloured black and the points in ROIs are in other colours. As a result, 14 spatial ROIs (Figure 4.7) are discovered by this conventional method.

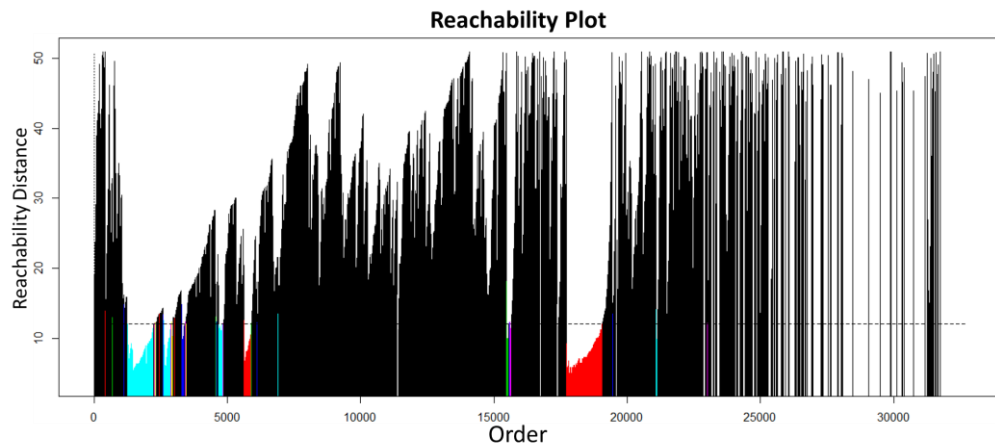


Figure 4.6 reachability plot showing the OPTICS clustering results

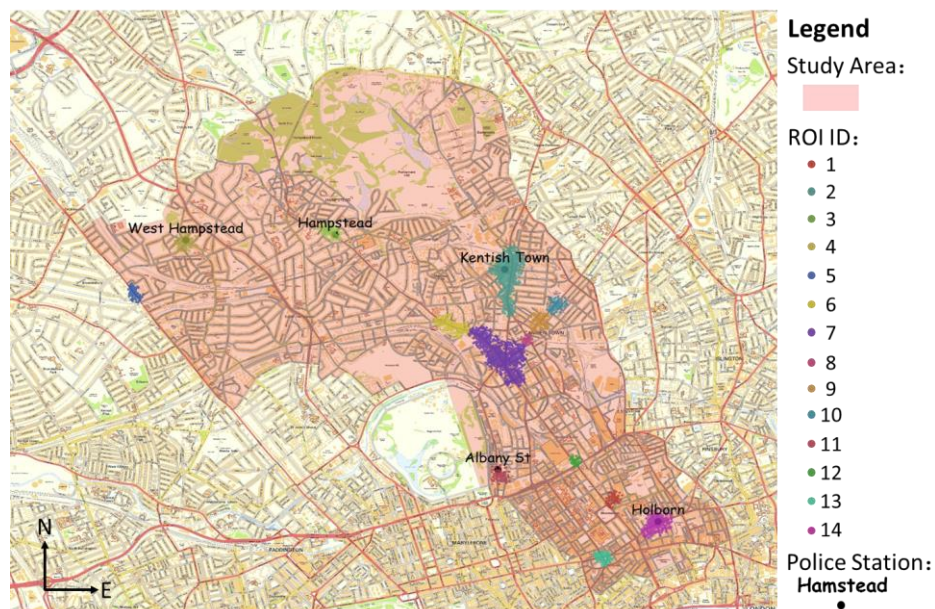


Figure 4.7 ROIs detected by conventional OPTICS in Camden APLS, February 2012

From the conventional result, we observed aggregations of stay behaviours near police stations. This phenomenon is a common reflection of police daily routines and office time and therefore provides little information. To discover more semantically significant ROIs other than police officer activities, all GPS records within a 150 m radius of police

stations are removed before the ROI detection module in our framework in the following chapters.

### 4.3.2 Statistical patterns of people's visits to ROIs

Because of the variation in police officers' person preferences and work types, different patterns can be found in the officers' movements and stops. Figure 4.8 shows the raw trajectories of two officers' daily trips in Camden BOCU in February 2012. The trajectories of the two officers are given different colours and each continuous polyline represents a trip. The officers' call signs have been anonymised as "102PO" and "619PO", from which it can be seen that different officers often visit different places at different times of the day, although they sometimes visit the same areas.

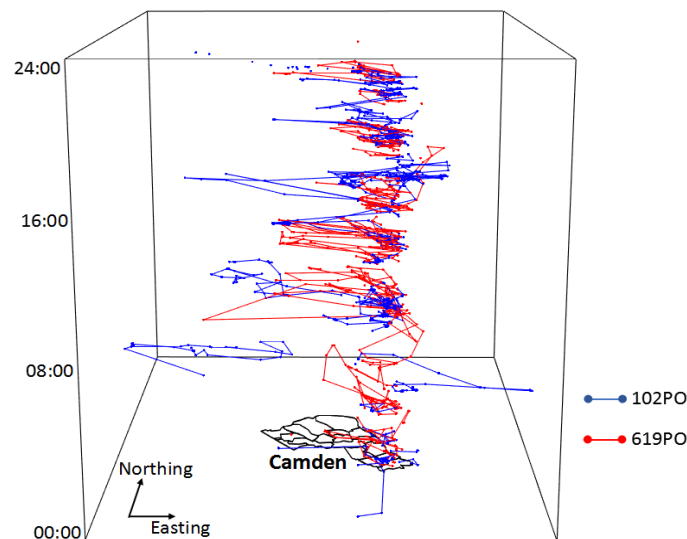


Figure 4.8 Visualisation of two officers' (102PO and 619PO) raw movement trajectories (as polylines) in a space-time cube

Figure 4.9 shows the total number of officers in patrol in every hour of February 2012. This visualisation is extracted from the APLS data. Officers keeping active updates in APLS within certain periods are considered as being on duty in such periods. The x-axis in the heat map represents the 24 hours of a day and the y-axis represents the 28 days in February 2012. We use the number of active officers as an indication of intensity of activities. The visualisation shows clear peaks of active intensity in the small hours at weekends and afternoons of weekdays. Additionally, there is a gap around 7:00 am to 9:00 am every day that shows scarce patrol activities. This period marks the officers changing shifts and starting the work of another working day.

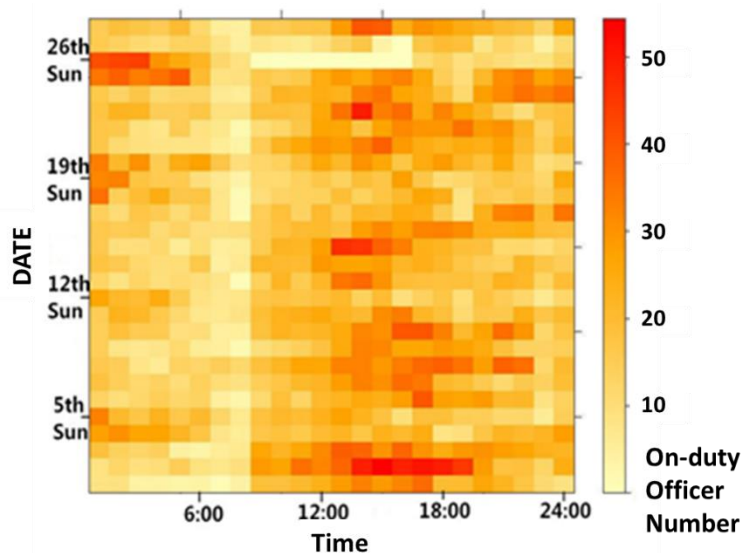


Figure 4.9 Heat map of hourly intensity of activities in Camden police patrol

Figure 4.10 demonstrates a temporal comparison of the two anonymous officers' activities in February 2012. The heat maps show both daily and weekly periodic patterns. It can be seen that "102PO" had far fewer days on duty in the study period, while "619PO" was active in most of the days. It also shows that "102PO" only worked in the afternoon, whereas "619PO" was active in the mornings of weekends.

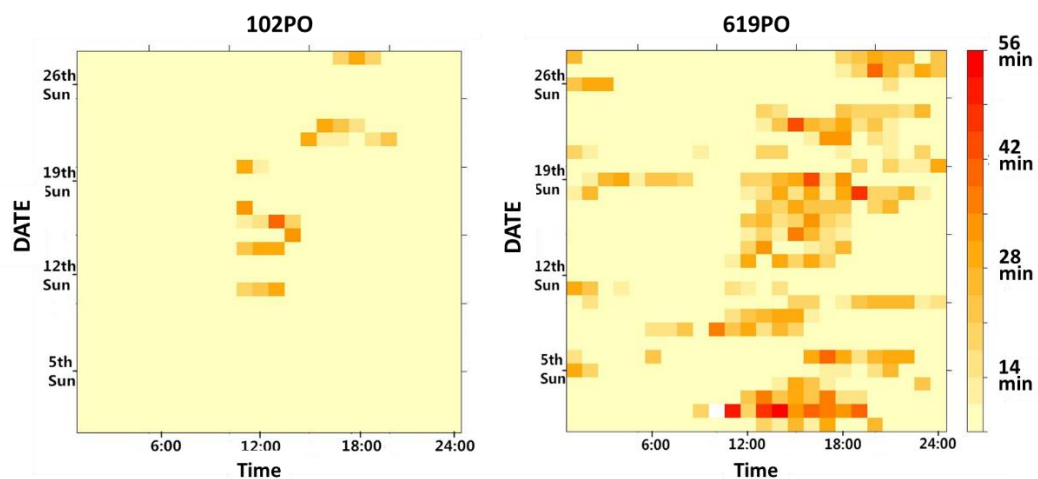


Figure 4.10 Heat map showing individual active intensity of two officers

Apart from the pattern differences in working hours, officers also paid attention to different places. Figure 4.11 shows the allocation of two officers' dwelling time in different ROIs over the month. Officer "102PO" always stays in ROI No.3 in working time, while officer "619PO" kept visiting multiple ROIs. This exploratory analysis shows a clear spatial aggregation of activities in ROIs and temporal patterns of individual officers'



movements. To deepen the understanding of the urban people's activity patterns, a more advanced framework that incorporates the analyses into temporal, spatial and semantic aspects of human dynamics is necessary.

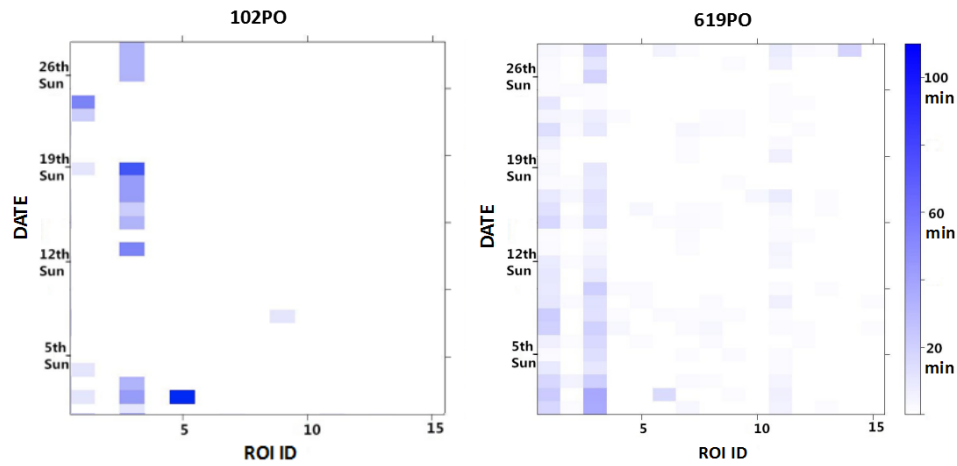


Figure 4.11 Heat map of dwelling time of officers "102PO" and "619PO" in different ROIs in February 2012

## Chapter 5

# The Euclidean paradigm

## 5 THE EUCLIDEAN PARADIGM<sup>2,3</sup>

### 5.1 INTRODUCTION

As first described in Chapter 3, our methodological framework consists of four modules to turn raw movement trajectories in the city into patterned activity groups in four steps. The modules are: space-time stay point identification, ST-ROI detection, semantic enrichment and aggregation of semantic profiles. This chapter describes in detail the methods used in the Euclidean paradigm of our frameworks.

In the Euclidean paradigm, all spatial objects are assumed to be in Cartesian space and all space related algorithms use Euclidean distance in calculations. In the studies of individual activities, the basic assumption is that people stop at a certain place to undertake various activities and then leave for the next place. Therefore, the stop episodes in an individual trip trajectory are of greater interest than the move episodes for the detection of interesting places (Palma et al., 2009). The first step of the paradigm therefore is a stay point identification module featured by a kernel-based temporal scanning window that we defined. The outputs of Module I, i.e. the stay points, are used as inputs to the ST-ROI detection module to find regions that attract people's high volume stopping behaviours in space and time. By doing this, the trajectories of an individual are simplified as a sequence of visited ST-ROIs by the person and information of the person's arrival, leaving and dwelling time in each ST-ROI during the trip. The (non-semantic) space-time profile (i.e. individual time budget allocation of ST-ROIs) of every individual can also be generalised after Module II. Before going on to describe Modules III and IV, the (non-semantic) space-time profiles are hierarchically clustered as a preliminary study and the intermediate outcomes are discussed to show the necessity of semantic enrichment in the following modules.

After the ST-ROIs are detected and the individual space-time profiles are generated, the semantic information of the POIs opening in the ST-ROIs are used to enrich the semantic meaning of the ST-ROIs in Module III. Specifically, the enrichment method uses the opening hours of the POIs to find the differences of semantic meaning of the same spatial region rather than look at the place's semantic meaning as a constant attribute. In Module IV, we use the semantically enriched ST-ROIs to turn the individual space-time profiles into semantic profiles for depicting an individual's activity pattern

---

<sup>2</sup> Part of this chapter was presented in: Shen, J. and Cheng, T., 2016. *A Framework for Identifying Activity Groups from Individual Space-time Profiles*. International Journal of Geographical Information Science, 30 (9), 1785-1805.

<sup>3</sup> Part of this chapter was presented in: Shen, J. and Cheng, T., 2015. *Clustering Analysis of London Police Foot Patrol Behaviour from Raw Trajectories*. Proceedings of GeoComputation 2015.

and use JSD (Jensen–Shannon Divergence) to define the dissimilarity of profiles for aggregative analysis of the activity patterns. The detailed description of methods in sections 5.2 to 5.5 are organised according to the four modules mentioned above. Together, they perform as four steps to transform raw trajectories into high level aggregation of the activity patterns.

We firstly apply the entire framework onto the 100 police officers’ movement data in one borough as the test case study to illustrate the detailed algorithms and outcomes of every module. After the outcomes of the single-borough test are discussed, an extended case study of police movements in the 12 (three) boroughs (Camden, Islington and Westminster) is designed to test the fully-fledged Euclidean paradigm. Compared with the single-borough case study in section 5.6, the extended case study show the framework’s potential to be scaled up for the analysis of larger datasets in larger areas and the advantages of semantic profiles over (non-semantic) space-time profiles.

Finally, we discuss the aggregation results and their accuracies at the end of this chapter, highlighting the areas that need to be addressed in the following chapter. These areas are then addressed by the network paradigm described in Chapter 6.

## **5.2 MODULE I: PRE-PROCESSING**

The pre-processing module includes two tasks: trip-segregation and stay point identification.

### **5.2.1 Trip segregation**

As explained in Chapter 3, trip segregation is a relatively simple task. We use a rule-based method here to segregate the sequence of individual GPS records into multiple continuous trips. In APLS and many other location logging systems, a person’s movement trajectories are collected by a GPS device, and his/her location and status information is updated and logged at a constant and regular sampling rate (every 5 minutes for APLS). Therefore, a long-time switch-off of the device can be regarded as the end of a trip and a later restart of continuous logging marks the beginning of a new trip. A travelling person may sometimes go into underground space or experience a poor GPS signal for many reasons; however, as long as the time of signal loss is not unreasonably long, the sequential records before and after the temporary signal loss will still be counted as one trip. Here we set the inactive time threshold to be one hour.

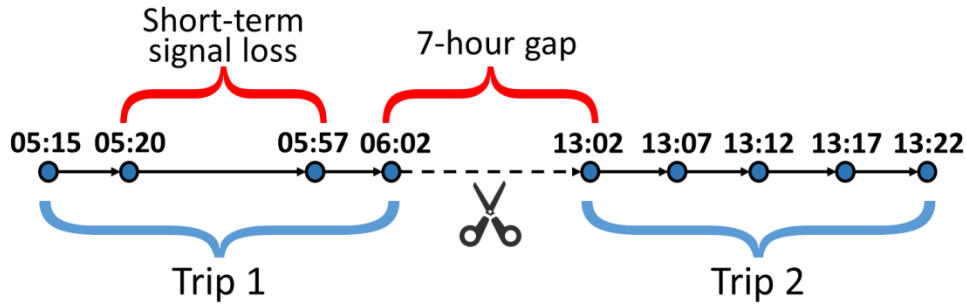


Figure 5.1 Trip segmentation example of one person's sequential location updates

An example of trip segmentation is shown in Figure 5.1. When the continuous location update of an officer pauses for a period shorter than one hour, the updates afterwards will still be regarded as records within the current trip. Such an update pausing period will therefore be considered as a stop episode within the trip and the last update before the pause will be considered as a stay point. If the update was missing for longer than one hour, the last point before the update stops will be marked as the end of a trip. When an police officer is at work, his location and communication device should always be on. According to the work of Cich et al. (2016), there are many “junk parts” of missing information in personal GPS records. The one-hour threshold is set to divided all the GPS records of one individual officer into multiple trips so that the missing location updates in the gaps between different trips cannot interrupt the correct calculation of dwelling time and the information of every trip is complete.

### 5.2.2 Stay point identification

Identifying stops in the trajectories is the first step in location-based activity studies (Alvares et al., 2007; Zheng et al., 2009). In this module, we aim to detect the stay point within every trip segmented by the previous module. A stay point, or stop episode, occurs when a person stops moving, stays stationary or moves slowly around a small area when the location updates continue in a trip. We have critically reviewed in Chapter 2 the existing stay point identification methods. The conventional (speed) threshold-based algorithm cannot detect “slowly-walk-around” stay behaviours or a complete stop with spatial noise (i.e. large positioning errors), while the conventional density-based methods ignore time and sometimes misidentify a place people visited at different times as a stop. However, erroneous locations cannot be avoided in GPS devices due to systematic errors and the effect of urban canyons. These errors can vary from some 1 m to 46 m in our case stay dataset. Because of these errors and the slight movements of a person's body, the updated coordinates can differ every time, even when the person holding the device completely stops moving his/her position.

Chapter 2 has described a variety of developed approaches for stop identification in a GPS track. Some of these methods are based on point density calculations, while the others set up thresholds of speed or distance between updates. Although most of the developed algorithms are simple and low cost in terms of computation, they cannot efficiently identify all stops, especially in an urban movement dataset with many positioning errors. They also fail to account for the structure of complex street networks. Another major limitation of density-based stop identification methods is relying solely on spatial clustering while ignoring the spatiotemporal nature of people's movements, which in turn leads to miscalculation of dwelling time and ignoring shorter stays.

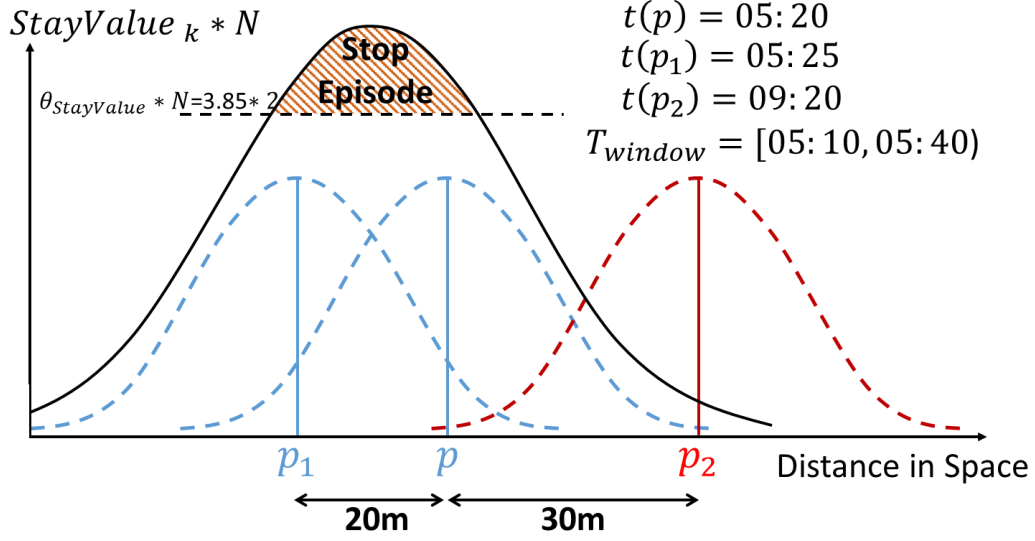


Figure 5.2  $p_1$  and  $p$  are in the temporal scanning window while  $p_2$  is not

Sila-Nowicka et al. (2016) introduced a temporal sliding window to kernel density based approach to integrate time information and spatial point density to overcome the above mentioned defects in stop identification. However, their method is only suitable for movement data with constant sampling rates as reviewed in Chapter 2. This feature prevented it from its implementation to trajectory records that often pauses or have changing sampling rates such as the APLS dataset. The APLS's location sampling and update can be done at any time on special demand of officers, which often means disruptions of the constant sampling process. In contrast to previous research, we attempt to apply a kernel-based temporal scanning window (KTSW) to make use of both spatial and temporal information of people's movements so that the accuracy of stop identification can be improved even in a movement dataset with changing sampling rate. Our approach is to use a 30-minute window to scan through every trip trajectory of people in time and undertake spatial kernel density estimations on the points within the temporal window. The size of the temporal window affects the performance of stop identification. If the size of the temporal window is too big, the sliding window will lose

its meaning of existence and misidentify a stop with another stop in the returning trip (see Trip 1 in Figure 2.5). If the window size is too small, the window will contain not enough of points for stop identification and misidentify a single stop episode with multiple stops or a move episode in case of large positioning errors (see Trip 2 in Figure 2.5). In our study, the size of the window is set as 30 minutes (i.e. one half of the trip segregation threshold), so that the window is smaller than the gaps between trips and can still contain 6 location points (if the device is working properly) for stop identification.

KDE is used to interpolate the points' adjacency to make a smooth and continuous value surface. We call the KDE value of every point on the smooth surface "stay value". This process is demonstrated in Figure 5.2, where the stay value of  $p$  is calculated.  $p_2$  is not included in the calculation of the stay value because it is not in the temporal window ranging from 05:10 to 05:40, although it falls in the spatial adjacency of  $p$ .

Because the APLS's EPE shows a skewed normal distribution in Chapter 4 (Figure 4.2), we use a Gaussian kernel (Equation 5.1) with a spatial searching bandwidth  $B$  of 66 m so that we can have a 99% confidence that the point with error is within the searching bandwidth of its actual location.

$$k(D) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{D}{B}\right)^2\right] \quad \text{Equation 5.1}$$

where  $D$  is the Euclidean distance between an arbitrary point in the temporal window and the kernel centre.

Given the Gaussian kernel function  $k$ , and time span  $T_{\text{window}}$  of the scanning window, the stay value (i.e. density estimate) of point  $p$  within a group of points  $p_1, p_2, p_3 \dots p_N$  in the temporal window can be expressed as Equation 5.2.

$$\text{StayValue}_k(p) = \frac{1}{N} \sum_{i=1}^N k(D_{p,p_i}), \quad t(p_i) \in T_{\text{window}} \quad \text{Equation 5.2}$$

where  $t(p_i)$  is the time stamp of  $p_i$  and  $D_{p,p_i}$  is the Euclidean distance from  $p$  to  $p_i$ .

When the temporal window slides through the point and the density estimated stay value at this point is larger than the threshold  $\theta_{\text{StayValue}}$  that we define, the point is marked as a stay point in the trip, as suggested in Figure 5.3. The erroneous noises that fall in the same high stay value area and the same scanning window as the detected stay points are also marked as stay points. It is noteworthy that the point with error is inside

the searching bandwidth with a 99% confidence. This means that  $\text{StayValue}_k(p)$  is a approximate value.

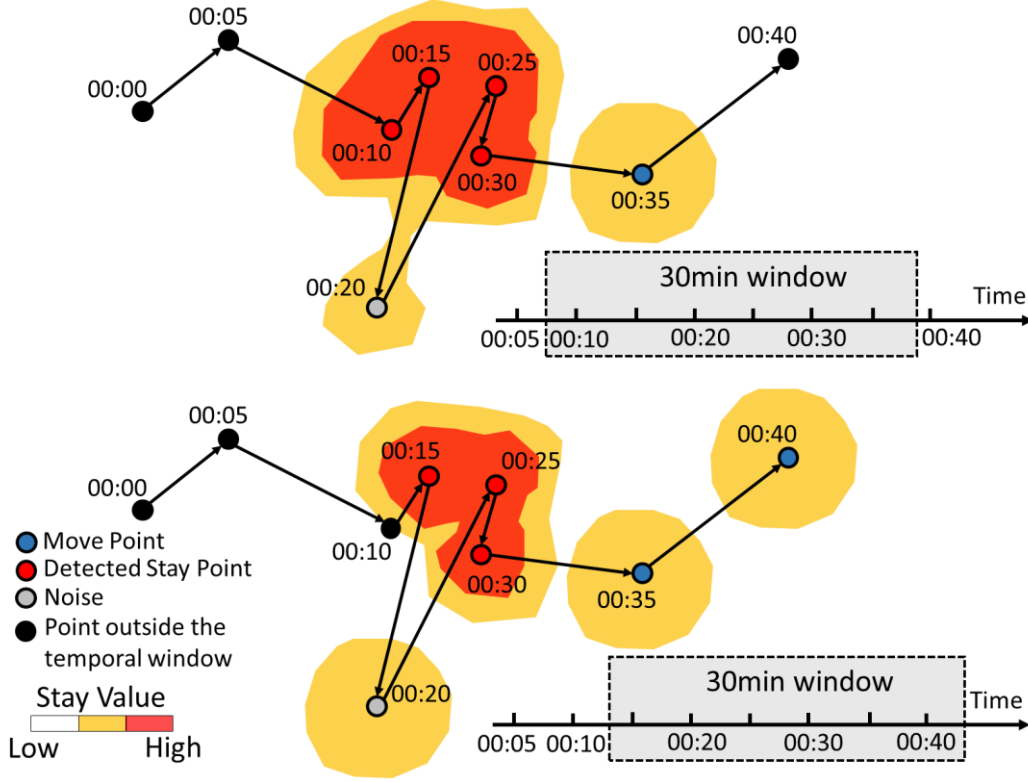


Figure 5.3 A kernel-based temporal window scanning through a trajectory

In the previous chapter, the recommended speed threshold for a threshold-based stop identification method is 0.2 m/s, which is equal to a 60 m displacement in 5 minutes.

We set our stay value threshold  $\theta_{\text{StayValue}} = \frac{k(0)+5*k(60)}{6} = 0.1854$  to make an equivalent stop identifying rule, so that the accuracy of our method can be compared with conventional methods in section 5.6. The denominator is  $6\theta_{\text{StayValue}}$  because there are normally 6 position update points of a trajectory in each 30min time window. Once all stay points are identified, we can then enter the next module to detect significant aggregation of these stay points in space and time.

### 5.3 MODULE II: ST-ROI DETECTION

All conventional approaches and models reviewed in Chapter 2 hold purely spatial views of the ROIs (hotspot areas) and completely ignore the temporal aspect of the activities in the region. These approaches failed to describe most places that only attract people's activities in a confined time period and places that change their functions or semantic meanings over time. In this module, we overcome this limitation by introducing Spatio-



Temporal Region of Interest for describing phenomena and activities with a time-varying view of places.

Evolving from the definition of ROI (introduced in Chapter 2), an ST-ROI is defined as a region intensively visited by people in a limited time period. In other words, an ST-ROI is a region having a dense aggregation of stay points in space and time. Every ST-ROI has its spatial boundaries, as well as emerge and perish times describing its time span.

### 5.3.1 ST-DBSCAN

Density-based clustering algorithm and its variations are the most commonly used methods for ROI detection. This widespread use arises because the working mechanism of the density-based clustering algorithm (DBSCAN) enables it to detect point clusters of arbitrary shapes without specifying the number of clusters in the data a priori. It also has a notion of noise and is tolerant to outliers. Moreover, because the algorithm can work directly with a database, the clustering process can be speeded up by optimising the query strategy in the database (Patwary et al., 2012). However, due to the purely spatial view of interesting regions in previous studies, all existing density-based methods for ROI detection only search for stay point aggregations in space as reviewed in Chapter 2. We therefore apply ST-DBSCAN (Birant and Kut, 2007) to detect ST-ROI and hence introduce a view of places with spatio-temporal ontology.

Among the many variations of the density-based approach to cater to different research purposes, ST-DBSCAN is an extension particularly developed to deal with space and time intervals comprehensively. Besides the advantages inherited from DBSCAN, ST-DBSCAN has features of its own to make it even more effective for detecting ST-ROIs. By extending the idea of traditional DBSCAN, the ST-DBSCAN not only sets up Maximum Reachable Distance (MRD) in space but also in time. Similar to the spatial search for neighbouring points in DBSCAN, a space-time searching cylinder scans through the adjacency of every point (Figure 5.4). Any stay point must satisfy the criteria of spatial maximum reachable distance (Spatial Eps) and temporal maximum reachable distance (Temporal Eps) simultaneously in order to be included in the spatio-temporal cluster. While other density-based methods can only use one MRD parameter for all types of variable no matter whether they share the same measurement units or not, ST-DBSCAN enables us to set spatial and non-spatial (temporal) MRD separately according to the nature of the moving data on which we are working.

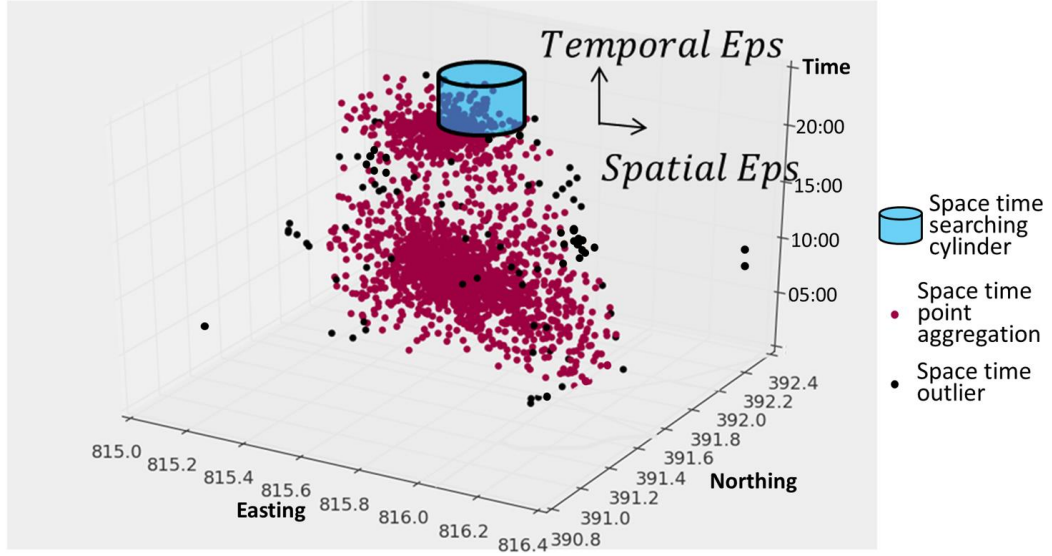


Figure 5.4 An example of ST-DBSCAN

As for determining the parameters in density-based clustering, many researchers have optimised the parameters by tuning according to the domain knowledge and aided by visual representation (Cortez et al., 2014). ST-DBSCAN has 3 parameters, namely Spatial Eps, Temporal Eps and MinPts (the minimum number of reachable points needed to form a new cluster). In a similar way, we firstly determine Spatial Eps and Temporal Eps according to the estimated GPS spatial error and time resolution in the APLS dataset. Then MinPts is defined in Equation 5.3, determined by calculating the neighbourhood of every point in the dataset, as proposed by Zhou et al. (2012).

$$\text{MinPts} = \frac{1}{n} \sum_{i=1}^n \text{nump}_i \quad \text{Equation 5.3}$$

where  $\text{nump}_i$  is the number of points in Spatial Eps and Temporal Eps neighbourhood of point  $p_i$ , and  $n$  is the total number of all the points.

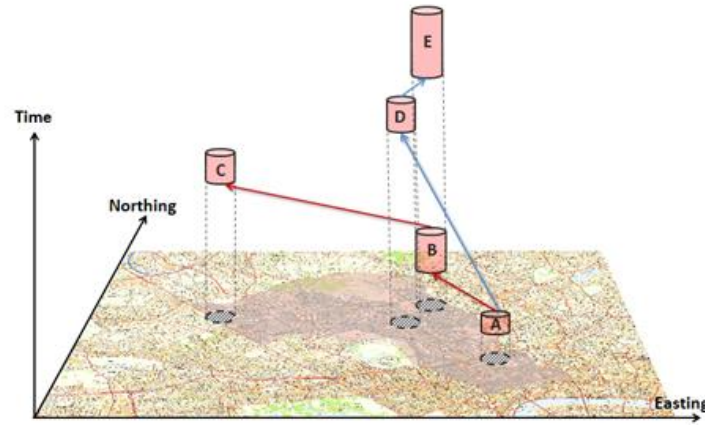
ST-DBSCAN is capable of clustering objects with a combination of both spatial and temporal measurements and detecting noise when different densities exist. These characteristics make ST-DBSCAN the best option to detect the location as well as the time span of ST-ROIs, revealing where ST-ROIs are, when they emerge and when they perish in a day. As far as the semantic meaning of the place is concerned, the time spans of ST-ROIs are of equal importance to their locations because interesting places are not always busy all day long and can become interesting for different reasons in different time periods. Therefore, it is possible for ST-DBSCAN to find places that are interesting for different groups at different times of the day. For instance, a district with bars and an underground station nearby can be busy in the morning peak because of commuters' intensive visits to the underground station, and then become lively again at midnight

when London underground trains stop their services and bars reach their business climaxes when people relax. The ST-ROIs can be visualised in a space-time cube (Andrienko et al., 2010) as shown in Figure 5.4.

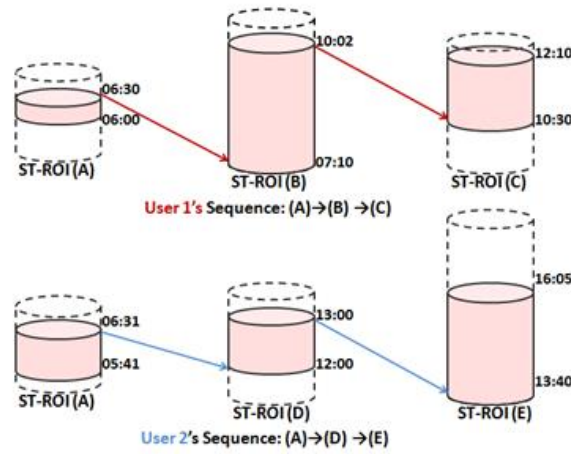
To ensure the speed of processing of large point datasets, a kd-tree (Wald and Havran, 2006) for space index acceleration is used in the ST-DBSCAN algorithm to optimise the neighbour searching strategy.

### **5.3.2 The simplified representation of trajectories**

Knowing the ST-ROIs and the stops and moves of every individual user, we can establish a model in which the time and spatial aspects are considered in a joint effort, similar to Parent et al.'s (2013) idea of modelling human movement by aggregated trajectories among a set of ROIs in a city. We describe an individual's trip process by noting when the user visits a particular ST-ROI and how long he/she stays before leaving for another ST-ROI and so on. The movement description of a user can be represented by the time he/she arrives at an ST-ROI and then leaves for another; thus, the whole movement of a user in a day is simplified as a series of ST-ROIs visited. As discussed in the literature review, temporal information is an important aspect of activity studies. In this way, we can not only preserve the information of location and sequence of location visits in the trips as Parent et al. (2013) (and many other Bluetooth movement studies) did, but also add time of visit and duration of stay (i.e. dwelling time) into the simplified tracks to fully describe the stop episodes in trips. This improvement, although not a fundamental contribution in our framework, bridges the gap between sequential study and time allocation study of activity patterns. Figure 5.5 shows how two persons' trajectories (Figure 5.5(a)) are simplified as two sequences of their respectively visited ST-ROIs (Figure 5.5(b)) with the time information recording when they arrive at and leave each ST-ROI. This simplified representation is also widely used in movement pattern studies (Zheng, 2011) but without taking the time span of interesting places into account.



(a)



(b)

Figure 5.5 The simplified representation of two example users' movements  
(a) the trajectory of two GPS users in space-time; (b) simplified movements with sequence of time-stamped ST-ROIs

Based on this simplified representation of individual movements, the daily behaviour routine of individuals in the study period can be expressed by how much time each user spends in different ST-ROIs. As in the example shown in Figure 5.5, A, B, C, D and E are the major ST-ROIs visited frequently by two users. The circular shadow of the ST-ROIs projected on the base map indicates their spatial boundaries. It can be noticed that B and E are spatially located at the same place but not at the same time; therefore, we see B and E as two independent ST-ROIs. In this example, user 1 keeps active from 06:00 to 12:10 in the day. He/she spends approximately 0.5 hours, 3 hours and 2 hours in ST-ROI (A), ST-ROI (B) and ST-ROI (C) respectively, while user 2 spends about 1 hour, 1 hour and 2.5 hours of the stopping time in ST-ROI (A), ST-ROI (D) and ST-ROI (E), respectively, from 05:41 to 16:05.

One of the advantages of this model over the purely spatial models is that it considers not only spatial information but also temporal and sequential factors so that more information can be discovered. In our model, although two ST-ROIs may be in the same spatial location, they can exist in different time periods with different time spans and have a clear temporal gap in between with no activity linking these two spatio-temporal entities as one. Therefore, the semantic meaning of these two ST-ROIs may be different. Taking Figure 5.5(a) as an example, ST-ROI (B) and (E) are in the same location, but user 1 visits ST-ROI (B) in the morning and user 2 visits ST-ROI (E) at night. ST-ROIs (B) and (E) locate at the same place but the purposes of visits can still be very different because of the differences in time.

### 5.3.3 Space-time profile and its similarity

In terms of similarity of activity patterns, it is assumed that individuals usually stop at places for certain objectives. Different social groups may have different preferences and habits that may lead to dissimilarities in their movement patterns and reactions to certain events (Chapin, 1974). Based on the patterns in which individuals stop at a series of places, various similarity metrics are proposed with emphases on different features of movements.

In our study, the stay durations and time budget allocations are the major concern. However, the information of sequences in which an individual visits different places can still be preserved. This is because the generated ST-ROIs include emerge and perish times and people can only visit an ST-ROI that exists early in the day first before visiting an ST-ROI that comes later. For example, a user can start his/her day in a coffee shop that is of great interest to lots of people in the early morning and then go to work in a business centre that becomes “interesting” afterwards.

In our proposed model, the basic assumption is that people of different socioeconomic compositions allocate time to different places and phases of the day for different pursuits in their everyday affairs. Not only the places, but also the time of the activity indicates the behavioural preference of the individual. For example, in Figure 5.5(a), B and E are ST-ROIs that are geographically located in the same place, but the reason why people visit them can differ at different times of the day.

With the model describing the behaviours of individuals as movements from one ST-ROI to another, the patterns of how users spend different percentages of their time in each of the ST-ROIs are acquired. Just like research that uses time allocation to indicate personal characteristics in behavioural studies (Kölbl and Helbing, 2003), we use the

profiles of time allocation in ST-ROIs (Figure 5.6), called ‘space-time profiles’, as a measure of activity features. The question now is how to quantify the pairwise similarity of the movement patterns based on these space-time profiles so that they can be used as a defined distance metric in the following clustering analysis. To satisfy the requirements of clustering analysis as well as the purpose of the behaviour comparison, discrete Jensen-Shannon Divergence (JSD) (Lin, 1991) is used to measure the dissimilarity of the time distribution profiles of two users.

