# The Dynamic and Multidimensional Context of Urban Mobility

*Thesis submitted in accordance with the requirements of the University of Liverpool for the degree of Doctor in Philosophy by*

**Yunzhe Liu**

Supervisors:

**Professor Alex Singleton**

**Dr Daniel Arribas-Bel**

Department of Geography and Planning

School of Environmental Sciences

University of Liverpool

March 2021

# Statement of Originality

I, Yunzhe Liu 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 text.

Signed: Yunzhe Liu

Date: 30/03/2021

# ABSTRACT

**The Dynamic and Multidimensional Context of Urban Mobility**

**Yunzhe Liu**

The process of urbanisation exacerbates tensions between people and urban resources, which are particularly evident in those challenges related to urban mobility planning. Meanwhile, the expanded availability of Big (urban) Data has enabled new avenues for urban analysts to draw perspectives that better describe and model the urban environment, and often with a finer spatiotemporal scale. Such data are also increasingly offering the potential to integrate studies at both spatial and temporal dimensions, better capturing the dynamic context of urban environments. These studies could offer more holistic insights into aspects of the form and function of cities, facilitating improved context-based decision making.

With the continuous advancement of information technology and the exponential expansion of computational power, it has become technically feasible to integrated dynamic and multidimensional variables into urban mobility research. Within this context, this thesis presents three published journal articles that address methodological and substantive research gaps in urban analytics by developing an improved analytical framework based on conventional geodemographic analysis to investigate aspects of the urban environment within the context of urban mobility. The framework presents a typical knowledge-discovery workflow that systematically links measures of both spatiotemporal dynamics and multidimensional urban contexts. It covers detailed steps from the initial data selection to classification and interpretation, contributing to a better understanding of space, time, and urban mobility contexts.

As the research progressed, the contents in the prototype framework were constantly improved and enriched, and the developed framework has been implemented with open data collected from the targeted areas to conduct mobility-related research in practice. The usefulness of this framework has been exemplified by case studies, and it may easily be applied to other urban settings. Moreover, its contributions to related literature have also been exemplified through peer-reviewed academic papers.

# ACKNOWLEDGEMENT

The year 2020 will surely stand as one of the most unforgettable years in human history. Since the first COVID-19 case was reported in December 2019, the novel coronavirus (SARS-Cov-2) has been rapidly spreading to more than 200 countries or territories around the world. At the time of submitting this thesis, the unprecedented global pandemic has infected over 207 million people and resulted in more than 4.3 million deaths.

Taking the opportunity made available to me, I would like to express my sincere condolences to people who lost their loved ones. I would also like to express my respect and heartfelt gratitude to those people, working tirelessly on the front line and behind the scenes in the fight against coronavirus.

Completing the PhD journey is a long and winding road, especially during this challenging and extraordinary period. Without the guidance and support of my primary supervisor Prof. Alex Singleton, this work would not have been possible. Alex was also my dissertation supervisor during my undergraduate study at the University of Liverpool, and he has been giving me many wise and valuable suggestions since then. I would like to sincerely thank him for encouraging and advising my PhD research and allowing me to fulfil my childhood dream – to become a scientist.

I would also like to thank Dr Daniel Arribas-bel, my second supervisor, for his helpful feedback and comments during my PhD study and paper publications. Moreover, I would like to thank Dr Les Dolega, Dr Francisco Rowe, Prof. Guanpeng Dong, and Daniel for letting me be their teaching assistant of many undergraduate and postgraduate courses and allowing me to participate in many collaborative research projects. I would also like to thank the rest of the members at the Geographic Data Science Lab (GDSL) for their support and suggestions. I want to thank Dr Les Dolega and Prof. Scot Orford for being my PhD viva's internal and external examiner and giving me many valuable suggestions about my thesis.

I want to thank my family for funding my PhD, for the encouragement and endless help despite the long distance. Particularly, I would like to thank my girlfriend, Dr Meixu Chen. Thanks for putting your dream in my reality, for being the one who inspired me a lot, for standing by me all the time. You give me wings to fly, cast all my fears into the sky. I will never forget the time we shared, and I hope you are willing always to be on my side.

Last but not least, my special thank would go to ‘Oxford’ and ‘Cambridge’, my little two adorable guinea pigs. Thanks for being so sweet and cute all the time, healing me through looking into your large round eyes.

# LIST OF FIGURES

# LIST OF TABLES

# NOMENCLATURE

| Abbreviations | Descriptions |
|---|---|
| ACS | American Community Survey |
| AHAH | Access to Healthy Assets and Hazards |
| AOIs | (Urban) Area of Interests |
| AVS-OAC | Automated Variable Selection OAC |
| BCSS | Between Cluster Sum of Squares |
| CDR | Call Detail Records |
| COWZ-UK | Classification of Workplace Zone for the UK |
| CTOD | Center for Transit-Oriented Development |
| DBSCAN | Density-Based Spatial Clustering for Application with Noise |
| DIKW | Data; Information; Knowledge; Wisdom |
| DSM | Digital Surface Model |
| Eps | Eps1: Spatial Epsilon; Eps2: Temporal Epsilon |
| FAR | Floor Area Ratio |
| FGWC | Fuzzy Geographically Weighted Clustering |
| FHVs | For-Hire Vehicles |
| GB | Great Britain |
| GD-DBSCAN | Grid and KD-tree DBSCAN |
| GIS | Geographic Information System/Sciences |
| GPS | Global Positioning System |
| H-DBSCAN | Hierarchical DBSCAN |
| H-K-means | Hierarchical-K-means Clustering |
| ICTs | Information and Communication Technologies |
| IDC | International Data Corporation |
| IUC | Internet User Classification |
| KNN | K-Nearest Neighbour |
| LADs | Local Authority Districts |
| LBS | Location-based Services |
| LiDAR | Light Detection and Ranging |
| LOAC | London Output Area Classification |
| LTDS | London Travel Demand Survey |
| MAUP | Modifiable Area Unit Problem |
| MHCLG | The Ministry of Housing, Communities and Local Government |

| | |
|---|---|
| MinPts | Minimum Points |
| MODUM | Multidimensional Open Data Urban Morphology |
| MST | Minimum Spanning Tree |
| MTA | Metropolitan Transportation Authority |
| NPIs | Non-Pharmaceutical Interventions |
| NYC | New York City |
| OA | Output Areas |
| OAC | Output Area Classification |
| OD | Origin-Destination |
| ONS | Office for National Statistics |
| OS | Ordnance Survey |
| OSM | OpenStreetMap |
| OxIS | Oxford Internet Surveys |
| PCA | Principal Component Analysis |
| PCs | Principle Components |
| PRIZM | Potential Rating Index for Zip Markets |
| RFID | Radio-Frequency Identification |
| RS | Remote Sensing |
| SCAFC | Smartcard Automated Fare Collection |
| SLD | Smart Location Database |
| SLR | Systematic Literature Review |
| SOM | Self-Organising Map |
| ST-DBSCAN | Spatial-temporal DBSCAN |
| TAZs | Traffic Analysis Zones |
| TCoL | Transport Classification of Londoners |
| TfL | Transport for London |
| TLC | Taxi and Limousine Commission |
| TOD | Transit-oriented Development |
| TWSS | Total Within-cluster Sum of Squares |
| UK | United Kingdom |
| UNDESA | United Nations Department of Economic and Social Affairs |
| US | United States |
| WCSS | Within Cluster Sume of Squares |
| WEF | World Economic Forum |
| ZoLa | Zoning and Land Use |

# TABLE OF CONTENTS

# 1.0 INTRODUCTION

This chapter serves as the introduction to this thesis. It begins by describing the background and motivations for the research. The research objectives are elaborated. A summary of the general structure of the thesis is presented, highlighting the main contributions from each chapter.

## 1.1 RESEARCH BACKGROUND AND CHALLENGES

### 1.1.1 URBANISATION AND URBAN MOBILITY CHALLENGES

The number of people who live in cities is growing. According to UNDESA (2018), 55% of the global population (i.e., 4.2 billion inhabitants) reside in urban settlements, with estimates suggesting that, when combined with the overall population growth, this proportion is likely to increase to approximately 60% by 2030 and close to 70% by 2050. Urbanisation is ‘a complex socioeconomic process that transforms the built environment, converting formerly rural into urban settlements, while also shifting the spatial distribution of a population from rural to urban areas’ (UNDESA, 2018, 3). In 2009 there was a tipping point where more than half of the global population resided within urban as opposed to rural areas.

The quantity and scale of cities are also increasing. In 2018, there were 33 megacities, with populations over 10 million, and 10 more cities were projected to join this community by 2030. Meanwhile, the number of cities between 500,000 and 1 million people is also expected to grow from 598 to 710 (UNDESA, 2018). The total land area of cities is expanding, with rates twice as fast in terms of land area as they are in terms of population, and this trend is estimated to be nearly tripled during the period between 2000 and 2030, adding 1.2 million km$^2$ of additional urban built-up area to the world in the next three decades (Angel et al., 2011; World Bank, 2020).

Although urbanisation may bring prosperity and stimulate economic development, the rapid transition from predominantly rural to urban living is levying historically unprecedented demand for urban resources, posing numerous risks to cities (WEF, 2019). As cities grow, the demand for urban mobility increases (Michel & Ribardière, 2017). Such trends have been observed in both the developed and developing countries, while supply (e.g., transport infrastructures and capacity), by comparison, has consistently lagged far behind demand (de Palma & Lindsey, 2001; Narayanaswami, 2016). Disequilibrium between the demand for mobility and the supply of the transport system results in many negative economic, environmental, and social effects. These negative impacts are pervasively embodied in urban mobility challenges associated with high private motor vehicle dependency, including but not limited to excessive greenhouse gas emission, public health issues, increasing levels of congestions and road accidents, air and noise pollution, and wellbeing-related social exclusion problems (Chavez-Baeza & Sheinbaum-Pardo, 2014; Ettema et al., 2016; Hickman & Banister, 2014; Hynes, 2017; Rodrigue, 2020; She et al., 2017).

### 1.1.2 COMPLEX URBAN ENVIRONMENT

The aforementioned challenges not only threaten the urban systems in terms of socioeconomic and environmental aspects but also require more effective and holistic decision-making from urban planners and policymakers with significant concerns for "inclusive, healthy, resilient, and sustainable" urban planning at local, regional, and national levels (Webb et al., 2018; World Bank, 2020). Cities are quintessential complex systems, involving various components that are constantly changing with multifaceted relationships and interactions, some of which are hard to explain and include nonlinear dynamics, feedbacks, and high interconnectivity and unpredictability (Alberti et al., 2018; Batty, 2013b; Haken & Portugali, 2021; McPhearson et al., 2016; Stevenson & Gleeson, 2019). Hence, 'they are full of contestations, conflicts, and contingencies that are not easily captured, steered, and predicted respectively' (Bibri, 2021a, 1). Urban decision making needs to be supported by knowledge discovered from insights into urban morphology and metabolism (Thakuriah et al., 2017; Webb et al., 2018). These complexities make the urban environment challenging to understand and to manage when moving towards more sustainable and resilient development pathways, for example, the smart city initiative (Betis et al., 2018; Kumar et al., 2020; Rossetti, 2015; UN, 2015).

The complexities of the urban environment are expanded by the heterogeneous pace of change among its components. Wegener (1995) has conceptualised the major components within the urban environment in his famous model, that is, a model of urban models. In his model, the idealised urban environment consists of eight constantly changing components that are interconnected with each other and differ in temporal rates of change. Transport networks and land use, for example, are slow-changing components; and mobility-related components, such as, transport and activities, are fast-changing (Rodrigue, 2020; Wegener, 1995). Along with the accelerating urbanisation process, urban environments are becoming even more complicated and changeable, with more and more layers of interactions between their inhabitants and diverse components (Andrienko et al., 2020; Batty, 2013). From the urban mobility perspective, these complexities could be reflected in the intensity and distribution of urban transport resources at different times and locations, representing the dynamic aspect of urban mobility (Yuan & Raubal, 2012). Due to the close association between urban mobility and multidimensional urban contexts, for example, land use and socioeconomic attributes (Cervero & Kockelman, 1997; Ewing & Cervero, 2001, 2010), the complexity of urban mobility is further amplified, raising a myriad of challenges to urban transport planning and management.

As Gershenson (2016) has stated, it is difficult to separately study factors in any complex system since ‘its future is partly but strongly determined by its interactions with other components and its environment’ (p. 1). Hence, comprehensively understanding the urban environment and its impact on urban mobility through a systematic evidence-based activity has become an urgent demand for policymakers, city governors and urban planners (Kourtit et al., 2020). This situation requires the enhanced involvement of scientists and researchers (e.g., urban scientists and analysts) in urban policy, planning, and management processes (McPhearson et al., 2016). As Webb et al. (2018) have discussed, ‘researchers can contribute through collaborative knowledge development with urban stakeholders, capturing and translating learning for decisionmakers in a more systematic way, and facilitating innovation, evolutionary co-design and adaptive management of our cities’ (p.57).

### 1.1.3 UNDERSTANDING THE URBAN ENVIRONMENT FROM DATA

It should be emphasised that the phenomenon of urbanisation is not new, nor are the urban challenges emerging from these processes. For example, traffic congestion attributed to urban population growth had been an issue for urban areas since the Roman Empire (van Tilburg, 2006). However, a principal distinction between contemporary and past urbanisation is there are new ways to understand these processes (Singleton et al., 2017). The emergence of new ways of understanding and managing cities include the emergence of various enabling and driving instruments that can monitor activities within or attributes of urban environments. The rapid implementation of these instruments in everyday practices is enabling the accessing and sharing of data (Bibri & Krogstie, 2018; Singleton et al., 2017). With the continuous advancement of information and communication technologies (ICTs), the exponential expansion of computational power, coupled with the broader availability of urban data (i.e., Big/Urban Data Deluge), there has never been a time in history with more abundant data and tools. These data and tools provide significant opportunities for urban analysts to deepen their understanding of a plethora of urban problems (Andrienko et al., 2020; Batty, 2019; Hao et al., 2015; Singleton et al., 2017).

The majority of early studies in urban mobility are primarily limited in terms of the following three perspectives. First, most early studies are confined to data obtained from traditional data collection approaches, such as questionnaires and surveys (see Cullen, 1972; Love & Chapin, 1976), which are labour-intensive, time-consuming, and error-prone (Yuan & Raubal, 2012). Second, due to the limited spatial and temporal granularity, data from the conventional approaches may not provide sufficient evidence enabling a comprehensive analysis of the characteristics of the whole urban environment (Kong et al., 2020). Third, in the early age of urban mobility study, it was difficult to investigate the complex mobility patterns underlying human travel behaviours in space, time, and other multidimensional aspects simultaneously due to the lack of necessary computational power (Shen & Cheng, 2016). However, owing to the abovementioned changes, as in other quantitative urban studies, the landscape of urban mobility study has transformed dramatically. According to Long and Liu (2016) and Kong et al. (2020), these changes can be summarised into the following trends:

> 1) ‘transformation in spatial scale from high resolution but small coverage or wide coverage but low resolution to wide coverage with high resolution; 2) transformation in temporal scale from static cross-sectional to dynamic consistent; 3) transformation in granularity from land-oriented to human-oriented; and 4) transformation in methodology from conventional research group to crowdsourcing’. (p. 295)

A more in-depth discussion of these changes brought about by the Big (Urban) Data Deluge is presented in the literature review (see Section 2.1 in Chapter 2).

## 1.2 RESEARCH OBJECTIVES

With the increased availability of Big (Urban) Data and the exponential expansion of computational power, it has become technically feasible to integrate dynamic and multidimensional variables into urban mobility research. However, gaps in the literature continue to pose challenges in the research field. These challenges are preventing urban planners and policymakers from gaining a more thorough understanding of the urban environment and its mobility and making evidence-based decisions. The details of this issue are discussed in Chapter 2, Section 3.3, 4.2, and 5.2.

Within this context, the main aim of this thesis is to address gaps in the existing literature by developing a knowledge-discovery methodological framework to analyse the urban environment and urban mobility from multi-sourced urban data. In order to systematically achieve this goal, the typical workflow of building a geodemographic classification is utilised as the methodological foundation for this thesis. This method is a well-established and frequently used contextual approach in urban analytics (specified in Section 2.4 in Chapter 2). The developed framework should comprehensively consider both dynamic and multidimensional urban contexts while incorporating spatial, temporal, and contextual dimensions of urban mobility. To achieve the primary aim, several research objectives need to be met and are defined as follows:

- ***Objective 1***: to summarise and improve the typical workflow of building a geodemographic classification from existing literature
- ***Objective 2***: to identify and select variables that are commonly used in both traditional and recent studies to build the classification through a comprehensive literature review
- ***Objective 3***: to handle the adverse effects caused by the high dimensionality in the dataset by using the dimensionality reduction method
- ***Objective 4***: to extract urban and human mobility patterns from multi-sourced urban data while concurrently considering dynamic and contextual urban contexts
- ***Objective 5***: to apply the developed framework in the target case study area to manifest its utilities and contributions to the existing literature.

## 1.3 THESIS STRUCTURE AND IMPACT STATEMENT

This thesis is organised into six chapters that collectively elaborate on the multidimensional and dynamic contexts of urban mobility and how the methodological framework proposed in this thesis combines space, time and urban contexts together to provide a holistic understanding of urban mobility. In this chapter (Chapter 1), the overview of the research background and the research objectives are introduced.

Chapter 2 contains a comprehensive literature review elaborating the required background knowledge for this research in greater depth. The review first provides an overview of the contemporary status of the Big Data deluge and its manifestation in urban studies, that is, the urban data deluge. This is followed by the introduction of a typical knowledge-discovery framework, that is, the data, information, knowledge, wisdom (DIKW) hierarchy and its workflow in dealing with urban data. The theoretical concept of urban analytics and the relationship between urban contexts and urban mobility are reviewed. The latter half of the literature review focuses on geodemographics and how the geodemographic classification is created since it significantly facilitates the construction of the proposed methodology framework of this thesis.

Based on the conventional geodemographic analysis, this thesis proposes an enhanced methodological framework to more comprehensively analyse urban mobility by linking measures of both spatiotemporal dynamics and multidimensional urban contexts. The framework systematically covers the procedures from the initial data selection to the automated feature selection, then moves to data wrangling, processing, cluster analysis, and finally to the resulting visualisations and interpretations. The conceptual diagram of the methodological framework developed in this thesis is presented in *Figure 1.1*.

The overall advantage of the framework developed in this thesis is that it presents a relatively detailed knowledge-discovery workflow that presents holistic insight into urban mobility through space, time, and urban context. This thesis is of beneficial use in urban environment and mobility studies, exemplified by the publication of several academic journal articles (see Chapter 3, 4 and 5). The major contributions and innovations of the three published journal articles are listed as follows, while the other minor contributions are specified in the articles.

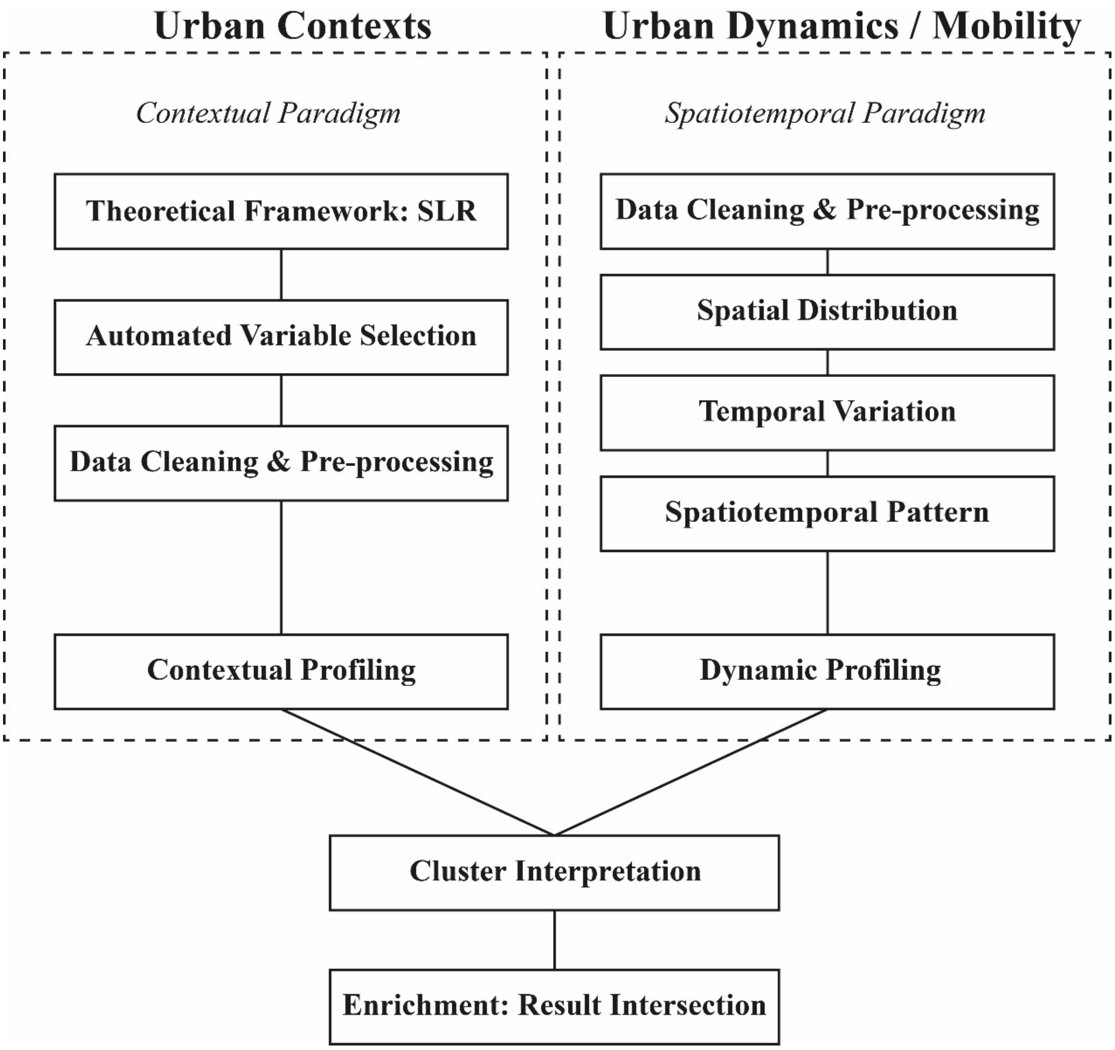


*Figure 1.1 Conceptual Diagram of The Proposed Methodological Framework*

In the first paper (Chapter 3), the main contributions are twofold. First, this paper contains a systematic literature review (SLR) of the existing urban mobility studies, aiming to identify commonly used variables capturing those urban contexts important to explaining urban mobility patterns. Findings from the SLR expand the traditional ‘three-Ds’ or ‘five-Ds’, which are concepts to capture a greater multidimensional perspective and categorise the identified candidate variables into four domains land use and built environment, location and accessibility, socioeconomic and demographic, and transit-related. These categories provide a theoretical framework guiding initial variable selection in building any mobility-related typologies. Second, the paper also contributes to the development of an analytical framework that considers the integration of both temporal dynamics and contextual domains in the study of urban mobility. Based on the framework, interaction and consistency between the context and use can be unveiled, with the interpretability of the temporal mobility pattern improved significantly due to contextual enrichment.