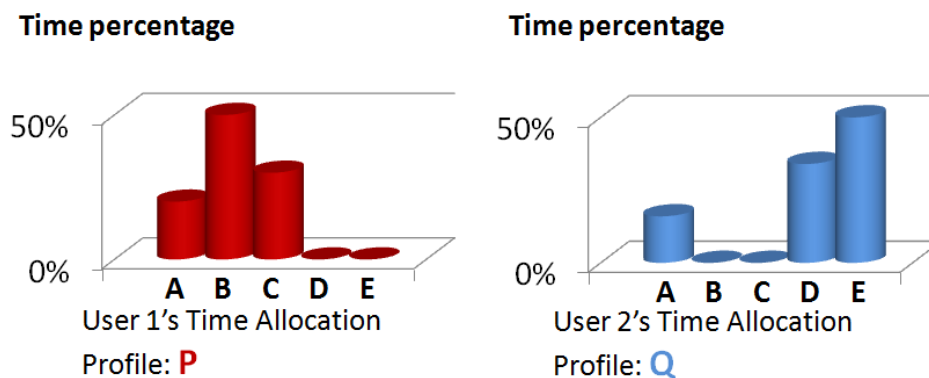


Figure 5.6 Histogram showing the percentage of the time two users allocate to ST-ROIs

Classic information theory concepts have the potential to be applied to new space-time data (Tsou, 2015). JSD, as demonstrated in Equation 5.5, is also known as the Information Radius. It is a popular method used in information theory and taxonomy in bioinformatics, measuring the dissimilarity of multiple distributions. In mathematical statistics, the Kullback–Leibler Divergence (KLD) (Kullback and Leibler, 1951) is a measure of how one probability distribution diverges from another expected probability distribution. In other words, it is the expectation of the logarithmic difference between two probability distributions. A KLD value of 0 indicates that we can expect similar, if not the same, behavior of two different distributions, whereas a KLD value of 1 indicates that the two distributions are so different that the expectation of difference approaches zero. JSD is an extension of the KLD, which is based on Jensen's inequality and the Shannon entropy. Some remarkable ameliorated properties of JSD make it especially suitable for our research:

- (1) Unlike the well-known Kullback divergences, JSD does not require the condition of absolute continuity of the distributions. It can be applied to discrete distributions just like the space-time profile shown as the percentage histogram in Figure 5.6.
- (2) Unlike many other similarity metrics used in information theory, the JSD between two distributions P and Q is symmetric, which means that  $JSD(P,Q)$  is equal to

JSD(Q,P). This symmetric characteristic is similar to the distances between objects, which enables it to act as a distance metric in clustering analysis.

- (3) The upper bound of the JSD has been proven to be no greater than 1 (Lin, 1991). These bounds are crucial for the definition of similarity.

The following equations show how the discrete version of the JSD is derived from the KLD (Equation 5.4). According to this equation, the JSD ranges between 0 and 1. The closer the JSD is to 1, the larger the difference between the two space-time profiles.

$$\text{KLD}(X||Y) = \sum_i X(i) \ln \frac{X(i)}{Y(i)} \quad \text{Equation 5.4}$$

where X and Y are two distributions to be compared by the KLD and X(i) is the i-th term in the distribution X.

$$\text{JSD}(P||Q) = \frac{1}{2} \text{KLD}(P||M) + \frac{1}{2} \text{KLD}(Q||M)$$

where P and Q are the two users' space-time profiles,  $M = \frac{1}{2}(P + Q)$ .

$$\text{JSD}(P||Q) = \frac{1}{2} \sum_i P(i) \ln \frac{2 \cdot P(i)}{P(i)+Q(i)} + \frac{1}{2} \sum_i Q(i) \ln \frac{2 \cdot Q(i)}{P(i)+Q(i)} \quad \text{Equation 5.5}$$

Whenever  $P(i)=0$ , the contribution of the i-th term to JSD is interpreted as zero because

$$\lim_{x \rightarrow 0} x \ln(x) = 0.$$

### 5.3.4 Hierarchical clustering of space-time profiles

With the JSD-based similarity metric (Equation 5.5), a dissimilarity matrix can be calculated. Each element in the matrix represents the pairwise dissimilarity of two users' profiles. This pairwise dissimilarity matrix can be processed by a hierarchical clustering algorithm for user segregation. The strength of hierarchical clustering is that any valid measure of distance can be used, including self-defined distance metrics. Furthermore, the observations themselves are not required: all that is used is a matrix of pairwise distances.

The number of clusters to be generated can be determined by the Dunn Index ( $DI_m$ ) (Dunn, 1973) that quantifies how well the dataset is separated. As defined in Equation 5.6, the Dunn index is the ratio of the minimal inter-cluster distance of m clusters to the maximal intra-cluster distance in each cluster:

$$DI_m = \frac{\min_{1 \leq i < j \leq m} \delta(C_i, C_j)}{\max_{1 \leq k \leq m} \Delta_k} \quad \text{Equation 5.6}$$

The Dunn index is also chosen as an evaluation metric to compare the hierarchical clustering results with other clustering methods in grouping users as shown in the preliminary single-borough case study (section 5.4). Grouping results can be demonstrated as a dendrogram (as shown in Figure 5.10) in a hierarchical structure. Besides histograms, the space-time profiles can also be visualised in space-time cubes as space-time point clusters to provide a more intuitive sense of difference (as shown in Figure 5.13).

#### **5.4 MODULE III: SEMANTIC ENRICHMENT OF ST-ROIS**

This module is a crucial step towards evolving Zhong's (2015) concept of "you are where you go" into "the place you go (Semantic ST-ROIs), when you go and how long you stay is who you are" that we propose. Chapter 2 reviewed situations of how a stop is related to the occurrence of outdoor activities, such as having a coffee break or shopping. Enriching the semantic meaning of the ST-ROI can help better answer "what the place/activity is about" and explain why individuals allocate their time differently among ST-ROIs. Moreover, more ST-ROIs will be generated if the framework is to be applied to a larger study area (e.g. all of London). The drastically increasing number of ST-ROIs means that the time allocation profile will contain many more variables than in the previous methodological framework, which will lead to the "curse of dimensionality" (Bellman, 1961) in the following clustering process. By applying the newly proposed method in this work, the number of dimensions of the profile can be limited to a small number of certain semantic categories and the problem of "curse of dimensionality" in large study areas can be avoided.

The semantic meaning of an ST-ROI is described by a summary of different POI categories' semantic contribution to ST-ROI. However, the quantity unbalance of POIs of different categories (see section 4.2) causes great bias in this description and should be mitigated. To achieve this, we annotate the POI semantic information to the ST-ROIs through the following steps:

- Firstly, the spatial boundary of each ST-ROI is defined by a 20m buffer zone of the convex hull covering all the stay points in the ST-ROI. The temporal boundary is the time span of the ST-ROI. Together, these two boundaries define the space-time boundaries of ST-ROIs.
- Secondly, the buildings and POIs that are located in the convex hull's buffer zone are extracted. The overlapping duration of every extracted POI's opening hours and the ST-ROI's time span is calculated. The quantity and the overlapping duration of



different categories of POIs decides different POI categories' raw semantic contribution to the ST-ROI.

- Lastly, the raw semantic contributions of various POIs are reweighted by a term frequency-inverse document frequency (TF-IDF) algorithm to generate different POI's categorical semantic contribution to the ST-ROI they are in and transform the ST-ROIs into semantic ST-ROIs.

#### 5.4.1 Space-time boundaries of ST-ROIs

The ST-ROIs are essentially aggregations of stay points and can be visualised with a 3D point distribution map in a space-time cube. Point distribution maps represent spatial distribution of geo-referenced data using points as a basic graphical element (Slocum, 2009). Every point represents a datum with geo-location information. The points can only show the density of an area, rather than depict the clear boundary and exact location of the area. We must therefore implement a method that allows us to define the spatial and temporal boundary of the ST-ROI before explaining the semantic meaning within the area.

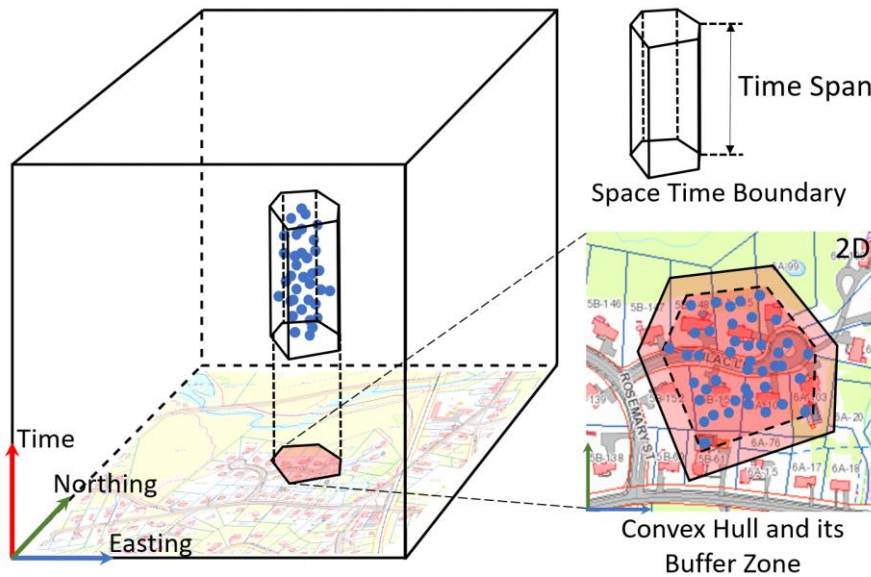


Figure 5.7 An ST-ROI and its space-time boundary

Convex hulls are usually used to turn point-based objects into spatial areas, and the minimal convex hull bounding method can create a polygon area enclosing all the stay points of each ST-ROI, making it the most ideal way to serve this purpose (Andrew, 1979). In this step, we used the parallel spatial retrieving method (Miller and Stout, 1988) to find the convex hulls that define the spatial boundaries of the ST-ROIs. To incorporate points with positioning errors, a 20m buffer zone is used to define the actual

coverage area of each ST-ROI. This buffer distance is set with the consideration that all the recorded GPS points have spatial errors and 20 m is the mean value of EPE in our APLS dataset. The time span of an ST-ROI is its temporal boundary. An example of an ST-ROI and its space-time boundary is showcased in Figure 5.7.

#### 5.4.2 POIs in the space-time boundaries

A POI dataset contains the information of all the public buildings (POIs) that can be summarised to interpret the semantic meaning of a place in which they exist (Alvares et al., 2007; Alves, 2011; Braun et al., 2010). In order to understand the staying behaviour within each ST-ROI area, we used POI data to depict the functional images of the ST-ROIs and enrich the semantic meaning of users' visits to these ST-ROIs. The POI dataset used in the case study contains the information of a wide range of finely categorised infrastructures and buildings that offer different services and utilities (Ordnance Survey, 2016). Similar to the hierarchical category structure of POI information used by Krüger et al. (2015) and Yan (2013), we made slight changes to the official Ordnance Survey POI classification scheme to make an 11-category classification scheme as shown in Table 4.2, and use it as our POI categories for the semantic enrichment of ST-ROIs. After the space-time boundaries of all ST-ROIs are found, we use a spatial query to find POIs that locate inside the expanded convex hull of every ST-ROI (Figure 5.8). If a POI is outside the spatial boundaries or its opening hours do not overlap with the ST-ROI's time span, the POI contributes no semantics to the meaning of the ST-ROI.

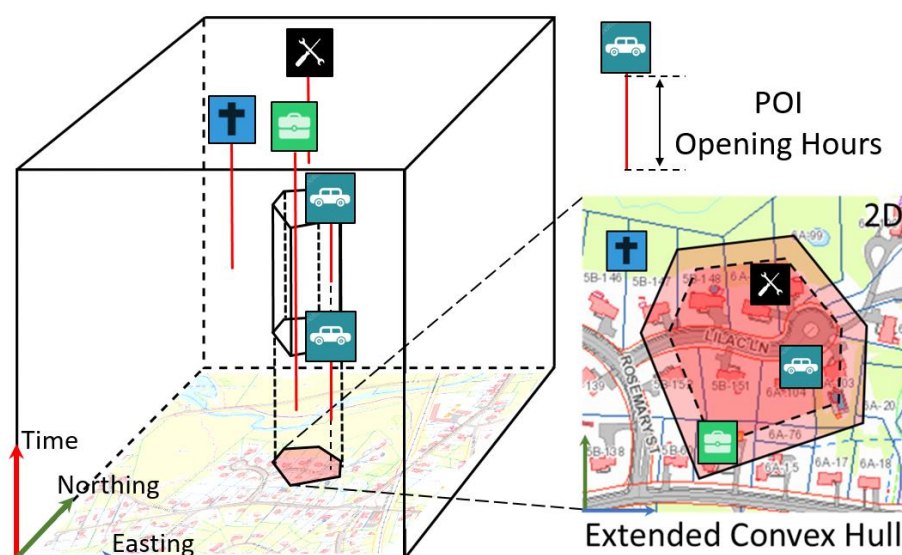


Figure 5.8 The space-time relationship between POIs and ST-ROIs

We assume in our model that the longer a POI's opening hours overlap with an ST-ROI, the more semantic meaning the POI will contribute to the ST-ROI. Besides, the more the number of POIs of the same category falling in the spatial boundary of a ST-ROI, the more the contribution that this category of POIs will have in the semantic meaning of the ST-ROI. It can be seen in Figure 5.8 that some POIs open multiple times in a day. We also take this situation into account and see the two periods of opening hours of a POI as two homogenous POIs that open at different times of the day and calculate their semantic contributions separately.

Hence, we revise Equation 2.2 and define the raw semantic contribution  $w$  of a POI sub-category in a ST-ROI with Equation 5.7.

$$w_{i,j} = \frac{\text{count}_{i,j}}{\sum_{i=1}^{52} \text{count}_{i,j}} * \frac{\sum \text{OverLap}_{k,j}}{\text{TimeSpan}_j} \quad \text{Equation 5.7}$$

where  $i$  is one of the 52 POI sub-categories,  $\text{count}_{i,j}$  is the number of POIs of sub-category  $i$  in ST-ROI  $j$ ,  $k$  is the POIs of sub-category  $i$  and  $\sum_{i=1}^{52} \text{count}_{i,j}$  is the total number of POIs in ST-ROI  $j$ .  $\text{TimeSpan}_j$  is the time span duration of ST-ROI  $j$ ,  $\sum \text{OverLap}_{k,j}$  is therefore the sum of the overlap times of all sub-category  $i$  POIs in ST-ROI  $j$ .

#### 5.4.3 Reweighting POIs for semantic annotation of ST-ROIs

However, quantities of different categories of POIs vary dramatically in urban space. For instance, a large number of iconic public telephones and 24-hour cash machines can be found throughout London, but they have a relatively small influence on the meaning or function of an area. On the contrary, if there is only one museum in the entire city, the influence of this museum upon the local region where it is located should be magnified to outrank the many telephone boxes nearby. Hence, directly using the quantity of POIs for semantic enrichment is not enough. The bias caused by unbalanced quantities of different POIs should be subdued.

A similar case can be found in text mining studies where article words like “the” and “a” appear far more frequently than the truly meaningful words in most sentences. The significance (semantic contribution) of a given word in one sentence increases proportionally to the number of times this word appears in the sentence, but is offset by the frequency of the word in the whole context. Likewise, the more sibling POIs (i.e. POIs that belong to the same major category) fall into an ST-ROI's expanded convex hull, the more impact they have on a place; however, the more sibling POIs exist in the entire district or other places, the less impact the POIs should have on the current place.

In other words, POIs distributed ubiquitously are less relevant to the meaning of the local area.

In information retrieval and text mining studies, Term Frequency–Inverse Document Frequency (TF-IDF) is designed to measure the semantic contribution (weighting factor) of a word to the meaning of the sentence it falls in (Salton and Buckley, 1988). TF-IDF can downplay the semantic contribution of a word if it appears everywhere in the entire article. Inspired by its function in semantic analysis of articles and documents, we introduce the TF-IDF method to reweight and readjust the weight of different sub-categories of POIs to each ST-ROI so that the negative effect of dominant insignificant POIs can be eliminated.

In our case of study, a double normalisation weighting scheme was chosen for term frequency (TF) calculation and a classic inverse document frequency (IDF) weighting scheme was chosen for IDF calculation, so that the original TF-IDF for text mining was amended to process the semantic POIs, as in Equation 5.8. All POIs in the study area are regarded as the corpus of an entire article in text mining. Every sub-category of POI was equivalent to a word in the article and the combination of the sub-categories of all POIs in one ST-ROI's space-time boundary was considered to be a document sentence in the article. The process of semantic enrichment of ST-ROIs is equivalent to inferring the semantic meaning of a document sentence in text mining. When using the TF-IDF for this purpose, the reweighted semantic contribution of each POI sub-category  $i$  to the ST-ROI  $j$  is:

$$\text{TFIDF}_{i,j} = (0.5 + 0.5 \frac{w_{i,j}}{\max_i w_{i,j}}) \cdot \log \frac{N}{n_i} \quad \text{Equation 5.8}$$

where  $i$  is one of the 52 POI sub-categories and  $n_i$  is the number of ST-ROIs that contain sub-category  $i$  POIs in their expanded convex hulls.  $w_{i,j}$  is the raw semantic contribution of  $i$  to ST-ROI  $j$ .  $\max_i w_{i,j}$  is the semantic contribution of the POI sub-category that contribute the largest  $w$  to ST-ROI  $j$ .  $0.5 + 0.5 \frac{w_{i,j}}{\max_i w_{i,j}}$  is called augmented frequency. It can prevent a bias towards longer documents (i.e. ROIs with more POIs inside their boundaries).  $N$  is the total number of ST-ROIs.  $\log \frac{N}{n_i}$  is called inverse document frequency. It is a measure of how much information the POI provides, that is, whether the kind of POI is common or rare across all ROIs. According to Equation 5.8, the minimum value of TFIDF is zero and there is no upper limit for TFIDF.

As shown in Equation 5.9, the semantic contribution of the 52 POI sub-categories was weighted by TF-IDF and summed to generate the semantic contribution of the 11 major categories according to the major category to which they belonged.

$$SC_{I,j} = \sum_{i \in I} TFIDF_{i,j} / \sum_{l=1}^{11} \sum_{i \in l} TFIDF_{i,j} \quad \text{Equation 5.9}$$

where  $I$  is one of the 11 major categories. One major category of POI's semantic contribution to an ST-ROI is the normalised sum of the semantic contribution of all the sub-categories within a major category.

## 5.5 MODULE IV: SEMANTIC PROFILING AND HIERARCHICAL CLUSTERING

In module II, users' behaviour profiles were generalised based on their time budget allocation (i.e. space-time profiles) in the ST-ROIs. In this way, the users' inclinations in different time periods and spatial locations were discovered. However, this method did not consider the semantic meaning regarding the generated ST-ROIs. Spending time in two different places does not necessarily indicate any differences in visit purposes because the two places may have homogeneous functions or semantic meanings, even though they are far apart. Moreover, the non-semantic space-time profile cannot work in large cities because more and more ST-ROIs will be detected as the study area expands dramatically, and the difference between users' time allocation profiles will be erased. By summarising ST-ROIs into a limited number of categories according to their semantic and functional meanings, the time budget allocation in ST-ROIs can be translated into semantic profiles that demonstrate the time spent on different semantic places instead of meaningless locations.

In order to understand how people act in different functional areas at different times of the day, and especially how officers performing different roles on patrol allocate their attention across different activities in our case study, the TF-IDF reweighted semantics of the ST-ROIs were added to the analysis. One officer's staying time in an ST-ROI was assigned to the 11 categories of POI according to each POI's semantic contribution to that specific ST-ROI, as Equation 5.10 describes. This process turns space-time profiles (i.e. time allocations across ST-ROIs) into semantic profiles (i.e. allocation across functional areas), so that the differences in higher activity levels can be revealed.

$$SP_{o,I} = P_{o,j} \cdot SC_{I,j} \quad \text{Equation 5.10}$$

where  $P_{o,j}$  is the space-time profile of officer  $o$ , and  $SP_{o,I}$  is the semantically-weighted profile of officer  $o$  in the 11 POI major categories.

After semantically profiling the users (officers), the pairwise differences among the officers' TF-IDF weighted profiles can be again quantified with Jensen Shannon Distance (JSD) (Lin, 1991) and similar profiles could be grouped together as a result of the hierarchical clustering method. Equation 5.11 shows how the JSD value between semantic profiles of two people,  $SP_P$  and  $SP_Q$ , is calculated.

$$JSD(SP_P|SP_Q) = \frac{1}{2} \sum_i SP_P \ln \frac{2 \cdot SP_P}{SP_P + SP_Q} + \frac{1}{2} \sum_i SP_Q \ln \frac{2 \cdot SP_Q}{SP_P + SP_Q} \quad \text{Equation 5.11}$$

Once the dissimilarity metric of semantic profiles is settled, the profiles can be hierarchically clustered by the same method as demonstrated in section 5.3.4 and people sharing similar semantic profiles can be aggregated. The modules are firstly demonstrated in the following single-borough case study to demonstrate the work flow in greater detail.

## 5.6 A SINGLE-BOROUGH CASE STUDY

This preliminary study took place in the Camden Borough (Figure 4.7), which lies to the north of central London, United Kingdom. Five major police stations are located in this region, namely West Hampstead, Hampstead, Kentish Town, Albany Street and Holborn. We use the same dataset of police movement as collected in February 2012. This is the same dataset used for basic data exploration in section 4.3.

We choose the APLS movement dataset of this time period as a preliminary test because this was the first movement dataset that the Metropolitan Police declassified and made available for us at the beginning of our research. Another reason is that there was a major security related incident in February 2012 and this strongly influenced the activity patterns of officers in Camden. This incident will be used to explain the results of Module II in section 5.4.4. We also choose the top 100 active officers with a high number of continuous and frequent GPS records as the members of the pilot study. This is because we want to show the clustering result of all officers' profiles in the pilot study, and the result of having too many officers is impossible to visualise. In the afterwards extended case study, however, the data of all officers with enough records in APLS will participate in the analysis.

### 5.6.1 Module I: Pre-processing

The methods for trip segmentation and stay point identification described in section 5.2 are applied to the APLS data sample of February 2012. Trips containing less than three records and trips with their spatial location more than 10 km away from the study area

are considered as errors and are hence removed. If the time gap between two contiguous GPS records of a user is longer than one hour, the two records are assigned to two separate trips.

Like many movement dataset, APLS data is automatically collected and does not include manual surveys of information of the moving individuals. Because of this, the stopping behaviours in the movement cannot be verified by the moving individuals themselves and the validation of the stop identification accuracy is not feasible based on the original APLS data. To test the increased accuracy of the stay point identification caused by the KTSW method, we can only rely on artificial trajectory data to provide synthetic ground truth, synthetic raw trajectories and simulated positioning errors of the movements. To this end, an artificial trajectory generator is designed with its working process described in section 8.1.1. The KTSW has demonstrated prominent higher accuracy than conventional methods. The detailed performance comparison of existing classic methods and the proposed methods in the thesis for stay point identification can be found in Chapter 8.

### **5.6.2 Module II: ST-ROIs in foot patrol activity**

Just as places can have different meanings to people at different times, a similar situation also applies to police patrol activities. Many ROIs of officers have their own life spans and the patrols in a day are divided into 3 shifts (i.e. early, late and night shifts), each lasting for about 8 hours to give each officer proper working hours. Therefore, a place can only be meaningful to the officers during their working hours and the officers can go to the same area to perform different tasks at different moments in time.

To detect the high-density aggregation of stays in space and time, we applied ST-DBSCAN to the stay points of all officers to detect their common ST-ROIs. In ST-DBSCAN, Spatial Eps is set to be equal to the sum of the mean and twice the standard deviation of EPE of APLS. This is to make sure that 95.4% of the positioning errors can be offset. Temporal Eps is set to be equal to the minimal sampling interval of an individual officer's device so that the algorithm can be agile enough to detect a single officer's intensive activity in one place and identify this place as an ST-ROI. Then, MinPts is set based on Equation 5.3 after Spatial Eps and Temporal Eps are set. With ST-DBSCAN, 28 clusters were detected as ST-ROIs (Figure 5.9). It can be seen that although the town centre is an interesting place, it does not always draw the officers' attention throughout the entire day. It also shows that some ST-ROIs are outside the boundary of Camden, and we will discuss the reason for this in section 5.6.4. The movement of each officer is then captured as the movements and stays from one ST-

ROI to another. Dwelling time allocation profiles (i.e. space-time profiles) generated from this representation are then compared pairwise to group the officers.

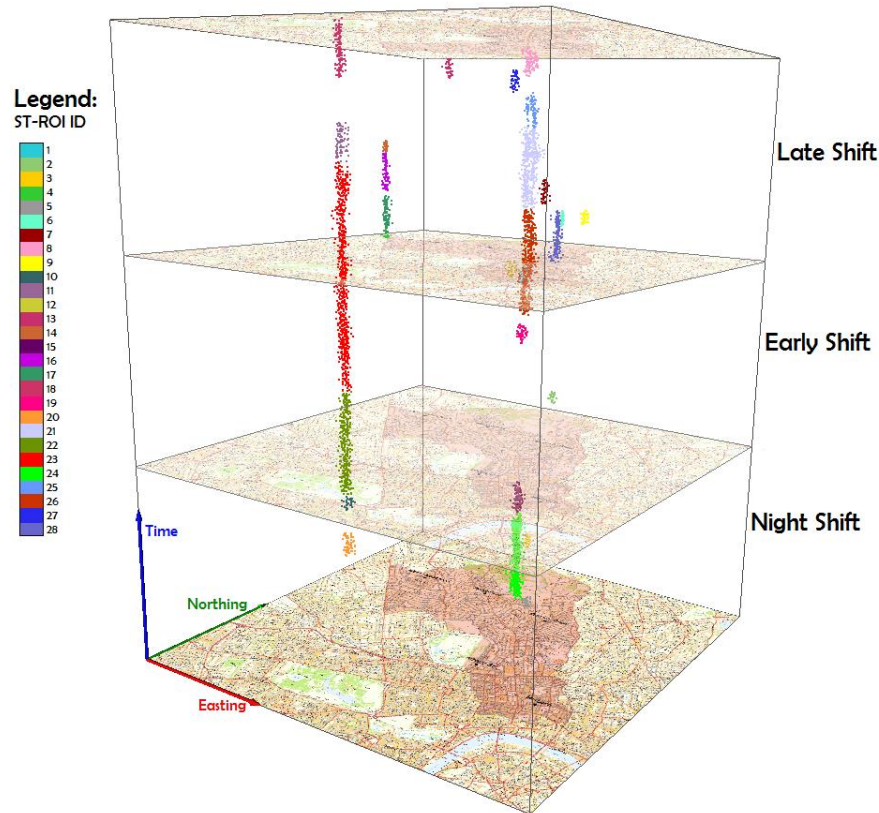


Figure 5.9 One typical working day of officers is separated into 3 shifts. 28 ST-ROIs of foot patrol officers are detected by ST-DBSCAN and are labelled with different colours (Shen and Cheng, 2016)

### 5.6.3 Intermediate outcomes: Space-time profiles

Using hierarchical clustering, the entire officer community in Camden was segregated into a dendrogram structure visualised in Figure 5.10, with the identification numbers representing each unique officer. The tree can be cut at certain places according to the condition that the researcher defines in order to separate the whole dataset into several clusters. In this research, we use Dunn Index as this condition.



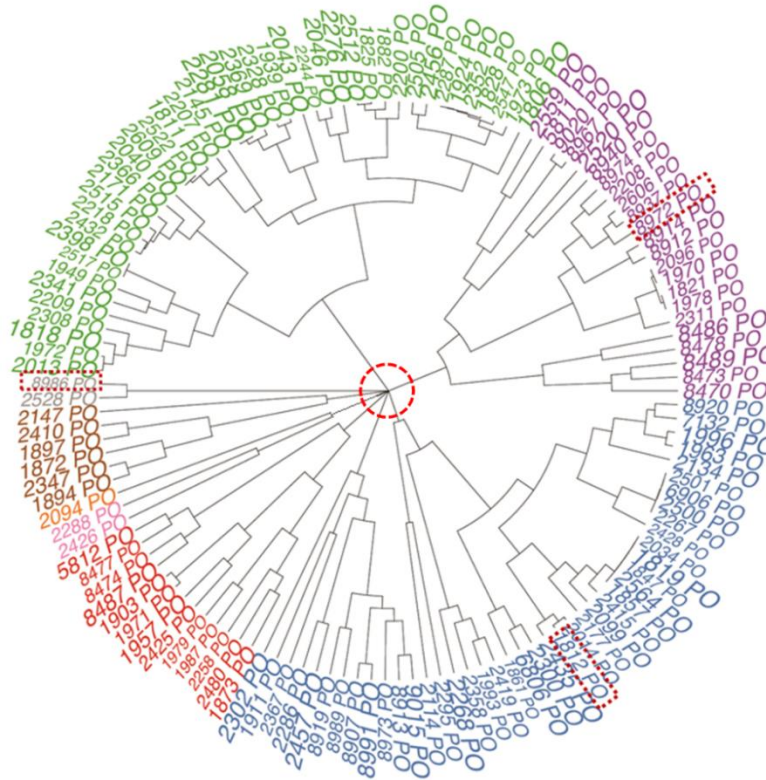


Figure 5.10 Dendrogram showing the clustering results of officers with different patrol patterns (Shen and Cheng, 2016)

To test the performance of our proposed similarity (Equation 5.5) based upon time allocation to ST-ROIs, we compared the hierarchical clustering results with those generated by using the similarity metric based on purely spatial ROIs (Equation 2.6) proposed by Zheng (2009). Figure 5.11 shows the performance comparison using the Dunn Index.

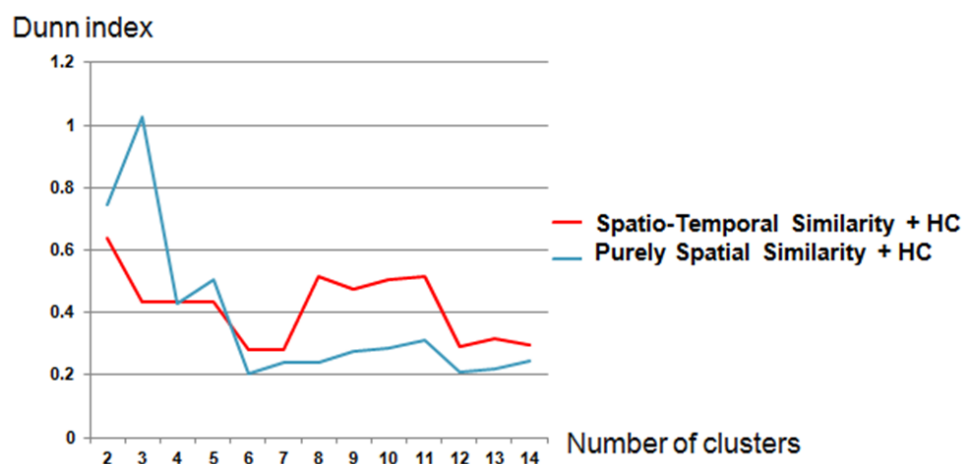


Figure 5.11 Evaluation of hierarchical clustering results based on two different similarity metrics (Shen and Cheng, 2016)

The similarity based on only spatial ROIs demonstrates better segregations when the cluster number is less than four, but it falls below the performance of our proposed metric based on space-time profiles with higher cluster numbers. This is partly because the number of detected spatial ROIs is much less than ST-ROIs and the distribution of each user's number of visits to each ROI is therefore much simpler and adapted to segregation of lower cluster numbers. However, a small-cluster-number clustering is not appropriate for semantic explanations of behaviours since a binary or ternary segregation will separate people into groups that are too simple to make sense. For instance, if the officers are only separated into two groups, i.e. one group that only moves inside the Camden border and another group outside, much potential valuable information will be subsumed. According to the cluster number determination method proposed by Salvador and Chan (2003) and the Dunn index evaluation, the number of officer subgroups generated by the hierarchical clustering based on the proposed similarity metric is set to be 8. The red dashed circle in Figure 5.10 is the place where the tree was cut so that the officers are segregated into 8 subgroups. The three officers highlighted by red rectangles in Figure 5.10 will be further discussed in section 5.6.7.

#### **5.6.4 Explanation of the outcomes**

It should be noted that effective segregation of the data does not necessarily indicate that the result will make sense in terms of having a reasonable semantic interpretation. To discover the semantic meaning of the generated cluster of space-time profiles, additional information and further study is required. By pinpointing the stay points of each cluster of officers and associating them with building and land use information, the semantic meanings of these differences are revealed. In this section, we try to explain the findings of Modules I and II by manually accessing the data of public POIs in the adjacency of the detected ST-ROIs.

For security reasons, we cannot present all the 8 clustered officer subgroups, although 4 of them were randomly chosen as examples to demonstrate the results. Figure 5.12 shows the average time percentage allocation to 28 ST-ROIs of the 4 chosen officer subgroups, subgroups I, II, III and IV. Each column in the histograms represents the percentage of time one officer subgroup has spent on one corresponding ST-ROI.

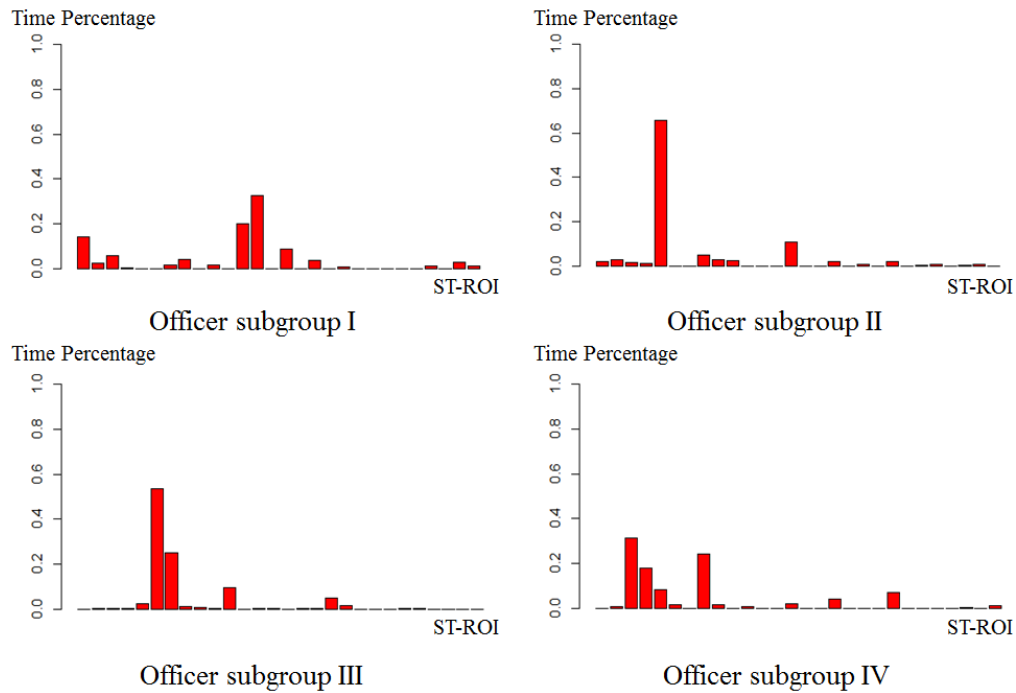


Figure 5.12 Histograms of average space-time profiles of different officer subgroups

For a more direct and concrete understanding of the discovered differences between the time allocation patterns of the subgroups, the stay points of the 4 example subgroups are visualised in space-time cubes in Figure 5.13 to show how different subgroups behave differently in both space and time. The base maps in Figure 5.13 depict the boundary of Camden and the circles marked with numbers and dashed circles on the base maps indicate the locations of the stay point clusters in space. By associating them with public points of interest data provided by the Ordnance Survey, we identified spatial region No. 1 as an underground station on a commercial street and spatial region No. 3 as an underground station in residential area. The location (spatial region No. 2) that exists in all the 3 graphs is the centre of Camden, which is a place with bars, restaurants, a large market and a busy London Underground station. It is a highly populated area and many crimes occur there. Some foot patrol officers spend a long time in Camden town centre because they believe that high visibility has a positive impact on public confidence and acts as deterrent to potential criminals. Region No. 4 is Belgrave Square, an embassy area outside Camden's border; the below mentioned Syrian embassy is located in this area.

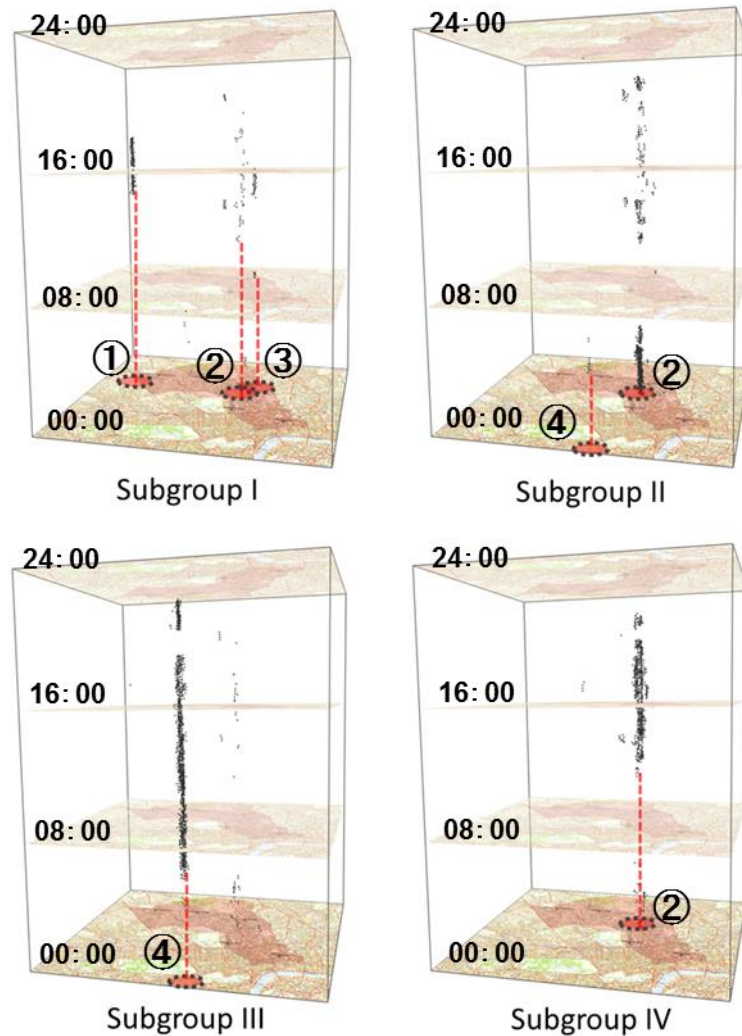


Figure 5.13 The stay points of 4 chosen generated officer subgroups.

(① Underground station; ② Central Camden; ③ Underground Station; ④ Syrian embassy in Belgrave Square)

It can be seen in Figure 5.13 that subgroup I has a special interest in regions No. 1, No. 3 and sometimes No. 2 during the afternoon peak periods. All of the three regions have underground stations. The interpretation of this behaviour pattern is that some officers are assigned to focus on peak locations at peak times for high visibility and crime reduction, and London Underground stations are often their typical targets.

The aggregation of police force, especially officer subgroup III in this study period, February 2012, was confirmed by the news that hundreds of violent protestors trying to get into the Syrian embassy clashed with the police, and that the police arrested several protestors overnight (Daily Mirror News, 2012). It was also explained by the metropolitan police that when there are big events in neighbouring boroughs and extra

manpower is needed, officers may be ordered to go out of their own borough to help. We can see that officer subgroup III spent most of their stay time from the early morning throughout the day in this embassy area while other subgroups spent very little, if any, time there.