The second paper (Chapter 4) presents an innovative methodological contribution in the form of a variable selection framework. The proposed framework is based on the principal component analysis (PCA) and automatically selects candidate contextual variables for a multidimensional indicator. However, innovations of this paper are not only confined to automating variable selection but also holistically optimise the data processing (e.g., strong correlation pairs) and the output quality (e.g., clustering quality). Implementing the proposed framework has been examined by comparing it to a benchmark geodemographic classification – 2011 Output Area Classification (2011 OAC). The overall comparison shows that the classification built, utilising the presented framework, outperforms the 2011 OAC in terms of the clustering quality, which is explicitly manifested by several statistical indicators.

Returning to the substantive domain of mobility, the third paper (Chapter 5) utilises the frameworks presented in the abovementioned papers to analyse a taxi GPS dataset to carry out the urban Area of Interest (AOI) detection. Based on the conventional three-phase workflow of urban AOI analysis, the research gaps embedded with each phase of this process as presented in previous studies are addressed. The proposed framework integrates space, time, and urban contexts to provide a comprehensive perspective of urban mobility and, thus, better depicts the highly dynamic and multidimensional urban environment. The advantages of our framework are threefold. First, the ST-DBSCAN algorithm is employed to detect urban travel hotspots across both spatial and temporal dimensions. This improvement overcomes the shortcoming of the widely used DBSCAN or its derivative algorithms that merely focus on space or time in isolation. Second, instead of using polygons to define the boundaries of AOIs, the urban street network is creatively utilised as the organic carrier of AOIs, formulating road-constrained AOIs. This adjustment not only increases the accuracy of AOI detection but also considers the reshaping influence of urban morphology on human mobility, which is often overlooked by relevant studies. Finally, both temporal pattern profiling and contextual enrichment techniques are adopted to analyse the identified AOIs in more depth. This integrated analysis attempts to answer questions about where the AOIs are located, when the AOIs appear, what latent attributes affect the configuration and characterisation of an AOI, and how the AOIs contextually relate to different travellers.

Chapter 6 closes the thesis with a concise summary, re-emphasising the research outcomes, highlighting the contributions from both theoretical and technical perspectives to the scientific literature related to this topic. This summary is followed by a discussion of the limitations and future research directions of the study.

## 1.4 PUBLICATION LIST

The primary research results of this thesis have been converted into three full papers respectively published in different high-impact peer-reviewed academic journals, which are coupled with several accepted conference papers (for formal oral presentation) derived from the process of getting the results. The publications are listed below.

### JOURNAL ARTICLES

- Liu, Y., Singleton, A. and Arribas-bel, D. (2020) Considering context and dynamics: a classification of transit-oriented development for New York City. *Journal of Transport Geography*. 85. 102711. https://doi.org/10.1016/j.jtrangeo.2020.102711
- Liu, Y., Singleton, A. and Arribas-bel, D. (2019) A principal component analysis (PCA)-based framework for automated variable selection in geodemographic classification. *Geo-spatial Information Science*. 22(4). 251-264. https://doi.org/10.1080/10095020.2019.1621549
- Liu, Y., Singleton, A., Arribas-bel, D., and Chen, M. (2021). Identifying and understanding road-constrained areas of interest (AOIs) through spatiotemporal taxi GPS data: a case study in New York City. *Computers, Environment and Urban Systems*. 86. 101592. https://doi.org/10.1016/j.compenvurbsys.2020.101592

### CONFERENCE PAPERS

- Liu, Y., Arribas-bel, D., Dong, G., and Singleton, A. (2017) Towards Automated Variable Selection in Geodemographic Analysis: A Case Study of New York City. *GISRUK 2017*. Manchester
- Liu, Y., Arribas-bel, D., and Singleton, A. (2019). Understanding the Dynamics and Context of New York Transportation Hubs. *GISRUK 2019*. Newcastle.
- Liu, Y., Arribas-bel, D., and Singleton, A. (2020). Identifying and Understanding 'Street of Interests' (SOIs) through Spatiotemporal Taxi Trip Data: Case Study in New York City. *GISRUK 2020*. London.
- Liu, Y., Arribas-bel, D., and Singleton, A. (2020). Supplementing Context-based Subway Station Area Transit-Oriented Development (TOD) Topology by Using Dynamic Open Data, A Case Study of the New York City. *AAG*. Washington.
- Liu, Y., Chen, M., Arribas-bel, D., and Singleton, A. (2021). Profiling the Dynamic Pattern of Bike-sharing Stations: a case study of Citi Bike in New York City. *GISRUK 2021*. Cardiff.

# 2.0 LITERATURE REVIEW

This chapter contains a comprehensive literature review. It begins by describing the research background elaborated in the previous chapter in greater depth, concentrating on topics related to the Big Data deluge and emerging urban data (Section 2.1 and 2.2). In Section 2.3, the DIKW hierarchy is reviewed and followed by one of its typical manifestations within the context of urban study, that is, urban analytics (Section 2.4). The latter half of the literature review focused on geodemographics and how the geodemographic classification is created step-by-step since it serves as the theoretical and methodological foundation of this thesis; and significantly facilitates the construction of the proposed methodology framework of this thesis.

***Objective 1*** has been fulfilled in this chapter.

## 2.1 BIG DATA DELUGE

The phrase Big Data is largely ambiguous and amorphous (Moorthy et al., 2015). Since various stakeholders view the Big Data phenomena from a variety of viewpoints, it is difficult to provide a precise definition. Although there are no standardised definitions in academia or industry, Big Data might refer to datasets that are too large or complex to be (efficiently) processed by conventional data processing methods and tools (Kaisler et al., 2013), or might be simply understood as 'any data that cannot fit into an Excel spreadsheet' (Batty, 2013a, 274). Compared with traditional datasets, usually referred to as 'small data', Big Data typically contain massive unstructured attributes requiring to be frequently updated, requiring real-time streamed analytics (Chen et al., 2014). Various studies have portrayed the major characteristics of Big Data by a series of 'V's, which have expanded over time (see Khan et al., 2014; Panimalar et al., 2017; Rani et al., 2019). These essentially include volume, velocity, and variety (Laney, 2001). Additionally, representative Vs denote veracity, validity, and volatility, with one special V, that is, value, which is the desired outcome of Big Data processing (Khan et al., 2014; Kitchin, 2014a).

There has been an exponential growth in the volume of data generated since the early 21st century. Moorthy et al. (2015) have stated that 'data that the human race has accumulated in the past one decades, far exceeds the data that was available to mankind during the proceeding century' (p.74). Hal Varian (cited in Smolan & Erwitt, 2012) has pointed out that more data are being produced every two days at present than in all of history prior to 2003. Zikopoulos et al. (2012) have reported that more than 800,000 petabytes ($2^{50}$ bytes) of data were stored worldwide in 2000. According to an article from International Data Corporation (IDC), these data figures increased to 1.8 zettabytes ($2^{70}$ bytes) in 2011 and were estimated to exceed 40 zettabytes in 2020 (Gantz et al., 2012). The recent forecasting figure in the IDC White Paper indicated there would be approximately 175 zettabytes of data created, replicated, and consumed worldwide by 2025 (Reinsel et al., 2018). The explosive growth of Big Data is referred to as the '(Big) Data Deluge' (Bibri & Krogstie, 2018; Kitchin, 2013; Kourtit et al., 2020).

### 2.1.1 ADVANTAGES OF BIG DATA DELUGE

The promise that Big Data deluge would revolutionise scientific discovery and technological innovation is now generally recognised. Big Data deluge has had a far-reaching impact on the evolution of many research disciplines, particularly urban studies and analytics. Recently, there has been much enthusiasm about the potentiality offered by the Big Data deluge for better understanding, monitoring, analysing, and planning contemporary cities to increase their contribution to the goals of smart, sustainable urbanism (Bibri, 2021b, 2021c; Bibri & Krogstie, 2018; Kitchin, 2013; Kong et al., 2020).

Kitchin (2013) has pointed out that Big Data deluge provides excellent opportunities for studies switching from 'data-scarce to data-rich; static snapshots to dynamic unfoldings; coarse aggregations to high resolutions; relatively simple hypotheses and models to more complex, sophisticated simulations and theories' (p.263). More recently, Kong et al. (2020) have systematically reviewed Big Data-based urban studies and listed the following advantages of utilising such data to conduct urban studies. First, Kong et al. (2020) have pointed out that Big Data can unveil individual-based human activities and mobility patterns in multiple dimensions, facilitating urban study toward a "people-oriented" perspective. For example, my previous work, i.e., Liu and Cheng (2020) have utilised the travel transaction histories extracted from the Oyster Card system in London to conduct a personalised travel pattern analysis. Passengers with similar travel behaviour were grouped together, resulting in several groupings reflecting various subway travel patterns. Second, Kong et al. (2020) have also pointed out that compared to traditional data, Big Data are more timely. This advantage provides urban analysts new opportunities to examine dynamic change within cities at more granular temporal scales. Therefore, Big Data-based urban research is more time-scalable, which means that they are simple to aggregate at various time scales such as hourly, daily, weekly, monthly, and annually, facilitating a variety of studies, ranging from nearly real-time traffic estimation to relatively long-term social mobility analysis (Meng et al., 2020; Wu et al., 2020). Third, the emergence of Big Data has made it possible to achieve high spatial resolution and wide geographic coverage simultaneously in the study of the urban environment and human mobility patterns. These advantageous characteristics have alleviated the contradiction caused by using most traditional datasets, that is, making trade-offs between spatial resolution and geographic coverage (Long & Liu, 2016). For example, Cai et al. (2017) have integrated data extracted from social media and nighttime light satellite imagery to analyse the structure of the three metropolitan areas in China and replaced the traditional urban administrative units with the redefined boundaries based on human activity distributions.

### 2.1.2 CHALLENGES OF BIG DATA DELUGE

Along with the advantages offered by Big Data, challenges of using them also exist. Kong et al. (2020) and Martin (2015) have highlighted the following three major potential problems or challenges incurred in applying Big Data to urban studies.

First, the data quality issue can be a significant concern in Big Data, which may be attributed to data bias, inaccuracies in the data, or a lack of coverage. The manner in which the data are collected, the degree of imputed data within the data source, or intentional obfuscation by providers due to privacy concerns can all lead to producing inaccurate data (Martin, 2015). For example, the map service providers in China are required by the law to use a specific coordinate system, that is, GCJ-02, to encrypt their GPS data, with the goal of improving national security. In this coordinate system, the original GPS data (based on the WGS84 system) are processed by an obfuscation algorithm that adds random offsets to both the latitude and longitude, which may cause substantial research problems in the cross-region analysis (Xu et al., 2016). Moreover, Big Data may be skewed toward a specific ethnic group, gender, or socioeconomic class due to data bias, particularly for those data from self-selecting groups or containing underrepresentation of particular population groups. For instance, Wan et al. (2018) have cautioned that the sociodemographic background of location-based service (LBS) data is more likely to be biassed towards the younger generation because such data are only collected from the smart devices, which are more available to the young generation than the elderly. Additionally, this data bias possesses spatial heterogeneity, which means that the spatial distribution of data is uneven. For instance, Chen et al. (2019) and

Hu et al. (2015) have pointed out that the spatial distribution of the LBS network (LBSN) data is much denser in the major urban areas (i.e., city or town centres) than in the suburbs or more remote areas. This data bias may frame the Big Data-based urban research overwhelmingly focus on the metropolitan areas where data availability is high, whereas more small towns and remote regions are neglected.

Second, with more granular data being utilised by practitioners and researchers, discussions about ethical issues are also arising for both industry and academia (Martin, 2015; Someh et al., 2019). Ethical challenges include but are not limited to the concerns of individual privacy, trust, and awareness, and social equity (Chang, 2021; Rubinstein, 2013; Someh et al., 2019). These ethical challenges may amplify the data quality concerns mentioned above, limiting the utility of using Big Data in urban studies. For instance, the abovementioned social and spatial data biases are somewhat related to the digital divide among the regions or communities, which have been well documented in many studies (Longley et al., 2006; Riddlesden & Singleton, 2014; Singleton et al., 2020; Üsküplü et al., 2020). Third, Big Data are challenging to coordinate and handle since they are massive in terms of their volume and diverse in terms of their structure. Given that there is always a lag between the ability to understand big data and the ability to produce and collect them, determining how to exploit valuable information from such massive and complex datasets and how to implement the findings for practical applications are also noticeable challenges (Kong et al., 2020). For example, the issue of high data dimensionality (discussed in more depth in Chapter 2.4.3 and 4.0) is one of the representatives of this type of challenge.

## 2.2 URBAN DATA

Within urban settings, the '(Big) Data Deluge' manifests in the growing availability of data generated in continuous streams from various types of sensors (Bibri & Krogstie, 2018). Examples of this data are individual travel transactions from the Smartcard Automated Fare Collection (SCAFC) systems, GPS trajectories from GPS-enabled vehicles and mobile devices, entry and exit counts from turnstiles in public transit stations, geotagged tweets or images from LBSNs or other volunteered data repositories, and light detection and ranging (LiDAR) point clouds and remote sensing (RS) imagery from satellite or aircraft. Such emerging data sources have significantly enriched the content of traditional data sources represented by population census and land-use and travel surveys, which collectively make up urban data.

Urban data can be categorised into different types based on various criteria, such as data source, data environment, data geometry, and so forth (Kong et al., 2020). Particularly, based on the intentionality of collection, urban data can be broadly classified into two categories: *purposeful (or designed) data* and *organic data* (Singleton et al., 2017). *Table 2.1* presents examples of the major data types and their possible sources.

| **Category** | **Type of Urban Data** | **Source** | **Example of Data Provider** | **Example of Dataset** |
|---|---|---|---|---|
| *Purposeful/Designed Data* | Census | National census | US Census Bureau | ACS |
| | Travel data | Travel survey | TfL | LTDS |
| | Building Land uses | Land use survey | NYC Planning | ZoLa |
| | Greenspace | Land survey | OS | OS Open Greenspace |
| | Real-estate records | Housing statistics | HM Land Registry | Price Paid Data |
| | LiDAR / RS Image | Aircraft / Satellite | Environment Agency | LiDAR composite DSM |
| | Air quality | Gas sensors | Londonair | Londonair Data |
| | GPS points / trajectories | GPS trackers | Esri | Esri Tracker GPS |
| | Internet usage | Interview | OxIS | OxIS survey data |
| *Organic Data* | Smart Card records | SCAFC | TfL | Oyster Card Data |
| | GPS traces | GPS-enabled vehicles | TLC | Taxi Trip Record |
| | Entry/Exit counts | Turnstile sensors | MTA | Turnstile Data |
| | Bikesharing Trips | Docking station sensors | Citi Bike | Citi Bike Trips |
| | LBSN | Crowdsourcing | Twitter | Tweeter Data |
| | Volunteered geodata | Crowdsourcing | OSM | OSM Data |

***Table 2.1 Urban Data Type: Purposeful Data and Organic Data.***

As its name implies, *purposeful or designed data* are collected through statistically robust collection schemes, providing critical insight into cities (Singleton et al., 2017). One of the most typical forms of *purposeful data* is the nationwide socioeconomic and demographic survey investigating topics such as public health, educational attainment, unemployment rate, or providing a count of the entire population and households (i.e., national census). Taking the national census for example, in the UK, the national statistical agencies, such as the Office for National Statistics (ONS)[1], have provided a nationwide census every ten years since 1801. This national census is regarded as “the most complete source of information about the population” in the UK (ONS, 2016). The latest published census is the 2011 Census, taken on 27 March 2011, covering the various characteristics of the whole usual resident population for the UK, including but not limited to socioeconomic, demographic, and built environment domains. As for the US, the traditionally decennial census has been replaced by the American Community Survey (ACS) since 2010. ACS, first launched in 2005, is a rolling survey program conducted by the US Census Bureau, gathering information from more than 3.5 million households across the country on a yearly basis, covering social, housing, demographic, and economic subjects (US Census Bureau, 2020).

The expense of collecting purposeful data is usually costly in terms of human, physical, and financial resources. In practice, national surveys often require the creation of a highly accurate register listing all known addresses (e.g. home or workplace) to avoid the problem of undercoverage, that is, failure to count legitimate households and populations (Leventhal, 2013; ONS, 2015; Singleton et al., 2017). Each address is contacted by delivering a questionnaire through various approaches, including post, phone, and the Internet. A force of field enumerators are then dispatched to those households that do not respond to the survey. According to the estimation, the 2011 UK Census costs approximately £432 million (Leventhal, 2013), as for the US, the census ‘cost about 42 per person counted’ (Singleton et al., 2017, 20).

*Organic data* are more “accidental” in nature, and their potential utilities can exceed the purposes other than the one for which they were originally intended (Arribas-Bel, 2014). For instance, passively collected smart card transaction data from the SCAFC systems can be employed to analyse passengers’ travel behaviour, whereas its collected purpose is to simplify the task of ticketing and revenue collection (Pelletier et al., 2011). Moreover, the GPS data extracted from GPS-enabled devices and vehicles are another example showing the “accidental” characteristics of organic urban data. Most GPS devices used in daily life are primarily designated to provide positioning and navigation for end-users. At the same time, the spatiotemporal coordinate data recorded in their background database can be further processed into detailed trajectories or origin-destination (OD) matrices that represent a reasonable proxy for individual mobility and hence facilitate many urban mobility studies (Tang et al., 2015; Wang et al., 2019; Wang

[1] ONS is responsible for the census in England and Wales; GROS is responsible for the census in Scotland; NISRA is responsible for the census in Northern Ireland

et al., 2015; Zhao et al., 2016). Accordingly, these urban data are regarded as a byproduct of some transactional processes of daily life within cities (Singleton et al., 2017). Compared with conventional *purposeful data*, *organic data* share more similar characteristics with big data because they usually contain detailed personalised information and are available at a high spatiotemporal resolution with wide geographic coverage. However, the potential challenges of using *organic data* are also shared with Big Data, which have been discussed in the previous section. Typically, similar to other Big Data, one of the most critical limitations of *organic data* is the lack of contextual information. This is especially pronounced in the transport-related dataset due to its "accidental" characteristics, such as GPS data, smartcard data, and data from other sensors (such as station turnstile). For example, in most vehicle-based GPS data, no extra information other than the spatiotemporal coordinates and essential identifiers are usually provided. As a result, many in-depth studies, such as travel purpose analysis or functional zone detection analysis, which are of importance in urban mobility analysis, cannot be carried out by solely using this data.

The difference between *organic* and *purposeful data* are vital as they present different utilities in different conditions. Typically, *purposeful data* are more useful for measuring (very) slow-changing components in the urban environment since they are usually rich with detailed contextual information capturing multidimensional characteristics of the observations. *Organic data*, on the other hand, are better utilised to capture fast-changing components since they usually contain finer spatiotemporal granularity enabling nearly real-time analysis (see more in Section 2.4.4). However, in order to provide a fuller image of the complex urban environment, it is necessary to combine these two types of urban data, with the essence being retained and the peripheral being discarded. For instance, urban analysts have sought to distil the general pattern of urban mobility from *organic data* and then use *purposeful data* to investigate the associations between them or enrich the result interpretability (Chen et al., 2019; Liu & Cheng, 2020; Wang et al., 2017b; Xu et al., 2019; Zhou et al., 2019). With the advancement of machine-learning techniques, it is possible to use the labelled but small-scale *purposeful data* to predict unlabelled and large-scale organic data through various predictive machine-learning models. For instance, Zhang et al. (2020) have applied a graph-based prediction model to the sociodemographic variables in the London Travel Demand Survey (LTDS) to estimate passengers' sociodemographic status in the massive unlabelled London Oyster card data.

## 2.3 DIKW PARADIGM

As discussed in previous sections, the emergence of urban (Big) Data has fundamentally transformed those existing research paradigms in urban science, enabling new perspectives on urban life than has even been available before. However, it should be noted that data in themselves do not convey insight, and ‘they only have utility if meaning and value can be extracted from them’ (Kitchin, 2014b, 100). As such, it is what is done with data that is crucial, not merely that they are generated.

The progressive relationships between data, information, knowledge, and wisdom are presented in the well-known DIKW pyramid, or hierarchy, in which **D** is the abbreviation for *data*, **I** for *information*, **K** for *knowledge*, and **W** for *wisdom* (Rowley, 2007). This pyramid illustrates the concept that ‘information is described in data terms, knowledge with respect to information, and wisdom in terms of knowledge’ (Aditya Shastry & Sanjay, 2020, 204). The cake metaphor, introduced by Gurteen (1998), can assist in understanding the difference between those components within the DIKW pyramid. According to Gurteen (1998), data are compared to molecular components of the cake; information to a list of ingredients; knowledge to the recipe (indicating how to make the cake); and finally, wisdom corresponds to know-why and for-whom to make the cake. As stated by Fricke (2019), ‘the DIKW suggests that there are more data than information in the world, more information than knowledge and more knowledge than wisdom’ (p.35).

*Figure 2.1* shows the adapted DIKW pyramid within the context of the urban environment, with each layer distinguished by a series of distillation processes, such as abstracting, organising, analysing, interpreting, and applying, which adds meaning and value by revealing relationships and insight (Kitchin, 2014b). The DIKW hierarchy is effective in demonstrating the process of turning raw urban data that are less useful into knowledge that is more informative for the end-users, for example, urban planners or policymakers (Nativi et al., 2020). Urban data are generated and collected from various sources, either organically or purposefully, acting as the abstraction of the real urban environment, which also can be regarded as the raw material to generate information. Information is considered as the added-value product made up of linked elements and meaningful patterns that results from the processing and organisation of accessible data. Knowledge can be viewed as a collection of well-organised information, which is gained from comprehending information through analysing and interpreting value patterns in information. Finally, at the summit of the DIKW pyramid, the wisdom is defined as a combination of multiple applied knowledge, which has a direct influence on decision-making.

The DIKW hierarchy provides an overall framework for transforming raw data into useful knowledge that is applicable for decision-making, manifesting a data-driven decision-making process. The investigation of deaths from cholera outbreaks in the Soho district of London, undertaken by Dr John Snow in 1854, is often regarded as the earliest recorded example of spatial data analysis and one of the most famous examples of the DIKW paradigm (Longley et al., 2015). The death cases of cholera collected in Snow's work are the raw *Data*; the map he made, that is, the famous John Snow's cholera map of Soho[1], showing the spatial distribution of these cases are *Information*; the correlation between the distance from the water pump and the case hotspots can be considered as *Knowledge*, revealing the transmission way of cholera; and the decision of removing the pump handle on Broad Street is the actual action referenced by the *Knowledge*, thus, can be considered as *Wisdom*.

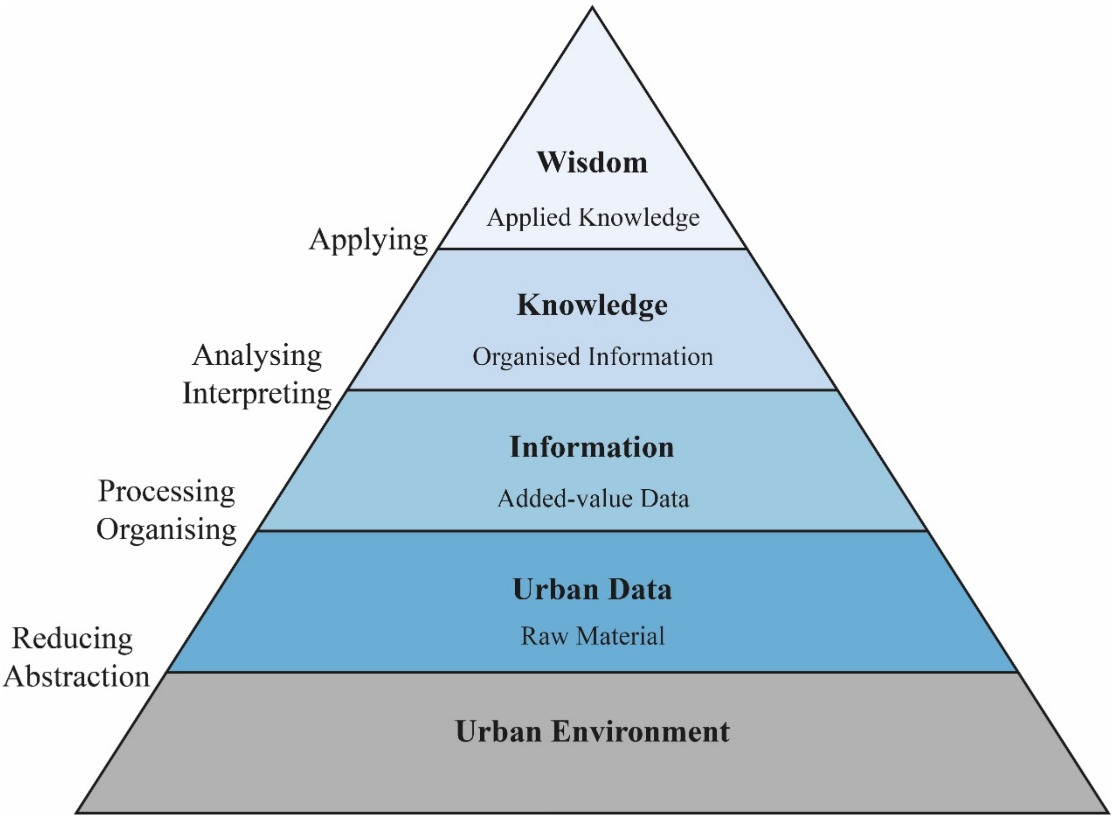


***Figure 2.1 DIKW Pyramid in the Context of the Urban Environment.***

Diagram adapted from Rowley (2007) and Kitchin (2014b).

[1] See https://www.bl.uk/collection-items/john-snows-map-showing-the-spread-of-cholera-in-soho-london-1855#

The data-to-knowledge transition is usually the responsibility of data analysts or scientists, whereas decision making (i.e., from knowledge to wisdom) necessitates more comprehensive considerations based on the knowledge from many perspectives. Since the primary aim of a decision is value creation (Nurulin et al., 2019), it is critical to emphasise the importance of multi-stakeholders in the decision-making progress as active contributors in order to move science beyond its traditional technical boundaries and translate knowledge into understandable and usable forms (Tumwebaze et al., 2021). In the context of the contemporary urban environment, making decisions has become more complex than ever due to the urban complexity and explosive availability of data (i.e., Big Data deluge) and the involvement of stakeholders with often-conflicting objectives and dynamic interactions at different phases of decision making. Decision-making in the modern urban system, as summarised by Eräranta & Staffans (2015), is not only a data-driven practice but also a collective learning procedure supported by advanced ICT-based technologies and visualisations of available data, constant processes, and local history and stories. This requires a higher involvement and collaborations of multi-stakeholders, such as citizens, end-users, planners, engineers, experts, elected representatives, in the different phases of the project (Dupont et al., 2015). Since the stakeholders involved preserve different information and knowledge of the problems, their complex decisions should be exposed to negotiation in various phases of the process, which can enable various perspectives to be considered and thus make a robust decision based on a shared vision of the city (Tran Thi Hoang et al., 2019).

## 2.4 URBAN ANALYTICS

Before the era of Big Data deluge, data about cities have been systematically collected, and various statistics have been presented for the purpose of influencing urban planning and policymaking since the late 19th century (Singleton et al., 2017). In the UK, for instance, Charles Booth undertook a massive survey of the socioeconomic and occupation conditions of the working-class population living in the inner London area in the late 1890s. Based on the data he collected, Booth innovatively built a poverty-related classification of households, which categorised those surveyed households into seven different clusters based on their conditions of poverty, aiming to provide 'a statistical record of impressions of degree of poverty' (Hennock, 1991, 190). Such classification was colour-coded and mapped out subsequently, known as the famous Booth's poverty maps, displaying the social class of inner London on a street-by-street basis; and it was regarded as one of the first attempts to map the large-scale sociospatial structure of London (Orford et al., 2002).

With the rapid development of ICTs and the widespread availability of large-scale urban data in the era of Big Data deluge, the term urban analytics “seems to effortlessly roll off the tongue as though we have used it all our lives” (Batty, 2019, 403). Apart from the benefits offered by the emerging ICTs in terms of improving quality of life and facilitating more sustainable resource management, they are also essential to comprehending cities as constantly changing complex systems configured by multi-layer networks that are self-organised and ‘embedded in space and enabled by various types of infrastructures, activities, and services’ (Bibri, 2021a, 2). Urban analytics can be generally defined as a set of urban research practices applying computational and statistical approaches to emerging urban data to develop an in-depth understanding of urban processes, which is ‘fast emerging as the core set of tools employed to deal with problems of Big Data, urban simulation, and geodemographics’ (Batty, 2019, 403).

Urban analytics can be considered as the synonym of the DIKW paradigm in contemporary urban research. Urban analytics aim to transform vast amounts of urban data into informative knowledge supporting evidence-based decision making and deepening understanding of the complex urban environment through the application of a range of data science-related techniques, such as data mining, machine learning, statistical analysis, database querying, or a combination of these approaches (Bibri & Krogstie, 2018). The basic assumption is that the analysis of emerging urban data can productively contribute to the solution to long-standing challenges in the urban environment (Kandt & Batty, 2021). The abundance of urban data, along with the computational and analytical power, opens up new avenues for urban analytics and planning in moving towards more sustainable and resilient development pathways. Hence, the importance of Big (urban) Data analytics are being emphasised more frequently in contemporary urbanism, as well as their novel implementations in enhancing and promoting sustainability. This trend is evinced by many studies carried out recently on smart cities, ecocities, and data-driven smart sustainable urbanism (e.g., (Bibri, 2021b, 2021c, 2021a; Bibri & Krogstie, 2020, 2018; Pasichnyi et al., 2019; Yigitcanlar & Cugurullo, 2020)).

Urban analytics might be argued as comprising two major research agendas: *urban contexts* and *urban dynamics*. Urban contexts can be regarded as the umbrella name of multidimensional attributes delineating the slow-changing components in the urban environment, represented by attributes from the built environment and neighbourhoods

(Rodrigue, 2020; Wegener, 1995). Urban analytics from this agenda typically utilise purposeful urban data, for example, census or land-use surveys, to capture the characteristics of urban contexts. Urban dynamics, as the counterpart of urban contexts, can be collectively referred to as human behaviour and mobility variables depicting fast-changing urban components (Rodrigue, 2020; Wegener, 1995). As mentioned before, *organic data* are more commonly utilised to conduct related analyses in recent years, while the usage of *purposeful data*, such as data from the travel demand survey, is still in evidence (see Jiang et al., 2012; Zhang et al., 2020). *Figure 2.2* presents the conceptual diagram, adapted from the urban model introduced by (Wegener, 1995), showing the relationship between urban contexts and urban dynamics in urban analytics. This diagram also shows the three major components in the urban environment: human mobility, neighbourhood attribute, and the built environment. Compared to fast-changing human mobility, neighbourhood attribute and built environment are classified as slow-changing categories, in which the built environment is viewed as the most 'static' component in the urban environment.

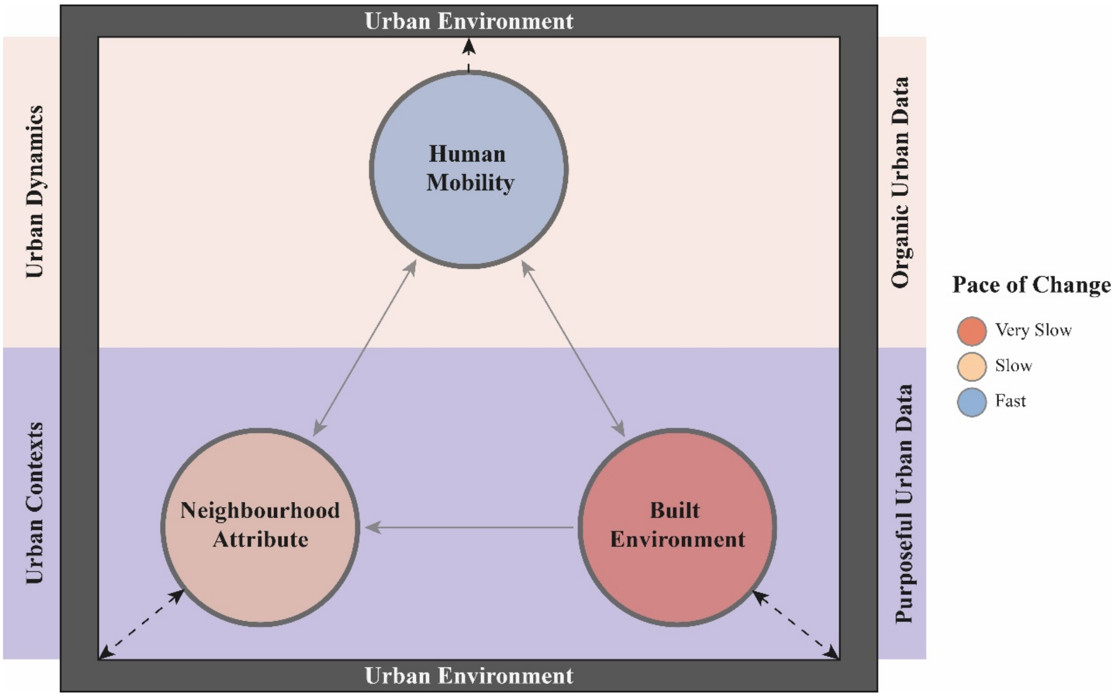


***Figure 2.2 Simplified Main Components in the Urban Environment***

Diagram adapted from (Wegener, 1995) 'a model of urban models' and (Rodrigue, 2020) 'dynamics of urban changes'.

## 2.4.1 URBAN CONTEXTS

Urban contexts are paramount for city life, which ‘refers holistically to the social, environmental, and economic settings within which we live our lives’ (Singleton et al., 2017, 79). Urban contexts have extensive subtle influences on people’s entire life, including but not limited to health and social wellbeing, education and occupation, access to opportunities (e.g., voting, new technologies, or the labour market), and are known to be influential for various behavioural patterns (Abreu & Öner, 2020; Gao et al., 2019; Krefis et al., 2018; Longley et al., 2006; Mouratidis, 2018; Singleton, 2010). Such enduring influences are usually referred to as the “neighbourhood effect”, which has a lengthy research history (see Mouratidis, 2018; Sampson, 2012; Thiele et al., 2016; Urban, 2009).

Particularly, numerous studies have discussed the impacts of urban contexts on people’s travel behaviour. In travel research, such influences have often been named with words beginning with D (see Chapter 3.3 for more details). For example, the well-known ‘three-Ds’ concept, namely, *density* in development, *diversity* in land use and urban *design* introduced by Cervero and Kockelman (1997) has emphasised the importance of environmental factors (i.e., built environment) in affecting people’s travel demand. The ‘three-Ds’ concept has been widely advocated by urban planning agencies around the world and is recognised as the theoretical basis of transit-oriented development (TOD) in strategic urban planning, which aims to encourage public transit use (e.g. Staricco & Vitale Brovarone, 2018; Sung & Choi, 2017; van Lierop et al., 2017; Xu et al., 2017). Furthermore, with the more profound insights into this field, the concepts of the original ‘three-Ds’ model established by Cervero and Kockelman (1997) has been expanded as ‘six-Ds’ since the introductions of more ‘D’ variables, i.e., *destination accessibility*, *distance to transit,* and *demand management* (Ewing & Cervero, 2001; Ewing & Cervero, 2010).

Kattiyapornpong and Miller (2009) have additionally identified that neighbourhoods’ demographic and socioeconomic characteristics (e.g., age, income, and life stage) have significant differential and interactive influences on residents’ travel behaviour in terms of travel preference and choice of travel mode. The profound influences of sociodemographic characteristics on people’s travel behaviour were also identified in the review conducted by Ewing and Cervero (2010) and are also much in evidence in a number of other international studies (e.g. Bajracharya & Shrestha, 2017; Dieleman et al., 2002; Gao et al., 2019; Ma et al., 2018; Syam et al., 2012). Thus, while not part of the built environment variables, Ewing and Cervero (2010) have recognised the sociodemographic variables as “the seventh D” (p.267).

It should also be noted that the influence of urban context is not only limited to residential neighbourhoods but also increasingly includes transitional spaces, such as areas occupied through travel (e.g., origins and destinations), and other collective locations such as workplaces (Cockings et al., 2020; Singleton et al., 2017).

### 2.4.2 APPROACHES TO MEASURING URBAN CONTEXTS

There are two major approaches usually employed by urban analysts to understand urban contexts, namely, the variable and contextual paradigms (Singleton et al., 2017). The variable approach typically focuses on answering questions about the influence of independent variables on a dependent variable. On the other hand, a contextual approach regards neighbourhoods as a complex mix of interrelated and difficult to separate attributes described through a multivariate rather than a univariate object (Webber & Burrows, 2018).

There are two approaches to composite metrics that are widely used in contextual analysis (Singleton et al., 2017). The first relates to the composite index, which provides a continuous measure of the constructs to be evaluated and takes and distils multiple input variables into a single number (i.e., a single metric). Such composite indicators are most commonly developed in the field of deprivation (i.e., hardship) of an individual’s socioeconomic context (Townsend, 1987). One successful attempt, for example, to capture various aspects of deprivation is the UK’s Index of Multiple Deprivation (IMD). This index is a single score of deprivation and has been utilised for many years to measure the relative deprivation of small geographical areas (e.g., lower super output areas, (LSOAs)) in the UK (MHCLG, 2010; Kinsella, 2007; Mclennan et al., 2019; Noble et al., 2006; Smith et al., 2015). The overall IMD is calculated as a weighted level aggregation of several constituent dimensions of deprivation. For instance, the 2019 IMD was calculated based on several attributes extracted from seven different dimensions of deprivation, namely, employment, income, education, crime, barriers to housing and services, and living environment (Mclennan et al., 2019). Another example of a composite index is the Access to Health Assets and Hazards (AHAH) index created by Daras et al. (2019), which presents a summary statistic of the health-related accessibility and built environmental characteristics of an area, with the goal of offering a holistic assessment of neighbourhood quality across the GB, based on how “healthy” they are.

An alternative approach is offered by the creation of geographic classifications, commonly referred to as geodemographics (Harris et al., 2005; Leventhal, 2016). This approach has been well-documented and has a lengthy history of research, which is discussed in more depth in the next sections. Compared to indices, one of the most distinctive characteristics of the classification approach is that the output is categorical and multidimensional and, therefore, does not rely on a single measure or index (Alexiou, 2016). Such output is more typically used to provide a qualitative summary portraying the most salient contextual characteristics of an area and has particular utility when reflecting nonlinear relationships between input measures (Singleton et al., 2017).

### 2.4.3 GEODEMOGRAPHICS

Geodemographics can be simply defined as “an analysis of people by where they live” (Sleight, 1997, 16). It is usually organised by classification of areas sharing similar multidimensional characteristics and aims to summarise multiple characteristics of socioeconomic, demographic, and built environment concerning a set of small geographic areas (Harris et al., 2005; Singleton et al., 2017). A geodemographic classification is constructed by utilising a clustering technique that organises each observation (i.e., geographic area) into a type of cluster according to the overall similarities concealed within the multidimensional attributes they shared. Since the observational areas targeted by geodemographic analysis are often relatively small, such as groups of postcodes or census tracts, many scholars tend to use “neighbourhood” as a shorthand term to describe this kind of small area zonal geography (Harris et al., 2005; Leventhal, 2016; Singleton & Longley, 2009).

The conceptual tenet of geodemographic classification relates to the notion of societal homophily, or the “birds of a feather flock together” phenomenon (Harris et al., 2005). In geographic terms, this relates to the tendency for people to be attracted to areas that comprise others with similar characteristics to themselves (Sleight, 1997), reflecting a fundamental axiom in human geography discipline, known as Tobler’s first law of geography, that is, ‘everything is related to everything else, but near things are more related than distant things’ (Tobler, 1970, 236). Although geodemographic analysis has developed significantly over decades, its theoretical foundation always adheres to the principle that ‘people tend to align themselves with the behaviour and aspirations of the local communities where they live’ (Alexiou, 2016, 21). Therefore, the objective when building a geodemographic classification is to partition a set of small areas into clusters that share similar attributes, with the output of such clusters offering a simplified and categorical representation of the salient multidimensional characteristics of the areas (Spielman & Singleton, 2015).