It is noteworthy that officer subgroups II and IV were both interested in region No. 2, the centre of Camden Town. Existing methods that are solely based on location history will not distinguish between them. However, the proposed similarity metrics still managed to distinguish these officers as two groups because their time of visit, length of stay and visit intensity to this common place differed. Officer subgroup II tended to pay frequent visits to central Camden Town at the beginning of the day from 00:00 to 04:00. This phenomenon can be explained by officers keeping an eye on alcohol related and recreational activities in this area at night. In contrast, officer subgroup IV preferred to appear at this area in the afternoon and to stay for a longer time for a different purpose, namely to maintain a visible presence in an area with large flows of citizens visiting the underground station and shops. Similar analyses can be carried out to explain the patterns of the other subgroups.

Besides, the information of the officer types contained in the APLS dataset can be used for validation as well. Most of the officers on duty are foot patrol officers (FP), community support officers (CSO) and senior officers (SO). The behaviour of different types of officers can be very different because different tasks they are asked to undertake are determined by their types. Figure 5.14 shows the percentages of these 3 officer types in the 4 generated officer subgroups. Officers in subgroup I focus on multiple ST-ROIs and they consist mainly of foot patrol officers and senior officers, while officers that are temporally seconded outside Camden to assist the security work in the embassy area are all foot patrol officers. We have also verified from the Metropolitan Police that FPs are interested in multiple places distributed in Camden and only they can be seconded to do the work outside Camden. The most interesting phenomenon is again seen in the comparison between subgroup II and IV. With the two subgroups concentrating their efforts in the same place (spatial region No.2 in central Camden) but in different time periods, the contribution percentage of foot patrol officers and community support officers within the two subgroups reversed. It is pointed out by the field expert in the Metropolitan Police that the nature of community CSOs' work is to help the FPs at peak places in peak periods and the CSOs do not have much work at night. Similar analyses can be conducted to explain the patterns of the other subgroups.

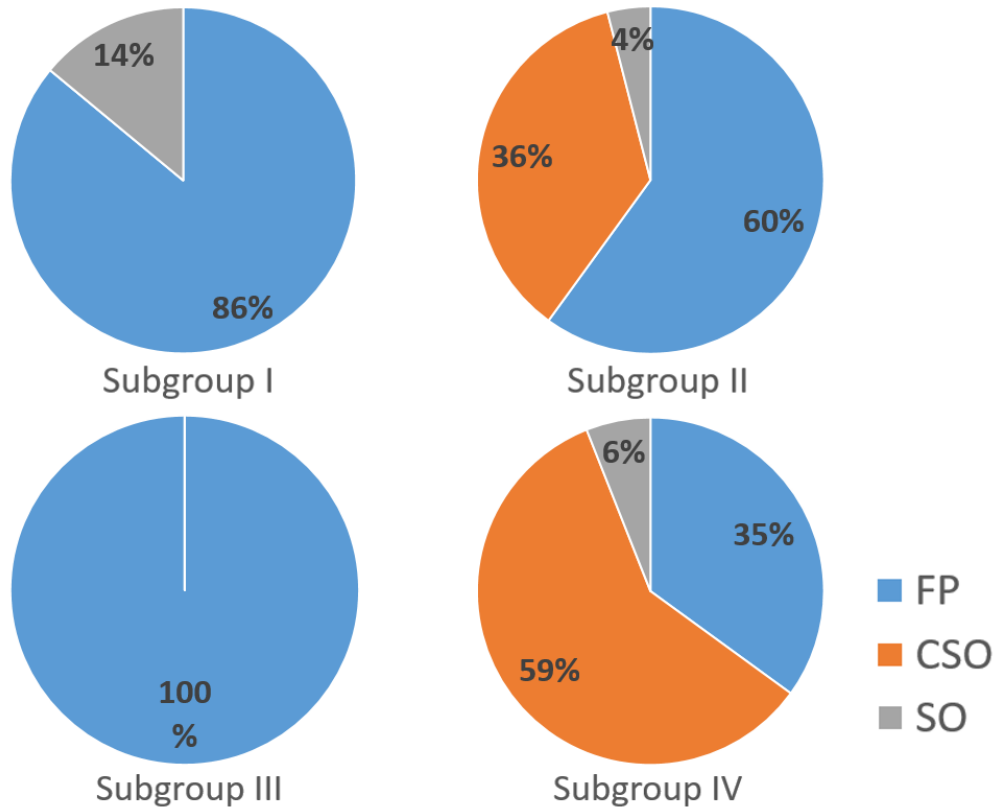


Figure 5.14 They stay points of 4 chosen generated officer subgroups

### 5.6.5 Discussion: The need for semantic enrichment

Modules I and II of our framework extend the traditional ideas of time budget allocation in behavioural studies and existing spatial-location-based user similarity definitions. It is possible to profile the activity patterns of people according to both space and time aspects by defining a new moving behavioural similarity metric. However, the semantic aspect of the places is not accounted for in these two modules. This means that similarity of individual space-time profiles does not directly translate into semantically similar activity patterns and we still need to manually ratify the meaning of places to police officers' activities as we did in section 5.6.4. Moreover, when we need to aggregate activity patterns in larger areas (e.g. a much bigger city with huge number of interesting places for clustering and grouping), there will be too many features in the (non-semantic) space-time profiles for the hierarchical clustering methods to achieve a clear and reasonable segregation.

Our solution for these problems is to look into the semantic meanings of places and summarise all the ST-ROIs into semantic ST-ROIs of a limited number of generic categories. After the semantic enrichment process in Module III, human dynamics data with more ST-ROIs in a larger area can be aggregated. Module III will also enable us to

detect similar activity patterns even though they happened in mutually far apart locations.

### 5.6.6 Module III: Semantic enrichment of ST-ROIs

After 28 ST-ROIs are detected by ST-DBSCAN in Module II, the space-time boundaries of all ST-ROIs are generated by the method described in section 5.4.2. Among all ST-ROIs, ST-ROI No.23 and its space-time boundary are shown in Figure 5.15 as an example.

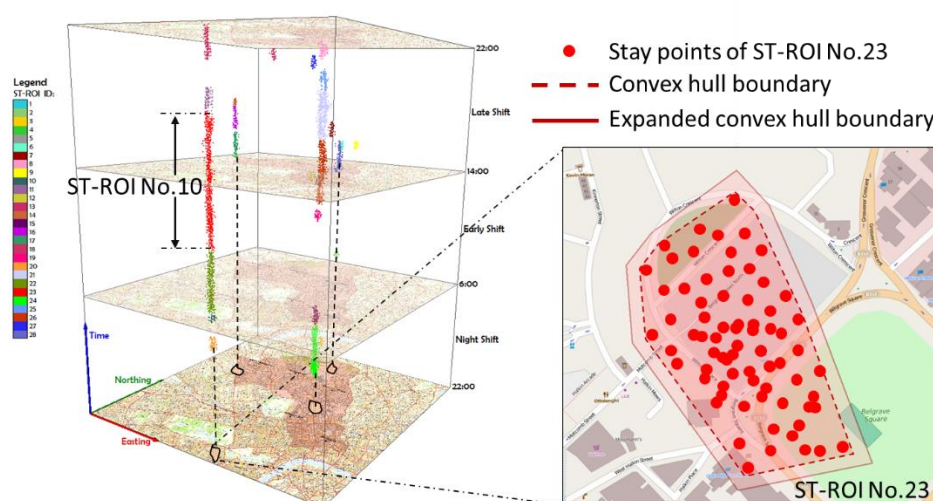


Figure 5.15 Spatial boundary of ST-ROI No.23 in the 28 ST-ROIs detected by Module II

We search for POIs in the expanded convex hull and find the overlapping periods of POIs with the time span of the ST-ROIs. Taking ST-ROI No.23 again as an example, this is a very special ST-ROI that locates outside Camden police’s mission area. It is detected because of the intensive police activities around the Syrian embassy for the safeguard of the embassy area against violent protesters (Daily Mirror News, 2012).

Nine POIs are found in the space-time boundary of ST-ROI No.23 as presented in Figure 5.16. Among these POIs, three belong to the “public infrastructure” major category, one is a POI of commercial services, two are “education” POIs and three POIs are related to a “government and organisation” major category. The surrounding of the Syrian embassy in February 2012 is one of the “government and organisation” POIs in this place. The simplest way of inferring the semantic meaning of ST-ROI No.23 is to directly use the quantity of sibling POIs as their semantic contribution index in an area, as did Krüger et al. (2013) and Polisciuc et al. (2015). As we know, however, public infrastructures, such as bus stops and phone boxes, can be found in large numbers



throughout the city and should be considered as less significant POIs in the semantic enrichment process. If we weight the semantic contribution of POIs solely by numbers, the three less meaningful “public infrastructures” that make up 33% of the POIs in ST-ROI No.23 will generate a considerable portion of semantic contribution to the ST-ROI and misrepresent the meaning of the ST-ROI. Even if we add the length of overlapping opening hours into the consideration, as Equation 5.7 describes, the contribution of “public infrastructures” is still dominant in the region. Obviously, the police officers were not interested in the “public infrastructures” in that particular area since they can be found everywhere in the city. Hence, the raw semantic contribution  $w$  does not translate into the truly significant contribution of police activities in this case.

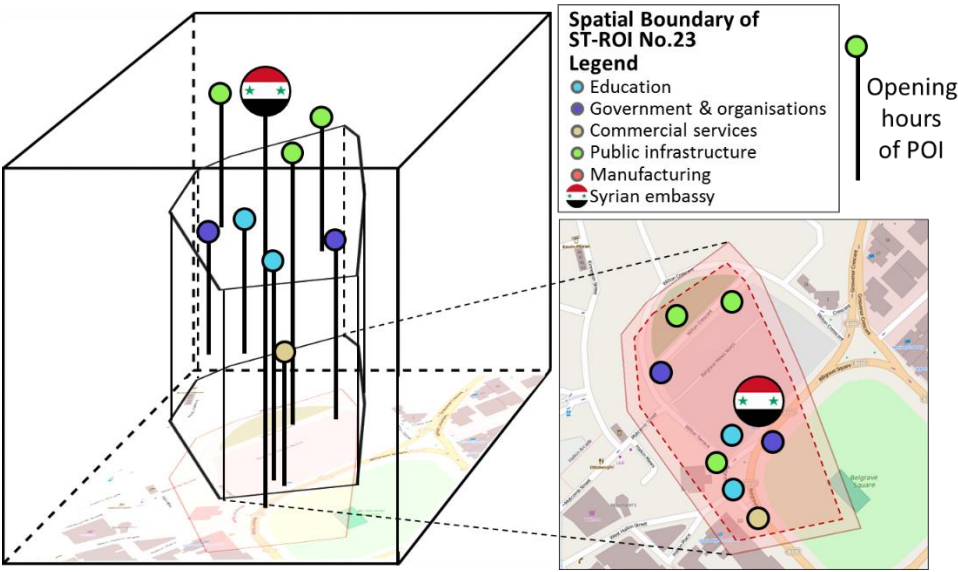


Figure 5.16 The Ordnance Survey POIs in the space-time boundary of ST-ROI No.23

Therefore, reweighting the POIs’ categorical semantic contribution and removing the negative influence of ubiquitous POIs in semantic enrichment is necessary. Table 5.2 shows that the semantic contribution of the major category “public infrastructures” calculated by Equation 2.2 and Equation 5.9. had been significantly weakened after the TF-IDF weighting process, from 0.333 to 0.043. In contrast, small-size categories, such as educational POIs, were emphasised and the weights of governmental POIs were reinforced after the significance is recalculated with TF-IDF, corresponding to the common sense that ST-ROI No.23 is an embassy area and most educational and cultural branches of foreign embassies are nearby their countries’ embassies.

Table 5.2 The semantic contribution of different categories of POIs in ST-ROI No.23 calculated with Equation 2.2 and Equation 5.9



	SC by count	TF-IDF weighted SC
Accommodation, eating, drinking	0	0
Commercial services	0	0
Attractions	0	0
Sport & entertainment	0	0
Health	0	0
Public infrastructure	0.3333	0.0434
Manufacture & production	0	0
Retail	0	0
Transport	0	0
Education	0.1667	0.3299
Government & organisations	0.5	0.6267

Table 5.3 shows the time span of each ST-ROI. For confidential reasons, the emerge times and perish times of most ST-ROIs are hidden. This table also includes the semantic contributions of the 11 categories of POIs in the 28 ST-ROIs reweighted by TF-IDF. The contributions are added to officers' space-time profiles to generate officers' semantic profiles (SP) in the next step.

Table 5.3 The TF-IDF weighted semantic contribution of different categories of POIs in each ST-ROI in Camden

ST-ROI No.	1	2	3	4	5	6	7	8	9	10	11	12	13	14
<b>Emerge Time</b>	15:30	Classified	Classified	Classified	Classified	Classified	Classified	Classified	Classified	Classified	Classified	Classified	Classified	Classified
<b>Perish Time</b>	16:05	Classified	Classified	Classified	Classified	Classified	Classified	Classified	Classified	Classified	Classified	Classified	Classified	Classified
Accommodation, eating and drin	0.023	0.1004	0.0964	0.1248	0.0271	0.0192	0	0.104	0.087	0.2867	0	0.0257	0.1204	0.0856
Commercial services	0.1657	0.2751	0.2494	0.2533	0.3181	0.2806	0.257	0.2712	0.2482	0.1193	0.389	0.7276	0.1957	0.2698
Attractions	0.2664	0.0103	0.0128	0.0238	0.0501	0.0711	0.1412	0.0136	0	0.0702	0.4033	0.1488	0.0083	0.0142
Sport and entertainment	0	0.1022	0.0962	0.064	0.0294	0.0416	0	0.0828	0.0993	0.0779	0	0	0.0756	0.0746
Health	0.0907	0.0628	0.0666	0.0266	0	0	0	0.0478	0.0438	0.0497	0	0.0338	0.0293	0.0533
Public infrastructure	0.031	0.0233	0.0269	0.0134	0.0137	0.0162	0.0103	0.0203	0.0201	0.0765	0.047	0.0087	0.0165	0.0206
Manufacturing and production	0.1179	0.0415	0.0234	0.0573	0.0456	0.0647	0	0.0441	0	0	0	0	0	0.0413
Retail	0.1677	0.2855	0.3304	0.3514	0.1476	0.195	0	0.3135	0.164	0.3197	0	0.0257	0.4586	0.3446
Transport	0.0796	0.0518	0.0576	0.0514	0.0228	0	0	0.0558	0.0578	0	0.1607	0.0297	0.0423	0.0534
Education	0.058	0.0317	0.0277	0.0248	0.0517	0.0733	0.1561	0.0298	0.0274	0	0	0	0.0367	0.0357
Government and organisations	0	0.0153	0.0126	0.0092	0.294	0.2384	0.4355	0.0172	0	0	0	0	0.0166	0.007
ST-ROI No.	15	16	17	18	19	20	21	22	23	24	25	26	27	28
<b>Emerge Time</b>	Classified	Classified	Classified	Classified	Classified	Classified	Classified	Classified	10:00	Classified	Classified	Classified	14:00	Classified
<b>Perish Time</b>	Classified	Classified	Classified	Classified	Classified	Classified	Classified	Classified	18:00	Classified	Classified	Classified	16:00	Classified
Accommodation, eating and drin	0.2339	0	0.1895	0	0.2471	0.1032	0.0573	0.1626	0	0	0.1448	0	0	0.0439
Commercial services	0.3543	0.7264	0.3659	0	0.2438	0.2509	0	0.5179	0	0.8689	0.404	0.5718	0	0.2244
Attractions	0.0187	0	0.0205	0.1525	0	0.0159	0	0	0	0	0	0	0	0.029
Sport and entertainment	0.0438	0	0	0.0751	0.0833	0.1043	0.2611	0	0	0	0.0872	0	0	0.0642
Health	0.0133	0	0.0145	0	0.0324	0.0525	0.1504	0.0184	0	0	0.028	0	0	0.0082
Public infrastructure	0.0306	0	0.0112	0.0111	0.0166	0.0207	0.0193	0.0047	0.0434	0	0.0287	0.0411	0	0.0084
Manufacturing and production	0.0792	0	0.0866	0	0	0.0332	0	0.0932	0	0	0.1079	0	0	0.0832
Retail	0.2145	0	0.2125	0	0.0861	0.3385	0	0.0921	0	0	0.1451	0.1811	1	0.4177
Transport	0.0116	0.2736	0.0741	0	0.2343	0.0346	0.132	0.1109	0	0.1311	0.0542	0.206	0	0.1051
Education	0	0	0	0.1685	0	0.0349	0.3801	0	0.3299	0	0	0	0	0.0158
Government and organisations	0	0	0.0252	0.5929	0.0563	0.0112	0	0	0.6267	0	0	0	0	0

### 5.6.7 Module IV: Aggregative analysis of semantic profiles

Most of the officers on duty are foot patrol officers (FP), community support officers (CSO) and senior officers (SO). The dwelling time that different types of officer allocate to ST-ROIs can be very different, because the different tasks they are required to undertake are determined by their types. According to the outcomes of Module II, the (non-semantic) space-time profiles of three typical police officers (i.e. 1812PO, 8971PO,

8986PO) can be seen in Figure 5.17; the identity call signs of the officers have been encrypted for security's sake.

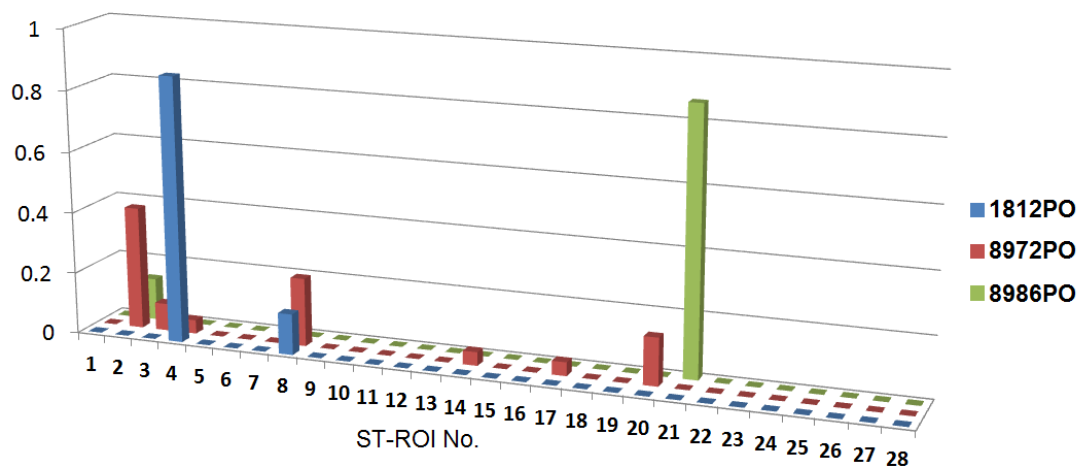


Figure 5.17 The space-time profiles of three police officers

The semantic enrichment of ST-ROIs in Module III allows us to transform ST-ROIs into semantic ST-ROIs. Similarly, the time allocation on ST-ROIs (i.e. space-time profiles) can also be transformed into time allocations on different semantic meaning categories after Module III. For instance, officer 8986PO spends 90% of his/her staying time in ST-ROI No.21 and 10% staying time in ST-ROI No.2. According to the semantic contributions in Table 5.3, 2.75% ( $0.2751 * 10\% + 0 * 90\%$ ) of 8986PO's total active time is assigned to commercial service places, whereas 34.5% ( $0.0317 * 10\% + 0.3801 * 90\%$ ) of 8986PO's total active time is assigned to educational venues. Figure 5.18 displays the comparison of the three chosen officers after their space-time profiles are turned into semantic profiles by Equation 5.10.

It is worth noting that the time allocations on ST-ROIs of officers "1812PO", "8972PO" and "8986PO" were by no means the same (Figure 5.17). After interpreting the activities of the officers with POI impacts, however, the profiles of officer "1812PO" and officer "8972PO" became quite similar to each other semantically (Figure 5.18). This is because they have visited semantically similar places in semantically similar times, despite the spatial locations they have visited being different. Their common interests in retail and commercially-related areas were revealed, whereas semantically interpreting the profile of officer "8986PO" made it prominently distinct from the other two officers. This shows the new method's capability to find users sharing semantically similar activity patterns, despite the fact that the places in which they stayed may be spatially far apart.

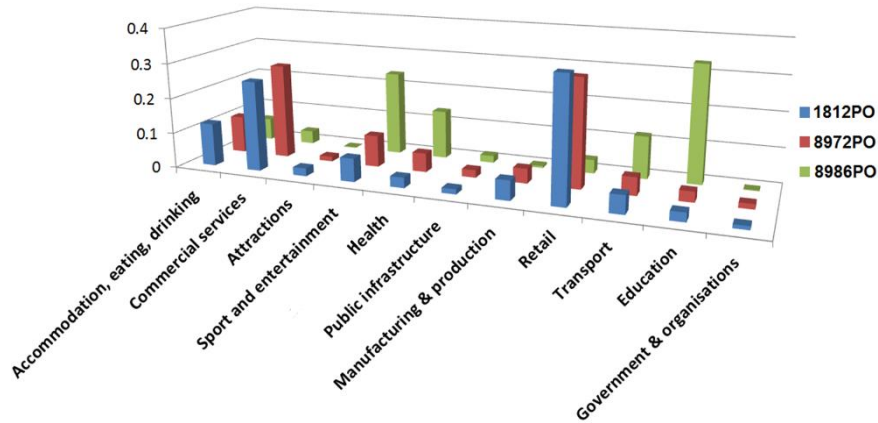


Figure 5.18 The summarised semantic profiles of three police officers

At the end of the case study, we hierarchically cluster the individual semantic profiles again with JSD-based dissimilarity (Equation 5.11). Figure 5.10 shows hierarchical clustering results of the non-semantic space-time profiles and Figure 5.19 shows the hierarchical clustering results with the consideration of semantic meaning of ST-ROIs. As suggested again by the Dunn index (Dunn, 1973), the officers with the newly proposed semantic profiles are divided into 5 groups. The call signs of officers have been encrypted and the three example officers are marked with dashed rectangles. In Figure 5.10, the three example officers are grouped into three different groups because of their differences in space-time profiles. In contrast, Figure 5.19 shows that the new method can find the semantic similarities between officers “1812PO” and “8972PO”, despite their location differences. The new method also generated clearer segregation and simpler grouping results than the old one in Module II. This is mainly because the number of summarised semantic categories in the semantic profile is far less than the number of the extracted ST-ROIs in a space-time profile.



### 5.7.1 Semantic ST-ROIs

ST-DBSCAN was first used in parallel with the police patrol activities in each of the 12 inner London BOCUs in August 2015. The input parameters of ST-DBSCAN are determined by Equation 5.3 in every borough. For simplicity of visualisation and security reasons, we demonstrated the results of the three inner London boroughs (City of Westminster, Islington and Camden) as examples. By trip segregation in the pre-processing modules, we found that 620 officers had more than 5 trips in the three selected boroughs in August 2015. The movements of these 620 officers are input to Module I for stop identification and the stay points are clustered by ST-DBSCAN in Module II. As for the parameters for ST-DBSCAN, the *Spatial Eps* =  $20 + 2\sigma = 36m$ , the *Temporal Eps* =  $5min = 300s$ , and the *minPts* is set to be 55 according to Equation 5.3.

Figure 5.20 shows all 54 ST-ROIs detected in the extended case study area. For security reasons, we cannot label all the ST-ROIs with place names and exact time spans, though we choose 5 places that people are familiar with to show their semantic meanings in the next step.

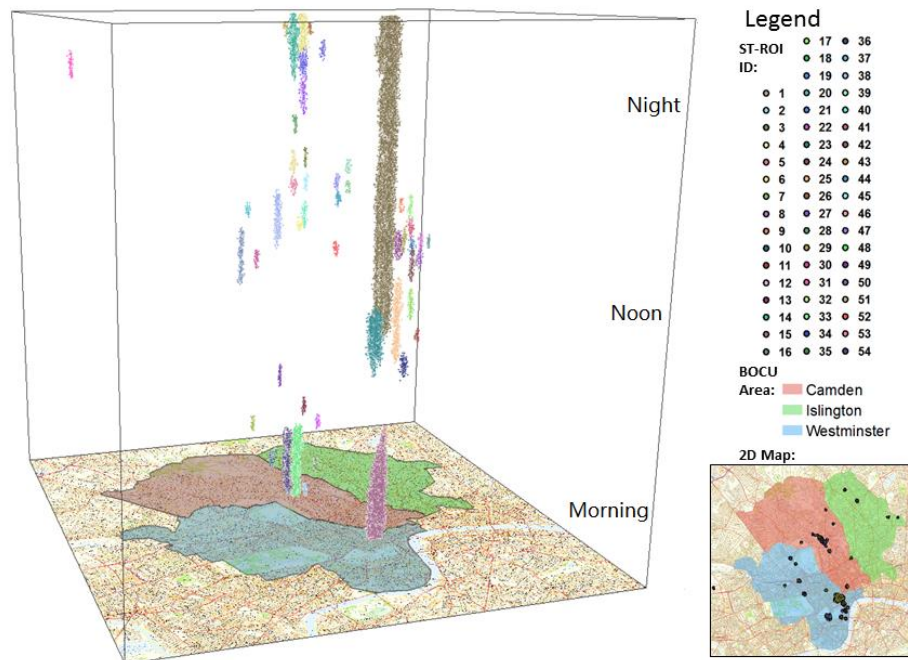


Figure 5.20 The 54 ST-ROIs in three BOCU areas chosen for demonstration

The second step in the extended case study was to import the POI data of the entire Greater London area and enrich the semantic meaning of the 54 ST-ROIs. Table 5.4 shows the semantic weights of POIs in the five ST-ROIs in Figure 5.20 as examples. It

shows that the TF-IDF results are in line with the citizens' common impression of the meaning of the places. For example, the detected ST-ROI No.37 is Buckingham Palace and its time span ranges from 10:35 to 12:07. This is exactly the time when the famous everyday changing of the guards at Buckingham Palace takes place and police officers gather there every day to block the road for the parade and safeguard this place of tourist attraction. This module also turns the semantic meaning into a time-varying attribute of a place. For example, ST-ROIs No.48 and No.21 in the extended case study are both in the area of Camden Town, yet the semantic contributions of different types of opening POIs change dramatically over time. Camden Town's dominant semantic meaning is "accommodation, eating and drink" at night, whereas "retail" POIs contribute most to its semantic meaning in the afternoon. This summarised POI semantic contribution information is used to turn the officers' time allocation to ST-ROIs into semantic profiles according to Equation 5.10.

Table 5.4 The TF-IDF semantic meanings and names of five chosen ST-ROIs in the extended case study

ST-ROI ID	37	13	25	48	21
Name of the Place	Buckingham Palace	Soho	Whitehall	Camden Town	Camden Town
Emergence Time	10:35	Classified	Classified	0:00	12:55
Perish Time	12:07	Classified	Classified	4:42	15:32
Accommodation, eating and drinking	0	0.3233	0.1375	0.6428	0.1041
Commercial services	0.0487	0.1676	0.1818	0.0945	0.2371
Attractions	0.7438	0.0375	0.0156	0	0.0118
Sport and entertainment	0	0.1339	0.0288	0	0.1047
Health	0	0.0411	0	0.1134	0.0812
Public infrastructure	0.0233	0.0387	0.0268	0.0284	0.0236
Manufacturing and production	0	0.0426	0	0	0.0267
Retail	0.0523	0.1412	0	0.0917	0.3027
Transport	0	0.0332	0.0357	0.0292	0.0375
Education	0	0.0296	0	0	0.0534
Government and organisations	0.1319	0.0114	0.5738	0	0.017

### 5.7.2 Semantic profiles and profile aggregation

After the semantic profiles were grouped via the JSD-based hierarchical clustering method, officers with similar activity patterns across all boroughs can be detected, even if they never belonged to the same BOCU area branch. 54 ST-ROIs are detected in the three boroughs and even more will be generated if the method is to be applied to the entire London area. This means that clustering space-time profiles containing too many ST-ROIs will lead to the "curse of dimensionality" (Bellman, 1961). By turning the space-time profile into a semantic profile, the number of dimensions of the semantic profile is limited to 11 (the number of POI major categories) and the following aggregative clustering process will not be undermined.



As the Dunn index test suggested, the optimal group number should be five in the hierarchical clustering for the extended case study. There are 2,064 highly active officers (20 times the number of officers who participated in the single-borough case study in section 5.6) patrolling in these three boroughs in the study period, and the visualisation of the hierarchical clustering results of all officers like Figure 5.19 is impossible to be presented properly on a printed page. Therefore, the average semantic profile of each officer group was summarised to show the representative pattern of their activities (Figure 5.21). It showed that the focus of officers on different semantic places varied greatly. Officers in Group 1 allocated their time more evenly than did others and paid more attention to commercial and retail streets. Group 2 preferred to stay near tourist attractions, whereas Group 3 focused on sport and entertainment events and Group 5 patrolled around both governmental and public infrastructures. Group 4 spent most of their time near hospitals and had far fewer activities than other groups. This demonstrates that the activity patterns of police officers show clear differences when the semantic meaning of places is brought into the profile clustering process, and each group has its own major interest.

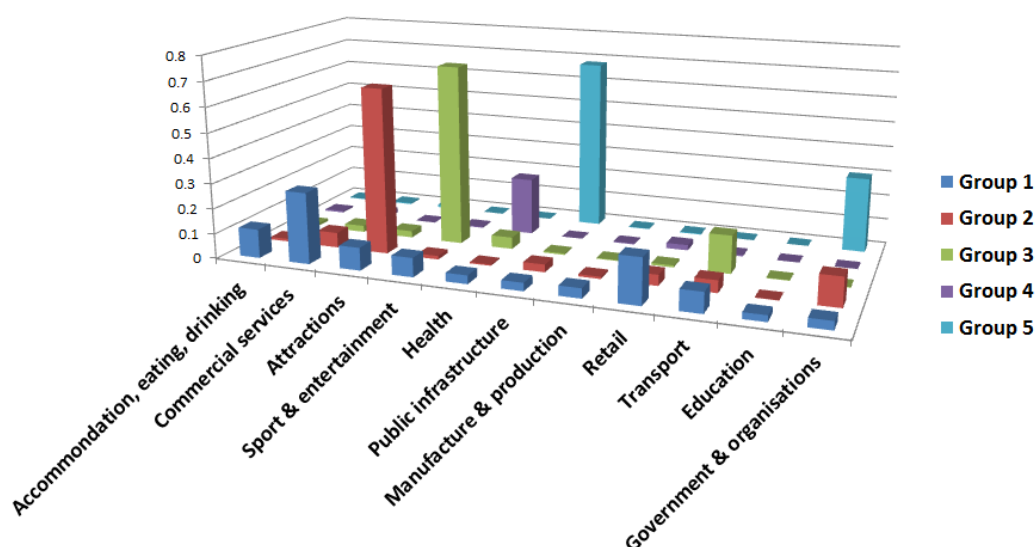


Figure 5.21 The average semantic profiles of five officer groups generated by the aggregative analysis

## 5.8 CHAPTER SUMMARY

In this chapter, we used a Euclidean paradigm of our methodology framework to demonstrate the work flow of our framework and resolved the four problems summarised in Chapter 1 with four modules. We also introduced the new concept, ‘the place you go, when you go and how long you stay is who you are’, to focus on the spatial,

temporal and semantic aspects of places rather than just spatial locations. Methodologically, the framework further extends the traditional ideas of time budget allocation in behavioural studies and existing spatial-location-based user similarity definitions to a semantic explanation of people visiting places. It can profile the activity patterns of people according to space, time and semantic aspects by defining the JSD similarity metric, which in reality is closer to people's place visiting purposes. Furthermore, after determining what the place is about semantically, the pattern differences of individuals' activities are better explained. We used police foot patrol data as case studies to represent kindred location-based applications. The semantic meanings of the places are extracted and measured based upon the space-time information of urban POIs. This evolution from Space-Time to Place-Time enables analysis of a large population with much higher heterogeneity and dynamism in a large city-scale area.

The modules of the Euclidean paradigm presented in this chapter, despite their generally positive evaluation compared with conventional methods, are constructed under the condition that all spatial distances are Euclidean, and require further awareness of the urban network's influence on people's movements and stops in order to provide a complete picture of their activity patterns.

Further work will turn this methodological framework into a street network-based version to further improve the space-time accuracy of activity extraction in police patrols and to further adapt all modules to the urban street environment. The next chapter will address the development of a network paradigm according to the unsolved limitations of the Euclidean paradigm. The stay identification, ST-ROI detection and semantic profiling modules developed during the next phase must integrate with a network representation of space. To this end, map-matching algorithms and a network-based space-time clustering method will be developed. The accompanying rise in computation cost also needs to be mitigated.



## Chapter 6

# The Network Paradigm

## **6 THE NETWORK PARADIGM**

### **6.1 INTRODUCTION**

Our methodology framework has been described and tested in the Euclidean paradigm in Chapter 5. The Euclidean paradigm has provided better solutions for the listed Challenges 1 to 4 in Chapter 3 than existing conventional approaches. Together, the four modules also formed a standard procedure for the aggregative analysis of human urban activity patterns. However, the spatial distance metric throughout the Euclidean paradigm is not perfect in depicting the true topological structure of a city, nor is it appropriate to act as the proximity indicator of physical entities (Meier, 2017) and movements in street segments. To apply our framework onto urban street networks where most human dynamics data were collected, adaptations and improvements in every module of our framework should be made.

In this chapter, we propose a new paradigm to specifically detect streets that attract intensive visits in certain time periods and analyse the semantic activity patterns of people in the complex street networks of a city. Here, we use street segments as the base level spatial unit in all the framework modules instead of polygon regions or grids in the planar space. Therefore, we detect streets and time periods with intensive human visits. Because the geometric units are linear segments instead of bounding areas, we extend the concept of ST-ROI to the street networks and define them as Spatio-Temporal Lines of Interests (ST-LOIs). Accordingly, the functional information of roadside POIs is then annotated to the street addresses contained in the ST-LOIs to enrich the meaning of the ST-LOIs. The semantic profiles in Modules III and IV are also summarised based on the detected ST-LOIs and POIs along the streets. These changes will bring three improvements: (1) The positioning errors in the raw trajectories are mitigated; (2) The boundary of ST-ROIs/ST-LOIs are better defined; and (3) The semantic enrichment is more precise.

To achieve all these improvements, changes of methods in the framework are as listed in sections 6.1.1 to 6.1.4.

#### **6.1.1 From trip trajectories to trip routes**

Most stop identification methods are based on a Cartesian expression of space and ignore the fact that the majority of people's movements in urban areas follow the segments of the interconnected road networks. Their problem settings, therefore, are not realistic for the study of urban movements.

Ignoring the structure of road networks can cause serious troubles for the analysis of moving trajectories in an urban context. An example can be seen in Figure 6.1. The two points are isolated from each other by the semi-underground rails in between, and they are by no means close to each other if the road network context is considered. One pedestrian would take more than 10 minutes to walk from one point to the other, although the straight line Euclidean distance is just 78 m. If the two points are two contiguous location updates in Module I of the Euclidean paradigm, Euclidean-based methods will miscalculate the distance and speed between the two points and identify wrong stop episodes and stay points. The comparison of stop identification accuracy of Euclidean-based and network-based methods can be found in Chapter 8. If the two points in Figure 6.1 are two stay points in Module II of the Euclidean paradigm to be clustered, Euclidean-based spatial clustering methods will fail to exploit the underlying network context and unreasonably take the two far away points in the network into one ROI, which can generate misleading point clusters in space.

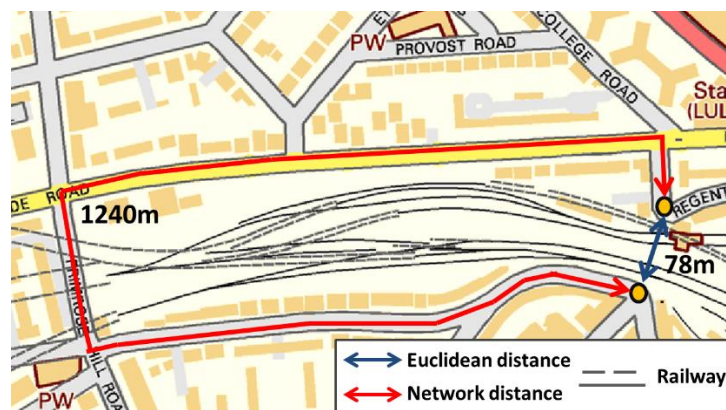


Figure 6.1. The Euclidean distance between two points is 78m, however, people need to move 1240m from one point to the other in the network

For the pre-processing module (Module I) of our framework, the accuracy of stay point detection can be improved with the awareness that the movements are along the streets. Map-matching is used as the technique to align the observed GPS positions with the road network on a given digital map. It acts as a fundamental pre-processing step to transform the point-based each trajectory in the Euclidean paradigm into a route consisting of a sequence of covered street segments in the network paradigm. All analyses of semantic places, times and activities in the network paradigm will be built on the basis of stop episodes and stay points along the map-matched trips routes.

### **6.1.2 From ST-ROI to ST-LOI**

Points generated by movements in road networks may appear to be sparser in the planar space, which makes it harder for Euclidean-based clustering methods to detect the aggregation, especially the margin of the aggregation. Besides, the Euclidean methods tend to generate large and imprecise ROI boundaries that cover unwanted areas where the moving objects have never actually visited or private spaces that do not contribute any semantic meaning to the public.

Since all the trips, especially the stop episodes, are snapped to the road segments in the network paradigm, the ST-ROIs detected in Module II are also space-time point aggregations along the street networks with linear shapes. To differentiate with the ST-ROI in the Euclidean paradigm, we call these network-based space-time point clusters Spatio-Temporal Lines of Interests (ST-LOI). Instead of covering a solid part of space regardless of whether it is a private space or a space of public semantic meanings like ROIs in most studies (cite), the ST-LOI only covers a partial set of urban streets with POIs on roadsides and preserves the interconnected segment structure inside. In Chapter 8, the advantages of ST-LOI representation over ST-ROIs and conventional ROIs are presented with a performance comparison.

### **6.1.3 From 3D scatter map visualisation to 3D wall map visualisation**

The 3D scatter map in a space-time cube shows a point aggregation in time and planar space. It is hard, however, for the 3D scatter map to inform us about the relationship between the stay points and the road links that they are in. Inspired by Tominski et al.'s (2012) 3D wall map method for massive trajectory data visualisation, we use the 3D wall map to visualise the space-time clusters of stop episodes in trajectories (i.e. ST-LOIs). This method is the most suitable option for visualising the ST-LOI that we proposed since it preserves and highlights the network structure in space. It allows us to use street segments as basic units in space-time visualisation and present results on a finer scale for people to understand.

### **6.1.4 Work flow of the network paradigm**

The network paradigm also follows the 4-step work flow of the general framework. Nevertheless, one essential difference between the network paradigm and the Euclidean paradigm is the increased computation burden accompanying network-based methods, especially in Modules I and II. Apart from the optimisation measures for the algorithms within each module, Modules I and II are separated into multiple work threads so that

the overall framework can be accelerated by parallel processing as described in Figure 6.2. When analysing the activity patterns in a large study area, the area is separated into multiple districts with buffer zones. Each district is home to a group of individuals and each work thread is in charge of the pre-processing and ST-LOI detection of the movement data within a district buffer zone. If the study area is inseparable (e.g. most individuals often move across different districts or the overall area is small), Modules I and II will still be operated in a single-stream manner. In our case study of London Police movements, the Modules I and II are executed for the officers and their movements in each BOCU and its surrounding areas in each independent work thread.