In accordance with other commonly used measures of socioeconomic stratification, geodemographic classification is not immune to criticism (Harris et al., 2005; Leventhal, 2016). By aggregating areas sharing similar multidimensional characteristics into clusters, the geodemographic clusters generally represent the average characteristics of an area where people live. However, using such averages from areal aggregation to infer individuals’ characteristics is to risk invoking the ecological fallacy (Dalton & Thatcher, 2015), that is, ‘an error of deduction that involves deriving conclusions about individuals solely on the basis of an analysis of group data’ (O’Dowd, 2003, 84). However, problems related to ecological fallacy or modifiable areal unit problem (MAUP) are certainly not limited to geodemographic analysis and appear more commonly in most socioeconomic phenomenon studies where disclosure is an important issue(Longley et al., 2015).

Booth’s mapping of London in the late 19th century, referred to in the previous section, has been argued as an early (albeit non-computational) geodemographic classification (Adnan, 2011; Alexiou, 2016; Leventhal, 2016; Webber & Burrows, 2018). However, a more robust intellectual heritage links back to the urban ecology studies of the Chicago School of Sociology in the 1920s and 1930s (Batey & Brown, 1995; Timms, 1971). In their contemporary methodological form, most geodemographics are strongly tied to the work of Richard Webber, who established a national classification of areas using the 1971 UK Census in the late 1970s (see Webber, 1978). Geodemographics developed parallelly in both UK and US and gained significant traction during the 1980s as a tool for commercial marketing (Reibel, 2011). It was used in *PRIZM* in the US and *Acorn* in the UK and sustained until the present day. (For an overview of these developments, see Singleton & Spielman, 2014). Geodemographic classification has an expansive and international lineage with utility for private and public sector applications within various geographic extents (Gale et al., 2016; Singleton & Longley, 2015; Singleton & Spielman, 2014).

Geodemographic classification can be utilised either as a general-purpose or an application-specific analytical tool. For instance, geodemographic classification can be used to depict general urban contexts within a national extent, for example, the 2001 and 2011 output area classification (OAC) (Gale et al., 2016; Vickers & Rees, 2007). Alternatively, it can be built to investigate general problems within the frame of a targeted geographic extent, for example, the 2011 London output area classification (LOAC) (Singleton & Longley, 2015). Moreover, it can be specifically constructed to analyse a bespoke substantive problem – for example, Internet User Classification (IUC) and a national bespoke educational geodemographic system (Alexiou et al., 2020; Singleton, 2016; Singleton & Longley, 2009).

## DATA USED IN THE CONSTRUCTION OF A GEODEMOGRAPHIC CLASSIFICATION

Predominantly but not universally, The majority of existing geodemographic classifications comprise input variables from the *purposeful urban data*, particularly the census (Alexiou, 2016; Harris et al., 2005). In the UK, for instance, each decennially released census has triggered a new generation of geodemographic classification fuelled by the latest presented results (see Charlton et al., 1985; Gale et al., 2016; Robinson, 1998; Vickers & Rees, 2007; Webber, 1975; Webber & Craig, 1978). Even given the Big Data deluge, the population census has always been “the most important and valuable source of geodemographic analysis” because it offers reliable, comprehensive, and coherent data on the sociodemographic characteristics of residents in each neighbourhood in the country (Leventhal, 2016, 7). Apart from sociodemographic attributes, other variables as part of the population census, such as the commuting flows between people’s residence and workplace, have also been employed to build a geodemographic classification (Hincks et al., 2018)**.**

In addition to those attributes extracted from the population census, variables captured from other purposeful urban datasets can be used to conduct geodemographic analysis. For instance, Alexiou et al. (2016) have created a classification of Multidimensional Open Data Urban Morphology (MODUM) to delineate the urban morphology of England and Wales by using built environment attributes from GIS and spatial datasets mainly provided by the national mapping and surveying agency Ordnance Survey (OS). Moreover, although some of the inputs employed in the IUC that was developed by Singleton et al. (2016) were partially from the UK 2011 Census, variables extracted from the Oxford Internet Survey (OxIS) primarily occupied a large share of the inputs in building such classification. The Transport Classification of Londoners (TCoL), a classification built by TfL (2017) to classify Londoners based on their travel mode choice and travel purpose, is another example of using purposeful urban data to create geodemographic classification, in which the majority of the input variables were extracted from the LTDS 2012-2015.

## GENERALISED WORKFLOWS FOR CREATING GEODEMOGRAPHIC CLASSIFICATIONS

Building a successful geodemographic classification can be a time-consuming and challenging process since it can be affected by various factors such as data availability, methodological choices, geographic coverage, and weighting schemes (Alexiou & Singleton, 2015; Liu et al., 2019; Openshaw et al., 1995; Webber, 1978).

Although there is no standard approach to creating a geodemographic classification, commonly used processes can be summarised into five stages, namely, deciding a geographic extent and scale, variable extraction and selection, variable preprocessing, cluster analysis, and result examination and interpretation.

### *DECIDE GEOGRAPHIC EXTENT AND SCALE*

The first stage of creating a geodemographic classification is to decide the geographic extent (i.e., the coverage) and the spatial resolution (i.e., the scale) that the classification will be created for. These decisions are not only determined by the research scope and purpose but are also based on data available to the classification builder at different geographic scales (Alexiou & Singleton, 2015). For example, most of the well-known geodemographic classifications in the UK, such as the 2001 and 2011 OAC, are based on the national census aggregated at the Output Area (OA) level, which is the lowest geographical level at which census estimates are provided. Hincks et al. (2018) constructed a new geodemographic classification of commuting flows for England and Wales by utilising the contextual variables characterising the commuters at Middle Layer Super Output Area (MSOA) level. Moreover, according to Singleton et al. (2016), the IUC, a purpose-specific geodemographic classification that focuses on measuring Internet use and engagement was created and released at the LSOA level.

Many geodemographic classifications have coverage for the whole extent of a country; however, some other bespoke classifications consider more localised extents. For instance, Singleton et al. (2016) used England as the geographic extent of the IUC and the Greater London area was adopted as the geographic extent of LOAC (Singleton & Longley, 2015).

### *VARIABLE SELECTION*

The second stage is to assemble a database comprising a series of variables that are considered important to the differentiation of areas. According to Murphy & Smith (2014), the overarching objective when creating any geodemographic classification is to develop a framework for selecting potential variables that are beneficial for producing meaningful and application-relevant classifications. There are differences of opinion concerning the optimal brevity of inputs. Some scholars have advocated that “the fewer the variables the better”, for example, Openshaw & Wymer (1995), whereas others, such as Harris et al. (2005), have claimed that a more meaningful classification is likely to be built through inputting more variables.

However, many studies have found that clustering performance can be significantly improved by reducing the number of input variables due to the phenomenon called the “curse of dimensionality” (see Iguyon & Elisseeff, 2003; Pacheco, 2015; Rojas, 2015; Tang et al., 2014). Therefore, the objective of variable selection might more effectively be framed by selecting the most parsimonious inputs, that is, to select the smallest subset of input variables that capture the most variation within the original dataset (Debenham, 2002; Gale et al., 2016; Harris et al., 2005). Apart from pragmatically considering the data availability, the goal of variable selection can usually be achieved by balancing both a theoretical and empirical rationale for variable inclusion (Spielman & Singleton, 2015). For instance, it is common to initially select a group of candidate variables that draw upon the references of comprehensive literature. This is accompanied by an input variable assessment process, which typically takes into account a wide variety of statistical factors of the candidate variables, including but not limited to their collinearity, spatial coverage, and potential impacts on the clustering results (Gale et al., 2016; Liu et al., 2019; Vickers & Rees, 2007).

In addition to the abovementioned compact approaches, dimensionality reducing methods, such as principal component analysis (PCA) (see Section 4.3 in Chapter 4) and self-organising map (SOM) (see Section 3.4.2 in Chapter 3), have been commonly (although not universally) integrated into some of the geodemographic classification products that corresponded to the decennially released census data (e.g. Charlton et al., 1985; Openshaw & Wymer, 1995; Robinson, 1998; Webber, 1975; Webber & Craig, 1978). The typical rationale of dimensionality reduction methods is to transform the high-dimensional variables into a low-dimension that is more appropriate for those mathematical algorithms applied in clustering (see Bara et al. 2018; Jolliffe, 1972; Miljkovic, 2017; Jiliang Tang et al., 2014). While the primary justification for dimensionality reduction methods in past studies many decades ago was to reduce the heavy computational burden (Singleton, 2016), such methods still present potential utility in the contemporary realm of building geodemographic classification, even within a computational intensive setting (Adnan, 2011; Liu et al., 2019).

### *VARIABLE PRE-PROCESSING*

After the assembly of selected variables, data preprocessing is typically conducted and involves a series of processes such as data manipulation, normalisation, standardisation, or weighting (Alexiou, 2016; Harris et al., 2005). The selected variables are seldom used in their raw formats. They are usually translated into measures such as percentages and ratios (e.g., density), which is followed by an examination of their various statistical attributes such as normality (kurtosis), variance, collinearity, and so forth. For example, classification builders are likely to use histogram or box plot to examine the distribution of the data, and to identify abnormal values and outliers, or to use a Minimum Spanning Tree (MST) algorithm to check the correlation between variable pairs (Alexiou, 2016; Harris et al., 2005; Liu et al., 2019).

In case the examined variables are not distributed normally, data normalisation techniques are commonly adopted by classification builders, for example, logarithm transformation, inverse hyperbolic sine, or Box-Cox transformation (see Box & Cox, 1964), are often applied with the aim of returning a more normal distribution. Furthermore, it is typically necessary to apply a universal scale of measurement to each of the variables prior to the clustering stage, ensuring that they are all measured in the same unit. Same unit measurement can be achieved by applying standardisation techniques to the dataset, for example, range standardisation, rank standardisation, z-scores, and interquartile range standardisation (Alexiou, 2016; Alexiou & Singleton, 2015; Gale et al., 2016).

### *CLUSTERING ANALYSIS*

After the data preprocessing is accomplished, cluster analysis typically follows to assess the area similarity or dissimilarity in terms of these selected variables, which is typically carried out by applying a clustering algorithm to the prepared dataset.

Cluster analysis as applied to geodemographics belongs to a branch of unsupervised machine learning. Typically, cluster analysis pursues the maximisation of the similarities between observations within the same group and minimises the similarities between different groups (Sinaga & Yang, 2020). Similarity or dissimilarity is usually determined by the "distance" between observations, which can be measured as Euclidean distance, weighted Euclidean distance, cosine distance, Mahalanobis distance, Manhattan distance, and many other measurements (Shirkhorshidi et al., 2015; Troy, 2017).

The choice of clustering algorithm can differ significantly, which is partially determined by the purpose of the classification as well as the nature of the data to be processed (Alexiou, 2016; Alexiou & Singleton, 2015). A geodemographic classification usually consists of a hierarchical series of aggregations that can be assembled either *top-down* or *bottom-up*. In a typical *top-down* clustering mode, the largest grouping is first constructed and is subsequently divided into smaller subgroups. The k-means clustering algorithm (see MacQueen, 1967) is frequently utilised for such purposes. For instance, Gale et al. (2016) have implemented the k-means clustering algorithm on selected variables from the 2011 UK Census data to build the 2011 OAC, classifying all OAs across the UK into eight Supergroups, which were further split into 26 Groups and further into 76 Subgroups. Similarly, Singleton et al. (2016) have created the IUC by employing k-means clustering with *top-down* implementation, generating a nested hierarchy of four Supergroups and 11 Groups classification, profiling the vulnerability of e-resilience for all LSOAs in England. Moreover, the top-down k-means clustering method was also used in the creation of COWZ-UK (Cockings et al., 2020).

Classifications built from the *bottom-up* create the most disaggregate level of the classification first, which are then aggregated successively based on their similarities into larger clusters. Such iteration steps can be repeated until only one cluster remains, or the distance between the two closest clusters above meets a predefined threshold, or until a specific number of clusters is reached. Such clustering process is usually carried out by hierarchical agglomerative clustering such as Ward's clustering algorithm (see Ward, 1963), and is more prevalent in commercial geodemographic systems (Alexiou & Singleton, 2015).

In addition, several other clustering algorithms, either with *top-down* or *bottom-up* implementations, have been used to create geodemographic classifications. For instance, the Self-Organisation Maps (SOM) (see Kohonen, 1998) was employed as an alternative classifier for the creation of the GB Profiles (Openshaw & Wymer, 1995). The implementations of such technique were also found in creating MODUM (Alexiou et al., 2016) and the geodemographic classification for NYC (Spielman & Thill, 2008). A Fuzzy Geographically Weighted Clustering (FGWC) algorithm was used to create a geodemographic classification with explicit consideration of geographical neighbourhood effects (Mason & Jacobson, 2006).

#### *CLUSTER REVISION AND INTERPRETATION*

The final stage of building a geodemographic classification involves an assessment of the clustering results, alongside descriptions of the classifications. This stage can be viewed as the optimisation stage, where the produced clusters are examined. For example, it is usual for the classification builder to check the cluster sizes to avoid creating clusters with too large or small a size. Typically, "pen portrait" descriptions are developed to describe the most salient characteristics of the areas represented by the clusters and accordingly assign shorthand names (Harris et al., 2005).

In summary, several analytical decisions require to be made in the pre, mid, and post stages of constructing a geodemographic classification. As claimed by Singleton et al. (2017), 'building a geodemographic classification is both an art and a science, and typically requires a degree of subjectivity in the exact choices of methods implemented' (p.92). For instance, choices include which variables should be selected to better differentiate between areas; which data standardisation or normalisation method should be adopted in preprocessing the data; which type of aggregation scheme and clustering algorithm to apply; how to set the optimal clustering parameters, such as how many clusters are appropriate to fit the scope of the analysis. In order to answer these questions, the experience of the classification builders is crucial. Classification builders typically rely on other established classifications, exploratory spatial data analysis, or other specific empirical evaluation before creating a classification (Harris et al., 2005).

### 2.4.4 FROM URBAN CONTEXTS TO URBAN DYNAMICS (HUMAN MOBILITY)

Urban dynamics, a multidisciplinary concept that originated from complex systems theory in system engineering, is the other key component in formulating the urban environment (Thériault & Des Rosiers, 2013). Urban dynamics are observed across a range of temporal scales from short-term or real-time to longer-term changes that arguably drive those patterns we observe through urban contexts. Understanding short-term urban dynamics is a well-established research topic in GIS, transportation planning, and behaviour modelling (Cullen, 1972; Huff & Hanson, 1986; Love & Chapin, 1976; Ravenstein, 1885). Such investigations have been framed within the literature through the use of various terminologies, including but not limited to human mobility, human dynamics, and urban human mobility (Kandt & Batty, 2021; Wang et al., 2019; Yuan & Raubal, 2012; Zhao et al., 2016). Generally, urban dynamics pertains to 'how people move in cities' (Zhao et al., 2016, 91).

Most of the early studies in this field were primarily limited to studies using data derived from traditional travel diaries, questionnaires, travel demand and household surveys, including detailed tracking, logging, and analysing of individual's life cycles (Cullen, 1972; Harvey & Taylor, 2000; Love & Chapin, 1976; Maat et al., 2005). Since the conventional data collection approaches suffer from limited spatial and temporal granularity, such datasets may not provide adequate evidence facilitating a comprehensive investigation of the characteristics of the whole urban environment. Additionally, due to a lack of powerful computational tools, it was challenging to simultaneously examine the intricate mobility patterns underlying human trips and behaviours from both integrated spatial and temporal perspectives (Shen & Cheng, 2016; Yuan & Raubal, 2012).

With the exponential advancement of ICTs and the increasingly prevalent application of mobile location-aware sensors, large-scale data collection of travellers' journeys has become technically feasible and economically affordable. Such unprecedented developments have replaced the previous challenges posed by data scarcity and a lack of computational power with a large amount of data containing more detailed and finer spatiotemporal granularity (Kandt & Batty, 2021; Kitchin, 2014b). Multi-source heterogeneous data acquired from deployed sensors, Call Detail Records (CDR), GPS devices, SCAFC transactions, WiFi or RFID access points, and LBSN, reveal unique opportunities for exploring underlying regularities of urban/human mobility, facilitating a deeper understanding of the urban environment and its metabolism (Cattuto et al., 2010; Chen et al., 2016; Chen et al., 2019; Liu & Cheng, 2020; Tang et al., 2015; Wang et al., 2019; Wang et al., 2015; Zhang et al., 2012). Notably, human mobility studies play a crucial role in addressing real-world challenges with a wide range of applications, such as urban transit planning, environmental protection, security, location-based services (LBS), migration studies, tourism, and epidemic control  (Badr et al., 2020; Bernard et al., 2014; Domínguez-Mujica et al., 2011; Hsieh et al., 2015; Kang et al., 2012; Liu & Cheng, 2020; Wang et al., 2017; Xia et al., 2016; Xiong et al., 2020).

According to Lew & McKercher (2006), the spatiotemporal distribution of human dynamics in the urban environment is uneven, meaning that some of the ‘popular’ urban areas exhibit much more crowded and denser human interactions than other places. As discussed in previous sections, existing studies have verified the close association between urban-scale mobility and urban contexts, including people’s sociodemographic attributes and urban spatial constraints (Ewing & Cervero, 2001; Ewing & Cervero, 2010). Explanations of such association can be twofold. First, it may suggest that different city areas preserve different inhabitants’ mobility patterns, for example, people who live proximal to the central business districts may be more active than those living in suburban areas in terms of travel intensity (Gordon et al., 1989; Yuan & Raubal, 2012). For instance, Chen et al. (2019) have used the hourly-aggregated intensity of cell phone usage to assess the regional travel demand in each of Beijing’s traffic analysis zones (TAZs). They found that travel demands vary between different urban districts. Workplaces and residential districts have more demands for travel than other districts. Similar aggregated mobility patterns have also been identified in the research conducted by Liu et al. (2020). Through applying clustering techniques to the hourly-aggregated subway passenger flow derived from NYC’s subway turnstile machines, Liu et al. (2020) have pointed out that subway stations situated on the outskirts of NYC are more likely to express a “double-humped” home-oriented mobility pattern, characterised by high inbound and low outbound passenger flow during the morning peak with the reverse during the evening peak. Secondly, as Wang et al. (2019) have stated, the association may also indicate that people who frequently co-occur in proximity may have some similarities in their respective sociodemographic attributes or mobility patterns. For instance, a study has found that well-off neighbourhoods attract people residing in areas of various levels of deprivation, whereas deprived areas are more likely to attract people who live in other deprived areas, indicating a social segregation effect (Lathia et al., 2012). Recently, Zhang et al. (2020) have posited the argument “you are how you travel” and attempted to predict people’s sociodemographic conditions by modelling urban human mobility patterns extracted from London Oyster Card transactions.

# 3.0 CONSIDERING CONTEXT AND DYNAMICS: A CLASSIFICATION OF TRANSIT-ORIENTED DEVELOPMENT FOR NEW YORK CITY

The research presented in this chapter is an adapted version of the publication:

- Liu, Y., Singleton, A., Arribas-Bel, D. (2020) Consideration context and dynamics: a classification of transit-orientated development for New York City. *Journal of Transport Geography*. 85. 102711. https://doi.org/10.1016/j.jtrangeo.2020.102711

While the article details the study's other innovations, its main contributions to this thesis are summarised below.

1. *Through a systematic literature review (SLR), a theoretical framework used for the initial variable selection for creating mobility-related classification is developed.*
2. *An analytical framework is proposed, integrating dynamic and contextual dimensions for the urban environment and mobility analysis.*

In this chapter, ***objective 2***, ***4***, ***5*** have been fulfilled.

## 3.1 ABSTRACT

Transit-Oriented Development (TOD) is a widely recognised planning strategy for encouraging the use of mass and active transport over other less sustainable modes. Typological approaches to TOD areas can be utilised to either retrospectively or prospectively assist urban planners with evidence-based information on the delivery or monitoring of TOD. However, existing studies aiming to create TOD typologies overwhelmingly concentrate input measures around three dimensions of: density, diversity and design; which might be argued as not effectively capturing a fuller picture of context. Moreover, such emphasis on static attributes overlooks the importance of human mobility patterns that are signatures of the dynamics of cities.

This study proposes a framework to address this research gap by enhancing a conventional TOD typology through the addition of measures detailing the spatiotemporal dynamics of activity at transit stations; implemented for the selected case study area, New York City.

## 3.2 INTRODUCTION

Transport Oriented Development (TOD) is considered as a type of sustainable urban development focusing on encouraging transit ridership through providing high density and mixed-use development within walking distance (e.g. 400–800 m; or 5–10-min walk) of public transport facilities (Thomas et al., 2018). The main objective of TOD advocates delivering a favourable environment consisting of urban forms that are highly compact, of mixed-use, pedestrian- and cycling-friendly, and develop neighbourhoods with the vicinity of public transport hubs (i.e. transit stations). Such influences are commonly within a framework referred to as the 'three-Ds': namely, high *density* in development, *diversity* in land use and good urban *design* (Cervero & Kockelman, 1997). This development pattern has been widely recognised and accepted as a leading planning strategy by most planning agencies around the world, exemplified by extensive cases in North American and European cities, China, South Korea and so forth (Staricco & Vitale Brovarone, 2018; Sung & Choi, 2017; van Lierop et al., 2017; Xu et al., 2017).

TOD principally aims to address common urban transportation challenges associated with automobile dependence, such as traffic congestion and parking difficulties, air quality and noise pollution, excessive greenhouse gas emission, public health and wellbeing-related issues (Chavez-Baeza & Sheinbaum-Pardo, 2014; Ettema et al., 2016; Hickman & Banister, 2014; Hynes, 2017; Rodrigue, 2020; She et al., 2017). Although urban planners have adopted a series of actions aimed at reducing the dependence of private automobile use through encouraging more sustainable alternatives including public transit and active travels (i.e. walking and cycling) (Lee et al., 2013; Winters et al., 2017), TOD presents a focus for more comprehensive planning solutions since it effectively integrates both urban land use and transport system planning (Lee et al., 2013; Papa et al., 2018; Taki et al., 2017).

Although TOD can be argued as consistent in its prescriptions for policy-making and planning, extensive studies have illustrated that for TOD to be successful, there is a necessity to be highly sensitive to local specificities. For the purposes of assisting urban planners in establishing new TOD or evaluating existing TOD, context-based TOD typologies have been implemented to differentiate various station catchment areas (Higgins & Kanaroglou, 2016; Kamruzzaman et al., 2014; Lyu et al., 2016; Papa et al., 2018). Existing studies have overwhelmingly differentiated TOD through measures related to the 'three-Ds' such as land use mix, residential and commercial density, and floor area ratio. However, other aspects of context, such as socioeconomic variables are neglected (Higgins & Kanaroglou, 2016). Moreover, such static attributes overlook the dynamic context of TODs, namely, human's mobility, which as others have shown also plays a vital role in the evolution of urban morphologies and functional regions (Wang et al., 2017; Xia et al., 2018).

It is within this context that we expand upon the existing literature to consider a more comprehensive definition of TOD through a broader range of multidimensional inputs, including their dynamic context. A new analytical framework is implemented here for the case study city of New York City, USA. The paper proceeds first to present a literature review of approaches used to build a TOD typology, followed by a Systematic Literature Review (SLR) designed to identify variables commonly considered as important drivers of differentiation between TOD contexts. General information about the case study area is presented in Section 3.4, followed by a discussion of the range of station catchment areas and a specification of data pre-processing of the 64 selected candidate variables. In Section 3.4.2, 472 subway stations are categorised into a four-category TOD typology through the implementation of a proposed methodology framework based on Self-Organising Map (SOM). The groups are named and described according to their salient characteristics. In Section 3.5, subway turnstile data are utilised to capture human mobility patterns, using the same framework to create Temporal Clusters featuring five featured travel patterns. In Section 3.6, the two produced clusters are integrated to explore the interaction between static and dynamic features of the TOD areas. Finally, the paper concludes with a discussion suggesting some future work and limitations to the approach.

## 3.3 LITERATURE REVIEW

There are multiple approaches to building a TOD typology, ranging from the qualitative ascription of idealised TOD contexts (Higgins & Kanaroglou, 2016; Lyu et al., 2016) to more quantitative frameworks utilising models of TOD catchments and associated measures drawn for within these areas (Higgins & Kanaroglou, 2016). TOD contexts within the urban environment have been characterised in the literature through various indicators and variables that are argued to have an effect on (or be a result of) the use of public transport. Given the variable definition of TOD extents, study objectives and locations, the specificity of criteria and indicators selected as influential to TOD characteristics vary between studies.

Following the development of TOD-related research, the concepts of the original 'three Ds' model established by Cervero & Kockelman (1997) has been expanded. For instance, Ewing and Cervero (2010) added Destination accessibility, Distance to transit, and an additional non-environmental variable, i.e. Demographics, to the family of 'D variables', formulating the 'five Ds' concept. These concepts were utilised within a Systematic Literature Review (SLR) to identify those TOD related measures used in the recent literature. We utilised the Scopus[1], Google Scholar[2], and Web of Science[3] referencing databases and looked for references published between 2009 and 2018. These databases were queried for research in the broadest sense, including journal articles, official documents, guidelines, and so forth; which either created a TOD typology (or indexed TOD features) for major transit stations or focused on analysing the relationship between multidimensional variables around stations and ridership of public transit more generally. Scopus returned 15 studies, the Web of Science and Google Scholar respectively identified 11 and 6240 results. The studies were checked for diversity, both in terms of geographic context (i.e. the location of the case study) and type (e.g. journal article, governmental documents/policies); and secondly, the quality of the reviewed studies was considered in terms of the influence of the academic studies (measured by the times cited) and the authority of the governmental documents/policies.

Through this process, 29 studies were identified, and from these, a set of common variables selected that are presented in *Table 3.1*. Although many align with the 'five Ds', most of the studies are not comprehensive in coverage of all domains of the 'five

[1] https://www.scopus.com/
[2] https://scholar.google.co.uk/
[3] https://wok.mimas.ac.uk/

Ds’. Moreover, some of the variables employed in these studies do not align with the ‘five Ds’, implying broader or context-specific considerations. The candidate variables could broadly be categorised into four domains, namely, Land Use and Built Environment, Location and Accessibility, Socioeconomic and Demographic, and Transit-related.

Input to the Land Use and Built Environment, and Location and Accessibility domains were drawn from a range of sources including the Census and other public survey data, but also Points of Interest (POI) databases either as supplements or alternatives to conventional land-use measures (see Lyu et al., 2016; Wang et al., 2017; Wang et al., 2016). These studies advocated that POI data may capture finer-grained and more up to date information depicting the land use composition and urban facilities. Moreover, other variables, such as the type of dwelling, type of tenure, building height, building age, and average travel time/distance to workplace/transit station, also take a relatively large share of commonly-used variables in these two domains; which may be as a result of their reasonably common availability and broadly understood definitions.

Within the Socioeconomic and Demographic domain, typical variables identified from the literature included the median household income, household vehicle ownership, educational attainment, and occupation type. As for demographic variables, the “seventh D”[1] in D-variables (Ewing & Cervero, 2010, 267), mainly including age composition and household size/type.

Transit-related attributes had high salience in the studies identified; and indicators included measures such as daily/weekly ridership, frequency of metro services, or peak passenger load/frequency in the transit station. Although of utility, such measurements were typically limited in temporal resolution (i.e., weekly ridership) or were somewhat static (e.g., morning peak ridership volume), and therefore only had limited account for actual periodic variation in patterns of use. Given that spatiotemporal data related to transit have become more prevalent, some of the studies, such as Qin et al. (2017), Wang et al. (2017), Wang et al. (2016) and Kim (2018), utilised attributes from trip transaction data extracted from a smart card system to calibrate more real-time measures. In addition to transit flow data, human mobility was also inferred by Wang et al. (2017) through mobile application data from an online mapping system.

Within the Location and Accessibility domain, travel distance/time from the transit nodes to the main working places are typically adopted by many of the reviewed studies.

[1] The “sixth D” in D-variables is considered as demand management (Ewing & Cervero, 2010, 267)

Moreover, proximity to activities is also a commonly used variable indicating the connectivity between transit nodes and the surrounding environment. Perceived attributes are also employed by some reviewed studies, such as cleanness and safety of the transit station. However, due to the difficulty of quantifying and data availability, most of the reviewed studies do not include these types of variables.

| **Domain** | **TOD Indicators used in the reviewed studies** | Atkinson-Palombo & Kuby (2011) | Austin et al. (2010) | Bhattacharjee & Goetz (2016) | CTOD (2013) | Chen et al. (2009) | Chorus & Bertolini (2011) | Dirgahayani & Choerunnisa (2018) | Guo et al. (2018) | Higgins & Kanaroglou (2016) | Huang et al. (2018) | Ivan et al. (2012) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Land Use & Built Environment | Average Block Size/ Length | | * | | * | | | | | | | |
| | Degree of Functional Mix | | | | | | * | * | | | | |
| | Housing Unit /Density | | | | | | | | | | | |
| | Land Cover | | | | | | | | | | | |
| | Mixed-ness of Land Use (Diversity/Entropy) | * | | * | | * | | * | * | * | * | |
| | Population /Residents Density | * | * | * | * | * | * | | * | * | * | * |
| | Job Density /Business Intensity | * | * | * | | * | | | | * | * | |
| | Floor Area Ratio (FAR) | | | | | | | * | | | | |
| | Property/Land Values | | | | * | | | | | * | * | * |
| | Street Network/Intersection Density | | | | | | | | | * | * | * |
| | Type of Dwelling/Tenure | | | | | | | | | | | * |
| | Year Structure Built (Building Age) | | | | | | | | | | | |
| Transit-related | Attributes of Transit Stations | | | | | | * | | | | | |
| | Frequency of Metro Services | | | | | | | | | | | |
| | Interchange to Other Transit Modes | | | | | | | | * | | | * |
| | Number of Directions Served (Bus/Subway) | | | | | | * | | * | | | * |
| | Number of Nearby Transit Hubs | | | | | * | * | * | * | | | * |
| | Parking Facility/Infrastructure | * | | | | | | | | | | * |
| | Utilisation of Transit (Passenger Load/Ridership) | | | | | * | | | * | | * | * |
| | Walkability/Pedestrian Networks/Cyclability | | | | * | | | * | | * | * | |
| Location & Accessibility | Accessibility to/ from Station | | * | | | | | * | | * | | |
| | Average Travel Time (to Work/Transit Stations) | | * | | | | | | | * | | |
| | Distance to City Centre/CBD | | * | | * | * | * | | | | | |
| | Perceived Attributes (e.g. Safety, Attractiveness) | | | | * | | | * | | | | |
| | Proximity of Activities/Amenities at Station | | | | * | | * | * | | * | | |
| Socioeconomic & Demographic | Ethnic/Age Composition | | | | * | * | | | | * | | |
| | Household Income | * | * | | | * | | | | * | | |
| | Household Type/Size | * | * | | | | | | | | | |
| | Occupation Type/Education Level | * | | * | | | * | | | * | | * |
| | Transport Mode to Work | | * | | | | | | | * | | |
| | Vehicles Ownership | | * | | * | | | | | | | |
| Case Study Area | | Phoenix US | 9 Cases US | Denver US | Allegheny County US | New York City US | Tokyo Japan | Jakarta & Bandung Indonesia | Tokyo Japan | Toronto region Canada | Arnhem–Nijmegen MA.<br>Netherlands | Ostrava Czech Republic |
| Number of Transit Stations (Cases) | | 27 | 9 | NA | NA | 468 | 99 | NA | 27 | 372 | 22 | 11 |
| Buffer Distance (metres) | | 800 | 800 | 800 | NA | NA | 700 | NA | 1500 | 800 | 800 | 700 |

| Jun et al. (2015) | Kamruzzaman et al. (2014) | Kim et al. (2017) | Kim et al. (2018) | Lee et al. (2013) | Lyu et al. (2016) | Monajem & Ekram Nosratian (2015) | Nasri & Zhang (2014) | Papa et al. (2018) | Pollack et al. (2014) | Singh et al. (2017) | Sohn (2013) | Song & Deguchi (2013) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | * | | | | * | | * | | | | | |
| | * | | | | * | * | | * | | | | |
| * | * | * | * | | | * | | | * | | | |
| | | | * | | | | | | | | | |
| * | * | * | * | * | * | | * | | | * | * | * |
| * | * | * | * | | * | * | * | * | | * | * | * |
| * | | * | * | | * | * | * | * | * | * | * | |
| | | * | | | | | | | | | | * |
| | | | | | | | | * | | | * | |
| * | | | * | | * | * | | | | * | | |
| * | * | | | | | | | | * | | | * |
| * | | | * | | | | | | | | * | |
| | | | | | * | | | | | | * | |
| | | | | | | * | | | | | | * |
| | * | | | * | | | | | | * | * | * |
| | | | | | * | | | | | | | * |
| * | | | * | * | | * | * | | | | * | * |
| | | | | | * | | | | | * | | |
| | | * | * | * | | * | | | | * | * | * |
| | * | | | | * | * | | | * | * | * | |
| | | | | * | * | | * | | | * | | |
| | * | | | | * | | | | | * | | |
| | * | | | | | * | * | | | | * | |
| | | | | | | | | | | * | | |
| | | | | | * | | | * | | * | * | |
| * | * | * | | | | | | | | | | * |
| * | * | * | | | | | * | | * | | | |
| * | | * | * | | | | * | | | | | |
| * | | * | * | | * | * | | | | * | | |
| | | | | | | | * | | * | | | |
| | * | | | | | | * | | * | | | |
| Seoul MA.<br>South Korea | Brisbane<br>Australia | Seoul MA.<br>South Korea | Seoul MA.<br>South Korea | Seoul MA.<br>South Korea | Beijing<br>China | Tehran<br>Iran | Washington D.C.<br>Baltimore MA. | Naples<br>Italy | Boston<br>US | Arnhem–Nijmegen MA.<br>Netherlands | Seoul MA.<br>South Korea | Tokyo<br>Japan |
| 442 | 1734 CCDs | 479 | 479 | 284 | 268 | 5 | 5 | 62 | 345 | 21 | 479 | 152 |
| 300,600,900 | 800 | 500 | 500 | 500 | 700 | 700 + 100 | 800 | 500 | 800 | 800 | 500 | 600, 1000 |

| Vale (2015) | Wang et al. (2016) | Wang et al. (2017) | Zemp et al. (2011) | Zhou et al. (2017) |
|---|---|---|---|---|
| * | | | | |
| | | | | |
| | | | | |
| | * | * | | * |
| * | | | * | |
| | | | * | |
| | | | | |
| | | | | |
| | | | | |
| | | | | |
| | | | | |
| * | | | * | |
| * | | | * | |
| * | | | * | |
| * | | | * | |
| * | * | | * | |
| * | * | | | |
| * | * | * | * | * |
| | | | * | |
| | * | | * | |
| | * | | * | |
| | * | | | |
| | | | * | |
| * | | | | |
| | | | | |
| | | | | |
| | | | | |
| * | | | | |
| Lisbon Portugal | Beijing China | Shanghai China | Switzerland | Wuhan China |
| 83<br>700 | 215<br>770 3800 | 588<br>500 | 1700<br>700 | 96<br>500 |

*Table 3.1 Variables Checklists from the Systematic Literature Review*

## 3.4 CONTEXTUALISING TOD: NEW YORK CITY

The case study area selected for this study is the New York City (NYC), which is the most densely populated cities within the US with an estimated 8.56 million residents distributed over a land area about 777 km$^2$ (US Census Bureau, 2019). The city is located at the southern tip of the state of New York on the US eastern seaboard, comprising five boroughs, namely, Brooklyn, Queens, Manhattan, Bronx, and Staten Island. The New York City Subway, first opened in 1904, is a rapid transit system that offers 24/7 service across four of the five boroughs of NYC (i.e. Manhattan, Queens, The Bronx, and Brooklyn), which is controlled by the Metropolitan Transportation Authority (MTA). The system spans 27 lines (665 miles of track) and 472 subway stations, facilitating a major transportation mode for residents and visitors to the city (MTA, 2016). According to the subway ridership statistics provided by MTA, in 2016, an average of 5.65 million passengers used the system daily on weekdays and about 5.75 million at the weekends, making it the largest rapid transit system in the US and the seventh busiest worldwide.

### 3.4.1 DEFINING STATION CATCHMENT AREA AND DATA PRE-PROCESSING

Fundamental to any TOD typology is a definition of the contextual area surrounding the transit stations. For this case study we selected an area of 800 m (approximately 0.5 miles) which mirrored the majority of studies conducted within the US (see Atkinson-Palombo & Kuby, 2011; Austin et al., 2010; Bhattacharjee & Goetz, 2016; Nasri & Zhang, 2014). Although most of these studies employed a Euclidean distance buffer (or circular buffer) to define the catchments of transit station areas, it can be argued that a network distance buffer is more suitable since it “more accurately representing the built environment as experienced by someone walking through it” (Oliver et al., 2007, 8). *Figure 3.1* illustrates 50 m-trimmed street network-based catchment areas (800 m walking distance) of the NYC subway stations. A zoomed-in inset map, on the upper left corner, shows an example circular buffer and a network buffer at Metropolitan Avenue Station. It is clear that a circular buffer area is less effective representation given the surrounding street density and available paths to walk. Additionally, in some other locations within a circular buffer, the walking distance is longer than 800 m due to more facilitating urban structure, such as block size. A more systematic discussion comparing the influence of these two types of buffer are detailed in Oliver et al. (2007).

Moreover, two inset maps located on the right side of *Figure 3.1* highlight a stretch of census blocks[1] along with their defined station catchment areas. Each catchment contains census blocks, for example, the catchment area of Mets-Willets Point station is formed by seven census blocks; Lefferts Blvd station catchment intersects with 25 census blocks, which are converted into proportion based on their area of overlap. The proportion is subsequently used as a weight ($w_i$) to calculate the weighted average value of selected variables. *Equation 3.1* illustrates how these weights were used to calculate values attributed to census block where they intersected with the catchment areas.

$$\bar{x} = \frac{\sum_{i=1}^{n}(x_i * w_i)}{\sum_{i=1}^{n} w_i} \tag{3.1}$$

***Equation 3.1** $\bar{x}$ **is the weighted mean;** $x_i$ **is an original value;** $w_i$ **is the weight (i.e. the proportion of the area occupied by a specific census block in the station catchment area).***

For other variables at the finer spatial resolution, particularly spatial points extracted from the NYCOD and NYCP (e.g. street trees, bus stops), these were first aggregated to the station catchment areas where they were located and either a density or percentage value calculated.

[1] Census blocks, the smallest geographic area for which the US Census Bureau collects and tabulates census data (US Census Bureau, 2019).

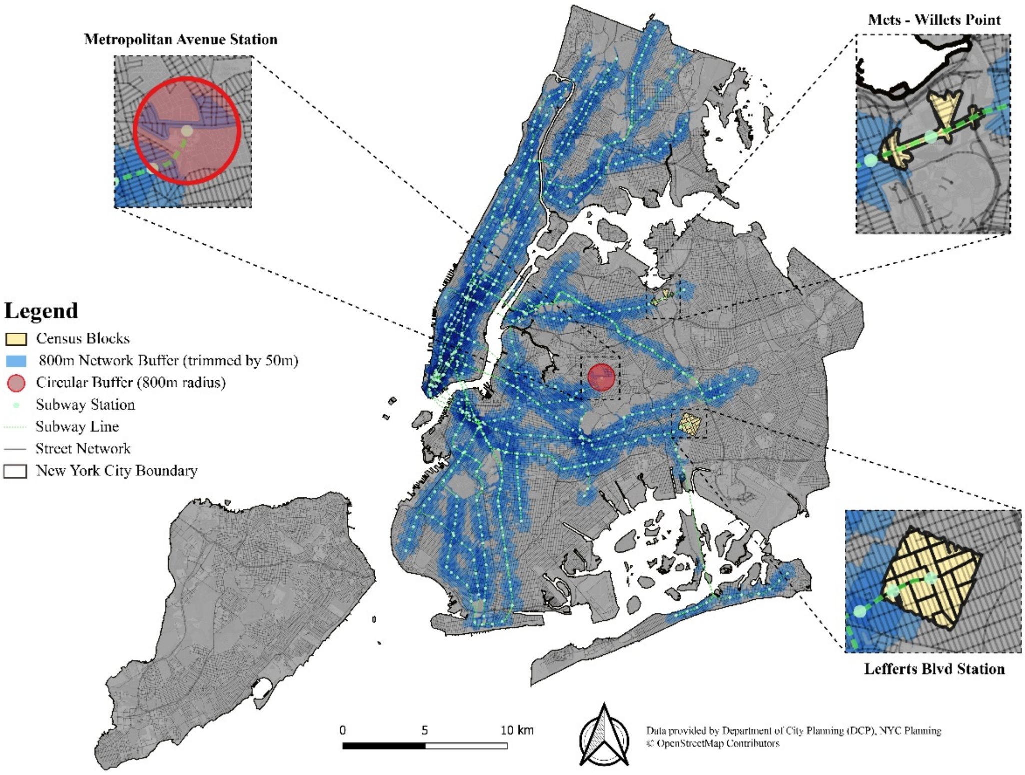

Metropolitan Avenue Station
Mets - Willets Point
Lefferts Blvd Station
Legend
Census Blocks
800m Network Buffer (trimmed by 50m)
Circular Buffer (800m radius)
Subway Station
Subway Line
Street Network
New York City Boundary
0
5
10 km
Data provided by Department of City Planning (DCP), NYC Planning
© OpenStreetMap Contributors

***Figure 3.1 The New York City Subway System and Catchment Areas (800m Walking Distance) With Highlights of Census Blocks***

The selection of variables was primarily guided by findings from *Section 3.3* alongside further consideration of the quality and availability of potential variables within the case study area. As such, variables selected for this study were categorised into the four previously identified domains: Land Use and Built Environment, Transit-related, Location and Accessibility, and Socioeconomic and Demographic. Variables representative of these domains were extracted from the following seven open data sources: American Community Survey (ACS)[1], National Walkability Index (NWI)[2], Smart Location Database (SLD)[3], NYC Open Data (NYCOD)[4], NYC Planning (NYCP)[5], Metropolitan Transportation Authority (MTA)[6]. *Table 3.2* presents the final 64 variables selected for this study alongside a brief description. After the selected variables were assembled for each of the subway station catchment, the Box-Cox transformation (*Equation 3.2*; Box & Cox, 1964) was adopted to transform non-normal