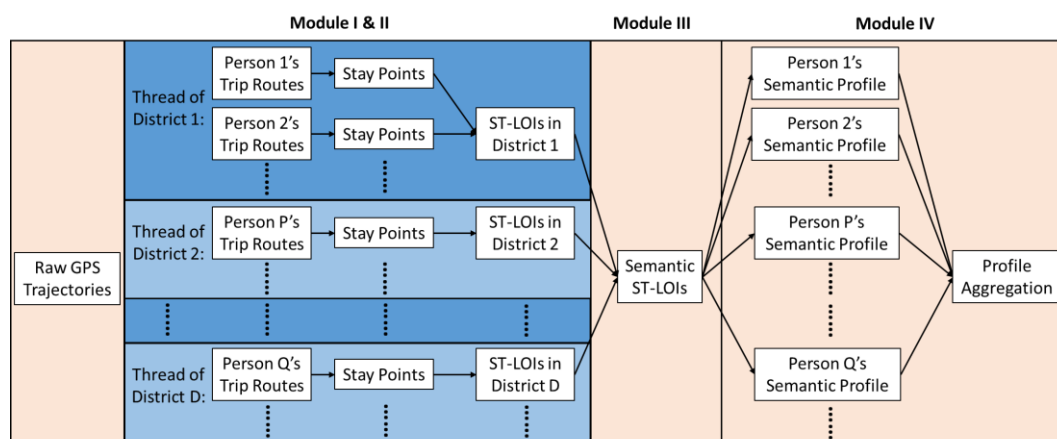


Figure 6.2 Work threads and work flow of the network paradigm for large study area

The detailed methodologies and case studies are presented in the rest of the chapter. All the network analysis methods and algorithms across modules in the network paradigm will require common tools such as network indexing and shortest network route calculation. Therefore, the common tool kit for the modules in the network paradigm is firstly explained in section 6.2 before the description of the modules. Sections 6.3 to 6.4 provide a step-by-step explanation of the proposed algorithms, demonstrating how the proposed methods can improve the ROI detection and semantic activity study in urban scenarios. The increases of computation complexity brought by the improvements in the network paradigm will be discussed at the end of every module. The framework is then tested with a case study of multiple boroughs in section 6.7. Finally, section 6.8 summarises the major findings and directions for further research.

## 6.2 BASIC TOOLKIT FOR SPATIAL NETWORK ANALYSIS

Here we introduce two general tools that will be frequently used across modules in the network analysis. These two toolkits are substantialised as Python packages in programing and are invoked by the module when required.

### 6.2.1 Shortest path finding tool

The network distance between two objects is defined by the length of the shortest path from one object to the other through the network. In the network paradigm, such a distance provides the spatial-closeness measurement for the map-matching process of Module I and our proposed ST-Net-DBSCAN clustering method in Module II. Nevertheless, replacing the commonly used Euclidean distance with network distance in Modules I and II involves shortest path computations, which causes a much higher complexity and inconstant cost for computation. Specialised shortest path finding algorithms are the solution to speed up this process. As reviewed in Chapter 2, A\* algorithm (Hart et al., 1968), which guides the query towards the destination with a heuristic function, is an informed search algorithm, or a best-first search, meaning that it solves problems by searching among all possible paths to the solution for the one that requires the lowest cost (i.e. least distance travelled). Among all these paths, A\* first inspects the ones that are likely to lead fastest to the solution. A\* constructs a tree of paths starting from a specific node in a weighted graph, expanding paths one step at a time, until one of its paths ends at the predetermined destination node (i.e. the solution). At each iteration of its main loop, A\* determine which of its partial paths to expand into longer paths based on an estimate of the cost (i.e. total weight in the path) still to go to the destination. A\* demonstrates a 40–60% saving of computational cost in a medium-scale network (Fu et al., 2006) as compared to the Dijkstra algorithm. Therefore, we use A\* algorithm as the shortest path finding tool throughout our network paradigm.

### 6.2.2 Range searching and nearest neighbour searching tool

Organising points by constructing a space-partitioning data structure can greatly improve the spatial query efficiency. K-d tree is a binary indexed tree that can split a hyperplane into two parts at every non-leaf node. Because of the tree index properties, the K-d tree can achieve efficient range searches and nearest neighbour searches by eliminating large portions of the irrelevant search space.

In Module II, we use K-d tree as a part of our space–time neighbour retrieval strategy as discussed in section 6.4.3. It can improve the overall speed of the neighbourhood query in street networks by narrowing down the search for the network-based space–time neighbour candidates in Euclidean space before the high-load network search begins. In Chapter 8, K-d tree is also used in speeding up KNN queries for performance evaluation of the paradigms.

## 6.3 MODULE I: PRE-PROCESSING OF MOVEMENT DATA IN URBAN NETWORKS

### 6.3.1 Trip segmentation

At the beginning of Module I, every trip with continuous short-term updates is identified and labelled. The rule of this trip segmentation is the same as for the trip segmentation process in the Euclidean paradigm.

### 6.3.2 Map-matching

To alleviate the spatial observation/positioning error of the GPS device and identify the network routes taken in the actual movement, we apply the ST-Matching algorithm proposed by Lou et al. (2009) as the map-matching method to snap the observations onto the streets. The original ST-Matching algorithm did not account for movements in complex transport networks with a large number of local and narrow pedestrian walkways. Therefore, we trim its parameters according to our case study in this module and apply it in the more complex ITN urban theme layer that contains local links of walking and cycling pathways. ST-Matching finds the most probable candidate point  $c_i$  on the nearby street sections within the searching bandwidth of each original GPS record  $p_i$  in each trip. The segment of streets covered by the shortest path from  $c_i$  to  $c_{i+1}$  is labelled as  $e_u$ . The shortest path is found and measured by the A\* algorithm mentioned in section 6.2. For every matched/confirmed trip, the route  $R$  is represented by a list of street paths ( $R: e'_1 \rightarrow e'_2 \rightarrow \dots \rightarrow e'_k$ ) between the matched candidates that the user has moved through chronologically. For all the observed GPS points  $p$  in the trajectory of one trip  $Tr: p_1 \rightarrow p_2 \rightarrow \dots \rightarrow p_i$ , the goal of map-matching is to find the  $R$  that fits  $Tr$  with the highest probability as shown in Figure 6.3. We choose this algorithm because it is more robust than conventional incremental and AFD algorithms when the sampling rate deteriorates, and it is specifically designed for the matching of GPS data with low sampling rates similar to our case (i.e. between 2-5 minutes). Another similarity between Lou et al.'s (2009) movement dataset and ours is that the positioning error in their dataset also follows a normal distribution, but with a different standard deviation of 20 metres.

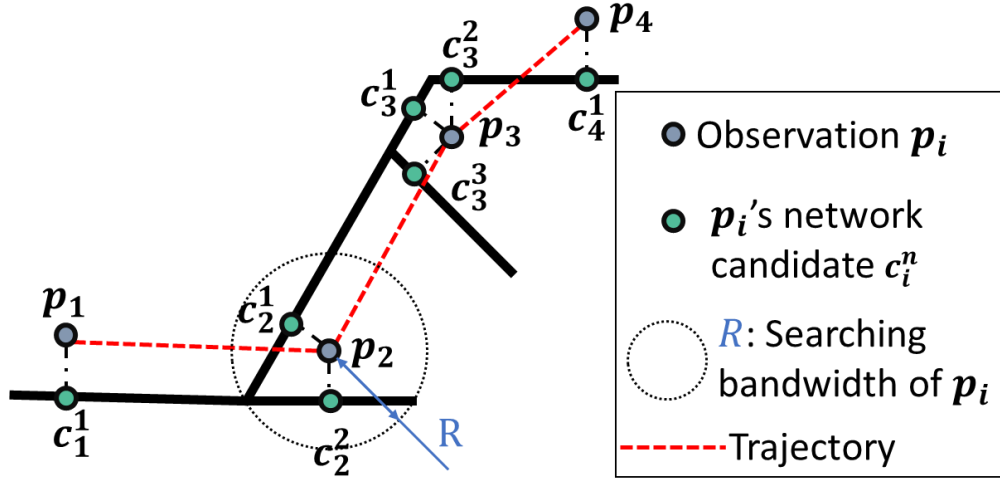


Figure 6.3 The task of map-matching

There are also differences between the two datasets. In Lou et al.'s (2009) experiment, GPS data are collected from moving vehicles in Beijing with an average speed of 13.88 m/s (50 km/h) and a sampling rate of 2 minutes. This means that the average Euclidean distance between every two observations/updates of a GPS device is 2,266 m in Lou et al.'s (2009) case study. As demonstrated in the data exploration of Chapter 4, the officers in the APLS dataset have an average Euclidean moving speed of 1.874 m/s and a GPS sampling rate of 5 minutes, which is equal to 562.2 m average Euclidean distance between every two observations/updates. This means that our dataset has far denser observations than Lou et al.'s (2009) experiment data, which is a positive contribution to the accuracy of map-matching. Besides, the ST-Matching method is originally tested for hybrid mode movements, most of which are vehicle movements. This is different from the police patrolling activities that include walking for most of the time, in terms of median speeds, traffic rules and the accessibility of narrow local streets. Moreover, the extent of the environment varies and the urban canyon effects differ between different cities.

In consideration of the differences between our dataset and the original dataset collected for testing the ST-Matching algorithm, corresponding changes are made according to the local situations of our study area. Details of these changes can be seen in the case study of section 6.7, where parameters are readjusted and the extended urban networks that include local pedestrian walkways are added into the experiment.

The mechanism of ST-Matching is as follows. The algorithm firstly defines the space-time transition probability (STTP) from every  $p_i$ 's candidate  $c_i^n$  to the candidate  $c_{i+1}^m$  of the next observation  $p_{i+1}$ . After that, a unilateral connected graph  $G_{Tr}$  is built to connect  $c_i^n$  with its next observation candidate's  $c_{i+1}^m$  in trip  $Tr$ . The space-time



transition probability (STTP) is used as weights on the segments between  $c_i^n$  to  $c_{i+1}^m$  as demonstrated in Figure 6.4.

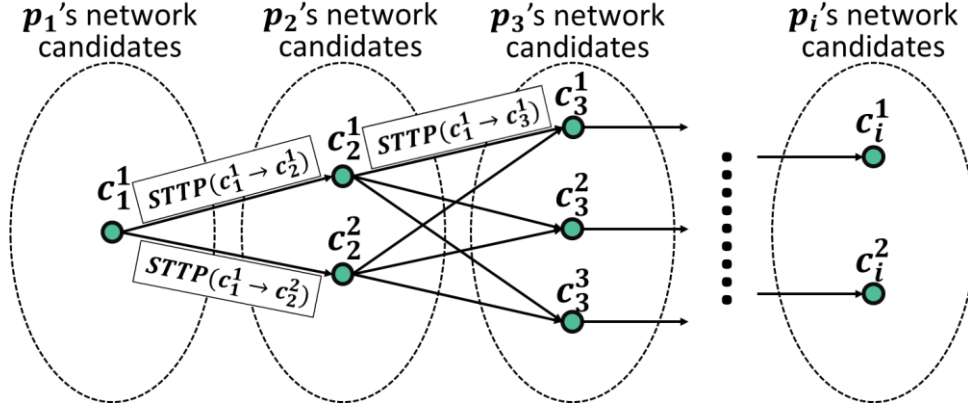


Figure 6.4 Candidate transition graph  $G_{Tr}$  of trip  $Tr$  (Lou et al., 2009)

The STTP is defined as the product of spatial transition probability (STP) and temporal transition probability (TTP). Here, we use Lou et al.'s (2009) description of STP in Equation 6.1. The closer a candidate is to their observation, the more likely they are the matched candidate of the observation. The closer the network distance between two candidates is to the Euclidean distance between their observations, the more likely the two candidates are matched candidates.

$$STP(c_i^n \rightarrow c_{i+1}^m) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(dist(p_{i+1}, c_{i+1}^m) - mean(EPE))^2}{2\sigma^2}} * \frac{dist(p_i, p_{i+1})}{shortpath(c_i^n, c_{i+1}^m)} \quad \text{Equation 6.1}$$

where  $dist(p_{i+1}, c_{i+1}^m)$  is the Euclidean distance between  $p_{i+1}$  and one of its candidates, and  $dist(p_i, p_{i+1})$  is the straight line Euclidean distance from  $p_i$  to  $p_{i+1}$ .  $\sigma$  and  $mean(EPE)$  are the standard deviation and distribution centre of EPE.  $shortpath(c_i^n, c_{i+1}^m)$  is the length of the shortest path (i.e. network distance) from  $p_i$ 's candidate  $c_i^n$  to  $p_{i+1}$ 's candidate  $c_{i+1}^m$ .

The temporal transition probability (TTP) can be described with Equation 6.2 (Lou et al., 2009). The idea of this is to match the assumed average speed between two contiguous candidates with the speed limit of the assumed covered streets. In reality, the closer the assumed speed is to the speed limit, the larger the temporal transition probability is between the two candidates.

$$TTP(c_i^n \rightarrow c_{i+1}^m) = \frac{\sum_u (speedlimit(e_u) \times \bar{v}(c_i^n \rightarrow c_{i+1}^m))}{\sqrt{\sum_u speedlimit(e_u)^2} \times \sqrt{\sum_u \bar{v}(c_i^n \rightarrow c_{i+1}^m)^2}} \quad \text{Equation 6.2}$$

where  $\bar{v}(c_i^n \rightarrow c_{i+i}^m)$  is the moving object's assumed average speed on the shortest path between two possible candidates and  $\bar{v}(c_i^n \rightarrow c_{i+i}^m) = \frac{\text{shortpath}(c_i^n, c_{i+i}^m)}{t(p_{i+1}) - t(p_i)}$ .  $\text{speedlimit}(e_u)$  is the legal speed limit on the street segment  $e_u$ . For local walkways that are not given a legal speed limit, we set the speed limit to be the average Euclidean moving speed of officers on patrol (i.e. 1.874 m/s).

$$\text{STTP}(c_i^n \rightarrow c_{i+i}^m) = \text{STP}(c_i^n \rightarrow c_{i+i}^m) * \text{TTP}(c_i^n \rightarrow c_{i+i}^m) \quad \text{Equation 6.3}$$

The STTP is the product of STP and TTP, as described by Equation 6.3. After the STTPs of every pair of contiguous observations are calculated, the goal of ST-Matching is to find the sequence of edges through graph  $G_{Tr}$  that maximise the global sum of the weights on the edges. Because there are so many observations that possess multiple candidates, and calculating the STTP of every pair of candidates involves intensive shortest path finding computations, we abandon the global weight maximum and use a sliding window to find the local maximum of the STTP sum of observations in the window at a given time. This improvement can greatly reduce the computation burden with a slight sacrifice of matching accuracy. The accuracy of the map-matching process is tested based on routes and trajectories simulated by an artificial trajectory generator that we design. The result is compared with other map-matching methods and discussed in Chapter 8. The map-matched routes and confirmed candidates are used for network-based stay point identification in the next step.

### 6.3.3 Network-based stay point identification

The map-matching process has provided us with the matched trip routes with speed and location information of higher accuracy (performance evaluation can be seen in Chapter 8). With these map-matched trip routes in the network, we can also identify the locations and dwelling times of stop episodes more accurately. Here we apply the kernel-based approach (i.e. KTSW), with the same kernel function (Equation 5.1) as described in section 5.2, to the map-matched trips. The major difference here is that we replace the Euclidean metric of  $D$  in Equation 5.1 with the network distance between map-matched GPS updates to define the kernel in spatial networks (Equation 6.4) for the network paradigm.

$$k(D) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2} \left(\frac{\text{shortpath}(p_i, p_j)}{B}\right)^2\right] \quad \text{Equation 6.4}$$

where  $\text{shortpath}(p'_i, p'_j)$  is the length of the shortest path between two arbitrary map-matched updates  $p_i$  and  $p_j$  in the temporal window.  $B$  is the present bandwidth of the kernel that decays along the street segments.

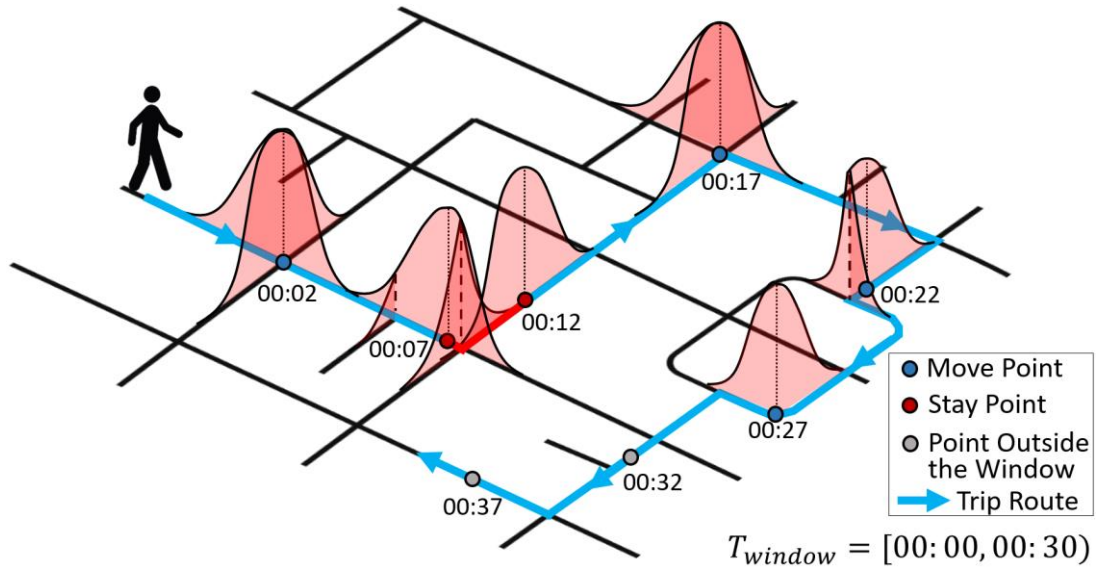


Figure 6.5 The kernel-based temporal scanning window based on network route length

After the spatial distance metric is changed, the calculation of stay value stays in the same way as defined by Equation 5.3 and the stay value is attached to the street segments instead of planar space (Figure 6.5). As the scanning window scans through time, only the points within the temporal window participate in the stay value calculation. Map-matched point updates with a stay value higher than  $\theta_{\text{StayValue}}$  are identified as map-matched stay points. Figure 6.6 shows an example of the difference of stay point identification results in the Euclidean paradigm and the network paradigm. A noticeable difference is that the precision of locations is improved by implementing map-matching before KTSW and conducting KTSW in street networks. This is because some actual stay point sequences with large positioning errors can have large displacements between observations. Module I of the Euclidean paradigm identifies these points as move episodes, whereas Module I of the network paradigm can correctly identify them as stops. The increased accuracy enables us to find more stay points correctly and more precisely estimate the actual time span of ST-ROIs. The accuracy comparison of conventional approaches and the KTSW methods in two paradigms are compared in Chapter 8 based on synthetic routes and trajectories. All identified stay points will be used as inputs to the ST-Net-DBSCAN method we proposed for ST-LOI detection in the next module.

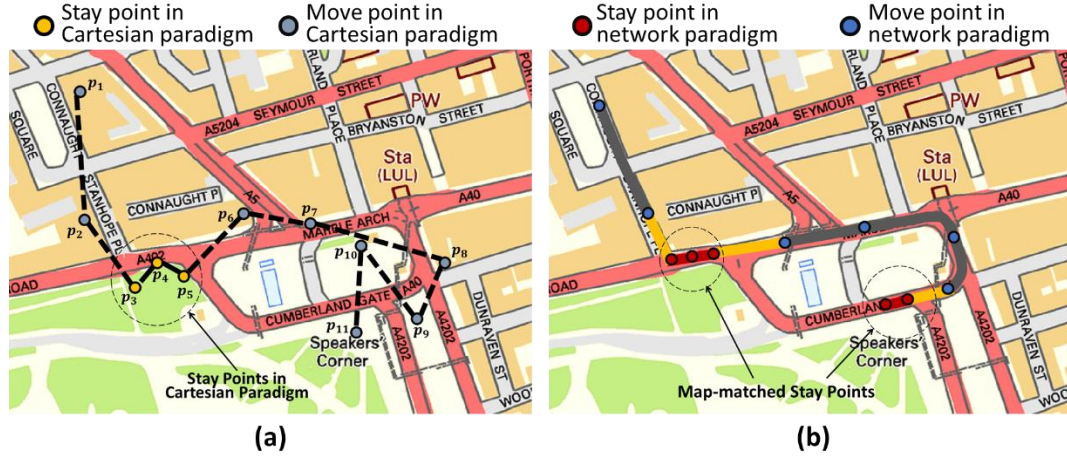


Figure 6.6 Comparison between the stay point identification in Euclidean paradigm and network paradigm in space

### 6.3.4 Complexity

The global ST-Matching algorithm has the time complexity of  $O(mk^2n \log n + mk^2)$  (Lou et al., 2009), where  $m$  is the number of GPS points generated,  $n$  is the number of segments in the street network and  $k$  is the maximum number of map-matched candidates.

Since each GPS observation only considers its side projection points on neighbouring road segments within its searching bandwidth as a set of candidates, the candidate set size is significantly smaller than the total size of road networks in real-life datasets. This makes the algorithm, besides having better matching quality, also more efficient, with linear complexity on the size of the GPS points  $O(n)$ . The parameters of map-matching (e.g. searching bandwidth  $R$  and the window size  $T_{\text{window}}$ ) are tuned in the experiment in section 6.7 according to the characters of the movements and street networks in our case study.

## 6.4 MODULE II: ST-LOI DETECTION

Like Module II in the Euclidean paradigm, ST-LOI is detected in this module and the individual space-time profiles are generated based on the detected ST-LOIs. As Chapter 5 has demonstrated, density-based clustering approaches based on Euclidean distance cannot fully represent the topological structure of interesting regions in street networks. Although DBSCAN and its variations are specifically designed to detect arbitrarily shaped point clusters, it is still impossible for DBSCAN to find the true shape of people's stay point aggregation when the input points fed into the clustering approach have

position errors and the true shapes of the point cluster are submerged in noise in the dataset beforehand. With the method used in Module II of the Euclidean paradigm, interesting regions such as line-shaped shopping streets are still detected as a polygon or circular region because the GPS observation points always fall around the suborbicular surroundings of their true locations with a normally distributed positioning error. In the network paradigm, however, most positioning errors are mitigated in Module I. Hence, we further take the advantage of the map-matched stay points provided by Module I of the network paradigm and apply our newly designed, ST-Net-DBSCAN algorithm (i.e. a network-based variation of ST-DBSCAN) to comprehensively detect interesting regions in time and at the urban street network, so that the shape of the interesting regions in the city can be represented in a finer scale. By inheriting the advantages of DBSCAN, which can detect arbitrary shapes, ST-Net-DBSCAN can detect regions of interest with shapes confined by road geometry. Hence, the proposed method enables spatio-temporal clustering of points with regard to the urban road network structure and generates road segments of interest instead of the approximate regions that attract people in spatial networks and time (i.e. ST-LOI). The following semantic enrichment module can also benefit from the better spatial boundaries of the ST-LOIs detected in this module.

As viewed in Chapter 2, most existing network-based spatial clustering methods are applied and tested on synthesised point data (Yiu & Mamoulis, 2004) or points representing independent issues such as crimes (Tompson et al., 2009). None of them are designed or used for the analysis of point data generated by consecutive movements. Therefore, before the clustering process there was no map-matching technique that could be used to guarantee that the movement points themselves are precisely located, the speed correctly calculated and the stay points correctly identified. Also, there is no existing network-based spatial clustering method that simultaneously aggregates points in spatial networks and time. So far, the network paradigm that we propose is the first to combine map-matching and space-time network clustering for movement analysis.

This module is the core of the entire network paradigm and involves the ST-Net-DBSCAN algorithm that we developed. ST-Net-DBSCAN is a space-time clustering method for urban network environments based on DBSCAN. We chose to develop our method based on DBSCAN due to its four major advantages:

- (i) DBSCAN can detect clusters of arbitrary shapes. Linear, ring-shaped and curve-shaped segments are ubiquitous in road network structures. By using network distance as the distance metric, DBSCAN can detect high-density clusters confined

in the shapes of any network structure. In contrast, it is difficult for other methods to detect clusters of noncircular shapes.

- (2) DBSCAN has the highest efficiency for clustering problems in road networks. Network neighbourhood queries and network-distance computations make up the largest number of calculations in the network space clustering process. Hierarchical clustering methods require that network distance between all points be precomputed and stored. Portioning-based methods, such as K-means and K-medoids, need to run for many iterations before converging into the final result, which also causes a huge number of network-distance calculations. Moreover, finding the centroid of points within the network is expensive and sometimes impossible, which makes portioning-based methods even more inappropriate. On the other hand, only partial distance computations are necessary when the network neighbours are queried in density-based clustering methods, which makes DBSCAN the algorithm of the lowest computation cost in the network space (Yiu & Mamoulis, 2004).
- (3) Unlike portioning-based methods and hierarchical methods, DBSCAN and its variants are not sensitive to noise. Outliers are not included in the generated clusters, and the locations and coverages of the detected interesting regions are therefore more accurate.
- (4) Unlike portioning-based methods and hierarchical methods, a priori knowledge is not needed by DBSCAN and its variants. The number of interesting regions does not need to be predefined.

On the other hand, the ST-Net-DBSCAN is different from normal DBSCAN in 3 aspects:

- (1) ST-Net-DBSCAN uses network distance as the spatial distance metric. Unlike most of the spatial distance metrics that can be expressed by equations, network distance is more complex and requires special algorithms to find shortest path within the network before the distance is calculated.
- (2) ST-Net-DBSCAN uses an extra temporal distance metric to work together with the spatial network distance. The stay point aggregations must be both spatially and temporally dense enough to be identified as an ST-LOI.
- (3) The network distance calculation and temporal dimension increase the intensity and complexity of computation. Therefore, ST-Net-DBSCAN needs to be accelerated by optimisation solutions to make the processing time tolerable.

### 6.4.1 Definitions

ST-Net-DBSCAN can be described with a series of definitions:

- Definition 1 **Map-matched stay points**: After using the ST-Matching algorithm to snap the GPS points of a journey onto the segments of the streets, the network version of KTSW is used to detect stay points in the map-matched trip routes. We call these stay points map-matched stay points.
- Definition 2 **Network\_Eps**: The network-distance bandwidth that defines the network-space neighbourhood of stay point  $s'$  along the road network. *Network\_Eps* is calculated from the network distance of the shortest path between stay points. All map-matched stay points within the *Network\_EPS* network distance from a given stay point  $s'$  are called the network space neighbours of  $s'$ .
- Definition 3 **Spatial\_Eps**: Same as the definition in Chapter 5, *Spatial\_Eps* is the Euclidean bandwidth that defines the neighbourhood in Cartesian space. All map-matched stay points within the straight line Euclidean distance *Spatial\_EPS* from a given stay point  $s'$  are called the Cartesian space neighbours of  $s'$ .
- Definition 4 **Temporal\_Eps**: The maximum time interval that defines the temporal neighbourhood of stay point  $s'$ . All map-matched stay points within the *Temporal\_EPS* period from a given stay point  $s'$  are called the temporal neighbours of  $s'$ .
- Definition 5 **Space-time network neighbours**: The space-time network neighbours of  $s'$  are the intersection set of its temporal neighbours and its network space neighbours.
- Definition 6 **Euclidean Space-time neighbours**: The Euclidean space-time neighbours of  $s'$  are the intersection set of its temporal neighbours and its Cartesian space neighbours.
- Definition 6 **MinPts**: The minimum number of stay points required to generate a new Net-ST-ROI.
- Definition 7 **Directly reachable**: A stay point  $s'$  is a core stay point if it has more than MinPts space-time network neighbours (including  $s'$  itself). Those neighbouring points are considered to be directly reachable from  $s'$ . By this definition, no points are directly reachable from a non-core point.
- Definition 8 **Reachable**: A stay point  $s''$  is said to be reachable from  $s'$  if there is a path  $s_1, s_2, \dots, s_k$  with  $s_1 = s'$  and  $s_k = s''$ , where each  $s_{i+1}$  is directly reachable from its previous point  $s_i$  in the path (i.e. all the stay points on the path must be core stay points, except for the last one  $s_k$ ).
- Definition 9 **Noise**: Points that are not reachable from any other point are noises.

- Now if  $s'$  is a core stay point, then the ST-LOI determination process of the ST-Net-DBSCAN can be described as follows.  $s'$  forms an ST-LOI (map-matched stay point cluster) together with all stay points (core or non-core) that are reachable from it. Each ST-LOI contains at least one core stay point. Non-core points can be part of a cluster, but they cannot be used to reach any further points, so they form the margin of the ST-LOI.

#### **6.4.2 Describing the algorithm**

The process of the ST-Net-DBSCAN algorithm is described with the pseudocode in Figure 6.7.



```

Algorithm pseudocode of ST-Net-DBSCAN
// Inputs:
//  $S = \{s'_1, s'_2, \dots, s'_x\}$  All map-matched stay points in all map-matched trips
// Net = {} Street network of the study area
// Network_Eps Maximum network reachable distance
// Spatial_Eps Maximum reachable distance in Cartesian space
// Temporal_Eps Maximum temporal reachable interval
// MinPts Minimum number of stay points within Network_Eps and Temporal_Eps
// Outputs:
//  $A = \{ST\_LOI_1, ST\_LOI_2, \dots, ST\_LOI_k\}$  Sets of spatio-temporal stay point clusters in street networks (ST-LOIs)

1: Establish A K-d tree for S
2: Current_ST_LOI_Label = 0
3: Network_Eps = Spatial_Eps //Set Network_Eps and Spatial_Eps to the same value
4: for u=1 to x
5:   if  $s'_u$  is not in a existing ST-LOI Then
6:     N = Retrieve_Euclidean_Space_Time_Neighbours ( $S'_u, S, Spatial\_Eps, Temporal\_Eps$ ) //Fast preliminary
retrieve of Euclidean space time neighbors with K-d tree
7:     NN = Retrieve_Space_Time_Network_Neighbours ( $S'_u, N, Network\_Eps, Temporal\_Eps$ )
// Retrieve of space time neighbors based on Network Distance among Euclidean space time neighbours
8:     if quantity (NN) < MinPts then
9:        $S'_u.label = outlier$ 
10:    else
11:      Current_Label = Current_ST_LOI_Label + 1

12:      for v=1 to quantity (NN)
13:        NN.label = Current_ST_LOI_Label
14:      end for

15:      seed.push(NN) // push all map-matched stay points in NN into the seed queue
16:      while not seed.isEmpty ()
17:        CurrentStayPoint = seeds.top ()
18:         $N' = Retrieve\_Space\_Time\_Neighbours (CurrentStayPoint, S, Spatial\_Eps, Temporal\_Eps)$ 
19:         $NN' = Retrieve\_Network\_based\_Space\_Time\_Neighbours (CurrentStayPoint, N', Network\_Eps,$ 
Temporal_Eps)
20:        if quantity (NN')  $\geq$  MinPts then
21:          for each StayPoint in NN'
22:            if (StayPoint.label != outlier || StayPoint.label != NULL) then
23:              StayPoint.label = Current_LOI_Label
24:              seeds.pop ()
25:            end if
26:          end for
27:        end if
28:      end while
29:    end if
30:  end if
31: end for
32: end algorithm

```

Figure 6.7 Pseudocode of ST-Net-DBSCAN

For example, Figure 6.8 shows the circular coverage area of stay point A with a 300 metre Euclidean radius. If  $Network\_Eps = 300$  m, map-matched stay points B and C are the network-space neighbours of A. D and E are A's Euclidean neighbours because they locate within the 300 metre radius, but they are not A's network-space neighbours. If  $MinPts = 3$ ,  $Network\_Eps = 300$  m and the time intervals from B and C to A are both less than  $Temporal\_Eps$ , then A will be a core point and the three points (A, B and C) can make up an ST-LOI.

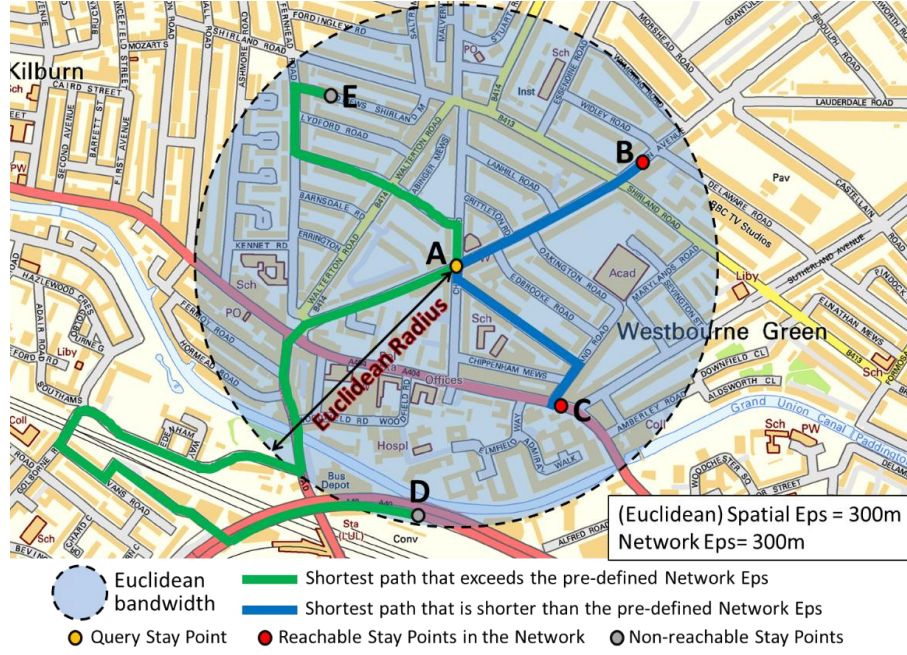


Figure 6.8 The Euclidean coverage of map-matched stay point A and its reachable points in the network

Like map-matching, ST-Net-DBSCAN also requires a large number of queries for network neighbours, which involves intensive shortest path computation. The strategy to search for space-time network neighbours therefore needs to be optimised.

#### 6.4.3 Space-time neighbour retrieving strategy

Neighbourhood retrieval is the most time-consuming and memory-consuming step in our network paradigm. By optimising the retrieving strategy, redundant and unnecessary distance computations and I/O operations for the network data can be avoided. An example can be found in reducing the time complexity of conventional DBSCAN: the spatial retrieving process is optimised with partitioning and indexing techniques (Abbasifard et al., 2014) so that not all distances between every pair of points need to be calculated. Similar but more complex improvements can also be made for ST-Net-DBSCAN.

Optimising the searches for space-time network neighbours plays a crucial role in the entire clustering process of ST-Net-DBSCAN, since retrieving neighbours and finding shortest paths with network distances involves complex and high-cost calculations. In our research, points that fall within the Temporal\_Eps interval of stay point  $s'$  are temporal neighbours of  $s'$ . Points that are within the Temporal\_Eps and (Euclidean) Spatial\_Eps from  $s'$  are Euclidean space-time neighbours of  $s'$ . Temporal neighbours of one point therefore always contain Euclidean space-time neighbours.

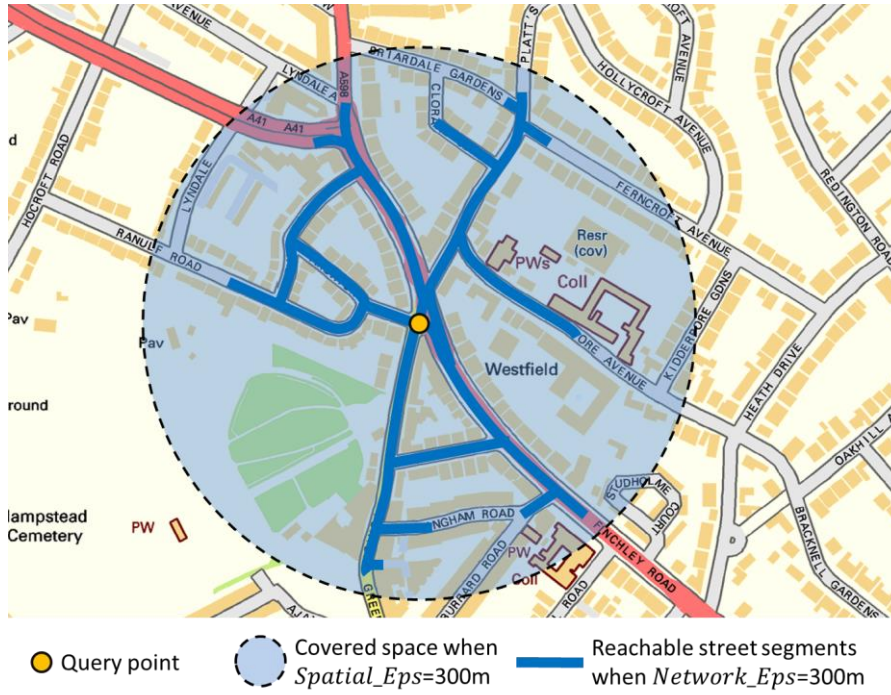


Figure 6.9 The Euclidean filter area of  $s'$  and the reachable street segments of  $s'$  when (Euclidian)  $Spatial\_Eps = Network\_Eps$

In ST-Net-DBSCAN,  $Network\_Eps$  is the network-distance parameter that determines whether a stay point is a space-time network neighbour of  $s'$  (i.e. whether a point is reachable from  $s'$  through the shortest path on the street network). The Euclidean neighbourhood area from  $s'$  is a planar and circular area of a  $Spatial\_Eps$ . When the length of  $Spatial\_Eps$  is equal to the length of  $Network\_Eps$ , this circular area always covers all the reachable street segments that  $Network\_Eps$  defines, as demonstrated in Figure 6.9. Consequently, the space-time network neighbours of a point are always the subset of the Euclidean space-time neighbours of this point.

On account of this, the space-time network neighbours are included in Euclidean space-time neighbours, and Euclidean space-time neighbours are included in temporal neighbours. We use a three-step query strategy to optimise the query by selecting the space-time network neighbours layer by layer:

- (i) Temporal selection: All space-time neighbours should satisfy the requirement defined by  $Temporal\_Eps$ . Also, time is a simple one-dimensional variable. Therefore, the first step of the neighbourhood query strategy is to determine the temporal neighbours of  $s'$  with a simple subset operation in time. The temporal neighbours of  $s'$  are used as inputs of the selection in the next stage.

- (2) Selection in Euclidean space: The second step is a fast Euclidean range query in planar space. In this stage, we search for Euclidean space-time neighbours of  $s'$  based on the results obtained in the temporal selection. When  $Spatial\_Eps = Network\_Eps$ , all space-time network neighbours of  $s'$  fall within the space defined by  $Spatial\_Eps$ , but not all stay points within  $Spatial\_Eps$  are space-time network neighbours of  $s'$ . We can therefore use the Euclidean range query as a filter to exclude the points that are too far away from further selection, therefore avoiding unnecessary network-distance calculations before searching for the space-time network neighbours in stage 3.

Since we are not searching in a data set that keeps changing, the K-d tree (Cormen et al., 2001), a relatively straightforward and memory-oriented spatial-indexing method, is constructed to speed up the massive Euclidean space search. During the search, the K-d tree is traversed for RNN queries (Reverse Nearest Neighbour queries, also called Euclidean range queries) from each stay point, finding the points within the  $Spatial - Eps$  (i.e. Euclidean space-time neighbours). The K-d tree greatly speeds up the Euclidean range query in 2D space because it avoids large global search spaces by partitioning and indexing. At the end of this stage, only points within the Euclidean radius filter are preserved as Euclidean space-time neighbours and input to the next selection stage.