[1] https://www.census.gov/programs-surveys/acs/
[2] https://catalog.data.gov/dataset/walkability-index
[3] https://www.epa.gov/smartgrowth/smart-location-mapping
[4] https://opendata.cityofnewyork.us/
[5] https://www1.nyc.gov/site/planning/data-maps/open-data.page
[6] http://web.mta.info/developers/turnstile.html

variables values to approximate a normal distribution. Furthermore, given that the assembled variables are measured on different scales, z-scores were implemented as a standardisation (*Equation 3.3*). This frequently used technique creates a transformed variable with a mean of zero and unit of standard deviation.

$$\boldsymbol{x_i^{'}} = \begin{cases} \dfrac{\boldsymbol{x_i^{\lambda} - 1}}{\boldsymbol{\lambda}} & , \quad \boldsymbol{if\ \lambda \neq 0}; \\ \boldsymbol{\log x_i} & , \quad \boldsymbol{if\ \lambda = 0}. \end{cases} \tag{3.2}$$

***Equation 3.2 where $x_i'$ is the transformed value; λ ranging from -5 to 5, which can be estimated using the profile likelihood function to achieve 'optimal value.'***

$$\boldsymbol{z_i = \frac{x_i - \mu}{\sigma}} \tag{3.3}$$

***Equation 3.3 where $z_i$ is the standardised value , $x_i$is an original value, μ is the mean of $x_i$, and σ  is the standard deviation from the mean.***

| Database | Code | Domain | Variables Title | Description |
|---|---|---|---|---|
| ACS | B01001 | Socioeconomic & Demographic | Age: 0-4 | % of Population Aged between 0 and 4 |
| | | Socioeconomic & Demographic | Age: 5-14 | % of Population Aged between 5 and 14 |
| | | Socioeconomic & Demographic | Age: 15-19 | % of Population Aged between 15 and 19 |
| | | Socioeconomic & Demographic | Age: 20-24 | % of Population Aged between 20 and 24 |
| | | Socioeconomic & Demographic | Age: 25-44 | % of Population Aged between 25 and 44 |
| | | Socioeconomic & Demographic | Age: 45-64 | % of Population Aged between 45 and 64 |
| | | Socioeconomic & Demographic | Age: 65&above | % of Population Aged 65 and above |
| | B08303 | Location & Accessibility | TTtW: < 5 mins | % of Workers whose Travel Time to Work is less than 5 minutes |
| | | Location & Accessibility | TTtW: 5-14 mins | % of Workers whose Travel Time to Work is between 5 and 14 minutes |
| | | Location & Accessibility | TTtW: 15-29 mins | % of Workers whose Travel Time to Work is between 15 and 29 minutes |
| | | Location & Accessibility | TTtW: 30-44 mins | % of Workers whose Travel Time to Work is between 30 and 44 minutes |
| | | Location & Accessibility | TTtW: 45-59 mins | % of Workers whose Travel Time to Work is between 45 and 59 minutes |
| | | Location & Accessibility | TTtW: > 60 mins | % of Workers whose Travel Time to Work is longer than 60 minutes |
| | B11016 | Socioeconomic & Demographic | HT: 1-person | % of 1-Person Household |
| | | Socioeconomic & Demographic | HT: 2-person | % of 2-Person Household |
| | | Socioeconomic & Demographic | HT: 3-person | % of 3-Person Household |
| | | Socioeconomic & Demographic | HT: 4+-person | % of 4 or more Person Household |
| | B15003 | Socioeconomic & Demographic | EA: No school | % of Population have no qualifications |
| | | Socioeconomic & Demographic | EA: Elementary school | % of Population attained kindergarten to 5th grade |
| | | Socioeconomic & Demographic | EA: Middle school | % of Population attained 6th to 8th grade |
| | | Socioeconomic & Demographic | EA: High school | % of Population attained 9th to 12th grade |
| | | Socioeconomic & Demographic | EA: College / Bachelor | % of Population attained College or Bachelor’s degree |
| | | Socioeconomic & Demographic | EA: Master / Doctorate | % of Population attained Master’s or Doctorate Degree |
| | B19013 | Socioeconomic & Demographic | Median Income | Household Median Income in the past 12 months |
| | B25003 | Land Use & Built Environment | Tenure: Owner | % of Housing Unit occupied by Owner |
| | | Land Use & Built Environment | Tenure: Renter | % of Housing Unit occupied by Renter |
| | B25024 | Land Use & Built Environment | US: Detached | % of Housing Unit categorised as detached |
| | | Land Use & Built Environment | US: Attached | % of Housing Unit categorised as attached |
| | | Land Use & Built Environment | US: Apartment | % of Housing Unit categorised as apartment (from 2 to 50 units) |
| | B25034 | Land Use & Built Environment | YB: 2010 / Later | % of Building built in 2010 or later |
| | | Land Use & Built Environment | YB: 2000 - 2009 | % of Building built in between 2000 and 2009 |
| | | Land Use & Built Environment | YB: 1980 - 1999 | % of Building built in between 1989 and 1999 |
| | | Land Use & Built Environment | YB: 1960 - 1979 | % of Building built in between 1960 and 1979 |
| | | Land Use & Built Environment | YB: 1940 - 1959 | % of Building built in between 1940 and 1959 |
| | | Land Use & Built Environment | YB: 1939 / Earlier | % of Building built in 1939 or earlier |
| | B24010 | Socioeconomic & Demographic | OT: M.B.S.A. | % of Workers in Management, Business, Science, and Art Occupations |
| | | Socioeconomic & Demographic | OT: S. | % of Workers in Service occupations |
| | | Socioeconomic & Demographic | OT: S.O. | % of Workers in Sales and office occupations |
| | | Socioeconomic & Demographic | OT: N.C.M. | % of Workers in Natural resources, construction, and maintenance occupations |

| | | | | |
|---|---|---|---|---|
| | | Socioeconomic & Demographic | OT: P.T.M. | % of Workers in Production, transportation, and material moving occupations |
| | B25044 | Socioeconomic & Demographic | VA: No-vehicle | % of Housing Units have no vehicle |
| | | Socioeconomic & Demographic | VA: 1-vehicle | % of Housing Units have 1 vehicle |
| | | Socioeconomic & Demographic | VA: 2-vehicle | % of Housing Units have 2 vehicles |
| | | Socioeconomic & Demographic | VA: 3+-vehicle | % of Housing Units have 3 or more vehicles |
| | B01003 | Land Use & Built Environment | Population Density | Population Density |
| NWI | D4a | Location & Accessibility | D4a | Distance from the population-weighted centroid to the nearest transit stop (meters) |
| SLD | D1c | Land Use & Built Environment | D1c | Job Density |
| | D2a_EpHHm | Land Use & Built Environment | D2a_EpHHm | Employment and Household Entropy |
| | D3a | Land Use & Built Environment | D3a | Road Network Density |
| | D4d | Location & Accessibility | D4d | Aggregate frequency of transit service per square mile |
| NYCOD | CSCL | Land Use & Built Environment | Intersection Density | Street Intersection Density Calculated from the Street Centreline |
| | STC | Land Use & Built Environment | Tree Density | Street tree density |
| | Bicycle | Land Use & Built Environment | Bike Facilities | Citi-Bike, Bicycle Routes and Parking Shelters density |
| | Bus | Land Use & Built Environment | Bus Facilities | Bus Stops Density |
| | Parking | Land Use & Built Environment | Parking Facilities | Parking meters/lots density |
| | POI | Land Use & Built Environment | POI | Point of Interest Data: contains seven land-use types |
| NYCP | MapPLUTO | Land Use & Built Environment | Landuse | Land Use: contains seven land-use types |
| MTA | Turnstile | Transit-related | Turnstile: Entry | Entry Counts for all turnstile data by every 4 hours per day |
| | | Transit-related | Turnstile: Exit | Exit Counts for all turnstile data by every 4 hours per day |

*Table 3.2 Final Variable Selection and Basic Description*

### 3.4.2 CONTEXTUALISING TOD

After the assembly of the normalised and standardised input data; similarity in the context of subway stations was explored by the application of a Self-Organising Map (SOM). The Self-Organising Map (SOM), also known as Kohonen Map, is a single layer feedforward artificial neuron network, which is trained by unsupervised, competitive learning as a tool for “visualisation and analysis of high dimensional data” (Bação & Lobo, 2010, 4). The SOM translates high-dimensional inputs into a low-dimensional space, also referred to feature map that is configured by the number of pre-defined neurons arranged on a regular lattice (e.g. a rectangular or hexagonal topology), through ‘fitting’ a grid of nodes to the data over a fixed number of iterations. The resulting map allows a graphical presentation of the data that can be easily interpreted by map-readers, which can be further classified by the machine learning techniques designed for low dimensionality (Bara et al., 2018; Natita et al., 2016; Spielman & Folch, 2015). Numerous studies have highlighted the utility of SOM for visualising complex, nonlinear statistical relationships within high-dimensional data (Bação & Lobo, 2010; Das et al., 2016; Miljkovic, 2017; Yin, 2008). The method is suitable for this application given the multiplex of measures assembled. Moreover, even after the application of Box-Cox transformation, some variables remained not normally distributed, which may have caused some problems if we directly adopted conventional feature extraction methods such as principal components analysis (PCA), since the underlying assumptions of these techniques are not satisfied (Das et al., 2016). Accordingly, Demartines & Blayo (1992) note that the SOM is not very sensitive to the normal distribution when the input data contain high dimensionality.

Several studies have highlighted the potential applications of SOM in terms of building typology for urban contexts (Arribas-Bel & Schmidt, 2013; Jain et al., 2018; Schäfer et al., 2018; Spielman & Thill, 2008), and specifically within the context of TOD: Sohn (2013) presented an application of SOM for Seoul, South Korea, metro station areas.

Several parameters need to be specified in advance when fitting a SOM, including the number of neurons (M), the range of the learning rate and its decline pattern (α), the shape/type of the neuron, and the type neighbourhood function (Spielman & Folch, 2015). The first step before training the SOM is to define an appropriate number of neurons that are used to configure the network. A small feature map (i.e. the number of observations far exceeds the number of neurons) results in a generalisation, whereas a large map allows a specific location in geographic space (subway stations, in this study) to be projected to a particular location in the corresponding attribute space, representing specific properties (Spielman & Thill, 2008). A useful 'rule of thumb' (*Equation 3.4*) suggested by Tian et al. (2014), is employed here to determine the number of neurons. Since the 472 stations configure the observations (N), 108 (M) neurons, projected on a 12 by 9 grid, are accordingly generated to structure the SOM. For the remaining parameters, these were set following an objective of maximising SOM quality through minimisation of the average quantisation error (QE) statistic. We tested the value of QE generated by using various combinations of different SOM parameters following (Natita et al., 2016). The results of these experiments are shown in *Table 3.3*, with the combination of a rectangular topology, the bubble neighbourhood function and a linear decline in learning rate (ranging from 1.0 to 0.01) resulting in the smallest average QE (3.41). Thus, this combination of parameters is eventually adopted to train the SOM network for creating spatial clusters.

| Test | Topology/Shape | Learning Rate Type | Neighbourhood Type | Average QE | Learning Rate range |
|---|---|---|---|---|---|
| 1 | Hexagon | Linear | Bubble | 3.51 | 1.0-0.01 |
| 2 | Hexagon | Inverse | Bubble | 3.59 | 1.0-0.01 |
| 3 | Hexagon | Linear | Gaussian | 4.21 | 1.0-0.01 |
| 4 | Hexagon | Inverse | Gaussian | 4.79 | 1.0-0.01 |
| 5 | Rectangle | Linear | Bubble | 3.41 | 1.0-0.01 |
| 6 | Rectangle | Inverse | Bubble | 3.69 | 1.0-0.01 |
| 7 | Rectangle | Linear | Gaussian | 4.18 | 1.0-0.01 |
| 8 | Rectangle | Inverse | Gaussian | 4.75 | 1.0-0.01 |

***Table 3.3 Result of SOM Parameter Settings for Building TOD Typology***

$$\boldsymbol{M \approx 5\sqrt{N}} \tag{3.4}$$

***Equation 3.4 Where M is the number of neurons, which is an integer close to the result of the right-hand side of the equation and N is the number of observations.***

To reduce the complexity of the computed SOM feature maps further, a hybrid hierarchical k-means (H-K-means) algorithm was applied to aggregate the neurons into groups sharing similar attributes (Chen et al., 2005; Kassambara, 2017). The procedure of this algorithm can be summarised into three steps: firstly, agglomerative hierarchical clustering is applied to the input data and generated tree (i.e. dendrogram), which is cut into k number of clusters; secondly, the cluster centroids (i.e. the mean value) are computed for each group; and finally, these cluster centroids are utilised as the initial centres for the k-means algorithm (Kassambara, 2017). To select an appropriate number of clusters, a clustergram was created that demonstrates a weighted mean of the first component of a PCA for each cluster centre across a range of tested k values, where the width of each line represents the number of observations (i.e. neurons in SOM). The detailed rationale of this technique has been discussed elsewhere (see Schonlau, 2002); but generally, the logic is to find the point where the centroids of the clusters are as dissimilar as possible (well-spaced). According to the clustergram shown in *Figure 3.2*, it is easily observed that when the number of clusters reaches four, the difference between cluster centroids (red dots) is maximised (after k = 4, these centroids are getting close to each other; when k = 5, two cluster centroids are nearly overlapped, indicating relatively bad clustering results).

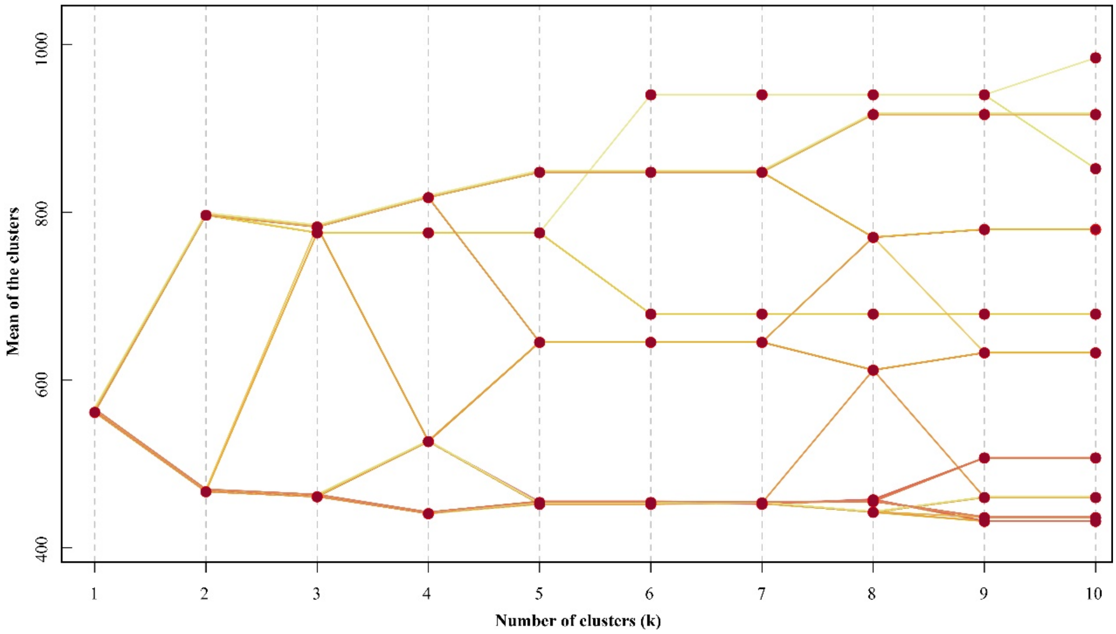


***Figure 3.2 Clustergram for Selecting Number of Clusters Differentiating TOD Typologies.***

The result from clustering the 108 neurons is shown in a 2-dimensional plane (*Figure 3.3*) and is mapped in *Figure 3.4* portraying the geographic distribution of TOD typologies for NYC. The geographic distribution follows a broadly concentric circle-shape, radiating away from the central area of Manhattan.

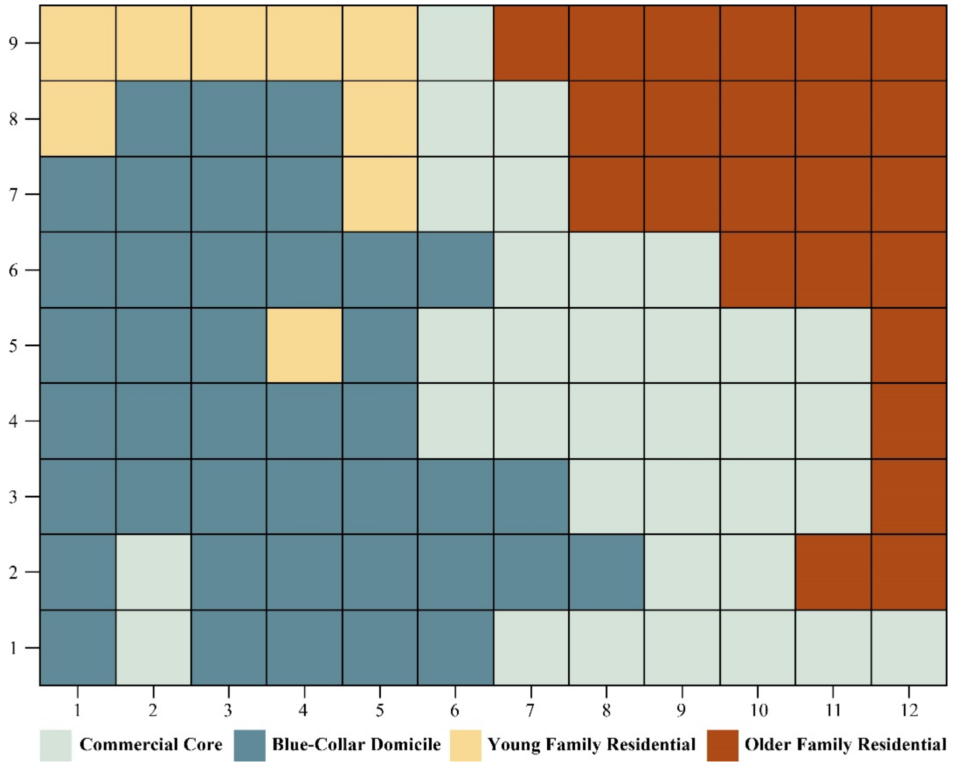


***Figure 3.3 TOD Typologies by SOM Nodes (presented on a 12*9 grid).***

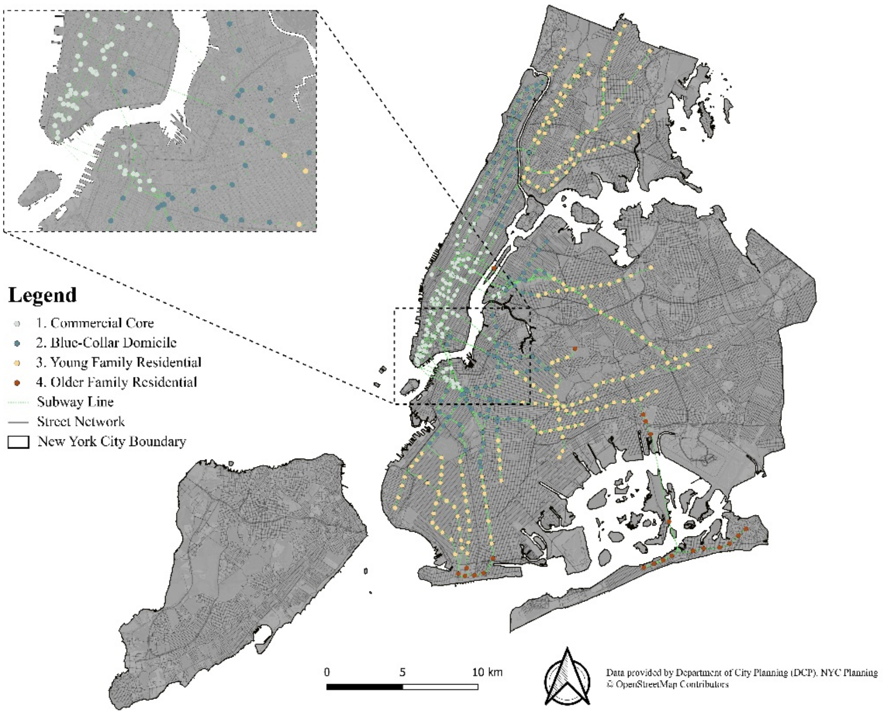

Legend
1. Commercial Core
2. Blue-Collar Domicile
3. Young Family Residential
4. Older Family Residential
Subway Line
Street Network
New York City Boundary
0
5
10 km
Data provided by Department of City Planning (DCP), NYC Planning & OpenStreetMap Contributors

***Figure 3.4 Geographic Distribution of TOD Typologies***

To ascertain the most salient characteristics of the clusters, index scores (i.e. $x/\bar{x}$ *100) were calculated for the input variables and displayed within each cluster in *Figure 3.5*. These scores indicate the (over-) underrepresentation of a target characteristic compared to the regional average value (i.e. a score of 100). An index score 50 would hence equate to a rate that is half the average, and 200 would be double. Utilising both the map and scores, descriptive profiles were generated.

### CLUSTER 1: COMMERCIAL CORE

This cluster is characterised by commercial areas with a highly educated (Master's or Doctorate) population, aged between 25 and 44, including many of those who are employed in well-paid management, business, science and arts occupations. Residents of such areas are more likely to live in apartments (built after the 1980s) consisting of one- or two-person households. These areas are characterised by an extremely high job density, high level of traffic permeability, and plenty of public services, commercial and mixed-use properties, accompanied with a mature infrastructure for cycling and a high level of accessibility of public transit.