- (3) Selection in spatial networks: The final step is to select the true space-time network neighbours from the Euclidean space-time neighbours. The network distance of the shortest path between Euclidean space-time neighbours generated in the previous step to  $s'$  is calculated by a data mining algorithm called A\*. In the search for the shortest path between two points in the network, the A\* algorithm (Nilsson & Raphael, 1968) heuristically guides the search from  $s'$  to each of its Euclidean space-time neighbours. In addition to the spatial index structure in step 2, the filters should also be able to reduce the search space for shortest-path calculations. That is to say, only street segments and nodes that fall within the circular filter areas of  $s'$  are preserved to participate in the network-distance and shortest-path calculations in step 3. They significantly reduce the input size of the network distance calculation and hence reduce the time consumption of the entire space-time network neighbour query in every work thread of Module II.

Steps 1 and 2 are expressed by lines 6 and 18 in the pseudocode (Figure 6.7) of ST-Net-DBSCAN. Step 3 is embedded in the spatial query process in lines 7 and 19. Experiments show that this three-step space-time filtering process can reduce 91.3–96.7% of pairwise network distance calculations, depending on the point distribution and network layout.

The result of the three-stage query is the reachability and neighbourhood status of each stay point, which determines whether the neighbours are directly reachable from  $s'$  and whether a Net-ST-ROI can be generated as defined in subsection 6.4.1.

#### 6.4.4 Visualisation of ST-LOIs

Visualising derived information in a suitable way is important for explaining and displaying the outcomes. Therefore, the task at the end of the module is to visualise temporally and spatially changing ST-LOI occurrences distributed over the study area. Additionally, the visualisation technique should be compatible with the spatial representation of the street in the network paradigm. The main goal is to support the exploratory understanding of the detected ST-LOIs. We have used a conventional 3D scatter map in the space-time cube to visualise the detected ST-ROIs in the Euclidean paradigm. If we continue to visualise the map-matched stay point clustering with the scatter map, additional projections are needed to show the points' associations with the street segments. In Figure 6.10, the vertical axis shows the time spans of the ST-LOIs, and the magnified detailed example shows the street segments covered by one ST-LOI located in Camden Market, London. In this way, the spatial coverages of Net-ST-ROIs are expressed by the projections of the space-time points on the streets in the map used as a basis.

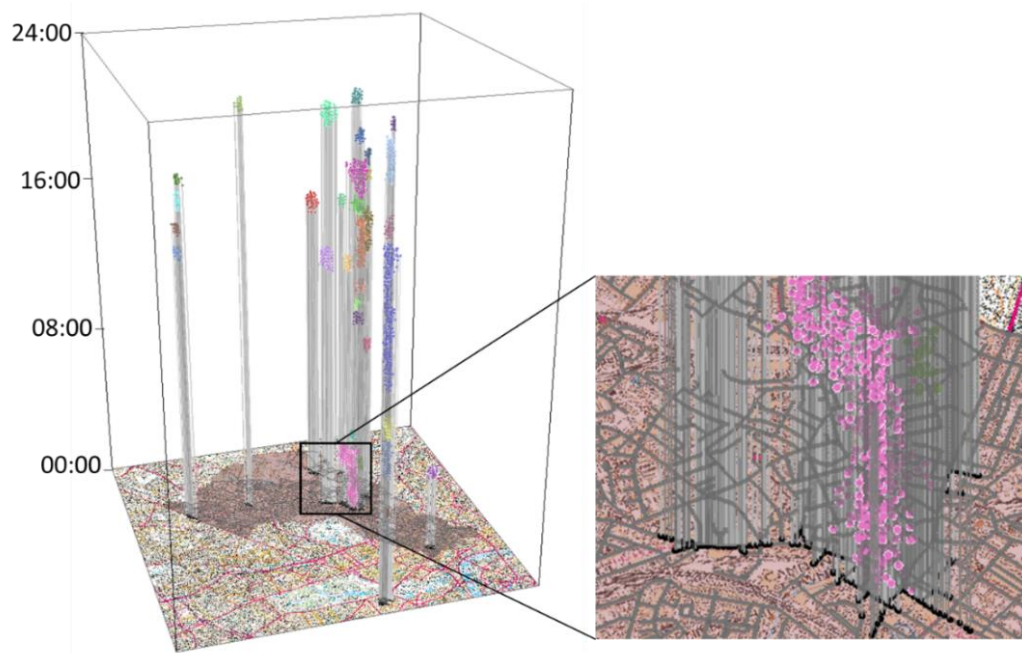


Figure 6.10 Detected ST-LOIs visualised in a space-time cube and their projections on the streets



Nevertheless, the scatter map is after all a point-based visual presentation and cannot reflect in a clear and tidy manner the relationship between the ST-LOIs and the street networks consisting of the interconnected line segments. It will be especially difficult to understand when there are too many points in the clusters and when there are clusters that are close to each other. The connectivity of streets in individual ST-LOIs is also impossible to be seen in a point cluster. Moreover, the semantic information can also not be delivered by the points. Therefore, a 3D wall map visualisation is designed to gain insights into the ST-LOIs' network geometry and additional semantic information. The 3D wall map is essentially a 2D network link map (Becker et al., 1995) stacked layer by layer along an additional time dimension for showing the temporal variations of the data on the links. As reviewed in Chapter 2, it was originally used for the visualisation of traffic data and trajectory speed on the street networks. Instead of using as originally a 3D wall map to visualise the data along the entire links, we use it to highlight only part of the segments in the ST-LOIs. We built a graphical user interface of a 3D wall map using R with *rgeos*, *rgl*, *shiny*, *leaflet* and *maptools* packages. From the 2D road network, we stacked a layer at a time. Each layer is equivalent to a 10 minute interval as the example describes in Figure 6.11. We call each 10 minute layer in each ST-LOI a time brick.

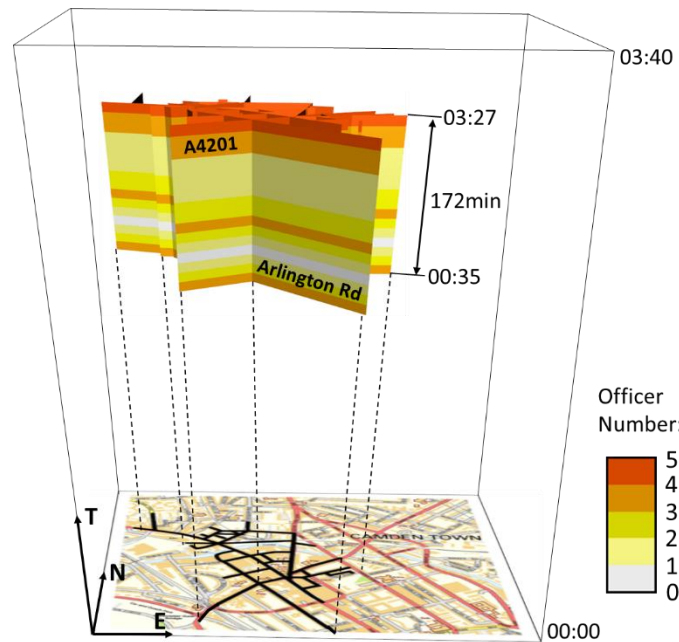


Figure 6.11 The 3D wall map visualisation of one ST-LOI in a space-time cube

The main goal for this visualisation is to support the joint expression and exploration of the ST-LOIs' temporal aspect, spatial (network) aspect and semantic aspect. Since we have only acquired the temporal and spatial information of the human's activities through Modules I and II, the first two aspects are included in the 3D wall map in this

section. Unlike the scatter map that can only roughly show the stay point density in a space-time cube, the 3D wall map enables us to add the heat map, similar to the data exploration heat maps in Chapter 4, onto the “walls” to display the ST-LOI’s exact visit intensity variation in time. The more officers are observed to have stayed within the 10-minute time brick of the ST-LOI, the hotter the time brick’s colour. As can be seen in Figure 6.11, some of the street names are marked out on the wall. This shows that the wall shapes also preserve the 2D geometry of street segments in the ST-LOI. Additional visualisation for the semantic information will be discussed in section 6.5 after Module III’s semantic analysis of places.

### 6.4.5 Complexity

The complexity of the traditional DBSCAN algorithm is  $O(m \log m)$ . ST-Net-DBSCAN does not change the runtime complexity of the algorithm’s ST-ROI determination process. However, the shortest-route search in the calculation of network distance between map-matched stay points, after the elementary spatial selection, has an extra run time of  $O(|vertices|)$  (Sedgewick & Vitter, 1986), where  $|vertices|$  is the number of end nodes and intersecting nodes that fall within the circular filter area of the elementary spatial selection.

Building the K-d tree has a time complexity of  $O(m \log m)$  when an  $O(m)$  median of the medians algorithm is used to select the median at each level of the nascent tree (Cormen et al., 2001). After construction of the K-d tree, using it for the Euclidean range search can bring an  $O(\sqrt{m} + |o|)$  worst-case time complexity (Ooi, 1987), where  $|o|$  is the number of output points of the Euclidean range query.

## 6.5 MODULE III: SPACE-TIME SEMANTIC ENRICHMENT IN ROAD NETWORKS

### 6.5.1 Network POIs

POIs in a city and the urban networks naturally bond to each other. On one hand, all POIs in a city can be accessed through streets of various levels and the majority of POIs possess geocoded street addresses when collected. On the other hand, POIs and waypoints are interchangeable synonyms in the urban navigation context (OpenStreetMap, 2017) and travellers in the city navigate themselves on a point-to-point basis through the urban network. Moreover, movements in the city following the geometry of streets and networks is a better representation of urban structures than raster-based methods when movements are involved. Therefore, generating semantic

ST-LOIs with POI information along the street segments can provide us with the semantic meaning of places at a finer scale.

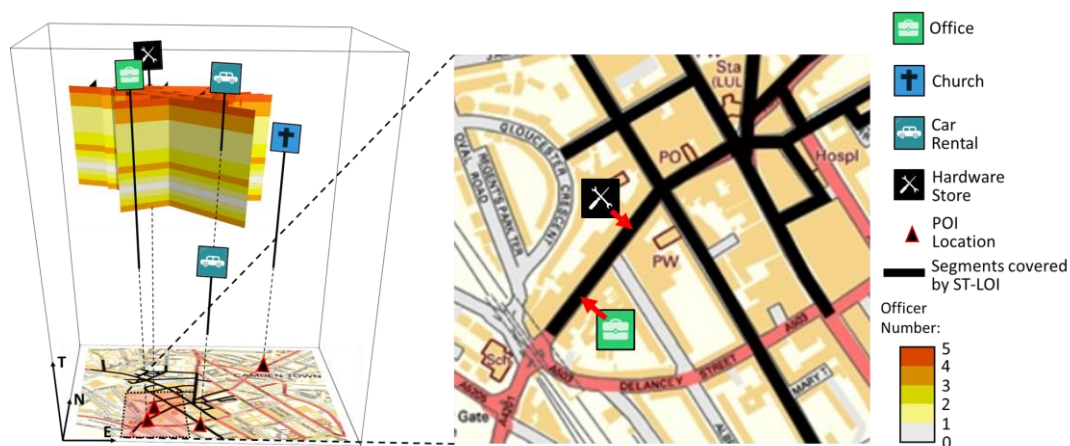


Figure 6.12 The relationship between ST-LOIs and the POIs along the streets

Figure 6.12 is a hypothetical example to demonstrate the method of ST-LOI semantic enrichment in street networks. The ST-LOI semantic enrichment method is based on an expression of space-time boundary different from ST-ROI. Its process can be divided into 3 steps:

- (1) Each ST-LOI's spatial boundary is defined as the union set of street segments to which the ST-LOI's stay points have been matched in Module I. The temporal boundary is the time span of the ST-LOI. The network space-time boundary is the combination of both.
- (2) The POIs are associated with the street segments containing the POIs' registered street addresses or streets closest to the POI. If an ST-LOI covers the associated streets of a POI, the POI is considered to be inside the ST-LOI's spatial boundary. POIs that fall inside a ST-LOI and, at the same time, have opening hours overlapping with such an ST-LOI, are considered to be inside the ST-LOI's network space-time boundary.
- (3) The TF-IDF algorithm is used to annotate the semantic information of ST-LOIs. All POIs in the study area are equivalent to the corpus. The POIs in the network space-time boundary of each ST-ROI are equivalent to a document, and the major categories of POIs are equivalent to the topics in text mining analysis. The semantic contribution of different (major) categories of POIs to ST-LOIs are weighted by the same method as used in the Euclidean paradigm.

As can be seen from Figure 6.12, the church is not located along the street segments within the ST-LOI's spatial boundary, so it is not considered in the semantic enrichment



process of this ST-LOI. The car rental service building has separated opening hours in a day. The earlier period of its opening hours do not overlap with the temporal boundary of the ST-LOI, so it also does not contribute any semantic meaning to the ST-LOI.

### **6.5.1 Visualisation**

The semantic enrichment module can quantify the semantic contributions of POI categories to each ST-LOI. To incorporate semantic information in the 3D wall map, each ST-LOI is semantically labelled with different colours according to the dominant semantic meaning (i.e. the major POI category with the largest semantic contribution) in its network space-time boundary. Besides, the shades of the colours can still be used to show the visit intensities, just like the 3D wall map for non-semantic ST-LOIs. The visit intensity and time space of semantic ST-LOIs vary dramatically. Thus, the visualisation should convey semantic information in addition to the spatio-temporal knowledge when viewed at different zoom levels and angles. An example of the visualisation result can be seen in the case study section of this chapter.

## **6.6 MODULE IV: PROFILE AGGREGATION**

Module IV is the aggregative analysis of the semantic profiles generated through the spatial and temporal analysis methods in Modules I, II and III. It does not involve any spatial operations or adaptive improvements for network analysis. Therefore, the procedure and methods in Module IV of the network paradigm are exactly the same as in Module IV of the Euclidean paradigm. After the individual space-time profiles are summarised based on ST-LOIs, they are transformed into semantic profiles according to Equation 5.10 and aggregated by the hierarchical clustering method with a JSD-based similarity metric (Equation 5.11). The results of hierarchical clustering are further evaluated and compared with the Euclidean paradigm and conventional approaches in Chapter 8.

## **6.7 CASE STUDY**

We tested the network paradigm with the same dataset as used in the extended multiple-borough case study in the previous chapter. We separate the ITN street network file of the study area into 12 parts according to the 2 km buffer zones of 12 boroughs, and consider police officers based in the 12 different BOCUs as 12 movement datasets. Modules I and II are organised in 12 workflows: each workflow processes the movement of officers in each BOCU in a spatial temporal way. For the intensive network computations in Modules I and II, we use UCL's Legion cluster computation platform

(University College London, 2017) to process the 12 work threads in parallel and significantly speed up the first two modules of the network paradigm. The outcomes of Modules I and II are then processed on a desktop PC for the methods of much lighter workloads in Modules III and IV.

In Module III, the outcomes (i.e. ST-LOIs and individual non-semantic profiles) of all workflows are combined and semantically enriched with the merged POI dataset of the entire study area. We execute Module III in a single workflow also because the text mining method requires a larger corpus to get high-quality results, which means a large enough number of POIs in all ST-LOIs of the study area should participate in the semantic enrichment process. In Module IV, all individual officers' semantic profiles are aggregated. This chapter presents the outcomes of every module, as well as the differences and improvements made to the network paradigm. For security seasons, we only showcase the results of three central London boroughs for demonstration.

#### **6.7.1 Map-matched observations**

Before all analyses, the raw movement observations are segregated into individual trips and the trip trajectories are turned into trip routes in the ITN urban theme layer by the ST-Matching algorithm (Lou et al., 2009). Figure 6.13 shows the observation points before and after snapping by the ST-Matching. To make the map concise, we choose the movement data of one week (August 1<sup>st</sup> – August 7<sup>th</sup>, 2015) in the three chosen boroughs to avoid too many points that cause messy visualisation. It can be seen that 138,041 observations during this week have been snapped onto the streets to recover the actual route taken by the officers. Afterwards, all the modules will be based on these post-snapping movements. The accuracy of the map matching for the movement in the case study and London's complex street networks cannot be directly tested with the APLS data because there is no ground truth of trip route recorded by the officers. We therefore design a synthetic trip route generator to mimic the true patrol routes and make up 200 artificial trip routes as well as simulated GPS observation errors to test the accuracy of the map-matching process in this module. The evaluation can be found in section 8.1 of the validation chapter.

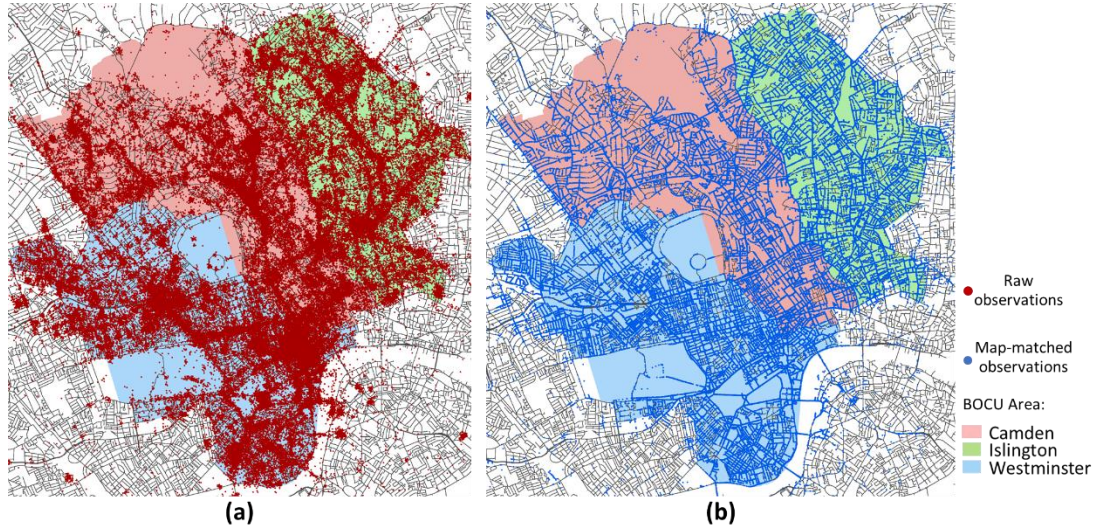


Figure 6.13 (a) Raw observations; (b) Map-matched observations

### 6.7.2 ST-LOIs and their visualisation

Here we use the same results of the ST-LOI detection in Camden, City of Westminster and Islington to demonstrate its merits over previous density-based clustering algorithms that use Euclidean distance as a distance metric. More than 1,800 officers working in the three chosen boroughs are recorded by the APLS. Among them, 620 officers generated more than five trips during August 2015, sufficient for the pattern analysis. The 620 outdoor active officers are selected and analysed. According to Equation 5.3, we set the  $Network\ Eps = 20 + 2\sigma = 36\ m$  and the  $Temporal\ Eps = 5\ min = 300\ s$ , and the  $minPts$  is set to be 55. Figure 6.14 is the visualisation of 67 ST-LOIs detected by the proposed method. The geographic names of some typical Net-ST-ROIs and their spatial coverages are also marked out in the figure.

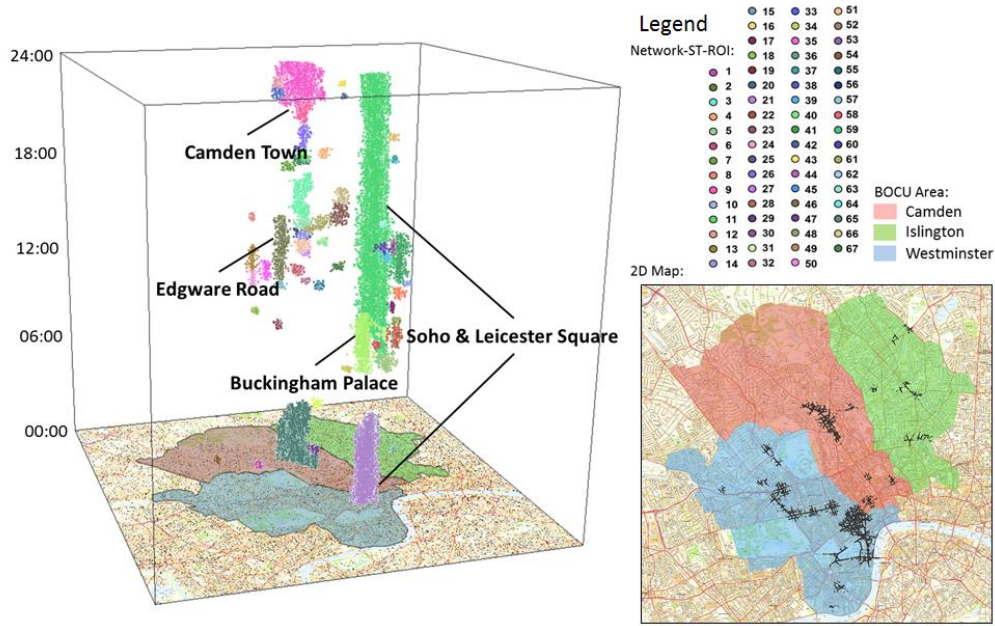


Figure 6.14 Space-time cube visualisation of the 67 Net-ST-ROIs detected

By comparing the 2D maps of the ST-LOIs (Figure 6.14) and the ST-ROIs (Figure 5.20) detected in the Euclidean paradigm, we can see that most ST-ROIs detected by conventional ST-DBSCAN are suborbicular clumps. Unlike normal DBSCAN methods, ST-Net-DBSCAN is able to preserve the geometry of the interesting regions, especially the line-shaped regions such as Oxford Street, the Mall near Buckingham Palace and even Westminster Bridge on the River Thames. It also allows some ST-LOIs that cannot be detected by normal DBSCAN to be detected in areas of low street segment density.

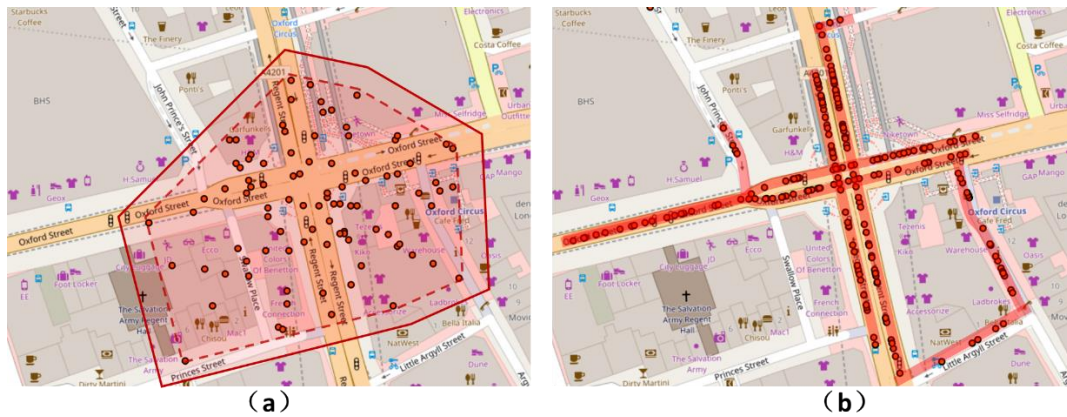


Figure 6.15 (a) The spatial boundary of an ST-ROI in Oxford Circus; (b) The network spatial boundary of an ST-LOI in Oxford Circus

Figure 6.15 shows an ST-ROI and an ST-LOI detected in the same area (i.e. Oxford Circus, London) in the afternoon with the same APLS dataset (i.e. August 2015). Figure 6.15 (a)

is the ST-ROI No.44 detected in the extended case study of the Euclidean paradigm in section 5.7, and Figure 6.15 (b) is the ST-LOI No.17 detected in the network paradigm. The comparison shows the differences between ST-ROI and ST-LOI in space. It can be seen that it is difficult to tell which stay point is in front of which POI or building in Figure 6.15 (a), whereas the ST-LOI is better ordered geographically. The ST-Net-DBSCAN is also able to discover stay points on some street segments not covered by ST-DBSCAN, and the Euclidean ST-ROI mistakenly covers some streets that officers never actually stopped at, probably due to GPS location error. It also shows that the stay points' relative locations to the POIs are also clarified because the map-matching process of the network paradigm can recover the true location and speed of the patrol activities with high certainty.

As illustrated in section 6.4.4, the 3D point clouds are not appropriate for visualising the detected ST-LOIs, whereas the 3D wall map can display the segment structures of ST-LOIs that we detected in Module II in space and time. Figure 6.16 is a sideview of the 3D wall map of the 67 detected ST-LOIs. The colours on the time bricks of the walls represent the visit intensity in every 10-minute interval. The hotter the colour, the more officers have stayed in an ST-LOI during the time brick's 10 minutes. In this way, not only the spatio-temporal information of ST-LOIs but also the time variation of officers ST-LOI visiting behaviours are visualised. It can also be seen that the high overall police visit intensity in the afternoon in Figure 6.16 also corresponds to the data exploration heatmap in Figure 4.9, although in comparison the 3D wall map is a great improvement. We incorporate this 3D wall map in a graphic user interface to allow zoom in, zoom out and rotate operations for the views in order to see and perceive the visualisation from different angles and different levels of detail. For example, a top-down angle of view allows the viewer to see the network structure of ST-LOIs in space, while a side profile can provide time span information of the ST-LOIs.



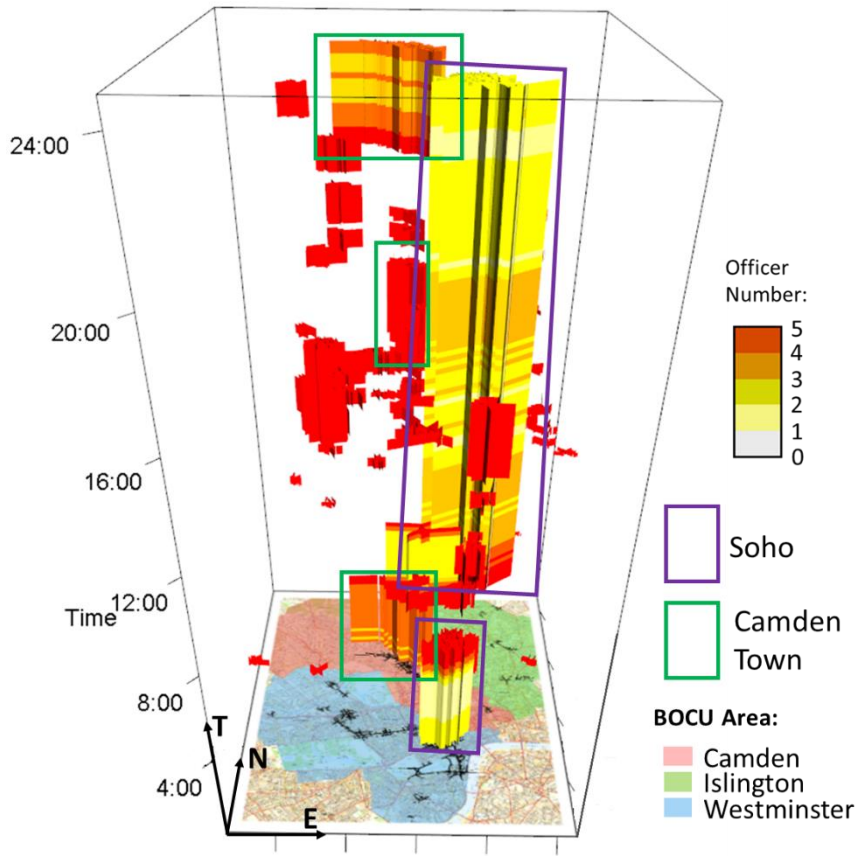


Figure 6.16 3D wall map visualisation of the 67 ST-LOIs

The ST-LOIs detected in Soho and Camden Town are also marked in rectangles in Figure 6.16. The spatial coverages of ST-LOIs are more clearly visualised by 3D wall maps. By observation, we find that the spatial boundaries of interesting places vary over time. The ST-LOIs detected in Camden Town before dawn and in the late evening show similar spatial boundaries, whereas the spatial coverage of the Camden Town ST-LOI in the afternoon shrinks back to the southeast side of the Regent's Canal. Similarly, the spatial boundary of the Soho ST-LOI in the early morning is evidently smaller than the Soho ST-LOI in other periods of the day. The possible semantic cause of this phenomenon will be discussed in the next sub-section.

### 6.7.3 Semantic enriched ST-LOIs

As illustrated in the methodological framework, we use the information of POI subcategories in every ST-LOI's and POI's opening hours as input to the TF-IDF algorithm to generate semantic ST-LOIs in a quantified manner. An ST-LOI's semantic meaning is described by the combination of the semantic contribution percentage of all POI categories. Among the 67 ST-LOIs detected by Module II, 7 of them are chosen to demonstrate the semantic enrichment results.

ST-LOI No.	2	40	1	4	3	17	13
Name of the place	Camden Town	Camden Town	Camden Town	British Museum	Parliament Square	Oxford Circus	HJE Hospital
Emergence Time	0:00	13:14	20:31	11:35	15:02	16:59	15:24
Perish Time	5:05	16:25	23:59	12:27	15:46	18:41	16:43
Accommodation, eating and drinking	0.3191	0.184600515	0.4068	0.1305	0	0.1759	0
Attractions	0.0371	0	0.0487	0.3027	0.4213	0	0
Commercial services	0.1067	0.237780257	0.1026	0.2778	0.0466	0.218	0.27
Education	0	0.039041175	0.0135	0	0	0	0
Government and organisations	0	0	0.0143	0.1719	0.2502	0.0073	0
Health	0.0577	0.08795296	0.0871	0	0	0.088	0.4921
Manufacturing and production	0	0	0.0064	0	0	0.0086	0
Public infrastructure	0.0966	0.025213847	0.0308	0	0.0335	0.0062	0
Retail	0.062	0.255141121	0.0644	0.1172	0.1091	0.3843	0.2379
Sport and entertainment	0.3207	0.025138659	0.2236	0	0	0.0212	0
Transport	0	0.145131465	0.0018	0	0.1594	0.0906	0

Figure 6.17 Semantic contributions of POI categories in ST-LOIs

The semantic contributions of POI categories in each of the 7 ST-LOIs in figure 6.17 generally fit people's understanding of the places. For example, retail POIs account for the largest contribution in Oxford Circus, and Health POIs stand out in the ST-LOIs near the Hospital of St John & St Elizabeth (HJE Hospital). Apart from this, Module III enables people to compare the semantic ST-LOIs by numbers and observe the temporal variation of a place's semantic meaning. For instance, ST-LOIs No.1, No. 2 and No.40 are all located in Camden Town and they reflect the busiest periods of this place; however, the semantic meaning keeps changing over time. It can be seen from Figure 6.17 that the transportation function of Camden is more significant during the day. The early morning ST-LOI (i.e. ST-LOI No.2) and late evening ST-LOI (i.e. ST-LOI No.1) of Camden Town attract visitors and police officers mainly because they are ST-LOIs with eating, drinking and entertainment related activities. A lot of POIs like bars and restaurants open during these periods, whereas ST-LOI No.40 shows a prominent function of retailing and shopping of Camden in the afternoon mostly because of Camden Market opening during the day time. This also explains why the spatial coverage of Camden Town ST-LOIs shrink in the daytime and expand at night. The market and most stores that are open in the daytime are distributed on the southeast side of the bridge over the Regent's Canal, whereas the bar and restaurants can be found on both sides of the canal.

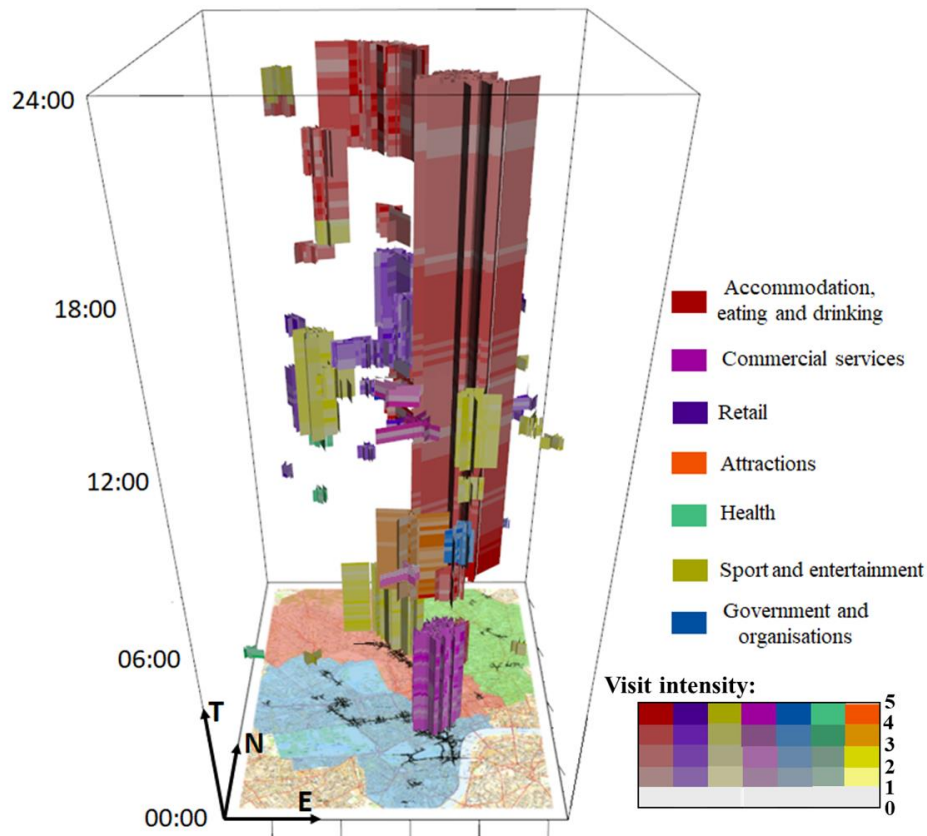


Figure 6.18 3D wall map visualisation of semantic ST-LOIs

In Figure 6.18, multiple colours are added into the 3D wall map to incorporate the semantic information. We colour the ST-LOIs differently according to the dominant POI category, i.e. the POI category that accounts for the largest share of semantic contributions in each ST-LOI. This improvement allows the viewer to intuitively compare the semantic meaning between places and observe the semantic meaning variation of a place over time. For example, the semantic meaning of Camden Town is dominated by the opening entertainment POIs before dawn and turns into a retail function in the daytime and then eating and drinking at night. Soho is a place of commercial services in the early morning before turning into a place for eating and drinking for the rest of the day.

#### 6.7.4 Semantic profiles and profile aggregation

The methods used in Module IV of the network paradigm for aggregative analysis of semantic profiles are the same as the ones in the Euclidean paradigm. After acquiring the individual dwelling time allocation in Module II and the semantic ST-LOI in Module III, we transform individual space-time profiles in into semantic profiles for hierarchical



clustering analysis. According to the Dunn index test (Dunn, 1973), it is most appropriate to separate the officers' activity patterns into 7 groups as shown in Figure 6.19.

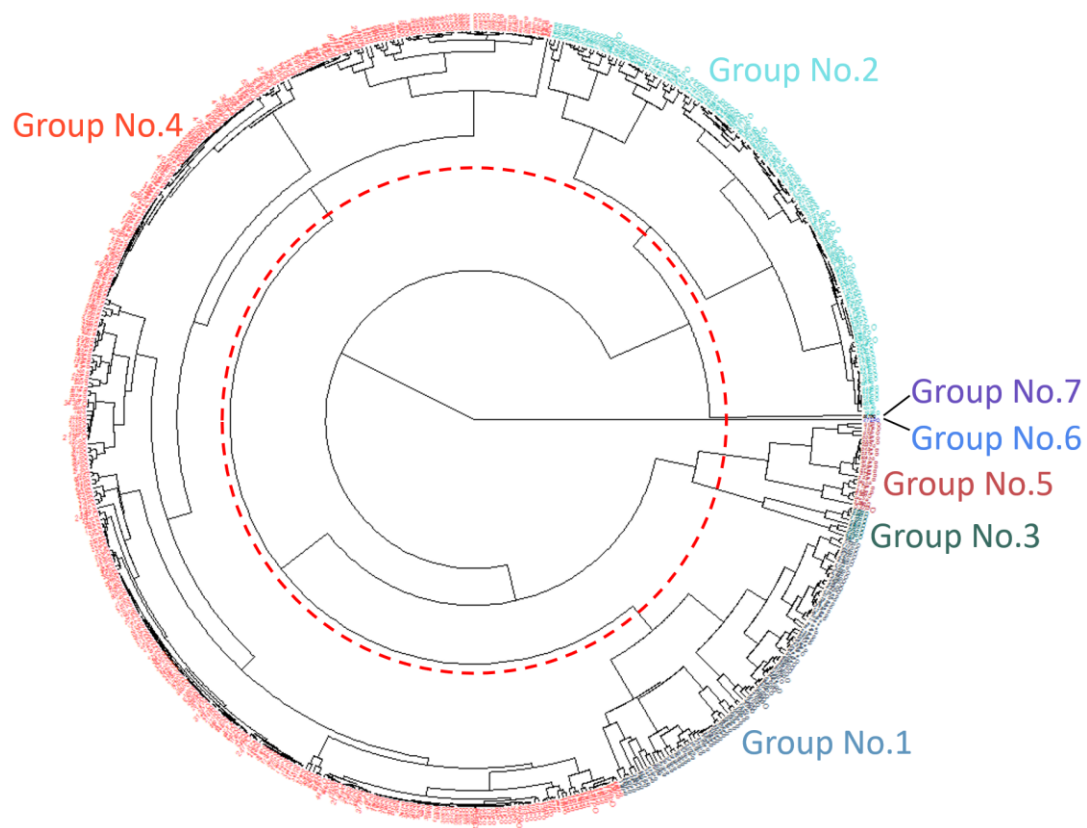


Figure 6.19 The 620 active officers separated into 7 groups

Because it is impossible to list all 620 officers' information and their profiles in the thesis, we present the officers in groups in Figure 6.20 by demonstrating the average semantic profile and the number of officers within each activity group for depiction of their activity patterns. The figure shows that different activity groups have different time allocation preferences. For example, Group No.1 spends more time in tourist attractions and government related areas than any other groups, while Group No.1 is more focused on streets of retail stores. Since our study area is near the city centre where no factories can be found, no one spent their time on ST-LOIs of a manufacturing type. The aggregative result also shows the methods ability to find behaviour outliers, i.e. officers that have semantic activity profiles different from all other officers. The only officer in Group No.6 has spent most of his/her working time in August 2015 on retail ST-LOIs and the only officer in Group No.7 spent much more time in public infrastructure and commercial services.

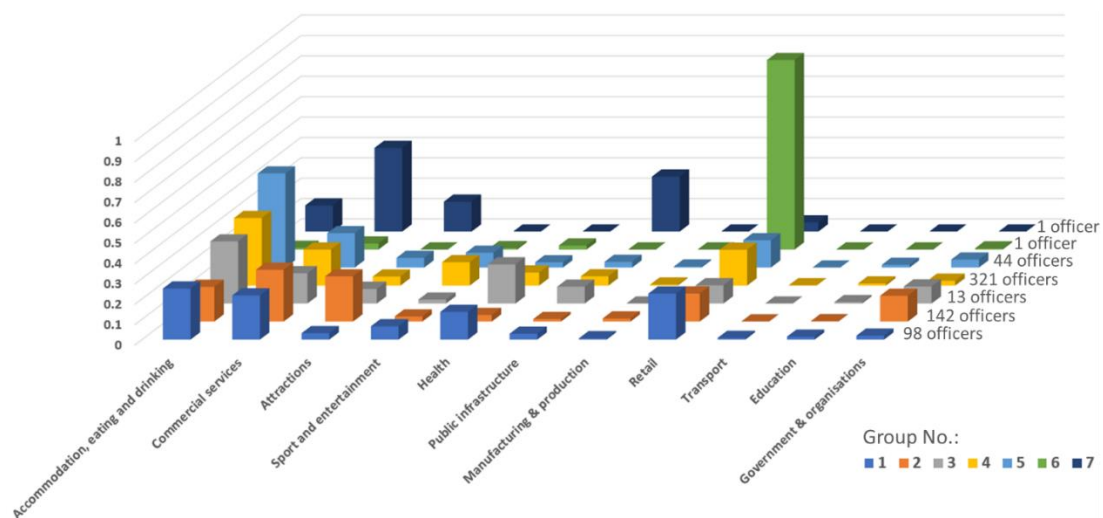


Figure 6.20 Average semantic profiles of officer subgroups

As introduced in section 4.2, the APLS data contain the work type information of officers, indicating the different sectors to which officers are attached. We also summarised the work type information of the grouped officers to look into the relationship between activity patterns and the officers' roles in the police force. Figure 6.21 is the pie chart showing the percentage of officer types in the 7 activity groups. Most community support officers are in Group No.2 and most special constables are in Group No.4. The section sergeants are very rare and they are not one of the common police work types listed in Table 4.1. The most interesting phenomenon is that all of the section sergeants are found in Group No.2 that has special interest in tourist attractions and governmental places. Due to the lack of ground truth behaviour records such as mission logs, we cannot analyse the cause of this phenomenon. Nonetheless, the grouping result shows that Module IV can provide quantified comparisons for the viewers to intuitively comprehend the activity patterns of officers and provide clues for more detailed inspections of officer behaviours.

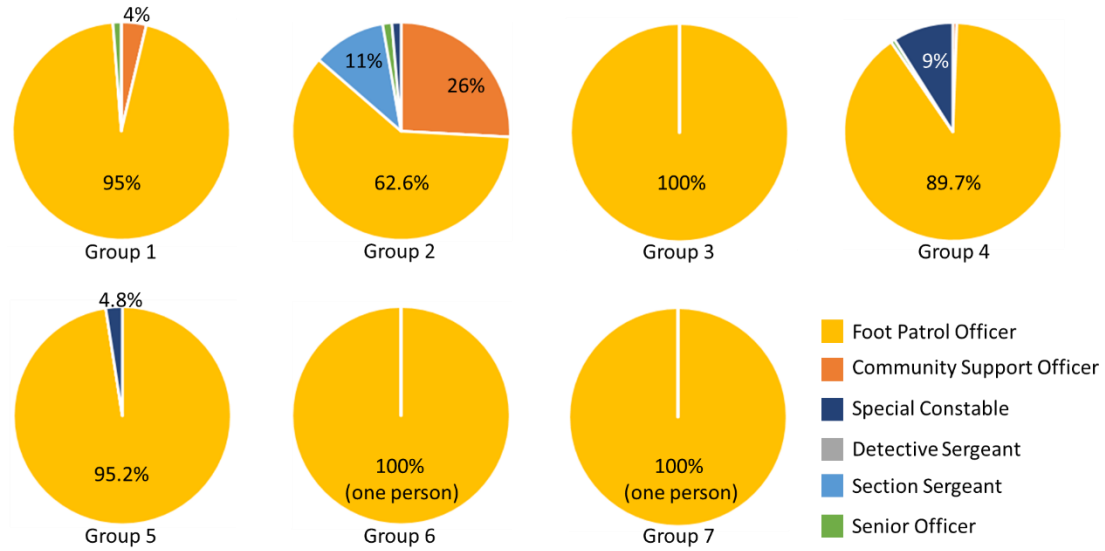


Figure 6.21 Work type composition within the officer subgroups

## 6.8 CHAPTER SUMMARY

In this chapter, we substitute the Euclidean distance metrics with the network distance and propose a network-based ST-DBSCAN algorithm (i.e. ST-Net-DBSCAN) to facilitate spatio-temporal clustering in urban road networks. We combine it with the existing map-matching algorithm to find ST-LOIs based on the movement trajectories in the city. To our knowledge, this is by far the first work to integrate map-matching and clustering method in networks for movement trajectory analysis. For the consequent increase of computation burden, we optimised the space-time retrieval strategy to speed up both the map-matching and the ROI-detection module. We also execute the workflow in Modules I and II in parallel to further accelerate the entire framework.

We tested it with real London's Metropolitan Police patrol data of 12 boroughs in London's complex and large urban network. Results showed that the algorithm can better pinpoint the street segments that officers are interested in and provide a more realistic tool for trajectory analysis in an urban context.

While the 3D scatter map gives a general view of stay point aggregation in space and time, the 3D wall map constrained by road networks avoids the information loss of the network structure in the visualisation. It also provides more precise and more sensible locations of stay points and ST-LOIs, and reveals how the visiting intensity changes in each street to provide comprehensive insights for the viewer.

The network paradigm generates better results than the existing Euclidean-based methods and overcomes the limitations of the previously proposed Euclidean paradigm

in terms of location accuracy, time sensitivity and legibility of the cluster boundaries in an urban context, and it is also scalable for large problems. Many previous studies on ROI detection have based their spatial metrics on the Euclidean distance and ignored the influences of location error and road networks. Other studies ignore the temporal dimension in human activities. The ST-Net-DBSCAN generates ST-LOIs that can fit into the structure of the road networks, truly reflect the patrolled street segments and consider the time span as an important aspect of the places. These modifications enable the method to pinpoint the locations and coverages of interesting regions with higher accuracy and hence help improve confidence in further semantic enrichment of the stays when needed. Detailed validations and comparisons of methods can be seen in Chapter 8.

The major limitation of the network paradigm is the accessibility of the POI data and the accuracy of the semantic enrichment process. The Ordnance Survey POI information we rely on for Module III is hierarchically categorised a priori. Nevertheless, a lot of POI data in other areas and countries are not that well organised. In the next chapter, we introduce a more advanced text mining algorithm that does not require a predefined hierarchical POI classification scheme and enables the semantic enrichment of places with higher accuracy.

**Chapter 7**

# **Improvements in Semantic Enrichment Module**

## 7 IMPROVEMENTS IN SEMANTIC ENRICHMENT MODULE

Chapter 6 has demonstrated a space time semantic enrichment approach based on TF-IDF algorithm and achieved satisfying outcome. However, this approach requires the POI dataset to be well organised and hierarchically categorised, whereas any real-world POI datasets are not hierarchically categorised (e.g. google places). Even though some POI datasets are hierarchically categorised, their global classification scheme do not necessarily fit the purpose of semantic enrichment in cases of particular cities. Thus, a semantic enrichment approach to serve the same purpose without a hierarchical POI classification scheme is needed. To this end, we introduce a semantic probability-based topic extraction model, Latent Dirichlet allocation (LDA) (Blei et al., 2003), to replace TF-IDF. LDA is an unsupervised generative model that can dig out hidden topics from a large collection of documents. As reviewed in Chapter 2, It is a more sophisticated and objective algorithm compared to conventional approaches based on word frequencies such as TF-IDF.

### 7.1 METHOD DESCRIPTION

The LDA works by statistically grouping words into potential topics by studying their occurrences across different documents/sentences and represent the meaning of documents as mixtures of topics with different probabilities. As a bag-of-words model, the main premise of LDA topic modelling in semantic enrichment application is that co-occurring same types of POIs in the same ST-LOIs are assumed to be related or bear similar semantic meaning for human activities, and are therefore more likely to be assigned to the same semantic category.

From the mathematical perspective, a topic possesses a semantic meaning or concept and expresses a series of related words with conditional probability; Each document containing different words can be seen as a probabilistic distribution of multiple topics. In short, the more correlated the word is with a certain topic, the greater the word's conditional probability is and vice versa. In our case of platial semantic enrichment, we regard all street segments in a ST-LOI as a document and the semantic category of the ST-LOI as a topic. In this way, the POIs located in the streets of ST-LOIs can be seen as words and terms. This means that, unlike TF-IDF in which each POI can only belong to one single category, each POI in LDA model can be given a corresponding probability of it belonging to different semantic categories to varying extents just like each word can be associated with different topics. Likewise, an ST-LOI can also contain multiple semantic meanings with different probabilistic weights just like a document possessing

various topics in text mining. This analogy between semantic enrichment and text mining is demonstrated in Table 7.1.

Table 7.1 Analogy from textual topics to semantic analysis of places

Text mining	Semantic enrichment of places
Corpus	→ All POIs
A word in a document	→ A POI in a ST-LOI
Documents/sentences	→ Combination of all POIs in a ST-LOI
Topic of a document	→ Semantic meaning of a ST-LOI
Topic assignment	→ Semantic enrichment

With this analogy in mind, the probability of each type of POI appearing in the space time boundary of a certain ST-LOI can be illustrated in a generative model as follows:

$$p(POI \text{ in } ST - LOI) = \sum_{semantic} p(POI \text{ in } semantic) * p(semantic \text{ in } ST - LOI)$$

Equation 7.1

This Equation can also be expressed in a matrix form as show by Equation 7.2.

$$\begin{bmatrix} p_{11} & p_{12} & \dots & p_{1n} \\ p_{21} & p_{22} & \dots & p_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ p_{m1} & p_{m2} & \dots & p_{mn} \end{bmatrix} = \begin{bmatrix} \phi_{11} & \phi_{12} & \dots & \phi_{1t} \\ \phi_{21} & \phi_{22} & \dots & \phi_{2t} \\ \vdots & \vdots & \ddots & \vdots \\ \phi_{m1} & \phi_{m2} & \dots & \phi_{mt} \end{bmatrix} \times \begin{bmatrix} \theta_{11} & \theta_{12} & \dots & \theta_{1n} \\ \theta_{21} & \theta_{22} & \dots & \theta_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \theta_{t1} & \theta_{t2} & \dots & \theta_{tn} \end{bmatrix} \quad \text{Equation 7.2}$$

As described in Equation 7.2,  $p_{mn}$  is the probability of the m-th POI in the n-th ST-LOI;  $\phi_{mt}$  is the probability of the m-th POI in the t-th semantic category; and  $\theta_{tn}$  represents the probability of the t-th semantic category in the n-th ST-LOI. Consequently, the  $p$  matrix represents the probability of each POI falling in each ST-LOI; the  $\phi$  matrix represents the probability of each POI falling in each semantic category  $c$ ; and the  $\theta$  matrix represents the contribution of each semantic category to the meaning of the ST-LOI. After detecting all ST-LOIs in the study area, a tokeniser can use all POIs in the space time boundaries of all ST-LOIs to generate a  $p$  matrix. The process of the LDA algorithm is to iteratively derive the corresponding  $\phi$  matrix and  $\theta$  matrix in the

process of training this  $p$  matrix with Gibbs sampling method (Griffiths & Steyvers, 2004).

LDA require two parameters to be determined before implementation.  $\alpha$  is the parameter of the Dirichlet prior on the per-document topic distributions.  $\beta$  is the parameter of the Dirichlet prior on the per-topic word distribution. A low  $\alpha$  value places more weight on having each document composed of only a few dominant topics, whereas a high  $\alpha$  value will generate many more relatively dominant topics. Similarly, a low  $\beta$  value places more weight on having each topic composed of only a few dominant words. In the implementation of LDA, different values of  $\alpha$  and  $\beta$  are iteratively tested with log likelihood. The combination of  $\alpha$  and  $\beta$  that generates the highest and most stable log likelihood is selected as the most suitable parameter input to the algorithm. The mathematical details of log likelihood will be discussed in Section 7.2.2 together with the evaluation method.

In the case study of Inner London, we implement the LDA semantic enrichment using the R text mining package. Subsequently, we illustrate the results of different semantic meanings with different colors in the same 3D wall map visualisation method in Section 6.4.4.

## **7.2 CASE STUDY**

### **7.2.1 The multiple-borough case**

Here we infer the semantic meanings of each ST-LOI in Inner London by applying a basic LDA algorithm instead of TF-IDF to the same dataset used in Section 5.7 and Section 6.7. To demonstrate LDA's ability to generate the semantic categories based on non-hierarchically categorised POIs, we abandon the OS POI dataset in this case study. Instead, we use the POI dataset and the POI classification scheme (see Appendix C) of Google Places in Inner London as the input corpus of the LDA algorithm.

The LDA model can firstly generate a “top words - topic” table for researchers to determine the thematic description for each “topic” (i.e. semantic category) intuitively. The “top words - topic” table generated for our case study is demonstrated in table 7.2. The top 8 probabilistically significant subcategories of POIs are listed for each semantic category. According to the contents of the “top words”, we manually annotate the 7 categories as food and eating, shopping, health and beauty, financial and public services, local life centre, entertainment and night life, and attractions. The higher the rank of a “word” (i.e. POI subcategory) in a semantic category's top word list, the more



contribution it makes to form the category. Some POI subcategories appears in multiple categories and make these categories more similar to each other than the rest of the categories. However, they can still be distinguished semantically by taking more “top words” into account. For example, the “store” POIs rank first in both Category 2 and Category 5. However, the other POI subcategories listed in these two categories in Table 7.1 shows that Category 2 is related to more specialised and luxurious shopping activities, whereas Category 5 focuses on the retail of daily necessities.

Table 7.2 Semantic categories summarised by LDA though the combination of popular POI subcategories

	Category 1	Category 2	Category 3	Category 4	Category 5	Category 6	Category 7
Annotat ion	Food & eating	Shopping	Health & beauty	Financial & public services	Local life centre	Entertain ment & night life	Attractio ns
1	Food	Store	Health	Finance	Store	Bar	Museum
2	Restaura nt	Clothing store	Beauty salon	Travel agency	Food	Night club	Park
3	Bar	Shoes store	Haircare	ATM	Grocery	Food	Restaura nt
4	Café	Jewelry store	Spa	Real estate agency	Homegoo d store	Restaura nt	Place of worship
5	Liquor store	Painter	dentist	Lawyer	Laundry	Movie theater	Store
6	Convenie nce store	Meal takeaway	Gym	Park	Café	Casino	Lodging
7	Meal takeaway	Shopping mall	Doctor	Church	General contracto r	Lodging	Florist
8	Book store	Pet store	Pharmac y	Local governm ent	Bakery	Car rental	Stadium

The same 3D wall map used in Section 6.7 is used to visualise the ST-LOIs semantically enrichment by LDA (Figure 7.1). The orange rectangles in Figure 7.1 mark out the ST-

LOIs located in Oxford Street. These ST-LOIs are identified by LDA as Category 2 semantic ST-LOIs and related to shopping activities. Compared with the outcomes of TF-IDF in Figure 6.18, the LDA can distinguish the major shopping street in a city with minor commercial areas in local districts. The LDA algorithm also generates similar results in some ST-LOIs as the TF-IDF method. For instance, the changes of semantic meanings in Camden Town and Soho are detected and the unique tourist attraction in Buckingham Palace is also correctly identified.

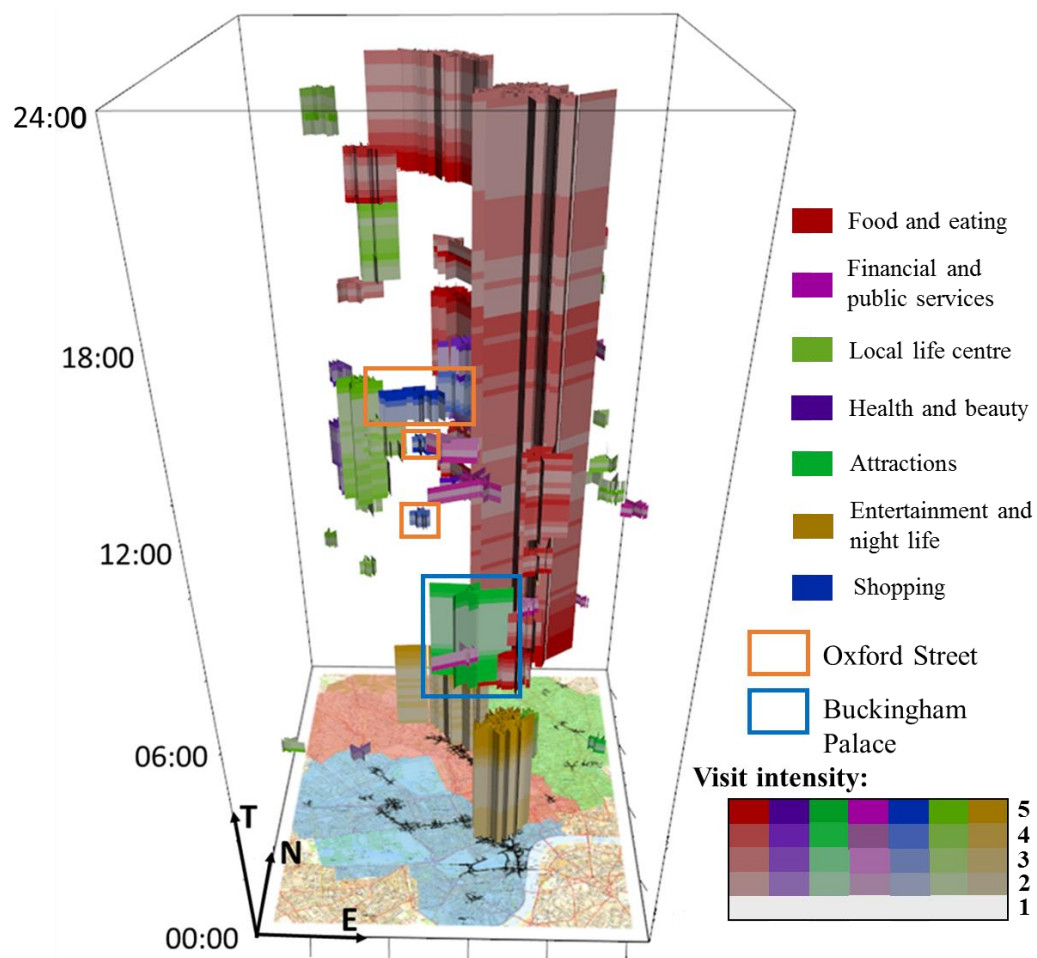


Figure 7.1 3D visualisation of the outcomes of LDA semantic enrichment

### 7.2.2 Model evaluation

The LDA model is based on the sampling of part of the “words” (i.e. POIs), therefore the results are inconsistent when the samples for training the model are changed in every iteration. Therefore, we use the goodness of fit to evaluate the LDA model. Log

likelihood (Griffiths & Steyvers, 2004) is the most commonly used indicator for this evaluation. For LDA, the whole input dataset of POIs in ST-LOIs are splitted into two parts: one for model training, the other for testing. The log likelihood in this case is calculated based on the sampled training set in iterations as illustrated in Equation 7.3.

$$\text{Loglikelihood}(\{d'\}|D) = \sum_{i=1}^N \sum_{j=1}^{\text{length}(d)} \log(p(d_{i,j}|D)) \quad \text{Equation 7.3}$$

where  $D$  denotes all the POIs in all ST-ROIs;  $\{d'\}$  is the sampled training set of ST-ROIs;  $N$  is the number of ST-ROIs;  $\text{length}(d)$  is the number of POIs in a ST-ROI and  $d_{i,j}$  is the  $j$ -th POI in the  $i$ -th ST-ROI. The larger the value of the log likelihood, the better the fit. The likelihood curve's stabilising speed and consistency across multiple runs indicate the convergence of the model (Griffiths & Steyvers, 2004) and the proper assignment of the semantic categories to POIs.

To demonstrate the relationship between the log likelihood with the size of our input data (i.e. POIs and ST-LOIs in London), we apply the LDA semantic enrichment in different scales in London. The scales of the input data vary from the ST-LOIs detected in August 2015 in one borough (i.e. Camden) to ST-LOIs all Inner London boroughs. The log likelihood curves of different tests are demonstrated in Figure 7.2. The curves show that as the scale and number of ST-LOIs increase, the log likelihood become higher and the curve is more likely to be a convergence after many iterations. The stability and goodness of fit of the LDA model are improved as the study area and the input data expand. This means that the LDA method is not suitable for a small study area because the results are unstable when the input number of POIs and ST-LOI are small. However, the increased and converged loglikelihood in experiments on large study areas demonstrates that this method is stable for large number of POI inputs. Therefore, the LDA semantic enrichment module is only appropriate for large study areas and the stability of the result should be verified before using the result as the final conclusion.

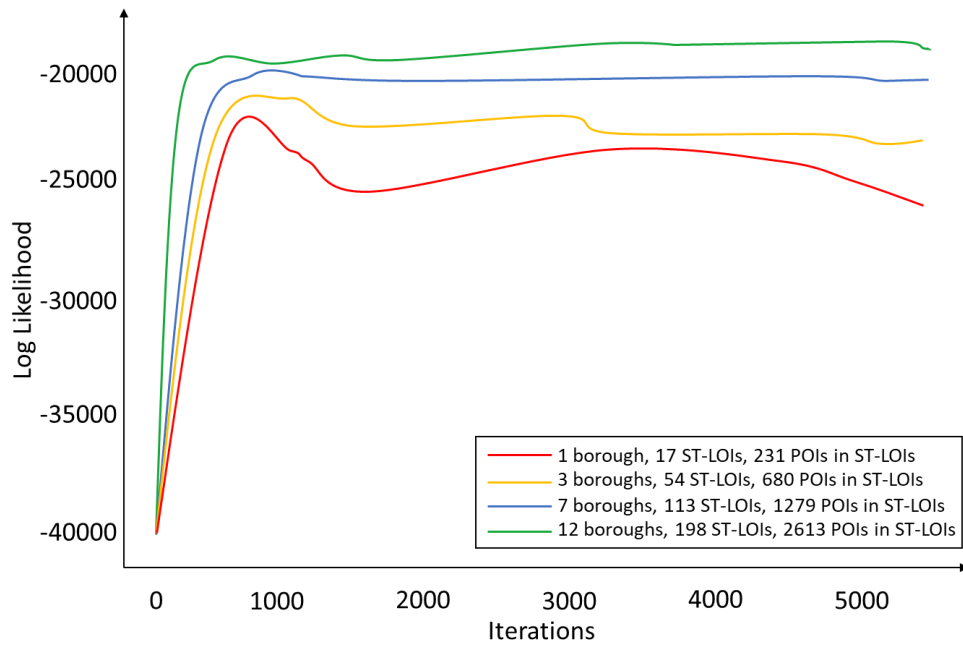


Figure 7.2 Loglikelihood values of LDA outcomes based on different scales of inputs

The differences and relationship between the outcomes of TF-IDF and the outcomes of LDA are visualised with a alluvial diagram in Figure 7.3. Although it should be bear in mind that the outcomes of the two approaches are generated based on different input POI data and therefore cannot be used to judge the performance of these two methods, the comparison shows how the choice of algorithm and source data can influence the perception of the meaning of places and time periods. For example, most of the ST-LOIs recognised as retail category and eating category by TF-IDF are also recognised by LDA as shopping and food. In contrast, not all health and beauty related places in the LDA approach come from the health related ST-LOIs discovered by TF-IDF. This is probably because the health and beauty category in LDA include wider range of facilities than the health category in TF-IDF. The most conspicuous difference between Figure 6.18 and Figure 7.1 is the ST-LOI in London's Soho district between 0:00 am to 4:00 am. The TF-IDF sees this ST-LOI as a place of commercial services, whereas the LDA identifies it as a place related to entertainment and night life. Based on the fact that the time period is late at night and Soho is famous for night entertainments, the LDA's judgement of the place's semantic meaning seem more reasonable in this case.

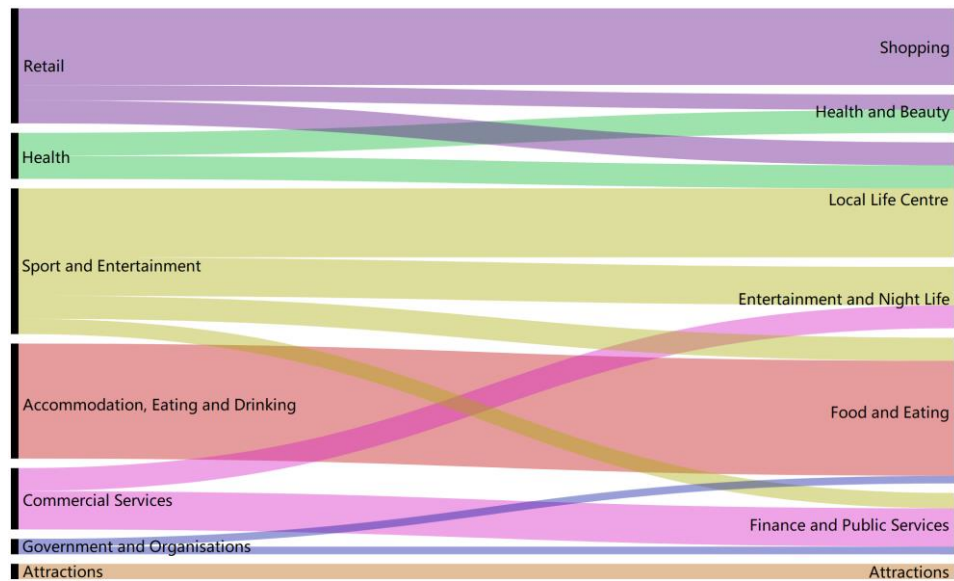


Figure 7.3 Alluvial diagram showing the outcome comparison of TF-IDF and LDA approaches in Camden, Islington and Westminster

### 7.3 CHAPTER SUMMARY

This chapter provides an alternative methodological option for Module III in the proposed framework. By applying the LDA algorithm, we can infer the semantic meaning of each ST-LOI based on simple and non-hierarchical POI classification schemes and looking into the contents that contribute to each semantic categories in greater details. It overcomes the shortcomings of predefined hierarchical POI classification schemes and can automatically searches for the semantic meaning among great amount of POIs.

To sum up, the LDA-based semantic enrichment module has four major features that distinguish it from conventional approaches:

1. This method does not need a hierarchical classification scheme of POIs as inputs, which means its applicability for different POI data sources is expanded.
2. This method can automatically summarise semantic “topics” and generate different POI classification scheme in different cities.
3. Each POI can belong to multiple semantic categories, which is more realistic.
4. The outcome of this method is not stable with a small dataset, but its performance increases with the size of the input dataset. This indicates that this method is more appropriate for the semantic enrichment of large POI datasets and large study areas.

## **Chapter 8**

# **Further Model Comparison and Validation**

## **8 FURTHER MODEL COMPARISON AND VALIDATION**

The drawbacks and advantages of single conventional approaches and paradigms of the proposed framework have been briefly discussed during the description of the methods and algorithms in Chapters 4, 5 and 6. In this chapter, multiple approaches and paradigms are compared to demonstrate the improvements achieved by the proposed methodological framework.

Here the comparisons are organised according to the four modules used in both Cartesian and network paradigms. Section 8.1 describes an artificial route and trajectory generator to provide simulated ground truth information so that the accuracy of the stay point identification and map matching methods in Module I can be evaluated against conventional approaches. Section 8.2 evaluates the results of space time clustering methods in the two proposed modules and conventional approaches with predefined spatial and temporal performance indicators respectively. Section 8.3 uses distance from stay point to POIs to measure how good the detected ROIs are in supporting semantic enrichment algorithms. Section 8.4 summarises the comparisons and emphasises the advantages of the network paradigm over other approaches.

### **8.1 MODULE I**

The major task of Module I in the proposed methodological framework in this thesis is to identify the stop episodes in individual trips with high accuracy so that the afterwards modules can be provided with reliable location and temporal information of the stops. Therefore, the accuracy of stop identification is needed in this module. The APLS dataset, however, was originally used for police operations and did not include ground truth records reporting the true and precise locations of the officers in patrol. This means that direct accuracy evaluations of stay point identification and map-matching based on the original APLS dataset is impossible. To overcome the lack of ground truth of trip routes, an artificial trajectory and route generator is designed in this chapter to simulate the APLS movement data with artificial positioning errors as well as the precise route taken during the artificial trips. Section 8.1.1 describes the working mechanism of the route and trajectory generator. Section 8.1.2 and Section 8.1.3 then use the simulated trajectories and trip routes to evaluate the performance of stay point identification and map-matching methods applied in Module I of the proposed framework.

### 8.1.1 Artificial route and trajectory generator

Our urban trajectory and route generator is developed on the basis of Kami et al. (2010). The trajectory generating process is as follows:

- (1) Random generation of 800 (x, y) coordinates on the road segments in Camden. From the 800 generated coordinates, randomly take out four points (A, B, C, D) as a group every time so that the 800 coordinates are divided into 200 groups. For each group of coordinates, a synthetic trajectory of a trip is represented by a polyline starting from A, moving through B and C and ending at D. 200 synthetic trips are generated in this step.
- (2) Calculation of shortest road network paths from A to B, from B to C and from C to D in each trip using igraph package in R. The synthetic route of each trip is generated by linking the network paths through A, B, C and D in each trip.
- (3) Transformation of trip routes into GPS point updates by adding points along each route with a regular spacing of 400 m – corresponding to a preferred walking speed of 1.33 m/s in London and the 5-minute sampling rate of APLS.
- (4) Randomly shifting the coordinates of each update in space to simulate the GPS positioning errors along routes. The shifting direction is random and the shifting distance of points follows the normal distribution of EPE according to Chapter 4, centring on 20m with a standard deviation of 8 m.
- (5) Selecting one to five data points per track randomly to be stop episodes. The dwelling time of each stop episode is set to follow a normal distribution centring on 40 minutes with a standard deviation of 8 minutes.
- (6) Generating additional stay points in the stop episode. According to the duration of each stop, add one stay point every 5 minutes at the stop episode's location with a simulated distance shift that follows a normal distribution of a 40m mean value and a 16m standard deviation. This distance shift is larger than the simulated EPE on the move because people are more likely to stop by a building or even enter the building and thus generate more imprecision in their location updates.

### 8.1.2 Accuracy of stay point identification methods

To demonstrate the advantages of the kernel-based temporal scanning window over conventional stop identification methods, the artificial route and trajectory generator in section 8.1.3 is used to simulate the ground truth trajectories for validation. It mimics



the features of the individual movements in the original APLS and generates 200 synthetic trips with error, speed and stop durations similar to the real APLS data. The artificial data contain the street segments covered by the route of each trip and the sequence of location updates with synthetic positioning errors. We test our pre-processing methods of Module I on the artificial trajectories to provide a reference of its performance in processing real urban movement data. We also compared the accuracy of several conventional (speed/distance) threshold-based and density-based stay point identification methods with ours. The results are listed in Table 8.1.

We define the stay point accuracy of stop identification algorithms as  $Accuracy_p$  in Equation 8.1.  $Accuracy_p$  measures how correct the algorithms are in labelling stay points.

$$Accuracy_p = \frac{Count(TP)}{Count(TP)+Count(FP)+Count(FN)} \quad \text{Equation 8.1}$$

where  $Count(TP)$  is the number of stay points that the algorithm identified correctly (i.e. true positive),  $Count(FP)$  is the number of points that are not real stay points, but the algorithm identifies as stay points (i.e. false positives), and  $Count(FN)$  is the number of stay points the algorithm fail to identify (i.e. false negative).

For the correctly identified stops, we define the dwelling time accuracy as  $Accuracy_t$  in Equation 8.2.  $Accuracy_t$  measures how precise the algorithms are in calculating the dwelling time for each correctly labelled stop.

$$Accuracy_t = \sum_0^{Count(TP)} \frac{2*|D_{true}(TP) \cap D_{identified}(TP)|}{|D_{true}(TP)|+|D_{identified}(TP)|} / Count(TP) \quad \text{Equation 8.2}$$

where  $D_{true}(TP)$  and  $D_{identified}(TP)$  are the true dwelling time period and the identified dwelling time period of each correctly labelled stop episode  $e$ , respectively.  $|D|$  is the length of the dwelling time.  $|D_{true}(TP) \cap D_{identified}(TP)|$  is therefore the overlapping time length between the identified and true stop episodes.

As shown in Table 8.1, our methods, KTSW and KTSW-network adaption, achieved significant higher accuracy in both stop identification  $Accuracy_p$  and dwelling time  $Accuracy_t$ , and KTSW- network adaption is slightly higher than the pure STKW in the

stop identification, but 6% higher in the dwelling time accuracy. Therefore, the method used in the network paradigm is more accurate in stop identification.

Table 8.1 Accuracy comparison of stay point identification methods

Algorithm	Stopher et al. (2005)	Schüessler and Axhausen (2009)	Thierry et al. (2013)	Kernel-based temporal scanning window (KTSW)	KTSW with network adaption
Type	Distance threshold-based	Speed threshold-based	Kernel density-based	Kernel density-based with time constraint	Integration of time and spatial network
Tested trips	200	200	200	200	200
TP number	6005	6223	5371	6192	6278
FP number	757	670	510	314	258
FN number	561	343	1195	374	288
Accuracy <sub>p</sub>	82%	86%	76%	90%	92%
Accuracy <sub>t</sub>	72%	79%	65%	84%	90%

### 8.1.3 Map-matching accuracy

The artificial route simulated by the trajectory generator proposed in Section 8.1.1 can also be used to evaluate the accuracy of map-matching techniques. We use the percentage of correctly matched records in the 200 simulated trip trajectories to evaluate the accuracy of map-matching algorithms. If one point in a trip is correctly matched to the road segment the simulated trip route had gone through, the point will be considered as correctly matched. Table 8.2 shows the matching accuracy of the simple strategy of snapping to the nearest segments, spatial incremental map-matching, and ST-matching. It also demonstrates that the ST-matching achieved highest map-matching accuracy.

Table 8.2 Accuracy comparison of map-matching methods

Algorithm	Snapping to the nearest segment	Global incremental map-matching (Yin & Wolfson, 2004)	ST-matching (Lou et al., 2009)
Tested trips	200	200	200
Accuracy	58.5%	87.1%	91.4%

## 8.2 MODULE II

Module II in the two proposed paradigms generates ST-ROIs and ST-LOIs with density-based space time clustering algorithms. As clustering results, the ROIs (Spatial ROIs, ST-ROIs, ST-LOIs) can be validated by the cohesion of the points within the clusters. Section 8.2.1 evaluates the cohesion of stay points in ROIs in space and Section 8.2.2 evaluates the cohesion of stay points in ROIs in time. The ST-LOI detection algorithm in Module II of the network paradigm is more complex than conventional approaches. Therefore, we also evaluate how the proposed network query strategy in Section 6.4.3 has decreased the time cost of the ST-network-DBSCAN algorithm.

### 8.2.1 Spatial cohesion of points in ROIs

One important indicator for evaluating the spatial clustering results is the cohesion of the points in each cluster. Here we measure the spatial cohesion by calculating the average Euclidean distance from each stay point in a detected ROI to its  $k$ -th nearest neighbouring points ( $k$ -NN). The smaller the distance, the closer the points in the ROIs are to their neighbouring points in space and the better the result. The average  $k$ -NN distance of a ROI is defined by Equation 8.3.

$$\text{Average } k\text{NN distance} = \frac{\sum_{ROI} \sum_{n=1}^k D(p, NN_n)}{\text{count}(ROI) \cdot k} \quad \text{Equation 8.3}$$

where  $p$  is a stay point in a ROI and  $ROI$  contains all stay points in that ROI.  $NN_n$  is the  $n$ -th nearest stay point to  $p$  in the ROI.

Figure 8.1 demonstrates the concept of  $k$ -NN with three examples of different chosen points' neighbours. Figure 8.1 (a) shows the  $k$ -NNs of point A when  $k = 3$ , while Figure 8.1 (b) shows the  $k$ -NNs of point A when  $k = 4$ . As demonstrated in Figure 8.1 (c), the distribution of point B and its neighbouring points are sparser than A and its

neighbouring points in Figure 8.1 (b) when  $k = 4$ , therefore, the average k-NN distance from  $B$  to its neighbours are larger than the average k-NN distance of  $A$  and its neighbours when  $k = 4$ . Hence, the spatial cohesion of  $A$  and its four nearest neighbours is higher than  $B$  and its four nearest neighbours. In comparing the ROIs detected by different methods, high spatial cohesion indicates better result.

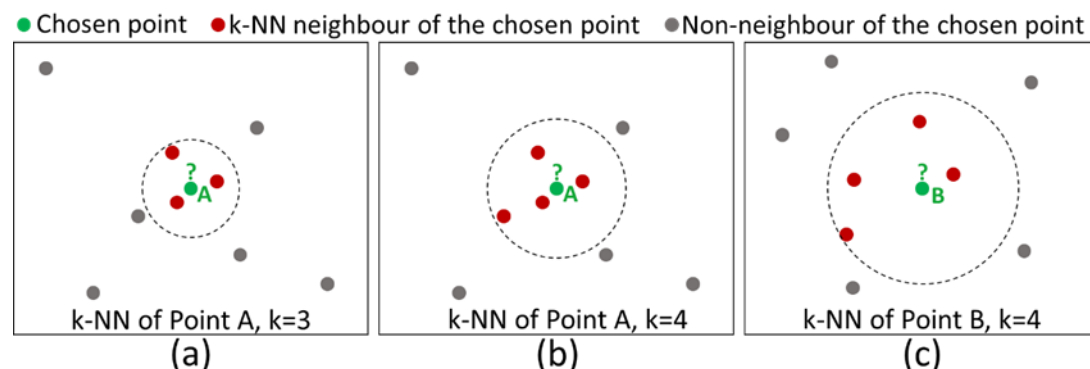


Figure 8.1 Three examples of k-NN queries of chosen points (a) the top 3 nearest neighbour points of point A; (b) the top 4 nearest neighbour points of point A; the top 4 nearest neighbour points of point B

The average k-NN distance of points in the Spatial -ROIs detected by conventional DBSCAN, ST-ROIs detected by the Euclidean paradigm and the ST-LOI detected by the network paradigm in the same APLS dataset of August 2015 are calculated and compared in Figure 8.2. The result shows that the points in Spatial-ROIs are the most aggregative, then is the ST-LOI and the ST-ROIs. This is because the conventional DBSCAN ignore time differences and counts stay points in the same place but at different times into one single ROI. Therefore, the stay points in a Spatial-ROI are far more than those in a ST-ROI or a ST-LOI, which causes the higher spatial cohesion. The lower average k-NN distances of the ST-LOIs than ST-ROI (when  $k \leq 8$ ) indicates the better outcome in the spatial cohesion of ROIs contributed by the network paradigm.

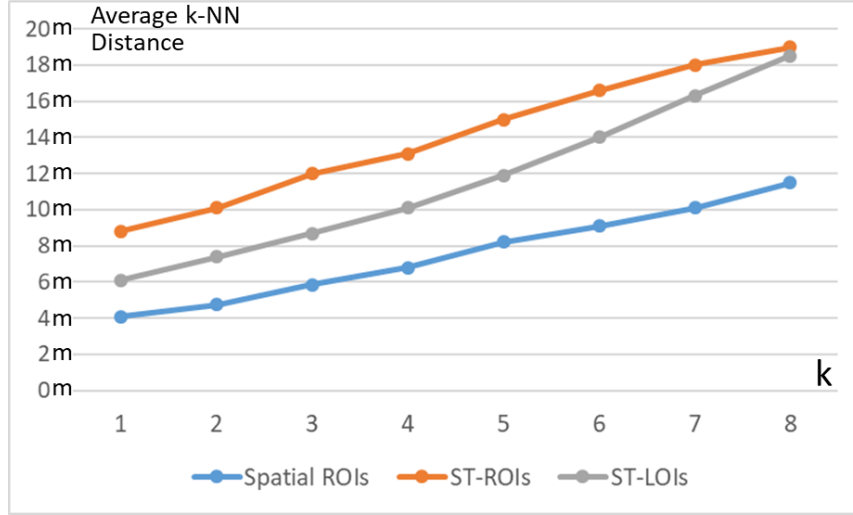


Figure 8.2 Average k-NN distance of the ROIs generated by different approaches

### 8.2.2 Temporal Cohesion

We define the average temporal density as the major indicator to express the temporal cohesion of the ROIs. The larger the density, the closer the points in the ROIs are temporally and the better the result. The average temporal density of ROIs is defined by Equation 8.4.

$$T\_density = \frac{\sum_{ROI} \frac{count(p)}{timespan(ROI)}}{count(ROI)} \quad \text{Equation 8.4}$$

where  $p$  is a stay point in a ROI (i.e.  $p \in ROI$ ),  $timespan(ROI)$  is the time difference between the first stay point and the last stay point within a ROI,  $count(ROI)$  is the total number of detected ROIs.