### CLUSTER 2: BLUE-COLLAR DOMICILE

Residents of this typology have an age distribution closer to the regional mean, who have increased prevalence to live with family in rented apartments that are situated in areas with high population density, forming a typical three-person household size. Many more of these residents are likely to have occupations within the areas of service and production, transportation, and material moving sectors. Additionally, the annual median income earned by residents of these areas is much lower. The physical environment is characterised by detached properties and apartment constructed in the 1940s and typically linked with adequate parking infrastructures.

### CLUSTER 3: YOUNG FAMILY RESIDENTIAL

These areas are characterised by residential occupants with (pre-)school-age children. Many residents live in the detached property located in boroughs outside Manhattan. Given the distance of travel to work (more than 60 min), the car dependency of these areas is higher than the regional average, also demonstrated by the high household vehicle availability.

## CLUSTER 4: OLDER FAMILY RESIDENTIAL

Populations living within this cluster can be broadly characterised by college-educated middle-age residents (aged between 45 and 64) who are likely to own a detached property built between the 1940s to 1970s and located on the periphery of NYC. Many residents live in relatively large households with dependent children (aged from 5 to 19). Residents of this group show high use of private automobiles for commuting, manifested by high levels of vehicle availability (two or more cars) at more than four times the regional average.

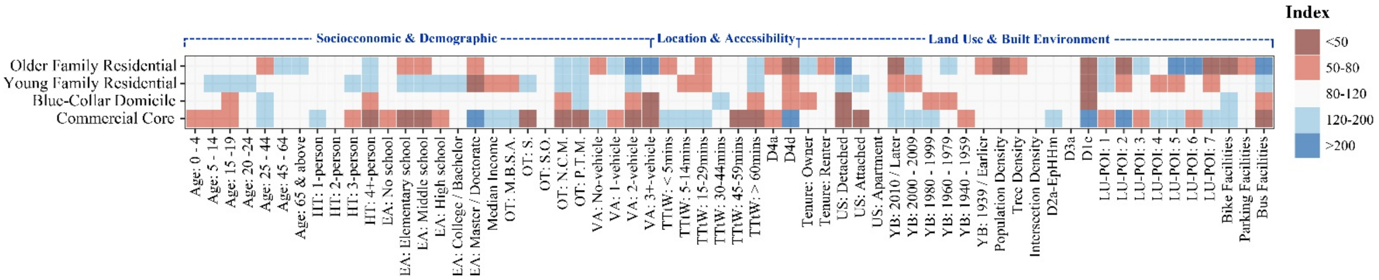


*Figure 3.5 Index Scores by Four TOD Typologies.*

## 3.5 CLASSIFYING TEMPORAL TOD DYNAMICS

The space–time dynamics of TOD localities were considered through subway turnstile data supplied by the MTA. This dataset provides a variety of information on subway station entries and exits, organised into four-hourly daily time bands (i.e. six intervals a day); with the period 2015 to 2016 selected. The turnstile data was aggregated by days of the week, and for each station (a station contains several turnstiles) created 30 variables (six-time bands, five working days) for entry counts and a further 30 variables for exits. To provide further insight into those stations sharing similar patterns of transit use, the analytical framework applied earlier to station contextual data was replicated for the subway turnstile ingress and egress. This included the data pre-processing (e.g. normalisation and standardisation), alongside SOM construction and clustering. A clustergram was again used to select an appropriate number of clusters (see *Figure 3.6*), with k = 5 selected. These 'Temporal Clusters' are mapped in *Figure 3.7*.

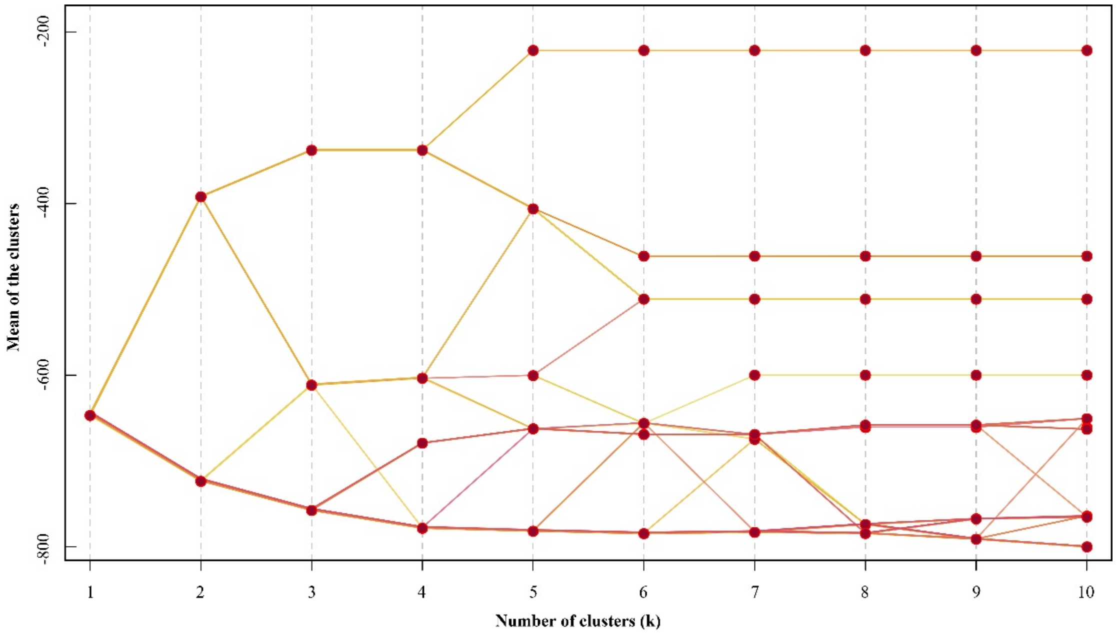


***Figure 3.6 Clustergram for Selecting Number of Clusters Differentiating Temporal Clusters.***

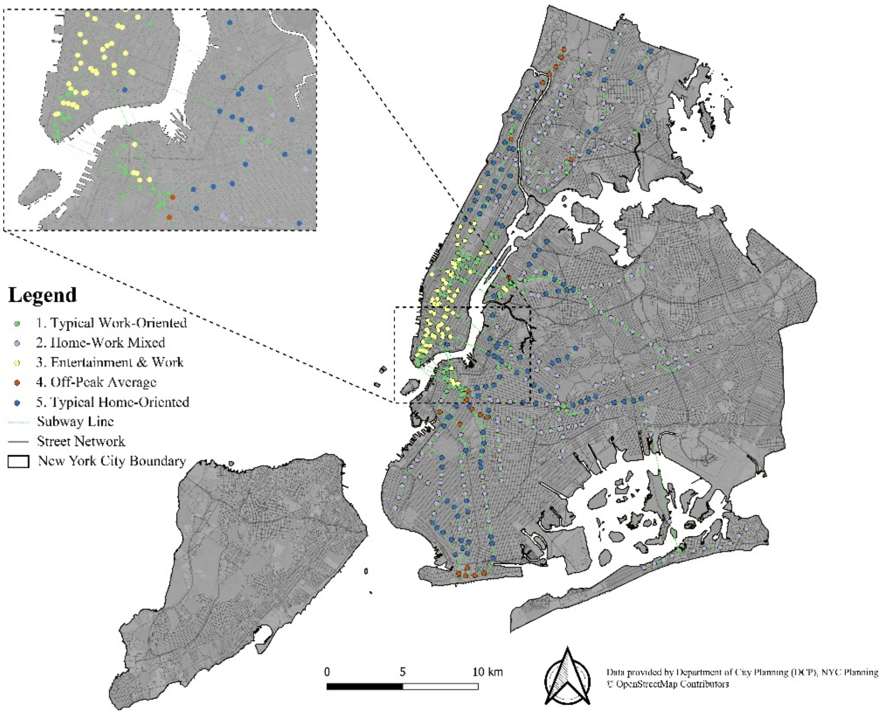


*Figure 3.7 Geographic Distribution of Temporal Clusters*

To ascertain the most salient characteristics of the clusters, a further set of index scores were created for the temporal clusters, using the method previously described in *Section 3.4.2*. A series of heatmaps show these scores in *Figure 3.8* for the five Temporal Clusters. Utilising both the map and scores, descriptive profiles were generated from these insights.

### CLUSTER 1: TYPICAL WORK-ORIENTED

Stations within this cluster are mainly located in the Lower and Midtown areas of Manhattan, downtown areas of Brooklyn and Long Island City. They feature a typical ‘double-humped’ (morning and evening) subway travel pattern associated with workplace-oriented usage. In the morning peak, low inbound passenger flow is identified accompanying with a high outbound flow; while during the evening peak, stations have high inbound flow and a low outbound flow. The role of these stations switches during workdays: from a ‘major destination’ in the morning to ‘major origin’ in the evening.

### CLUSTER 2: HOME-WORK MIXED

Stations classified by this cluster are mainly located outside Manhattan, featuring a mixed subway travel pattern. During the morning peak and even earlier, stations exhibit a high volume of inbound passenger flows and a high volume of outbound flows.

### CLUSTER 3: ENTERTAINMENT AND WORK

Stations within this cluster are predominantly located in either Downtown or Midtown Manhattan, occupying more than half of subway stations in Manhattan. These stations meet a low inbound and high outbound passenger flow during the morning but reverse this pattern during the evening. Moreover, there is additionally a large volume of inbound flows during the midnight-to-late-at-night period, which may e a result of these destination being the popular place of departure from evening events.

### CLUSTER 4: OFF-PEAK AVERAGE

Stations of this group are distributed reasonably randomly across New York. This cluster also consists of stations exhibiting moderate levels of passenger flow, which are very close to the average. More generally, these are less popular stations and experience fewer passengers during commuting peak periods.

### CLUSTER 5: TYPICAL HOME-ORIENTED

Although a fraction of stations from this group can be identified in the northern part of Central Park, most stations are located outside Manhattan (especially in the periphery of NYC). These stations also experience the ‘double-humped’ travel pattern, however, high inbound and low outbound passenger flow during the morning peak, with the reverse during the evening.

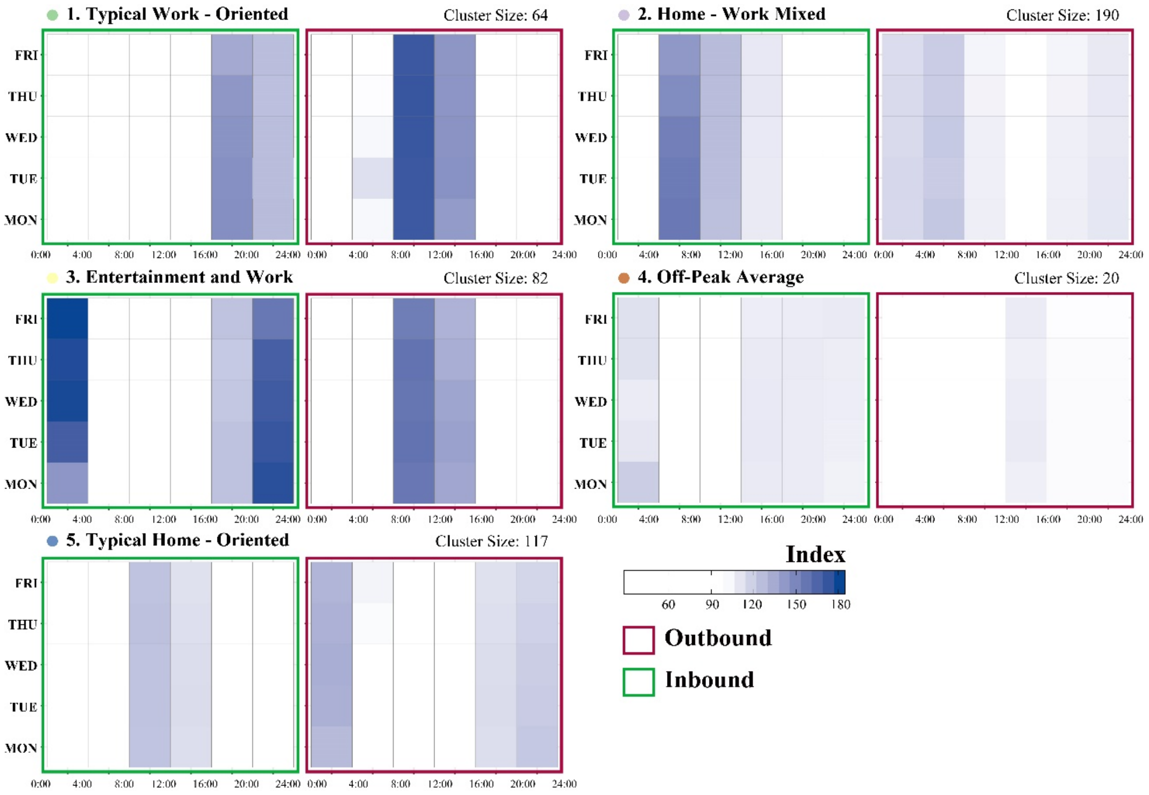


*Figure 3.8 Inbound and Outbound Index Value of Five Temporal Clusters*

(presented in a 'weekly travel profile' manner). In order to achieve better visualisation result, all values less than 100 (less than the mean value) are presented by white.

## 3.6 INTEGRATING CONTEXT AND SPACE-TIME DYNAMICS

An overarching purpose of this work has been to extend an existing framework for the creation of TOD typologies to examine both context and dynamics. As such, in this section, we explore the intersection of our two created classifications. *Figure 3.9* presents an alluvial diagram showing the proportion of subway stations categorised at the intersection of these two classifications for the NYC extent. There is reasonable consistency between these two classifications with some emerging differences.

As might be expected, stations with their context classified as 'Commercial Core' predominantly correspond to 'Typical Work-Oriented' and 'Entertainment & Work' temporal clusters, manifesting typical workplace-oriented function of these TOD areas. Similar temporal patterns can be observed in those stations categorised as 'Blue-Collar Domicile' which splits between 'Typical Home-Oriented' and 'Home-Work Mixed' which might be expected given more residential-oriented usage.

TOD areas categorised as ‘Young Family Residential’ unsurprisingly predominantly correspond with the temporal cluster ‘Typical Home-Oriented’ and ‘Home-Work Mixed’. Stations are both major origins and destinations during peak times, which may be a result of proximity to local employment centres or schools. Given that many of the residences of this cluster are students, the high volume of (early-) morning peak flows may partially be explained by educational establishment opening times.

Within TOD stations classified as ‘Older Family Residential’, there is correspondence to the temporal clusters ‘Home-Work Mixed’ and ‘Off-Peak Average’. There are likely demographic drivers of these patterns alongside a higher rate of private vehicle ownership as a result of their more suburban locations.

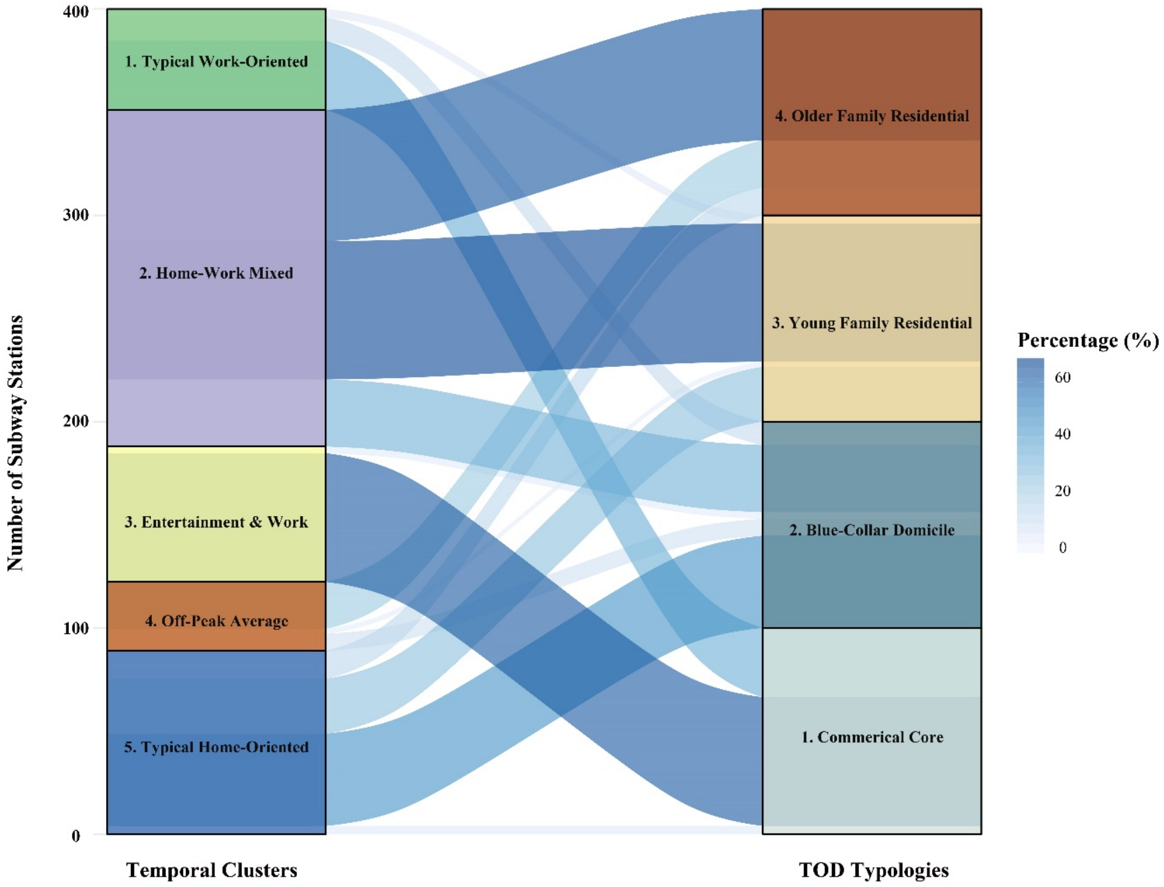


*Figure 3.9 Alluvial Diagram: Percentage Strata by Cross-Tabulating TOD Typologies and Temporal Clusters*

## 3.7 CONCLUSIONS

Transit-Oriented Development (TOD) is a widely recognised planning method for tackling transport-related challenges. Typological approaches to TOD can be utilised either retrospectively or prospectively to assist urban planners with evidence-based information on the delivery or monitoring of TOD. However, most existing studies creating TOD typologies have overwhelmingly relied upon inputs selected alongside the ‘three Ds’ or the ‘five Ds’ principles, which might be argued as not capturing effectively multidimensional aspects of context alongside dynamics of such areas through human mobility. This study proposed and implemented an analytical framework to address this research gap by enhancing a conventional TOD typology with a wider array of contextual data, while also considering the spatiotemporal dynamics of activity at transit stations.

Our presented contextual TOD typology was implemented with candidate data inputs gather through systematically reviewing 29 recent studies related to TOD typology. The ‘five-Ds’ principles were enriched through various measures that were broadly categorised across four domains: Land Use and Built Environment, Transit-related, Location and Accessibility, and Socioeconomic and Demographic. Four salient TOD clusters were generated for the case study city of NYC, by applying a methodology framework formed by the combination of Self-Organising Map (SOM) and hierarchical k-means clustering (H-K-means) to the multidimensional input data. These clusters were further named, described and mapped.

The spatiotemporal dynamics of activity at transit stations was considered through subway turnstile data from the MTA. Through the second application of the proposed framework to the temporal dataset, 472 subway stations were classified into five unique clusters respectively representing different types of travel activity.

The contextual TOD typology was then enhanced through linkage with the classification of aggregate space-time dynamics to illustrate the interaction between context and use. Through cross-validation there was much consistency unveiled, for example, the work-oriented stations are mainly corresponding to the stations located in major employment centres.

One of the main limitations of this study relates to the temporal resolution of the subway turnstile data. The 4-h temporal interval adopted by MTA to aggregate the passenger flows is limited in granularity and may mask valuable details should more disaggregate data be made available. In other contexts, researchers such as Liu and Cheng (2020) and

Mahrsi et al. (2014) have utilised smart card data to conduct the travel pattern analysis, which brings finer resolution for both boarding and alighting information. However, such data or similar products were not publicly available from MTA. Secondly, due to data limitations of this contextual area, this study did not consider multimodal journeys which may also offer insight as the ability to interchange to other transit modes has been spotlighted by many of the reviewed studies (Chorus & Bertolini, 2011; Dirgahayani & Choerunnisa, 2018; Zemp et al., 2011). Although the present study employed variables, such as bus stop density, parking facilities, and bike facilities, to attempt to represent the intermodal connectivity, these variables were relatively ‘static’ compared to the data that could infer mode swapping. None-the-less, despite such caveats, this paper has demonstrated a new and powerful technique that implements an innovative methodology to extend a TOD typology to represent both context and dynamics; and will likely be a useful framework for application within other urban contexts.

# 4.0 A PRINCIPAL COMPONENT ANALYSIS (PCA)-BASED FRAMEWORK FOR AUTOMATED VARIABLE SELECTION IN GEODEMOGRAPHIC CLASSIFICATION

The research presented in this chapter is an adapted version of the publication:

- Liu, Y., Singleton, A. and Arribas-bel, D. (2019) A principal component analysis (PCA)-based framework for automated variable selection in geodemographic classification. *Geo-spatial Information Science.* 22(4). 251-264. https://doi.org/10.1080/10095020.2019.1621549

While the article details the study’s other innovations, its main contributions to this thesis are summarised below.