Table 8.3 Average temporal density of ROI detect by different approaches

	Spatial ROIs	ST-ROIs	ST-LOIs
$T\_density$	4.73 points/hour	16.50 points/hour	18.11 points/hour

Table 8.3 shows the average temporal density of the result ROIs detect by conventional DBSCAN, the Euclidean paradigm and the network paradigm respectively. It demonstrates that the temporal cohesion of Spatial ROIs is significantly poorer than ST-

ROIs and ST-LOIs because of the absence of temporal dimension in the conventional DBSCAN approach, and the better outcome in the temporal cohesion of ROIs contributed by the network paradigm.

### 8.2.3 Optimisation of ST-network-DBSCAN

In Section 3.3.3, we proposed a fast network query strategy to speed-up space time clustering in networks by reducing redundant distance calculations. Therefore, the percentage of avoided unnecessary distance calculations can be used as the quantitative indicator to measure how much the map-matching algorithm has been accelerated. Without the space time neighbour retrieving strategy, network distances between all pairs of stay points should be calculated. This means that if  $n$  stay points are input in to ST-network-DBSCAN without a query strategy, distances between all  $n!$  pairs of points should be calculated. Equation 8.5 expresses how much speed improvement is contributed by the space time neighbour retrieving strategy.

$$CalculationsAvoided = 1 - \frac{count(AC)}{n(n-1)} \quad \text{Equation 8.5}$$

where  $count(AC)$  is the number of pairwise network distance calculations that are undertaken after applying the 3-step space time query strategy.  $n(n-1)$  is the number of pairwise network distance calculations need to be undertaken without the help of any query optimisations. Table 8.4 below shows the percentage of distance calculations between stay points avoided by the space time query strategy in three boroughs.

Table 8.4 The basic network information and the percentage of avoided distance computations.

BOCU Area	Number of Edges	Number of Nodes	Number of GPS Point Records	Number of Stay Points	Distance Calculations Avoided
Camden	5049	6018	253526	71220	95.0%
Islington	4202	5125	198044	49788	91.3%
City of Westminster	5302	6409	308466	104542	93.6%

### 8.3 MODULE III

Module III is the semantic enrichment module from the framework. As the semantic meaning is related to human perception of places, there is no ground truth to support its validation. We can only evaluate the accuracy of the spatial coverage of ROIs to make sure that the spatial boundaries of the detect ROIs covers the correct POIs related to the stops of people.

#### 8.3.1 Closeness of ROIs to POIs

As has been discussed in the literature review, many studies use distance from the staying location to a POI to measure the semantic contribution of the POI to the staying behaviour (Krueger et al., 2015; Spinsanti et al., 2010). Therefore, the closer a POI is to a person's stay point, the more likely the POI and the person's activity to be related and the semantic enrichment to be meaningful for the ROIs. The short distance from a stay point to its nearest POI helps clarifying the meaning of the stay. Figure 8.3 shows the difference of the distances from a stay points to its nearby POIs. If a stay point is very close to its nearest POI and relatively far away from other nearby POIs, the semantic meaning of the stay easy to be distinguished. If each stay points in a ROI are close to its own nearest POI and relatively far away from other POIs, the POIs covered by the ROI's spatial boundary is more likely to be the right POIs to be put into the semantic enrichment algorithms. The closeness of a detected point-based ROI to the contextual POIs can be defined by the average of each stay point's distances to k-th nearest POI in this ROI. The mathematical description of a ROI's closeness to the k-th nearest POIs is demonstrated by Equation 8.6.

$$Closeness_{ROI} = \frac{\sum_{ROI} D(p, KNP_p)}{count(p)} \quad \text{Equation 8.6}$$

where  $p$  is a stay point,  $p \in ROI$ ,  $count(p)$  the is number of stay points in a ROI,  $KNP_p$  is the k-th nearest POI to  $p$ .

We compare the closeness of a ROI to its nearest POI and the closeness of such ROI to other nearby POIs to evaluate how good the approaches are in supporting the semantic enrichment processes. The larger the difference is in a ROI's closeness to different nearby POIs, the better the ROI is for semantic enrichment.

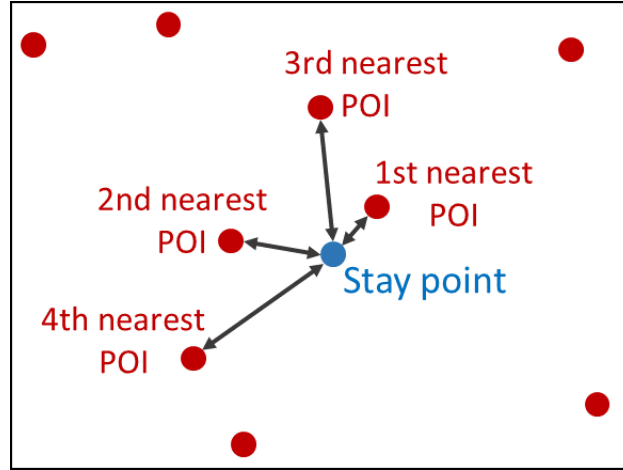


Figure 8.3 The different distances from a stay point to its nearby POIs

Figure 8.4 shows the example of spatial distribution of stay points in one ST-ROI and one ST-LOI detected in the same place and time period. Figure 8.4 (a) is a typical case of the relative location of stay points in a ST-ROI to the surrounding POIs and Figure 8.4 (b) is a typical case of the relative location of map-matched stay points.

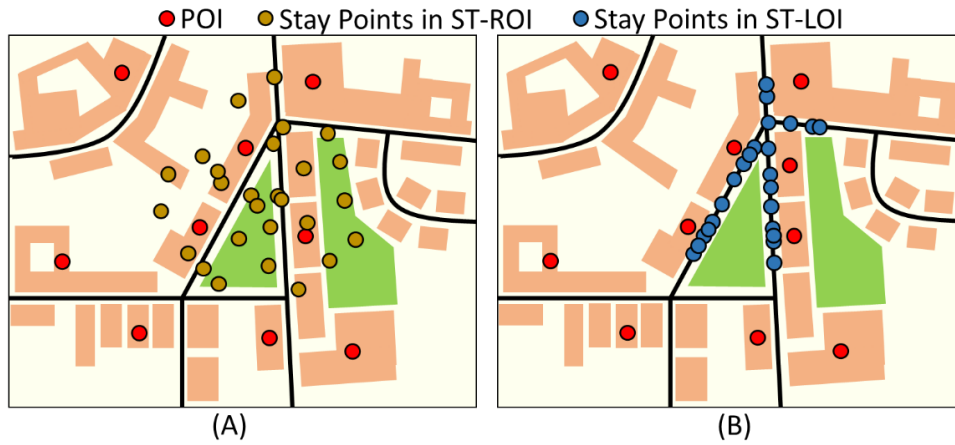


Figure 8.4 The relative location of stay points to POIs:  
(a) stay points in an ST-ROI; (b) map-matched stay points in an ST-LOI

The line chart in Figure 8.5 shows the mean value of all in-ROI points to their own  $k$ -th nearest POIs (i.e. ROI's closeness to  $k$ -th POIs) calculated based on the APLS data in August 2015. It shows that the map-matched stay points in ST-LOIs are closer to their closest POIs and second closest POIs than stay points in the ST-ROIs are when  $k < 3$ . The map-matched stay points are also further away to  $k$ -th nearest POIs than the ST-ROI stay points when  $k > 3$ . On the other hand, the changes in the distances of  $k$ -th



nearest POIs to the stay points in ST-ROIs are far less significant when  $k$  increases. That is to say, the nearest POI to a stay point in an ST-LOI is easier to be spatially distinguished from all other POIs. As reviewed in Chapter 2, the closer a stay point is to its nearest POI and the further away a stay point is to other POIs, the easier the semantic meaning of the stay can be inferred. Therefore, the ST-LOIs outperform the ST-ROIs in clarifying the relative locations of stay points to the nearby POIs and supporting the semantic enrichment module of the framework.

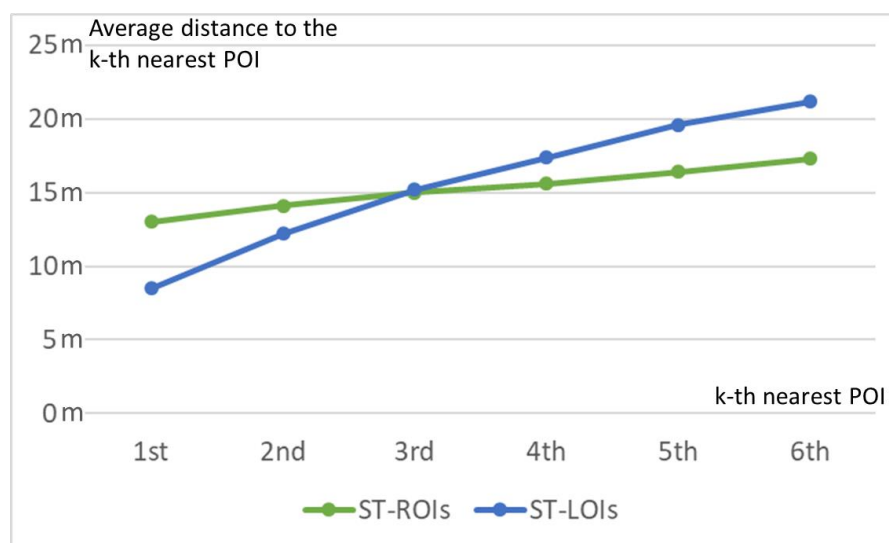


Figure 8.5 Average closeness of ROIs to their  $k$ -th nearest POIs

## 8.4 CHAPTER SUMMARY

This chapter demonstrated the performance of the framework modules developed in this thesis. The performance of methods applied in each module is assessed based on the predefined indicators and compared with existing conventional approaches. To test the accuracy of the methods in Module I, an artificial trajectory generator is design and applied due to lack of ground truth moving routes. We have also compared the Euclidean paradigm and the network paradigm with real movement datasets to assess their suitability for ROI detection in urban scenarios. We have then presented the validation results for different modules of the framework.

Through comparison and validation in Section 8.1, the network paradigm shows greater accuracy in pinpointing the true movement trajectories, identifying stops, and

estimating dwelling durations than Module I of the Euclidean paradigm and other non-spatial-temporal methods. Since Module I serves as the pre-processing module of the geographic process in the entire framework, the improved accuracy of Module I in the network paradigm demonstrates distinct advantages over conventional approaches in guaranteeing the data quality of the input to the afterwards semantic and knowledge discovery processes.

The spatial and temporal cohesion tests in Section 8.2 show better performance of spatio-temporal ROI detection methods over purely spatial methods from the perspective of clustering evaluation. These also show that the spatio-temporal clustering methods can take account of aggregations of points in both space and time and therefore reflect the highly dynamic nature of urban activities. As for the computational burden caused by the spatio-temporal clustering algorithm in networks, Section 8.2.3 demonstrates the effectiveness of the space time retrieving strategy in improving the speed of network clustering algorithm in the network paradigm.

Section 8.3 of the chapter discussed how the network paradigm clarifies the relative location between stay points and POIs and hence provides better spatial boundaries to support the semantic enrichment of places.

In summary, the comparisons in this chapter firstly demonstrate the higher spatio-temporal accuracy and cohesion of space time clustering methods over conventional approaches in stay point identification and ROI detection. Secondly, the network paradigm is more suitable for urban activity pattern analysis than the Euclidean paradigm. These results demonstrate that, by incorporating the spatial, temporal and network aspect of the urban movements and activities, the network paradigm is able to provide a more advanced methodological framework for urban activity pattern studies.

## **Chapter 9**

# **Conclusion and Future Work**

## 9 CONCLUSION AND FUTURE WORK

This thesis has sought to introduce a fresh perspective on the subject of space-time activity pattern analysis for human dynamics. Although the area has a long research history and various options of conventional approaches, our proposed works and theories that we present in this thesis demonstrate that there remains considerable room for improvement in this field of study. Through the combination of the abundant human dynamics data and the state-of-the-art data mining tools in the digital age, we are able to build a methodological framework of human activity pattern analysis that contributes to a clear advancement on conventional study approaches.

In this final chapter, we revisit the research objectives and challenges previously discussed in Chapter 1, while introducing how this thesis overcame each of these challenges. It also generalises the findings of this research on current theories and approaches of activity pattern studies, as well as their implications. This is followed by a discussion of the applications and policy influences. Then, the limitations and a list of further research work to further improve and extend our proposed framework in this thesis is summarised. The chapter is finalised by concluding the achievements and contributions of the efforts in this thesis to research and to provide applications.

### 9.1 REVIEW OF FINDINGS IN RESEARCH

The stated overall aim of the project, as outlined in the introduction chapter, was to develop a methodological framework to automatically profile and aggregate people's space-time activity patterns based on space-time human mobility data in urban semantic backgrounds. Accomplishing this aim required the completion of four objectives listed in Chapter 1. In this section, we describe the findings in the research and extend the discussion to the broader implications associated with each objective.

**Objective 1:** *Review existing approaches and methodologies for activity pattern analysis based on mobility data from both geography and other disciplines. Summarise and critically assess the fundamental theories, experiment settings and methodologies in these approaches.*

Objective 1 was achieved during the literature review in Chapter 2. The review explored conventional approaches towards the profiling and summary of human activity patterns emerging from both traditional activity pattern studies and data driven research domains enabled by the advancement of modern information technology.

In assessing the state-of-the-art in activity and behaviour patterns and human dynamic studies, we started the review from the early theories and traditional understandings of activity studies from temporal, spatial and semantic aspects. By collating the changes and developments of the research methods, we highlighted the modern approaches that stem from the principles of these three aspects of activity pattern studies. Intertwined within these three aspects, and thus implemented within many subsequent models and approaches proposed by researchers, lies the assumption that stops between movements are associated with the possible occurrences of activities. Therefore, all of the works on activity pattern studies put emphasis on the stop episodes over move episodes. Based on this theory, some researchers consider the sequence in which the moving individual stops at various places as the indicator of the individual's travel/activity pattern. Some researchers look into the time duration of stops at places for the description of activities. Others use the series of visited semantic locations for aggregative analysis of activity patterns. This theory, and accompanying modelling approaches, normally focus on the revealed activity patterns through the single aspect or the combination of two aspects of the travel behaviours of multiple individuals. A number of authors were identified to have raised concerns around the deficiency of the incomplete description of activities inherent within these approaches, and advocated the establishment of new approaches to combine the spatial and temporal analysis of activities in a joint effort, and emphasise the inseparability of people's movements and the semantic context and environment in which they move for human dynamics studies.

Although convincing and having a reference value with respect to this thesis, a range of limitations were identified with these approaches to activity pattern analysis. One significant limitation lies in the lack of tests and practical implementations of these principles and approaches with the awareness of the urban context. No works were found linking the activities and movements to the city streets in the real-world where most of the mobility data were collected. The second limitation is that no existing activity studies have considered the time varying nature of the interesting places in the city and their semantic meanings. In reality, the attractiveness of a place to human activities changes over time and the meaning of a place is not permanent, which again raises questions as to the conventional approaches' applicability within the highly dynamic urban context. In view of this, section 2.5 of the literature review introduced literature pertaining to general theories, methods and algorithms of spatial network analysis. The third limitation was highlighted in that many of the experiments in which these approaches were tested utilised only small-scale samples of travel/movement data, usually conducted within controlled or virtual environments. All of these deficits should be overcome by the new methodological framework that incorporates spatio-temporal and semantic aspects of the activity patterns in an urban context.

In the process of accomplishing Objective 1, we outlined the literature review that highlighted both the obstacles in conventional methods for a comprehensive activity pattern study and the inspiration of existing frameworks in this thesis. It is on the foundation of the findings from this thorough review of literatures that the methodological framework, presented in Chapter 3, was inspired and developed. The rest of the thesis is also organised according to the research problem and limitations summarised in the review.

**Objective 2:** *Develop an ST2P (Space-Time to Profile) framework of multiple modules. Each module is designed to address part of the limitations listed in section 1.2.*

Objective 2 was achieved through the conception of the general methodological framework in Chapter 3. Based on the summarised limitations of existing approaches and the ideas inspired during the literature review, a methodological framework was designed together with a combination of methods for its implementation. The methodological description in Chapter 3 and validations in Chapter 8 have shown that the proposed methodological framework can harnesses technological advances to extract activity pattern subgroups from low-level raw GPS trajectories. In this thesis, we proposed the theory of ***‘the place you go (ST-ROIs types), when you go and how long you stay is who you are’*** to provide a more complete and realistic picture than the ideas of ***‘you are where you go’*** in traditional activity studies. The framework constructed according to the new theory further extended the traditional time budget allocation analysis in activity studies and existing spatial-location-based similarity definitions of individual profiles. The methodological framework consisted of four modules to address the “ST2P” task in four steps: pre-processing of movement data, ST-ROI detection, semantic enrichment of detected places, and aggregative analysis of semantic activity profiles. Each module generated intermediate outcomes to be input to the next module. Combining the modules, the framework was capable of aggregatively analysing the activity patterns of people according to spatial, temporal and semantic aspects by defining a new way for profile description and new similarity metric for profile comparison. The final clustering analysis based on this similarity metric explains the semantic meaning of various behaviours more reasonably than competing methods. Our contribution also provided a set of computational and visual techniques to human dynamics researchers who may be interested in the variety of individual moving behaviours and helped location-based businesses to better understand the characteristics of their customers.

**Objective 3:** *Design and build two paradigms according to the structure of the framework in Objective 2 to incorporate the spatial, temporal and semantic information for activity profiling and aggregation in Cartesian space and urban networks.*

The completion of Objective 3 was described in full in Chapters 5 and 6, where a Euclidean paradigm and a network paradigm of the proposed framework were utilised in addressing the limitations identified during the process of fulfilling Objective 1. The two paradigms respectively incarnated the framework with different representations of space: a Cartesian view of space and a network view of space.

The implementation of the framework in Cartesian space demonstrated how the framework turned raw trajectories into ST-ROIs, ST-ROIs into semantic profiles, and then semantic profiles into semantic activity subgroups.

During the case study of police movements in London, it was discovered that the hotspots or interesting places were not just a concept of space. The significance and attractiveness of each place varied through time and its semantic meaning for human activities differed from other places and even from itself at different times of day. Results of the experiment showed that the conventional static expression of interesting regions was inadequate for depicting places and a temporal dimension should be added for the study of movements and activities in highly dynamic cities. The example of patterned police activities detected in embassy areas in section 5.6.3 shows that major changes of activities can be revealed with the proposed space-time profiles for activity description. It is also found in section 5.6.6 that the unbalanced number of POIs of different categories was the major problem for semantic enrichment of places, and the ideas of rebalancing term and word frequency weights in text mining studies could be an important reference in addressing such issues.

In the experiments of the Euclidean paradigm, we were able to find that a Cartesian representation of space had negative impacts on the precision of all modules in the framework. A network variation of the framework was therefore designed, tested and compared in Chapter 6. In this process, the network paradigm demonstrates great suitability to urban street networks and a new method for visualising hotspots (i.e. ST-LOIs) in network space and time was created.

**Objective 4:** *Compare the performance of the two paradigms proposed in Objective 3 and the existing mainstream approaches. Examine and demonstrate our framework's suitability on real people's large-scale movement trajectories in urban road networks from various aspects.*

The completion of Objective 4 was detailed during Chapter 8, where the full evaluations and comparisons were made of methods used in the two paradigms of the framework, as well as the conventional approaches. The comparisons are outlined to align with the work flow of modules in the framework. Of particular focus during the execution of this objective was the accuracy of the proposed methods against conventional approaches. As such, during the course of the evaluation, the indicators of accuracy in spatial, temporal and semantic aspects of the methods in modules are respectively proposed. We tested the methods with real-world movement data and designed a trajectory generator to simulate synthetic ground truths that were unable to be included in the real-world data. It can be found in the results that the methods used in the Euclidean paradigm improved on conventional methods in stop episode and ST-ROI detection. It also shows that the network paradigm was a more accurate option for patterned activity analysis in cities with complex transportation networks, but comes with a sacrifice in computation complexity.

As outlined above, during the thesis a more detailed representation of individual activity patterns was developed, aiming to better reflect the spatial, temporal and semantic aspects of human activities in urban space. As a result of these findings, it may be concluded that the limitations summarised in the completion of Objective 1 have been fulfilled by the development of the Euclidean paradigm and the further improvements made in the network paradigm.

## **9.2 THEORETICAL CONTRIBUTIONS AND TECHNICAL INNOVATIONS**

During the description in Chapter 3 of the methodological framework on which the thesis would be based, we have outlined the methods that we designed and novel applications of existing methods as solutions to the limitations in conventional approaches. Among them, a number of elements in the functioning modules of our methodological framework were identified as major contributions of this thesis to research.

As our framework has overcome all the limitations listed in Chapter 3, we can come to the conclusion that these contributions have been successfully delivered. The theoretical contributions have been summarised and listed below, with reference to the point at which the works and algorithms related to each contribution were described in the thesis. In addition to these theoretical contributions, we also made technical innovations by making improvements to or new applications of existing methodologies. These technical innovations, having played auxiliary but equally significant roles in the completion of the thesis, are also outlined below along with their positions in the thesis.



### Conceptual/theoretical contributions:

- Chapter 3: Integrating spatial, temporal and semantic dimensions of activities into the analysis and developing the concept of “the place you go (ST-ROIs types), when you go and how long you stay is who you are. Formalising the procedure to extract the profile and activity patterns from raw movement trajectories.
- Section 5.3.2: Development of a model for the simplified representation of individual trips and activities that preserves the temporal, sequential and spatial information in trips.
- Section 5.4.2: Improvement of the semantic enrichment method’s awareness of the time varying semantic meanings of places by adding POI opening hours and ST-ROIs’ time boundaries into the semantic enrichment model. This improvement turns the conventional and static definition of semantic places into a time varying and dynamic one.
- Section 5.5: Development of a time allocation profiling approach for the description of individual activity patterns. This profile representation jointly implemented the theories of Chapin’s (1974) time budget allocation for activities and Zhong et al.’s (2015) spatial activity profile of individual activities.
- Chapter 6: Implementation of a network representation of urban activity space in the framework to extend the spatial dimension from planar space to network space which in line with the realistic urban environment.
- Section 6.4: Proposing the concept of ST-LOI to incorporate the network geometry into the representation of ST-ROIs in urban areas.

### Technical innovations:

- Section 4.2.2: Merging a POI dataset containing well-organised semantic category information with a POI dataset containing opening hours information to create a more complete POI dataset. The merging method filled the merged dataset with well-rounded spatial, temporal and semantic information of POIs and supported a time-sensitive semantic enrichment method that we proposed to detect the semantic variation of places in time.
- Section 5.2.2: Development of a kernel-based scanning window to implement spatial threshold and temporal confines together for high-accuracy stay point/stop episode identification.
- Section 5.3.1: Novel application of ST-DBSCAN for the identification of interesting regions, as well as the regions’ interesting times that attract visitors. Proposing the concept of ST-ROI to incorporate the concept of temporal boundaries into the conventional spatially defined interesting regions.

Designing an optimised retrieving strategy to avoid redundant and unnecessary distance computations and I/O operations in the network analysis.

- Section 5.4: Novel application of text mining and topic modelling algorithms for semantic enrichment of places with POI data.
- Section 5.5: Novel application of JSD (i.e. information radius) as the similarity metric of individual activity patterns for aggregative analysis.
- Section 6.3.2: Application of existing spatio-temporal map-matching methods to extended street networks that include both major traffic links and minor local pedestrian walkways
- Section 6.3.3: Application of the newly proposed kernel-based temporal scanning window in a network environment to pinpoint stay points with higher accuracy.
- Section 6.4: Development of the ST-Net-DBSCAN algorithm for interesting region detection within urban networks.
- Section 6.4.4: Novel application of a 3D wall map to comprehensively visualise the spatial structure, time span, semantic meaning and activity intensity of ST-LOIs in one model.
- Chapter 7: Novel application of LDA topic modelling algorithm to increase the applicability of the semantic enrichment process.

## **9.3 APPLICATIONS AND POLICY IMPLICATIONS**

### **9.3.1 Application perspectives**

The overall advantage of the framework developed in this thesis is that it provides a useful toolkit to automatically extract activity patterns from GPS data and make sense out of these GPS-based activity/travel logs by analysing the staying behaviours in a semantic environment. This toolkit can be used to depict the time-varying semantic meaning of places for human activities so that urban planning authorities can be provided with well-rounded, dynamic and time sensitive information of the hotspots in the city, which helps policy makers make better informed decisions. It can also be applied to the emerging location-based social networks, linking users' social features in virtual space with their real-world activities to provide a more accurate and complete profile for the users. The implementation of the framework can aggregate users sharing similar semantic activity profiles and space-time preferences in activities, and facilitate smarter friend recommendations in the social network applications.

### 9.3.2 Policy implications

Apart from the findings in research, some policy implications are also revealed during completion of the research process in the thesis. They include improvement in data collection, spatial representation for urban contexts, and future development in Geographic Information Science.

For organisations and governmental geographic surveying agencies who are in charge of geographic data collection and rectification, it is suggested that the future POI data collections should take into account the temporal information of POIs, as it is as significant as spatial location information in describing the accessibility of POIs. No matter that it be for research purposes or daily navigation, people need to know not only where the place is but also when it is available, especially in a highly dynamic urban environment. By replacing a purely spatial expression of POIs with POIs that incorporate temporal information, the changes and differences in the time dimension can be discovered and hence the semantic meaning of places can be greatly enriched. The POI dataset of Ordnance Survey UK is well organised with precise location and categorical information. Its high-quality laid the foundation for the semantic study of places in the thesis; however, the lack of temporal data is its major deficiency. As described in Chapter 4, we had to mitigate this problem by referencing other datasets that are not as well collected but come with temporal information. The authorities can improve on this by updating the POI dataset with a thorough survey of POI opening hours, or by opening up an online platform to be compatible with the voluntarily uploaded temporal information. The quality of the VGI temporal data source may not be guaranteed but it can broaden the applicability of the POI data for the booming spatio-temporal data mining research.

Another implication for POI data collection is that the current location information in POI datasets cannot perfectly reflect the true logical relationship between POIs and street segments with high accuracy. The current POI dataset only uses the coordinates on map and geocoded addresses as location descriptions of POIs. Nonetheless, for spatial network analyses that involve POIs, such as the methods used in Module III of the proposed framework, the conventional location information cannot pinpoint the true location and accessibility of POIs within the network. Admittedly, researchers can associate the POI with the nearby street segments according to the registered street address, like the work in this thesis, or simply use the nearest segment as the POI's location in the networks. These simple solutions, however, ignored the fact that the POI can be related to multiple streets or the entrance of the POI building may not be on the nearest street. For example, some large POIs, like shopping centres near a road

intersection, can have multiple entrances on more than one side of the building and can be located through multiple segments in the street network. This means that the correct network location should not be confined to a single location on one street segment. As network analysis has been a well-developed and popular realm in geo-spatial analysis, it has become a necessity that the network expression of urban POI locations is added into the future geographic information systems. The research in this thesis has also demonstrated that stop and move episodes in an urban area, as well as the related activities, can be considered in finer scales by transforming the raw movement trajectories into routes along the streets. Therefore, the implications of this research (and similar initiatives) on GPS-based activity and travel pattern studies would be to switch to a network representation of public spaces and movements in urban areas.

This study has also provided evidence showing that the effectiveness and accuracy of activity and travel pattern analyses can be improved by taking spatial and temporal information together and bringing spatial-temporal thinking into every step and method in the framework. The view of places and activities with spatio-temporal ontology in this thesis also suggests that the semantic meaning and many other attributes of places should not be considered static, since human activities forge the highly dynamic nature of cities and places. More and more human dynamics data will be generated and collected in the foreseeable future with the continuous advancement of information technology and location-based web applications. The trend of spatio-temporal analysis to replace spatial analysis in human activity studies would be even more evident if organisations/working-groups took an initiative to instrumentalise and standardise the developing theories of spatial phenomena with spatio-temporal ontology. The emergence of similar spatio-temporal data mining applications in the future may give birth to a new geographic information platform by which more up-to-date, time varying and dynamic spatial phenomena, especially human activities, can be comprehensively analysed.

#### **9.4 CRITIQUE OF LIMITATIONS**

The novel contributions made by the work in this thesis towards the current research and literature base have been presented the previous chapters. A number of criticisms may, however, be drawn with respect to a number of limitations encountered in the construction of the proposed methodological framework. These limitations are highlighted here and they need to be considered in evaluating the outcomes of the corresponding modules.

### 9.4.1 Data limitations

Dataset limitations include the quality of the POI dataset and the homogeneity in the human movement dataset.

Firstly, the opening hours information, spatial information and semantic information are collected by different organisations with different standards. About 11% of POIs in the Google places dataset do not match with the Ordnance Survey POIs. Secondly, not all POIs possess temporal information as they should. 21% of the matched POIs in the merged dataset do not include opening hours data. This is probably due to the irregular opening patterns of the POIs and the reluctance of the POI owners to register temporal information. The imperfections of the POI data impose negative influences on the quality of semantic enrichment results and the outcome of afterward modules.

Secondly, although more and more human movement data are being generated every day, the access to large-scale and high-precision human movement datasets is still limited, mostly because of privacy issues. The movement dataset that this thesis is based on is the patrol movements of police officers when they are at work. This means that all moving individuals share a certain homogeneity in their behaviours because of their common occupation in this dataset, even though officers of different work types focus on different tasks and objectives. The fact that the GPS location sensors only work during the working hours of the officers indicates that the APLS dataset only describes parts of the officers' life cycles instead of all of them. This limitation is not a major problem for activity pattern study itself as the objective of this is not focused on life patterns. However, if the developed framework were to be applied to the analysis of activities and trips of individuals of heterogeneous personal backgrounds and higher diversity of activities, changes and adaptations should be made to some modules accordingly. For instance, for individuals that tend to visit unpopular places for others, the study should focus on space-time aggregations of personal stay points instead of ST-ROIs (or ST-LOIs) that are commonly visited by multiple individuals.

Last but not least, the APLS data were originally collected for law enforcement operational purposes and did not include an interactive status logging mechanism. Ground truth data, such as the actually covered trip routes, actually dwelling times, reasons for stops and the detailed relationship between stops and nearby POIs, were not recorded in the movement dataset in this thesis. The evaluation of the results, therefore, can only be based on assumptions and synthetic movement data simulated by a simple rule-based trajectory generator. The lack of ground truth raises challenges to the validation of the proposed framework.

### **9.4.2 Limitation in the hybrid spatial representation of places**

In Chapter 6, a novel representation of urban places is introduced during the development of the network paradigm. The network representation of the spatial boundaries of the ST-LOIs, derived from the topological structure of Ordnance Survey's extended ITN urban theme layer, was intended to incorporate the geometry of streets that people actually travel in. This way of representation is particularly designed for an urban area with a dense street network. Nonetheless, pedestrians do not necessarily follow the path along the network links in urban open fields such as parks, grasslands and squares, although the ITN urban theme layer dataset includes minor walkways in these places. That is to say, there is a small portion of urban space that does not follow the network representation of space. This is not a major limitation for semantic analysis in the network paradigm because all urban POIs can be reached through the segments in the extended ITN urban theme layer. However, a hybrid paradigm that applies network spatial representation for an area of high density of streets and Cartesian spatial representation for open fields can still be helpful in improving the accuracy of stay point identification and ROI detect in areas of hybrid landscapes such as rural-urban fringe zones.

### **9.4.3 Information losses in ROI detection**

As the ROI detection module in look for significant aggregation of people's stay points in space and time, some relatively insignificant hotspots will be ignored. No matter how the parameters (i.e. Eps and minPts) were set in density-based space-time clustering algorithms and many other algorithms for similar purposes, there will always be point clusters under the aggregation threshold. Some place may have personal meanings for individual activities but unpopular for others, but information of these may be lost in the ROI detection process because they are not visited with high enough overall intensity. For the activity pattern analysis on people that do not share many commonly visited ROIs, personal ROIs should be detected and semantically enriched on a personal basis instead of applying ROI detection methods on all individuals' trajectories altogether.

## **9.5 FUTURE WORK**

Aiming at the limitations listed in the previous section and considering the progress made so far, several research areas are clearly open to further exploration. The future work contents are summarised in this section and aligned with the four modules in the methodological framework developed in this thesis.

### 9.5.1 Data pre-processing

- Improving data collection

Clearly, one important next development will involve the improvement of data quality. The new movement dataset with ground truth status and activity information specifically collected for validation use will be of great help in explaining the patterns revealed through the implementation of the proposed framework and will provide more reliable performance measures in the evaluation of results and comparison of models and methods. To achieve this, an experiment and data collection can be organised specifically for activity pattern studies to replace the existing datasets. During the new experiment, an interactive activity data collection system can be developed and integrated into the portable GPS device. The system should be able to detect stop episodes in real time during the trip, and make queries to the carrier about the possible activity information on the scene or right after the end of the trip to avoid lapses and errors in personal memories. The queries should be designed to be short and brief to avoid over-intervention of the activities per se. This system should also lighten the burden of the experiment participants compared with complex surveys and traditional questioners.

- Transport network matching

Further improvement can also be made to the map-matching technique in the framework, as not all transportation in big cities is related to road networks. In our case study, London police officers hardly use rail transportation at work because places and vehicles relating to public transportation are task areas of other law enforcement agencies (i.e. British Transport Police). Rail and underground transportation, however, do play important roles in common people's life cycle, especially for urban commuters. Hence, if the framework is used to analyse the patterned activities of common citizens, the raw movement trajectories should not only be matched to street segments, but also to railway lines and underground stations. This means that map-matching should be replaced by hybrid network matching in the pre-processing modules in order to mitigate positioning errors for trips containing multiple transportation modes.

- Extending the framework for LBS data

Although a reasonable volume of continuous GPS movement data is used in the research of this thesis, large-scale discrete check-ins and geotagged information are more widely

seen in location-based applications such as Foursquare, Twitter and Instagram. These datasets contain spatial, temporal and semantic information altogether and are directly related to activities, which is a great advantage when describing individual activity patterns. The results of an experiment based on these datasets can also be directly turned into applications and be tested by the user because LBS applications are often associated with social media accounts. To extend the proposed framework for LBS datasets, adaptations are needed. For instance, Module I can be exempted and the semantic enrichment works in Module III can be directed based on the textual information contained in the LBS datasets instead of external environmental information sources.

- Trajectory and route generator

An agent-based modelling (ABM) method can be used to replace the simple and random route generator. It can produce synthetic routes of moving individuals that more closely resemble routes generated by real people with a decision-making process and preferences in route choices. This improvement will increase the credibility of performance evaluations of Modules I and II based on ABM-simulated trajectories.

### 9.5.2 ROI detection

- More advanced density-based clustering algorithm

The parameters of ST-DBSCAN are determined for each borough area in this thesis; however, the density of point aggregation in space and time still varies within each borough and some ST-LOIs may still be ignored. It is therefore necessary to tailor the parameter for space-time point clusters of different densities. Another possible solution is the development of a more advanced ST-ROI/ST-LOI detection algorithm. OPTICS (Ankerst et al., 1999) is a variation of DBSCAN that addresses one of DBSCAN's major weaknesses, i.e. the problem of detecting meaningful clusters in data of varying density. OPTICS also incorporates the advantage of hierarchical clustering algorithms, so that parameter determination like DBSCAN is exempted. These two major advantages make OPTICS a promising algorithm for future improvements of ROI detection. At present, OPTICS is a purely spatial clustering algorithm and much work still needs to be done to develop an algorithm that can detect spatio-temporal clusters based on OPTICS. Besides, based on the parallel improvements of OPTICS (Patwary et al., 2013), the spatio-temporal version of OPTICS would be less time-consuming than the proposed ST-Net-DBSCAN algorithm in this thesis. The development of a novel space-time clustering algorithm will be our major direction of research in the near future.



### **9.5.3 Semantic enrichment**

- Using VGI data

The semantic enrichment of places is based on rigorously collected POI data in this thesis. There are also many other data sources that can serve similar purposes. Geotagged tweets, photographs on social media and VGI data are less precise but more easily accessible and are continuously refreshed, which means they can reflect the most up-to-date semantic aspect of places and ongoing activities. Semantic enrichment based on these datasets collected by web-based approaches would provide a common ground on which the activities in real-life can be linked with the virtual internet space.

### **9.5.4 Aggregative analysis of activity profiles**

- More advanced aggregative analysis method

In the aggregative analysis of semantic activity profiles, we have applied a hierarchical clustering algorithm and determine the resulting subgroup number via the Dun index test (Dunn, 1973). Its advantage over other traditional clustering methods is that it does not require the number of clusters to be determined or estimated before running the algorithm. However, the cluster number still needs to be determined after the clustering is done. The ongoing progresses in data mining techniques will provide solutions to this limitation. For example, affinity propagation (AP) is a state-of-the-art clustering method proposed by Frey and Dueck (2007) and takes advantage from the growth in computational capabilities of CPUs. It has been successfully applied to broad areas of computer science research because it has much better clustering performance than traditional clustering methods. By applying the AP algorithm, setting subgroup numbers before or after the clustering process would be unnecessary and the algorithm can determine the most appropriate subgroup number in an iterative manner by itself.

## **9.6 FINAL CONCLUSION**

The work carried out in this thesis has demonstrated a methodological framework that seeks to generalise and aggregate the activity patterns from people's GPS movement records within the urban realm. The pre-processing, ROI detection, semantic enrichment and profile aggregation modules that we developed in the framework have achieved higher accuracies than conventional approaches. The map-matching technique was applied, and combined with the kernel-based scanning window that we

developed, to achieve a superb overall accuracy of nearly 95% for identifying stay points in the trips. It also paved the way to a network representation of urban space in the implementation of the framework to achieve a better overall fit with the real-world scenario. As parts of the framework, a network-based spatio-temporal clustering method has been designed for the detection of regions of interest in urban street networks and time. Advanced algorithms are borrowed from text mining to describe the time-varying semantic meaning of places and patterned activities. It is hoped that the framework, as well as the methodological advances introduced during this thesis, will have wider applications in other urban movements, and trip studies providing an understanding of the local data *per se*.

## REFERENCES

- Abbasifard, M. R., Ghahremani, B., & Naderi, H., 2014. A survey on nearest neighbor search methods. *International Journal of Computer Applications*, 95 (25).
- Adams, B., Phung, D., & Venkatesh, S., 2006. Extraction of social context and application to personal multimedia exploration. *Proceedings of the 14th annual ACM international conference on Multimedia*, 987-996.
- Agrawal, R., Faloutsos, C., & Swami, A., 1993. Efficient similarity search in sequence databases. 69-84, Springer Berlin Heidelberg.
- An, L., An, L., Tsou, M. H., Crook, S. E., Chun, Y., Spitzberg, B., Gawron, J. M., & Gupta, D. K., 2015. Space-time analysis: concepts, quantitative methods, and future directions. *Annals of the Association of American Geographers*, (ahead-of-print), 1-24.
- Andrienko, G., Andrienko, N., Demsar, U., Dransch, D., Dykes, J., Fabrikant, S. I., & Tominski, C., 2010. Space, time and visual analytics. *International Journal of Geographical Information Science*, 24 (10), 1577-1600.
- Andrienko, G., Andrienko, N., Hurter, C., Rinzivillo, S., & Wrobel, S., 2011. From movement tracks through events to places: Extracting and characterizing significant places from mobility data. *Proceedings of IEEE Conference on Visual Analytics Science and Technology*, 161-170.
- Andrienko, N., Andrienko, G., Fuchs, G., & Jankowski, P., 2015. Scalable and privacy-respectful interactive discovery of place semantics from human mobility traces. *Information Visualization*.
- Ankerst, M., Breunig, M. M., Kriegel, H. P., & Sander, J., 1999. OPTICS: ordering points to identify the clustering structure. *ACM Sigmod Record*, 28 (2), 49-60.
- Alt, H., Efrat, A., Rote, G., & Wenk, C., 2003. Matching planar maps. In *Proceedings of the fourteenth annual ACM-SIAM symposium on Discrete algorithms*, 589-598. Society for Industrial and Applied Mathematics.
- Alvares, O., Bogorny, V., Kuijpers, B., de Macedo, J. A. F., Moelans, B., & Vaisman, A., 2007. A model for enriching trajectories with semantic geographical information. *Proceedings of the 15th annual ACM international symposium on Advances in geographic information systems*, 22, 1-8. doi:10.1145/1341012.1341041
- Andrienko, G., and Andrienko, N., 2011. From Movement Tracks through Events to Places: Extracting and Characterizing Significant Places from Mobility Data. *Symposium on Visual Analytics Science and Technology*, October 23 - 28, Providence, USA. 163-165.
- Ashbrook, D. & Starner, T., 2003. Using GPS to Learn Significant Locations and Predict Movement Across Multiple Users. *Personal and Ubiquitous Computing*, 7 (5), 275-286.
- Ashish, N., & Sheth, A., 2011. *Geospatial Semantics and the Semantic Web: Foundations, Algorithms, and Applications*. Springer Berlin Heidelberg.
- Axhausen, K. W., & Gärling, T., 1992. Activity - based approaches to travel analysis: conceptual frameworks, models, and research problems. *Transport reviews*, 12 (4), 323-341.

- Baglioni, M., Fernandes de Macêdo, J. A., Renso, C., Trasarti, R., & Wachowicz, M., 2009. Towards a semantic interpretation of movement behaviour. *Proceedings of 12th AGILE, Lecture Notes in Geoinformation and Cartography*.
- Batty, M., Xie, Y., & Sun, Z., 1999. Modeling urban dynamics through GIS-based cellular automata. *Computers, environment and urban systems*, 23 (3), 205-233.
- Becker, R.A., Eick, S.G., Wiiks, A.R., 1995. Visualizing network data. *IEEE Transactions on Visualization and Computer Graphics* 1 (1), 16-21.
- Biljecki, F., 2010. Automatic segmentation and classification of movement trajectories for transportation modes. In *Workshop on Modelling Moving Objects 2010*. Informatics Institute, University of Amsterdam.
- Birant, D. and Kut, A., 2007. ST-DBSCAN: An algorithm for clustering spatial-temporal data. *Data and Knowledge Engineering*, (60), 208-221.
- Blei, D. M., Ng, A. Y., & Jordan, M. I., 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3, 993-1022.
- Bolbol, A., Cheng, T., Tsapakis, I., & Haworth, J., 2012. Inferring hybrid transportation modes from sparse GPS data using a moving window SVM classification. *Computers, Environment and Urban Systems*, 36 (6), 526-537.
- Bozkaya, T., & Yazdani, N., 1997. Matching and indexing sequences of different lengths. *Proceedings of the sixth international conference on Information and knowledge management*, 128-135.
- Braun, M., Scherp, A., & Staab, S., 2010. Collaborative semantic points of interests. *The Semantic Web: Research and Applications*, 365-369.
- Buchin, K., Cabello, S., Gudmundsson, J., Löffler, M., Luo, J., Rote, G., & Wolle, T., 2009. Detecting hotspots in geographic networks. In *Advances in GIScience*, 217-231. Springer Berlin Heidelberg.
- Cai, G., Lee, K., & Lee, I., 2016. A Framework for Mining Semantic-Level Tourist Movement Behaviours from Geo-tagged Photos. In *Australasian Joint Conference on Artificial Intelligence*, 519-524. Springer International Publishing.
- Cai, Y., & Ng, R., 2004. Indexing spatio-temporal trajectories with Chebyshev polynomials. *Proceedings of the 2004 ACM SIGMOD international conference on Management of data*, 599-610.
- Cao, X., Cong, G. and Jensen, C. S., 2010, Mining significant semantic locations from GPS data. *Proceedings of the VLDB Endowment*, 3, 1009-1020.
- Car, A., & Frank, A., 1994. General principles of hierarchical spatial reasoning-the case of wayfinding. In the *Proceedings of the 6th International Symposium on Spatial Data Handling*, 646-664.
- Chapin, F. S., 1974. *Human activity patterns in the city: Things people do in time and in space*. Canada: Wiley-Interscience.
- Chawathe, S. S., 2007. Segment-based map matching. In *Intelligent Vehicles Symposium*, 2007 IEEE, 1190-1197.
- Chen, B. Y., Yuan, H., Li, Q., Shaw, S. L., Lam, W. H., & Chen, X., 2016. Spatiotemporal data model for network time geographic analysis in the era of big data. *International Journal of Geographical Information Science*, 30 (6), 1041-1071.

- Chen, L., Özsu, M. T. and Oria, V., 2005. Robust and fast similarity search for moving object trajectories. *Proceedings of the 2005 ACM SIGMOD International conference on Management of data*, 491-502.
- Cheng, T., Tanaksaranond, G., Brunsdon, C., & Haworth, J., 2013. Exploratory visualisation of congestion evolutions on urban transport networks. *Transportation Research Part C: Emerging Technologies*, 36, 296-306.
- Cheng, T., Tanaksaranond, G., Emmonds, A., & Sonoiki, D., 2010. Multi-scale visualisation of inbound and outbound traffic delays in London. *The Cartographic Journal*, 47 (4), 323-329.
- Cheng, Y., 1995. Meaning shift, mode seeking, and clustering. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 17 (8), 790-799.
- Chon, Y., Lane, N. D., Li, F., Cha, H., & Zhao, F., 2012. Automatically characterizing places with opportunistic crowdsensing using smartphones. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, 481-490. ACM.
- Chu, W., & Wong, H., 1999. Fast time-series searching with scaling and shifting. *Proceedings of the eighteenth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, 237-248.
- Cormen, T. H., Leiserson, C. E., Rivest, R. L., & Stein, C., 2001. *Introduction to algorithms* second edition.
- Cortez, M., & Salas, H. A., 2014. Quality Metrics for Optimizing Parameters Tuning in Clustering Algorithms for Extraction of Points of Interest in Human Mobility. *1st Symposium on Information Management and Big Data*, 14.
- Cullen, G., 1972. Space, time, and the disruption of behaviours in cities. Paper Submitted to the Research Group on Time Budes, Brussels.
- Daily Mirror News, 2012. Syria embassy break-in by protesters leads to six arrests [online]. Daily Mirror UK News. Available from: <http://www.mirror.co.uk/news/uk-news/syria-embassy-break-in-by-protesters-leads-674577> [Accessed 27 June 2015].
- Delafontaine, M., Versichele, M., Neutens, T., & Van de Weghe, N., 2012. Analysing spatiotemporal sequences in Bluetooth tracking data. *Applied Geography*, 34, 659-668.
- Demsar, U., Buchin, K. A., Cagnacci, F., Safi, K., Speckmann, B., Weghe, N. D., Weiskopf, D., & Weibel, R., 2015. Analysis and visualisation of movement.
- Demsar, U., & Verrantaus, K., 2010. Space-time density of trajectories: exploring spatiotemporal patterns in movement data. *International Journal of Geographical Information Science*, 24 (10), 1527-1542.
- Demsar, U., & van Loon, E., 2013. Visualising movement: The seagull. *Significance*, 10 (5), 40-42.
- Demsar, U., Buchin, K., van Loon, E., & Shamoun-Baranes, J., 2015. Stacked space-time densities: a geovisualisation approach to explore dynamics of space use over time. *Geoinformatica*, 19 (1), 85-115.
- Dillenburg, J. F., & Nelson, P. C., 1995. Improving search efficiency using possible subgoals. *Mathematical and computer modelling*, 22 (4-7), 397-414.

- Dodge, S., Weibel, R. & Forootan, E., 2009. Revealing the physics of movement: Comparing the similarity of movement characteristics of different types of moving objects. *Computers, Environment and Urban Systems*, 33 (6), 419-434.
- Dodge, S., Weibel, R., & Lautenschütz, A. K., 2008. Towards a taxonomy of movement patterns. *inf visual* 7(3-4):240-252. *Information Visualization*, 7 (3-4), 240-252.
- Doherty, A. R., Gurrin, C., Jones, G. J. F., & Smeaton, A. F., 2006. Retrieval of similar travel routes using gps tracklog place names. *Proceedings of the 3rd Workshop on Geographic Information Retrieval*, 41 (5), 133-142.
- Downs, J. A., Horner, M. W., & Tucker, A. D., 2011. Time-geographic density estimation for home range analysis. *Annals of Gis*, 17 (3), 163-171.
- Dunn, J. C., 1973. A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters, *Journal of Cybernetics*, 3 (3), 32-57.
- Edwards, D. C., Griffin, T., Hayllar, B. R., Dickson, T., & Schweinsberg, S. C., 2009. Understanding Tourism Experiences and Behaviour in Cities: An Australian Case Study.
- Ester, M., Kriegel, H. P., Sander, J., & Xu, X., 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. *Kdd*, 96 (34), 226-231.
- Ester, M., Kriegel, H. P., Sander, J., & Xu, X., 1998. Clustering for mining in large spatial databases, 12, 18-24.
- European Commission, 2010. Regional GDP per capita in the EU in 2010: eight capital regions in the ten first places [online]. Available from: <http://ec.europa.eu/eurostat/en/web/products-press-releases/-/1-21032013-AP> [Accessed 30 August 2017].
- Fu, L., Sun, D., & Rilett, L. R., 2006. Heuristic shortest path algorithms for transportation applications: state of the art. *Computers & Operations Research*, 33 (11), 3324-3343.
- Goldin, Q., & Kanellakis, C., 1995. On similarity queries for time-series data: constraint specification and implementation. *Principles and Practice of Constraint Programming-CP'95*, 137-153.
- Goodchild, M. F., & Li, L., 2012. Assuring the quality of volunteered geographic information. *Spatial statistics*, 1, 110-120.
- Greenfeld, J. S., 2002. Matching GPS observations to locations on a digital map. In *Transportation Research Board 81st Annual Meeting*.
- Griffiths, T. L., & Steyvers, M., 2004. Colloquium paper: mapping knowledge domains: finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America*, 101 (1), 5228.
- Gunopulos, D., Gunopulos, D., & Das, G., 2004. Rotation invariant distance measures for trajectories. *Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 707-712. ACM.
- Güting, R. H., de Almeida, T., & Ding, Z., 2006. Modelling and querying moving objects in networks. *International journal on very large data bases*, 15 (2), 165-190.
- Guttman, A., 1984. R-trees: a dynamic index structure for spatial searching, 14 (2), 47-57. ACM.

- Hägerstrand, T., 1970. What about people in regional science?. In Papers of the Regional Science Association, 24 (1), 6-21.
- Haklay, M., 2010. How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordnance Survey datasets. *Environment and planning B: Planning and design*, 37 (4), 682-703.
- Hariharan, R., & Toyama, K., 2004. Project Lachesis: parsing and modeling location histories. *Geographic Information Science*, 106-124, Springer Berlin Heidelberg.
- Hart, P. E., Nilsson, N. J., & Raphael, B., 1968. A formal basis for the heuristic determination of minimum cost paths. *IEEE transactions on Systems Science and Cybernetics*, 4 (2), 100-107.
- ITO world blog, 2009. Visualising Transport Data for Data.Gov.Uk, (Accessed 29 September 2017).
- Jankowski, P., Andrienko, N., Andrienko, G., & Kisilevich, S., 2010. Discovering landmark preferences and movement patterns from photo postings. *Transactions in GIS*, 14 (6), 833-852.
- Jain, A. K., 2008. *Data Clustering: 50 Years Beyond K-means*. Machine Learning and Knowledge Discovery in Databases. Springer Berlin Heidelberg.
- Jenkins, A., Croitoru, A., Crooks, A. T., & Stefanidis, A., 2016. Crowdsourcing a collective sense of place. *PloS one*, 11 (4), e0152932.
- Jing, N., Huang, Y. W., & Rundensteiner, E. A., 1996. Hierarchical optimization of optimal path finding for transportation applications. In *Proceedings of the fifth international conference on Information and knowledge management*, 261-268. ACM.
- Joh, C. H., Arentze, T., & Timmermans, H., 2001. A position-sensitive sequence-alignment method illustrated for space - time activity-diary data. *Environment & Planning A*, 33 (2), 313-338.
- Joh, C. H., Arentze, T., Hofman, F., & Timmermans, H., 2002. Activity pattern similarity: a multidimensional sequence alignment method. *Transportation Research Part B*, 36 (5), 385-403.
- Joh, C. H., Arentze, T., & Timmermans, H., 2005. A utility-based analysis of activity time allocation decisions underlying segmented daily activity - travel patterns. *Environment & Planning A*, 37 (1), 105-125.
- Johnson, N., & Hogg, D., 1996. Learning the distribution of object trajectories for event recognition. *Image & Vision Computing*, 14 (8), 609-615.
- Kahveci, T., & Singh, A., 2001. Variable length queries for time series data. In *Data Engineering, 2001. Proceedings of 17th International Conference*, 273-282.
- Kami, N., Enomoto, N., Baba, T., & Yoshikawa, T. 2010. Algorithm for Detecting Significant Locations from Raw GPS Data. In *Discovery Science*, 221-235.
- Keogh, E., Chakrabarti, K., Pazzani, M., & Mehrotra, S., 2001. Dimensionality reduction for fast similarity search in large time series databases. *Knowledge & Information Systems*, 3 (3), 263-286.
- Kölbl, R. and Helbing, D., 2003. Energy laws in human travel behaviour. *New Journal of Physics*, 5 (48), 1-12.
- Kriegel, H. P., Kröger, P., Sander, J., & Zimek, A., 2011. *Density - based clustering*. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 1(3), 231-240.

- Krueger, R., Thom, D., and Ertl, T., 2015. Semantic enrichment of movement behaviour with foursquare – a visual analytics approach. *IEEE transactions on visualization and computer graphics*, 21 (8), 903-915.
- Kuijpers, B. and Vaisman, A., 2007. A data model for moving objects supporting aggregation. *23rd International Conference on Data Engineering*, 546 - 554.
- Kullback, S. and Leibler, R. A., 1951. On Information and Sufficiency. *Annals of Mathematical Statistics*, 22 (1), 79-86.
- Kwan, M. P., 2004. GIS methods in time-geographic research: geocomputation and geovisualization of human activity patterns. *Geografiska Annaler: Series B, Human Geography*, 86 (4), 267-280.
- Kwan, M. P., Xiao, N. and Ding, G., 2014. Assessing activity pattern similarity with multidimensional sequence alignment based on a multi-objective optimization evolutionary algorithm. *Geographical Analysis*, 297-320.
- Laube, P., Dennis, T., Forer, P., & Walker, M., 2007. Movement beyond the snapshot – dynamic analysis of geospatial lifelines ☆. *Computers Environment & Urban Systems*, 31 (5), 481-501.
- Lee, I., Cai, G. and Lee, K., 2013. Mining points-of-interest association rules from geo-tagged photos. *Proceedings of the Annual Hawaii International Conference on System Science*, 1580-1588.
- Lee, S. L., Chun, S. J., Kim, D. H., Lee, J. H., & Chung, C. W., 2000. Similarity search for multidimensional data sequences. In *Proceedings of 16th International Conference on Data Engineering*, 599-608. IEEE.
- Lew, A., & McKercher, B., 2006. Modeling tourist movements: A local destination analysis. *Annals of tourism research*, 33 (2), 403-423.
- Li, Q., Zheng, Y., Xie, X., Chen, Y., Liu, W., & Ma, W. Y., 2008. Mining user similarity based on location history. *Proceedings of the 16th ACM SIGSPATIAL international conference on Advances in geographic information systems-GIS '08*.
- Li, X., Han, J., Lee, J. G., & Gonzalez, H., 2007. Traffic density-based discovery of hot routes in road networks. In *International Symposium on Spatial and Temporal Databases*, 441-459. Springer Berlin Heidelberg.
- Liao, L., Fox, D., & Kautz, H., 2008. Location-based activity recognition using relational Markov networks. *International Joint Conference on Artificial Intelligence*, (26), 773-778.
- Lin, B., & Su, J., 2008. One way distance: for shape based similarity search of moving object trajectories. *Geoinformatica*, 12 (2), 117-142.
- Lin, J., 1991. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory*, 37 (1), 145-151.
- Little, J. J., & Gu, Z., 2001. Video retrieval by spatial and temporal structure of trajectories. *Photonics West 2001 - Electronic Imaging*, (4315), 545-552.
- Loecher, M., & Jebara, T., 2009. CitySense: multiscale space time clustering of GPS points and trajectories. *Joint Statistical Meeting*.
- Lou, Y., Zhang, C., Zheng, Y., Xie, X., Wang, W., & Huang, Y., 2009. Map-matching for low-sampling-rate GPS trajectories. In: *Proceedings of the 17th ACM SIGSPATIAL international conference on advances in geographic information systems*, 352-361.



- Lukasczyk, J., Maciejewski, R., Garth, C., & Hagen, H., 2015. Understanding hotspots: a topological visual analytics approach. In *Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems*. 36. ACM.
- Luo, W., & MacEachren, A. M., 2014. Geo-social visual analytics. *Journal of spatial information science*, 2014 (8), 27-66.
- Mazimpaka, J. D., & Timpf, S., 2017. How they move reveals what is happening: Understanding the dynamics of big events from human mobility pattern. *ISPRS International Journal of Geo-Information*, 6 (1), 15.
- Mcardle, G., Demsar, U., Spek, S., & Mcloone, S., 2013. *Interpreting Pedestrian Behaviour by Visualising and Clustering Movement Data*. Web and Wireless Geographical Information Systems. Springer Berlin Heidelberg.
- Mcardle, G., Demsar, U., Spek, S., & Mcloone, S., 2014. Classifying pedestrian movement behaviour from gps trajectories using visualization and clustering. *Annals of GIS*, 20 (2), 85-98.
- Mckenzie, G., 2014. *Activities in a new city: itinerary recommendation based on user similarity*. Spatial Knowledge and Information (SKI), Canada.
- Meier, S., 2017. Enhancing Location Recommendation Through Proximity Indicators, Areal Descriptors, and Similarity Clusters. In *Progress in Location-Based Services 2016*, 273-291. Springer International Publishing.
- Metropolitan Police, 2015. Structure of the London Metropolitan Police Service [online]. Available from: <https://www.whatdotheyknow.com/request/257162/response/635673/attach/html/3/Data.xls.html> [Accessed 23 May 2017].
- Miller, H. J., 2005. A measurement theory for time geography, *Geographical Analysis*. 37 (1), 17-45.
- Misra, P., & Enge, P., 2006. *Global Positioning System: Signals, Measurements and Performance Second Edition*. Massachusetts: Ganga-Jamuna Press.
- Monajemi, P. P. Z., 2013. *A Clustering-Based Approach for Enriching Trajectories with Semantic Information Using VGI Sources* (Doctoral dissertation, ITC).
- Nanni, M., & Pedreschi, D., 2006. Time-focused clustering of trajectories of moving objects. *Journal of Intelligent Information Systems*, 27 (3), 267-289.
- Nara, A., Izumi, K., Iseki, H., Suzuki, T., Nambu, K., & Sakurai, Y., 2011. Surgical workflow monitoring based on trajectory data mining. *New Frontiers in Artificial Intelligence*, 283-291.
- Nasraoui, O., Frigui, H., Krishnapuram, R., & Joshi, A., 2000. Extracting web user profiles using relational competitive fuzzy clustering. *International Journal on Artificial Intelligence Tools*, 9 (4).
- Nishida, K., Toda, H., Kurashima, T., & Suhara, Y., 2014. Probabilistic identification of visited point-of-interest for personalized automatic check-in. *ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 631-642. ACM.
- Niu, N., Liu, X., Jin, H., Ye, X., Liu, Y., Li, X., & Li, S., 2017. Integrating multi-source big data to infer building functions. *International Journal of Geographical Information Science*, 1-20.

- Okabe, A., Okunuki, K. I., & Shiode, S., 2006. The SANET toolbox: new methods for network spatial analysis. *Transactions in GIS*, 10 (4), 535-550
- Oliver, D., Bannur, A., Kang, J. M., Shekhar, S., & Bousselaire, R., 2010. A k-main routes approach to spatial network activity summarization: A summary of results. In *Data Mining Workshops (ICDMW)*, 2010 IEEE International Conference, 265-272. IEEE.
- Ooi, B. C., 1987. Spatial kd-tree: A data structure for geographic database. In *Datenbanksysteme in Büro, Technik und Wissenschaft*, 247-258. Springer Berlin Heidelberg.
- OpenStreetMap, 2017. Point of Interest [online]. Available from: [http://wiki.openstreetmap.org/wiki/Points\\_of\\_interest](http://wiki.openstreetmap.org/wiki/Points_of_interest) [Accessed 2 September 2017].
- Ordnance Survey, 2015. <https://www.ordnancesurvey.co.uk/business-and-government/products/itn-layer.html>
- Ozer, M., 2001. User segmentation of online music services using fuzzy clustering. *Omega*, 29 (2), 193-206.
- Palma, A., 2008. A Clustering based Approach for Discovering Interesting Places in Trajectories. *Symposium on Applied Computing*, March 16-20, Fortaleza, Brazil. 863-864.
- Palma, A., Bogorny, V., Kuijpers, B., & Alvares, L. O., 2009. A clustering-based approach for discovering interesting places in a single trajectory. *2nd International Conference on Intelligent Computing Technology and Automation*, (3), 429-432.
- Parent, C., Spaccapietra, S., Renso, C., Andrienko, G., Andrienko, N., Bogorny, V., and Theodoridis, Y., 2013. Semantic trajectories modeling and analysis. *ACM Computing Surveys (CSUR)*, 45 (4), 42.
- Patwary, M., Palsetia, D., Agrawal, A., Liao, W. K., Manne, F., & Choudhary, A., 2012. A new scalable parallel DBSCAN algorithm using the disjoint-set data structure. *International Conference for High Performance Computing, Networking, Storage and Analysis (SC)*, 1-11.
- Patwary, A., Palsetia, D., Agrawal, A., Liao, W. K., Manne, F., & Choudhary, A., 2013. Scalable parallel optics data clustering using graph algorithmic techniques. *International Conference for High Performance Computing, Networking, Storage and Analysis (SC)*, 1-12.
- Polisciuc, E., Alves, A., & Machado, P., 2015. Understanding urban land use through the visualization of points of interest. In *Proceedings of the Fourth Workshop on Vision and Language*, 51-59.
- Rajasekaran, S., 2005. Efficient parallel hierarchical clustering algorithms. *IEEE Transactions on Parallel & Distributed Systems*, (6), 497-502.
- Ranacher, P., & Tzavella, K., 2014. How to compare movement? A review of physical movement similarity measures in geographic information science and beyond. *Cartography and geographic information science*, 41 (3), 286-307.
- Ratcliffe, H., 2012. The Spatial Extent of Criminogenic Places: A Change-point Regression of Violence around Bars. *Geographical Analysis*, 44 (4), 302-320.
- Renso, C., Spaccapietra, S., & Zimányi, E., 2013. *Mobility Data Modeling, Management, and Understanding*. Cambridge University Press.

- Ren, F., & Kwan, M. P., 2007. Geovisualization of human hybrid activity-travel patterns. *Transactions in GIS*, 11 (5), 721-744.
- Salvador, S. and Chan, P., 2003. Determining the number of clusters/segments in hierarchical clustering/segmentation algorithms. Florida, USA.
- Samet, H., 1990. The design and analysis of spatial data structures, (199). Reading, MA: Addison-Wesley.
- Sankoff, D., & Kruskal, J., 1983. The Theory and Practice of Sequence Comparison. Addison-Wesley, Reading, Massachusetts.
- Schönfelder, S. Li, H., Guensler, R., Ogle, J., & Axhausen, K. W., 2006. Analysis of Commute Atlanta Instrumented Vehicle GPS Data: Destination Choice Behavior and Activity Spaces. Washington, DC, USA, Transportation Research Board 85th Annual Meeting
- Schuessler, N., & Axhausen, K. W., 2009. Processing raw data from global positioning systems without additional information. *Transportation Research Record Journal of the Transportation Research Board*, (2105), 28-36.
- Sedgewick, R., & Vitter, J. S., 1986. Shortest paths in Euclidean graphs. *Algorithmica*, 1 (1-4), 31-48.
- Shaw, S. L., Tsou, M. H., & Ye, X., 2016. Human dynamics in the mobile and big data era. *International Journal of Geographical Information Science*, 30 (9), 1687-1693.
- Shen, J., & Cheng, T., 2014. Group Behaviour Analysis of London Foot Patrol Police. The 23rd GIS Research UK.
- Shen, J., & Cheng, T., 2016. A framework for identifying activity groups from individual space-time profiles. *International Journal of Geographical Information Science*, (9), 1-21.
- Sherman, W., & Weisburd, D., 1995. General deterrent effects of police patrol in crime "hot spots": a randomized, controlled trial. *Justice Quarterly*, 12 (4), 625-648.
- Shi, J., Mamoulis, N., Wu, D., & Cheung, D. W., 2014. Density-based place clustering in geo-social networks. In *Proceedings of the 2014 ACM SIGMOD international conference on Management of data*, 99-110. ACM.
- Shine, J. A., 2007. Discovering and Quantifying Mean Streets: A Summary of Results Mete Celik, Shashi Shekhar, Betsy George, James P. Rogers, and.
- Shoval, N. and Isaacson, M., 2007. Sequence alignment as a method for human activity analysis in space and time. *Annals of the Association of American geographers*, 97 (2), 282-297.
- Shoval, N., and Isaacson, M., 2009. Tourist mobility and advanced tracking technologies. Routledge.
- Sideridis, S., Pelekis, N., Theodoridis, Y., 2015. From Trajectories to Semantic Mobility Networks. UNIPi-InfoLab-TR-2015-02, Technical Report Series.
- Siła-Nowicka, K., Vandrol, J., Oshan, T., Long, J. A., Demšar, U., & Fotheringham, A. S., 2016. Analysis of human mobility patterns from GPS trajectories and contextual information. *International Journal of Geographical Information Science*, 30 (5), 881-906.

- Sims, G., Jun, S., Wu, G., & Kim, S., 2009. Alignment-free genome comparison with feature frequency profiles and optimal resolutions. *Proceedings of the National Academy of Sciences of the United States of America*, 106 (8), 2677-2682.
- Sizov, S., 2010. Geofolk: latent spatial semantics in web 2.0 social media. In *Proceedings of the third ACM international conference on Web search and data mining*, 281-290. ACM.
- Spaccapietra, S., Parent, C., Damiani, M. L., de Macedo, J. A., Porto, F., and Vangenot, C., 2008. A conceptual view on trajectories. *Data & knowledge engineering*, 65 (1), 126-146.
- Spinsanti, L., Celli, F. and Renso, C 2010. Where you stop is who you are: Understanding people's activities by places visited. *CEUR Workshop Proceedings*, 678, 38-52.
- Szalai, A., 1966. Trends in comparative time-budget research. *The American Behavioral Scientists*, 9, 9.
- Tardy, C., Falquet, G., & Moccozet, L., 2016. Semantic enrichment of places with VGI sources: a knowledge based approach. In *Proceedings of the 10th Workshop on Geographic Information Retrieval*, 6. ACM.
- TFL, 2010. Measuring Public Transport Accessibility Levels [online]. Transportation of London. Available from: <https://data.london.gov.uk/dataset/public-transport-accessibility-levels> [Accessed 11 July 2017].
- Thériault, M., Claramunt, C., Séguin, A. M., & Villeneuve, P., 2002. Temporal GIS and statistical modelling of personal lifelines. In *Advances in Spatial Data Handling*, 433-449. Springer, Berlin, Heidelberg.
- Thierry, B., Chaix, B., & Kestens, Y. 2013. Detecting activity locations from raw GPS data: a novel kernel-based algorithm. *International journal of health geographics*, 12 (1), 14.
- Timmermans, H., Arentze, T. and Joh, C. H., 2002. Analysing space-time behaviour: new approaches to old problems. *Progress in Human Geography*, 175-190.
- Tominski, C., Schumann, H., Andrienko, G., Andrienko, N., 2012. Stacking-based visualization of trajectory attribute data. *IEEE Transactions on Visualization and Computer Graphics* 18 (12), 2565-2574.
- Tompson, L., Partridge, H., & Shepherd, N. (2009). Hot routes: Developing a new technique for the spatial analysis of crime. *Crime Mapping: A Journal of Research and Practice*, 1(1), 77-96.
- Tsou, M. H., 2015. Research challenges and opportunities in mapping social media and Big Data. *Cartography and Geographic Information Science*, 42 (1), 70-74.
- Tsui, S. Y. A. & Shalaby, A. S., 2006. Enhanced System for Link and Mode Identification for Personal Travel Surveys Based on Global Positioning Systems. *Transportation Research Record: Journal of the Transportation Research Board*, 1972, 38-45.
- UCL, 2017. Research Computing Services [online]. Available from: <http://www.ucl.ac.uk/research-it-services/research-computing> [Accessed 30 August 2017]
- Ukkonen, E., 1983. On approximate string matching. *Foundations of Computation Theory*. Springer, 487-495.
- UK legislation, 1963, Inner London Boroughs [online]. Available from: <http://www.legislation.gov.uk/ukpga/1963/33> [Accessed 10 July 2017].

- Vazquez-Prokopec, G. M., Bisanzio, D., Stoddard, S. T., Paz-Soldan, V., Morrison, A. C., Elder, J. P., & Kitron, U., 2013. Using GPS technology to quantify human mobility, dynamic contacts and infectious disease dynamics in a resource-poor urban environment. *PloS one*, 8 (4), e58802.
- Vlachos, M., Gunopulos, D., & Kollios, G., 2002. Robust Similarity Measures for Mobile Object Trajectories. *International Workshop on Database and Expert Systems Applications*, 721-728. IEEE Computer Society.
- Wald, I., and Havran, V., 2006. On building fast kd-trees for ray tracing, and on doing that in  $O(N \log N)$ . *Symposium on Interactive Ray Tracing*, 61-69.
- Ward Jr, J., 1963. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58 (301), 236-244.
- Wilson, C., 2001. Activity patterns of Canadian women: application of ClustalG sequence alignment software. *Transportation Research Record*, 1777 (1), 55-67.
- Wilson, C., 2006. Reliability of sequence-alignment analysis of social processes: monte carlo tests of clustalg software. *Environment & Planning A*, 38 (1), 187-204.
- Wilson, C., 2007. Experiments with activity pattern classification: alignment versus non-alignment methods. *International Association for Time Use Research Annual Meeting*, 1-14.
- Wilson, C., Harvey, A., & Thompson, J., 1999. ClustalG: Software for Analysis of Activities and Sequential Events.
- Webb, N. F., Hebblewhite, M., & Merrill, E. H., 2008. Statistical methods for identifying wolf kill sites using global positioning system locations. *Journal of Wildlife Management*, 72 (3), 798-807.
- Xiang, Z., Liu, R., Hu, Q., & Shi, C., 2012. Applied research of route similarity analysis based on association rules. *Transnav the International Journal on Marine Navigation & Safety of Sea Transportation*, 6.
- Xiao, X., Zheng, Y., Luo, Q., & Xie, X., 2010. Finding similar users using category-based location history. *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 49, 442-445.
- Xiao, X., Zheng, Y., Luo, Q., & Xie, X., 2014. Inferring social ties between users with human location history. *Journal of Ambient Intelligence and Humanized Computing*, 5 (1), 3-19.
- Xie, Y., 2001. Web user clustering from access log using belief function. *International Conference on Knowledge Capture*, 202-208.
- Xu, B., Zhang, M., Pan, Z., & Yang, H., 2005. Content-Based Recommendation in E-Commerce. *International Conference of Computational Science and ITS Applications Proceedings*, (3481), 946-955.
- Yan, Z., & Chakraborty, D., 2014. *Semantics in Mobile Sensing (Synthesis Lectures on the Semantic Web: Theory and Technology)*. Morgan & Claypool Publishers.
- Yan, Z., Chakraborty, D., Parent, C., Spaccapietra, S., & Aberer, K., 2011. SeMiTri: a framework for semantic annotation of heterogeneous trajectories. In *Proceedings of the 14th international conference on extending database technology*, 259-270. ACM.
- Yanagisawa, Y., Akahani, J., & Satoh, T., 2003. Shape-based similarity query for trajectory of mobile objects. *Lecture Notes in Computer Science*, (2574), 63-77.

- Yang, X., Sun, Z., Ban, X., & Holguín-Veras, J., 2014. Urban freight delivery stop identification with GPS data. *Transportation Research Record: Journal of the Transportation Research Board*, (2411), 55-61.
- Ye, M., Janowicz, K., Mülligann, C., & Lee, W. C., 2011. What you are is when you are: the temporal dimension of feature types in location-based social networks. *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information System*, 102-111.
- Yin, H., & Wolfson, O., 2004. A weight-based map matching method in moving objects databases. In *Scientific and Statistical Database Management, 2004. Proceedings. 16th International Conference on*, 437-438. IEEE.
- Ying, J. J. C., Lee, W. C., and Tseng, V. S., 2013. Mining geographic-temporal-semantic patterns in trajectories for location prediction. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 5 (1), 2.
- Yiu, M. L., & Mamoulis, N., 2004. Clustering objects on a spatial network. In *Proceedings of the 2004 ACM SIGMOD international conference on Management of data*, 443-454. ACM.
- Yuan, J., Zheng, Y., & Xie, X., 2012. Discovering regions of different functions in a city using human mobility and POIs. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, 186-194. ACM.
- Yuan, M., Mark, D. M., Egenhofer, M. J., & Peuquet, D. J., 2004. Extensions to geographic representations. *A research agenda for geographic information science*, 129-156.
- Zhang, F., Zhu, X., Guo, W., Ye, X., Hu, T., & Huang, L., 2016. Analyzing urban human mobility patterns through a thematic model at a finer scale. *ISPRS International Journal of Geo-Information*, 5 (6), 78.
- Zhao, X. and Xu, W., 2009. A clustering-based approach for discovering interesting places in a single trajectory. *2nd International Conference on Intelligent Computing Technology and Automation, ICICTA 2009*, 429-432.
- Zheng, Y., Chen, Y., Xie, X., & Ma, W. Y., 2009. GeoLife2.0: a location-based social networking service. *10th International Conference on Mobile Data Management: Systems, Services and Middleware*, IEEE Computer Society, 357 - 358.
- Zheng, Y., & Zhou, X., 2011. *Computing with Spatial Trajectories*. USA: Springer Science + Business Media
- Zhong, H., Zhang, S., Wang, Y., Weng, S., & Shu, Y., 2014. Mining users' similarity from moving trajectories for mobile e-commerce recommendation. *International Journal of Hybrid Information Technology*, 7.
- Zhong, Y., Yuan, N. J., Zhong, W., Zhang, F., & Xie, X., 2015. You are where you go: Inferring demographic attributes from location check-ins. In *Proceedings of the eighth ACM international conference on web search and data mining*, 295-304. ACM.
- Zhou, C., 2004. *Discovering Personal Gazetteers: An Interactive Clustering Approach*. Geographic Information Science, November 12-13, Washington DC, USA. 267-268.
- Zhou, H., Wang, P., & Li, H., 2012. Research on adaptive parameters determination in dbscan algorithm. *Journal of Information & Computational Science*, 9 (7), 1967-1973.

Zimmermann, M., Kirste, T., and Spiliopoulou, M., 2009. Finding stops in error-prone trajectories of moving objects with time-based clustering. *Intelligent interactive assistance and mobile multimedia computing*, 275-286.

## APPENDICES

### APPENDIX A: DATA SAMPLE OF APLS<sup>4</sup>

CALL SIGN	EASTING	NORTHING	TIME STAMP	WORK TYPE	STATUS	EPE	OTHER OPERATIONAL INFO
302PF	529723	185255	03/08/2015 12:39	FP	On patrol	32	
302PF	529590	184349	03/08/2015 14:44	FP	On patrol	17	
302PF	529594	184349	03/08/2015 14:49	FP	On patrol	12	
302PF	529591	184350	03/08/2015 15:54	FP	On patrol	15	
302PF	xxx	yyy	ttt	FP	On patrol	10	
302PF	xxx	yyy	ttt	FP	On patrol	47	
302PF	xxx	yyy	ttt	FP	On patrol	24	
302PF	xxx	yyy	ttt	FP	On patrol	34	
302PF	xxx	yyy	ttt	FP	On patrol	23	
847PF	529794	187919	03/08/2015 13:06	CSO	With vehicle	26	
847PF	529564	187290	03/08/2015 14:58	CSO	With vehicle	26	
847PF	529370	187524	03/08/2015 15:03	CSO	With vehicle	18	
847PF	xxx	yyy	ttt	CSO	With vehicle	16	
847PF	xxx	yyy	ttt	CSO	With vehicle	23	
847PF	xxx	yyy	ttt	CSO	With vehicle	15	
847PF	xxx	yyy	ttt	CSO	With vehicle	16	
847PF	xxx	yyy	ttt	CSO	Refreshment	10	
847PF	xxx	yyy	ttt	CSO	Refreshment	38	
34PF	524560	187941	03/08/2015 15:03	SP	Limited	21	
34PF	xxx	yyy	ttt	SP	Limited	17	
34PF	xxx	yyy	ttt	SP	Limited	15	

FP = Foot Patrol Officer

CSO = Community Support Officer

SP = Special Constable

<sup>4</sup> This sample dataset is an anonymised example for demonstration. The call sign, location and time data in this table are not true.



## APPENDIX B: MERGED POI DATASET

Name	Street Address	Post Code	Latitude	Longitude	Category Code	Major Category	Sub-category	Open Time	Close Time
St Margaret's School	18 Kidderpore Gardens, London	NW3 7SR	51.52889	-0.2187	1031	Education	Secondary School	08:00	17:00
Top Shop	70 Berners St, Fitzrovia, London	W1T 3NL	51.52927	-0.2181	0846	Retail	Clothing Store	10:00	19:00
Strauss Photography	31 Ranulf Rd, London	NW2 2BS	51.52850	-0.2171	0208	Services	Media	09:30	17:00
Malorees Junior School	Christchurch Ave, London	NW6 7PB	51.52882	-0.2173	1031	Education	Primary School	08:00	16:45
Sainsbury's Local	165 Ladbroke Grove, London	W10 6HJ	51.52892	-0.2188	0847	Retail	Multi-item Retail	08:00	23:00
Barbados High Commission	1 Great Russell St, London	WC1B 3ND	51.52875	-0.2180	1133	Organisations	Government	10:30	15:00
North London Appliance Repair	25 Purley Ave, London	NW2 1SH	51.52831	-0.2172	0213	Services	Repair Service	10:30	18:30
SBF Fitness Ltd	Jack Straws Castle, N End Way, London	NW3 7ES	51.56275	-0.1800	0424	Sport & Entertainment	Sports complex, gym	09:00	21:45
Quex Road (Stop K)	North Maida Vale, London	NW6	51.54024	-0.1944	0959	Transport	Bus Transport	00:00	23:59

## APPENDIX C: POI CLASSIFICATION OF GOOGLE PLACES

accounting	embassy	museum
airport	finance	night_club
amusement_park	fire_station	painter
aquarium	florist	park
art_gallery	food	parking
atm	funeral_home	pet_store
bakery	furniture_store	pharmacy
bank	gas_station	physiotherapist
bar	general_contractor	place_of_worship
beauty_salon	grocery	plumber
bicycle_store	gym	police
book_store	hair_care	post_office
bus_station	hardware_store	real_estate_agency
cafe	health	restaurant
campground	home_goods_store	roofing_contractor
car_dealer	hospital	school
car_rental	insurance_agency	shoe_store
car_repair	jewelry_store	shopping_mall
car_wash	laundry	spa
casino	lawyer	stadium
cemetery	library	storage
church	liquor_store	store
city_hall	local_government_office	subway_station
clothing_store	locksmith	synagogue
convenience_store	lodging	taxi_stand
courthouse	meal_delivery	train_station
dentist	meal_takeaway	transit_station
department_store	mosque	travel_agency
doctor	movie_rental	university
electrician	movie_theater	veterinary_care
electronics_store	moving_company	zoo