1. *A PCA-based variable selection methodological framework is developed, aiming to automate the process of selecting the most ‘meaningful variables for creating a geodemographic classification*
2. *Better performance from this framework has been identified by comparing its output to a benchmark geodemographic classification (i.e. 2011 OAC).*

In this chapter, ***objective 3*** has been fulfilled.

## 4.1 ABSTRACT

A geodemographic classification aims to describe the most salient characteristics of a small area zonal geography. However, such representations are influenced by the methodological choices made during their construction. Of particular debate are the choice and specification of input variables, with the objective of identifying inputs that add value but also aim for model parsimony. Within this context, our paper introduces a principal component analysis (PCA)-based automated variable selection methodology that has the objective of identifying candidate inputs to a geodemographic classification from a collection of variables. The proposed methodology is exemplified in the context of variables from the UK 2011 Census, and its output compared to the Office for National Statistics 2011 Output Area Classification (2011 OAC). Through the implementation of the proposed methodology, the quality of the cluster assignment was improved relative to 2011 OAC, manifested by a lower total within-cluster sum of square score. Across the UK, more than 70.2% of the Output Areas (OAs) occupied by

the newly created classification (i.e. AVS-OAC) outperform the 2011 OAC, with particularly strong performance within Scotland and Wales.

## 4.2 INTRODUCTION

A geodemographic classification aims to summarise the multidimensional socioeconomic and built characteristics of small area zonal geography, and are often referred as “neighbourhood” classification (Harris et al., 2005). Geodemographic analysis relates to the application of such classifications and is positioned within a history of analytical frameworks that have aimed to explore the comparative context of urban areas (Bassett & Short, 1980; Timms, 1971). The theoretical tenet of geodemographic classification relates to the principle of homophily, which in geographic terms is the tendency for individuals to be attracted to areas that contain others with similar characteristics to themselves (Sleight, 1993; Webber & Craig, 1978). As such, the methodological objective when creating a geodemographic classification is, therefore, to sort a set of small areas into clusters that share similar characteristics, with the output of such groupings providing a simplified and categorical representation of the overarching multidimensional geography (Spielman & Singleton, 2015).

In general terms geodemographic classifications are created in a series of stages that include the gathering of input variables that describe various characteristics of a given set of small areas; potentially normalising these inputs and then standardising the measures onto the same scale. Due to the high dimensionality of contemporary geodemographics, computational methods are implemented to examine the similarity between areas. This is most commonly achieved through an implementation of cluster analysis which refers to a family of computational methods that will typically have the general goal to maximise within-group similarity and between-group difference through various optimisation strategies (Adnan, 2011; Everitt et al., 2011). Outputs may typically be presented as a hierarchy, with larger and coarser groupings being split into smaller more specific nested groups; with such structure again created through various clustering or partitioning strategies. After this process is complete, it is typical that the characteristics of the assembled clusters are described by looking at which input variables are over- or under-represented within them; and these are then used to build written “pen portraits” and illustrative graphics.

As a methodological approach, geodemographic analysis has a lineage of application across the public and private sectors, spanning multiple decades and geographic contexts

(Bassett & Short, 1980; Paul Longley, 2005; Paul Longley & Goodchild, 2008; Singleton & Spielman, 2014). However, the utility of a geodemographic classification for a given application is substantially determined by those methodological choices made during construction (Openshaw et al., 1995). For example, it may be pertinent to align geodemographic classification inputs to those drivers of a small area differentiation within the context of a particular application (Singleton & Longley, 2009) or, for analysis of specific localities, a classification may be enhanced by considering inputs derived for a focused rather than national extent (Singleton & Longley, 2015). Furthermore, standardisation algorithms (e.g. z-scores, range, and inter-decile range) can have various impacts upon classification shape and performance (Gale et al., 2016). Given the impact of methodological choice, it is typical that great care is taken into the testing and evaluation of different approaches along with their outputs, and this is acute within the context of those national classifications released by official statistical bodies where extensive stakeholder consultation and ratification are typically implemented as part of the construction process (Gale et al., 2016; Vickers & Rees, 2007).

A primary task when building any geodemographic classification is to develop a framework for the selection of specific variables that will produce meaningful and application-relevant clusters (Murphy & Smith, 2014). Those debates about the brevity of geodemographic classification inputs have been rehearsed for a long time. Openshaw et al. (1995) advocated that "the fewer the variables the better", whereas Harris et al. (2005) state that a more meaningful classification is likely to be constructed through inputting more variables, unless these variables are not "reliable, robust, and adding new information". The dimensionality of inputs (i.e. the number of zones multiplied by the number of variables) also has an interaction with the effectiveness of clustering methods to find salient structure from the data. Clustering performance can be hugely improved through the reduction of the number of variables due to this "curse of dimensionality" (Guyon & Elisseeff, 2003; Pacheco, 2015; Rojas, 2015; Jiliang Tang et al., 2014). Taking such perspectives into consideration, a typical objective of variable selection is therefore to achieve input parsimony, that is, the identification of the smallest subset of input variables that capture the most variation within the original dataset (Debenham, 2002; Gale et al., 2016; Harris et al., 2005). This will typically be achieved by balancing both the theoretical and empirical rationale for variable inclusion (Spielman & Singleton, 2015). For example, it is common that initial inputs are presented within a framework that draws upon wider literature, guiding the type and balance between different potential influences upon or outcomes of area differentiation. More empirically, this will usually ensue a process of initial candidate input variable evaluation, typically considering a range of factors about the individual candidate variables including their correlation, distribution or spatial coverage.

The remaining sections of this paper are presented as follows. In Section 4.3, we introduce the Office for National Statistics 2011 Output Area Classification (2011 OAC) as an example geodemographic that has an open and reproducible methodology; and focus particularly on the variable selection method adopted to create it. This is followed by a consideration of alternative methods that have been used to select geodemographic inputs in some other past national classifications built for either the UK or Great Britain. We then consider the use of Principal Component Analysis (PCA) as an alternative methodology for automating variable selection within Section 4.4, alongside results of the developed methodology in Section 4.5. In Section 4.5.2, we compare and contrast the results of a cluster analysis using the variable selection method with 2011 OAC. The limitations of this research are discussed in Section 4.6, alongside some plans for further work and extension.

## 4.3 SELECTING VARIABLES IN NATIONAL CLASSIFICATIONS

The 2011 OAC is a UK census-only geodemographic, which was released in 2014 by the Office for National Statistics (ONS) (Gale et al., 2016). This followed a similar classification created for the UK from the 2001 Census (Vickers & Rees, 2007). Both the 2001 and 2011 OAC have an open methodology and data inputs, which enable reproducibility, and furthermore provide a useful framework upon which comparative studies can be designed (Gale et al., 2016). The 2011 OAC presents a three-tiered hierarchy, comprising eight supergroups, 26 groups, and 76 subgroups. Each output cluster presents a shorthand name and “pen portrait” (description) depicting the most salient multidimensional characteristics (Bates, 2015; Gale et al., 2016).

The initial variable selection for 2011 OAC only considered those non-redundant census variables that were consistently provided by the three different UK census agencies (England and Wales, Scotland, Northern Ireland); and as a result of public consultation, was also guided by the 2001 OAC inputs. In 2011 OAC, 166 prospective variables (including 94 variables that were referenced by the 2001 OAC) and a derived variable of the standardised illness ratio (SIR) were tested. Moreover, the suitability of these initial variables was also scrutinised by the ONS (Gale et al., 2016). The initial variables were rationalised with two main objectives. The first was to obtain a variable mix that represented the general characteristics of the UK's neighbourhoods, meanwhile, also distinguishing salient characteristics that varied geographically. A second requirement was to minimise the number of strongly correlated census variables, thus limiting any potential weighting effect that may be caused by collinearity. According to Gale et al. (2016), these requirements were achieved through two empirical approaches. The first was to examine the correlations of candidate variables, and specifically identifying those variable pairs for further consideration where the correlation was greater than ± 0.6. A second technique implemented cluster-based sensitivity analysis, which aimed to identify those variables that had the greatest impact, either positive or negative, on cluster formation. This method assessed the total within-cluster sum of squares and the total between-cluster sum of squares statistics after including or excluding different variables from a clustering run. After further evaluation including examination of statistical distributions and mapping, 60 variables were eventually retained to build 2011 OAC, which were broadly organised into three domains: demographic, housing and socioeconomic (Gale et al., 2016).

However, the OAC (both 2001 and 2011 OAC) approach to variable selection deviates from those methods implemented by academics building geodemographic classifications for pre-2001 censuses in the UK. The very different computational contexts of the past made more sophisticated multidimensional processing much slower or impossible (Adnan, 2011; Singleton, 2016). Prior to the 2001 OAC, dimensionality reducing methods such as principal component analysis (PCA) were commonly (although not universally) integrated into some of the classification products that corresponded to the decennially released Census (e.g. Charlton et al., 1985; Robinson, 1998; Webber, 1975; Webber & Craig, 1978). When a PCA is calculated for a dataset, a set of new orthogonal variables (i.e. principal components) are created which are the linear combination of the original variables. The principal component that accounts for the largest variance is called the first principal component, the second principal component that accounts for the second-largest variance as the second, and so forth (Jolliffe, 1972; Pacheco, 2015). Clustering a set of principal components reduces the overall number of inputs to a geodemographics, making the clustering process either possible or much faster to complete; which in the past had been a key constraint given more limited computational power/availability. However, as the data handling and processing capacity of computers have increased, the necessity for PCA in this context has been reduced. Furthermore, some scholars have also argued that use of PCA to create inputs may erase interesting patterns, and particularly those which are spatially heterogeneous (Harris et al., 2005; Leventhal, 2016; Tang et al., 2014).

However, there are some contemporary implementations of PCA when building geodemographic classifications. For instance, Santeo et al. (2016) employed PCA as an inspection tool that determines whether a linear relationship exists between candidate variables; Adnan (2011) adopts PCA as a standardisation technique in the progress of producing real-time geodemographics. Although not a necessity in terms of computation, and as illustrated by Debenham (2002), PCA can have a useful role as a tool that guides variable selection. Although, what is under-researched is how such a process could be automated, taking account of both the overall importance of input variables to cluster formation, but also those sensitivities of the extent to which such relationships may hold between different localities. One of the overarching objectives of this paper is therefore to re-examine the potential for PCA within a computationally intensive setting, where the benefits of PCA for the identification of variables that explain the main variance within a dataset can be integral to an automated variable selection process. We present this new methodology in the context of a UK census-based geodemographic, contrasting the output against the 2011 OAC.

## 4.4 AUTOMATED VARIABLE SELECTION USING PCA

As discussed in the previous section, an overarching objective of the variable selection stage of building a geodemographic classification is to identify the smallest possible subset of variables that can represent the main variance within a universe of potential inputs being considered, which may also be informed by theoretical or practical rationale. Although accepting of arguments that PCA can have an adverse effect when used to create inputs to a geodemographic classification (Harris et al., 2005; Leventhal, 2016; Tang et al., 2014), we would argue that PCA can still have utility as a tool in the identification of appropriate input variables; which is the basis of the method we introduce in the remainder of this section.

The flowchart presented in *Figure 4.1* illustrates an approach that is comprised of five main stages. The first stage generates a set of principal components (PCs) from the input variables. Meanwhile, by summing up of the squared factor scores for the PC, the eigenvalue associated with each of the PCs can be calculated, which is utilised to define the range of iteration tests at stage 2. Additionally, the contribution of the variable to each component can be obtained by calculating the ratio between the squared factor score for a variable and the eigenvalue associated with that component. The value of a contribution is between 0 and 1. Generally, the larger the value, the more a variable contributes to the component (Abdi & Williams, 2010; Pacheco, 2015).

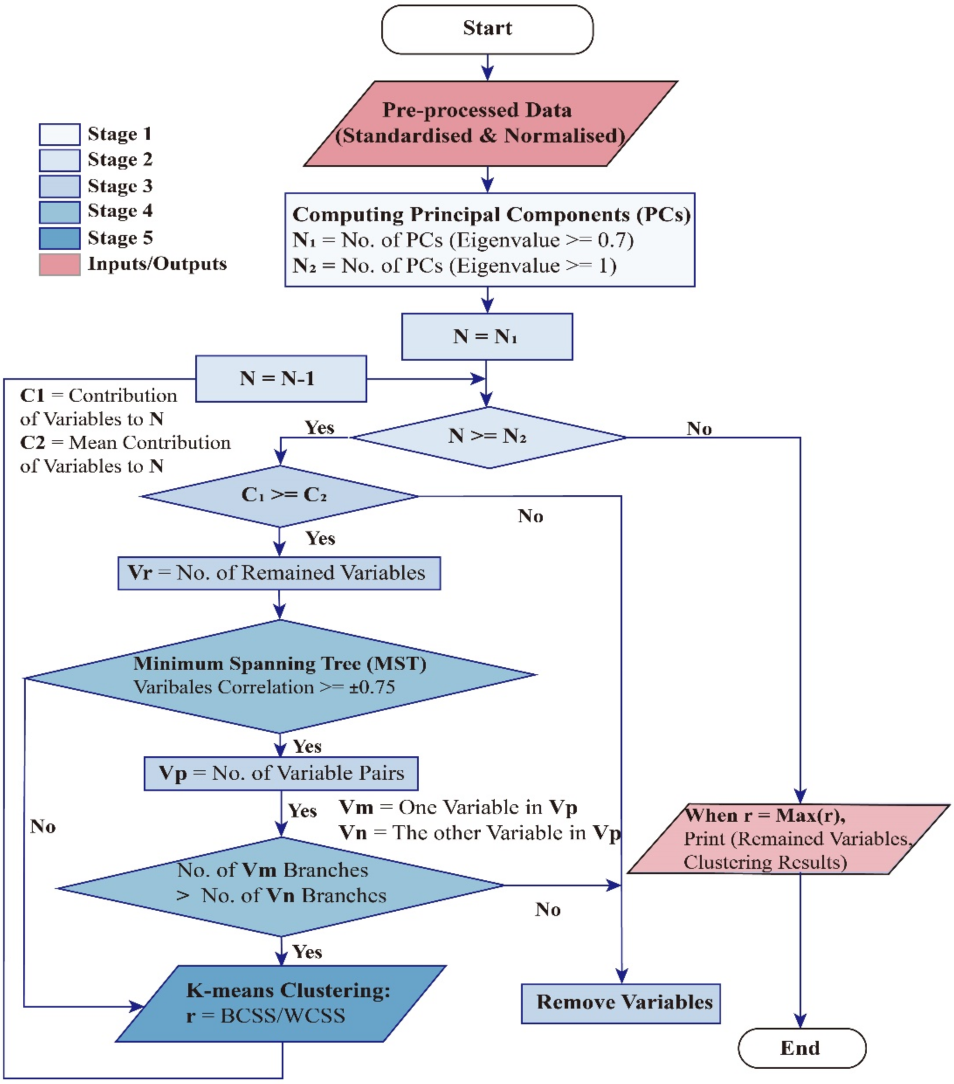

Start
Pre-processed Data (Standardised & Normalised)
Stage 1
Stage 2
Stage 3
Stage 4
Stage 5
Inputs/Outputs
Computing Principal Components (PCs)
$N_1$ = No. of PCs (Eigenvalue >= 0.7)
$N_2$ = No. of PCs (Eigenvalue >= 1)
$N = N_1$
N = N-1
C1 = Contribution of Variables to N
C2 = Mean Contribution of Variables to N
Yes
$N >= N_2$
No
$C_1 >= C_2$
No
Yes
Vr = No. of Remained Variables
Minimum Spanning Tree (MST)
Varibales Correlation >= ±0.75
Yes
Vp = No. of Variable Pairs
No
Yes
Vm = One Variable in Vp
Vn = The other Variable in Vp
When r = Max(r),
Print (Remained Variables, Clustering Results)
No. of Vm Branches > No. of Vn Branches
No
Yes
K-means Clustering:
r = BCSS/WCSS
Remove Variables
End

*Figure 4.1 Proposed Automated Variable Selection Method Workflow*

Stage 2 defines a threshold for the number of iterations to test between a “harsh” and a more “liberal” cut-off point. The maximum, i.e. “harsh”, threshold value is defined by a strict cut-off point that is generated from the commonly used Kaiser’s rule, namely, the eigenvalue of a meaningful PC is greater than or equal to 1 (Jolliffe, 2002; Kaiser, 1960; Pacheco, 2015). The minimum, i.e. “liberal”, threshold value is determined by adopting a cut-off point that is suggested by Jolliffe (1972), namely, the eigenvalue of a meaningful PC is greater than or equal to 0.7.

Stages 3–5 are iterative, with each run successively removing a PC from the set identified at stage 2. In every iteration at stage 3, the contribution of the variables to the retained PCs is quantified by taking the sum of their individual contributions multiplied by their respective eigenvalues in each PC (Pacheco, 2015). Abdi & Williams (2010, 437) suggest the use of “larger than the average contribution” as a heuristic cut-off when identifying variables with high contributions. Similarly, in this stage, where a variable contribution is greater than the average of these summed scores, it is retained for stage 4.

Stage 4 explores the correlation between the retained variables using a Minimum Spanning Tree (MST), which re-examines the level of data redundancy. Any highly correlated pairs are highlighted by the tree, defined as having a correlation coefficient of greater than or equal to ± 0.75, which is commonly cited as the “rule of thumb” indicating a high correlation (Santeo et al., 2016; Udovičić et al., 2007). In these instances, those highly correlated variables (i.e. nodes in the MST) with the fewest branches (i.e. less connected) were removed from the candidate variable list since they are considered of lower importance (Financial Network Analytics, 2012). Although automated in this instance, it follows similar methods implemented when building some commercial geodemographics (Harris et al., 2005).

At stage 5, the filtered variables are then clustered using the K-means algorithm with a user-specified number of clusters. This was optimised by running 10,000 times which is necessary given that the starting seeds used to initialise cluster partitioning are stochastic, and as such, there can be slight differences in outcomes between each run of the algorithm. Of the 10,000 runs that are generated for each iteration of Stage 5, the result with the lowest Total Within-cluster Sum of Squares (TWSS) statistics is extracted; representing a solution with overall more compact clusters. For this selected optimised run, two statistics were calculated as a measure of overall clustering quality: The between cluster sum of square (BCSS) and within-cluster sum of square (WCSS) statistics.

At the end of each Stage 5 run, the WCSS and BCSS are then stored in association with the currently tested PC cut-off defined at Stage 3. The ratio between the WCSS and BCSS is used to monitor the impact of the specific PC selection. Generally, the larger the ratio, the better clustering results. The iteration stops when the minimum/maximum (depending on removing/adding) number of PC defined by Stage 2 is met.

## 4.5 CASE STUDY APPLICATION

The automated variable selection process presented in the previous section was implemented in an example of building a UK census geodemographics that would be broadly comparable to 2011 OAC. As discussed earlier, the open methodology and data used to create this geodemographics make it a useful candidate for comparison; and in drawing parallels between the classifications we can illustrate broad comparability and the utility of the presented technique. It should be noticed that the objective of this application was, therefore, to retain broad comparability with the 2011 OAC, and to this end, the methods of standardisation, normalisation and clustering were mirrored. Thus, input data were normalised using an Inverse hyperbolic sine, and then range standardised onto a 1–0 scale. For the K-means implementation, only the most aggregate level of hierarchy in 2011 OAC was considered in this comparison, so k was defined as 8 for this model, although future work might consider further levels of disaggregation or a range of different k values might be tested. The rationale for the specific methodological choices in 2011 OAC can be found within Gale et al. (2016); however, the key point of departure in the presented methodology relates to how the final variables are selected for input into the clustering process.

In this application of the generally applicable methodology outlined in the previous section, we considered nearly all variables contained within the Key Statistics (KS) and Quick Statistics (QS) tables for the UK, which included the 167 initial variables considered for inclusion in 2011 OAC. Although, given that some tables contained duplicated or near identical topics, only one of these tables were included. For instance, both tables KS104 and QS108 concerned living arrangements; tables KS102 and QS102 detailed age structure. Finally, like 2011 OAC, the initial inputs also included computation of a Standardised Illness Ratio. The full initial variable specification is listed in Table S1 of the online Supplementary Materials.

After running a PCA on the input data, a total of 53 meaningful PCs were identified by examination of the eigenvalue thresholds which are plotted against the cumulative variance explained in *Figure 4.2*. If we had applied the Kaiser rule (Eigenvalue ≥1), only 30 principal components would have been selected which cumulatively accounted for about 72.4% of the variance being retained. However, it can be seen that by altering the cut-off value from 1 to 0.7, the filtering process identified 52 principal components, which cumulatively accounted for approximately 83.6% of the variance contained in the original data.

A summary of outcomes from the iteration tests is shown in *Figure 4.3*. The highest quality clustering results were identified (an objective function of maximising the ratio between BCSS and WCSS) when the first 51 PCs were used to identify input variables to the cluster analysis. Of the 86 variables firstly identified, 12 pairs were highly correlated (Correlation Coefficient ≥ ± 0.75); and as such, utilising the minimum spanning tree (*Figure 4.4*), 12 variables were removed.

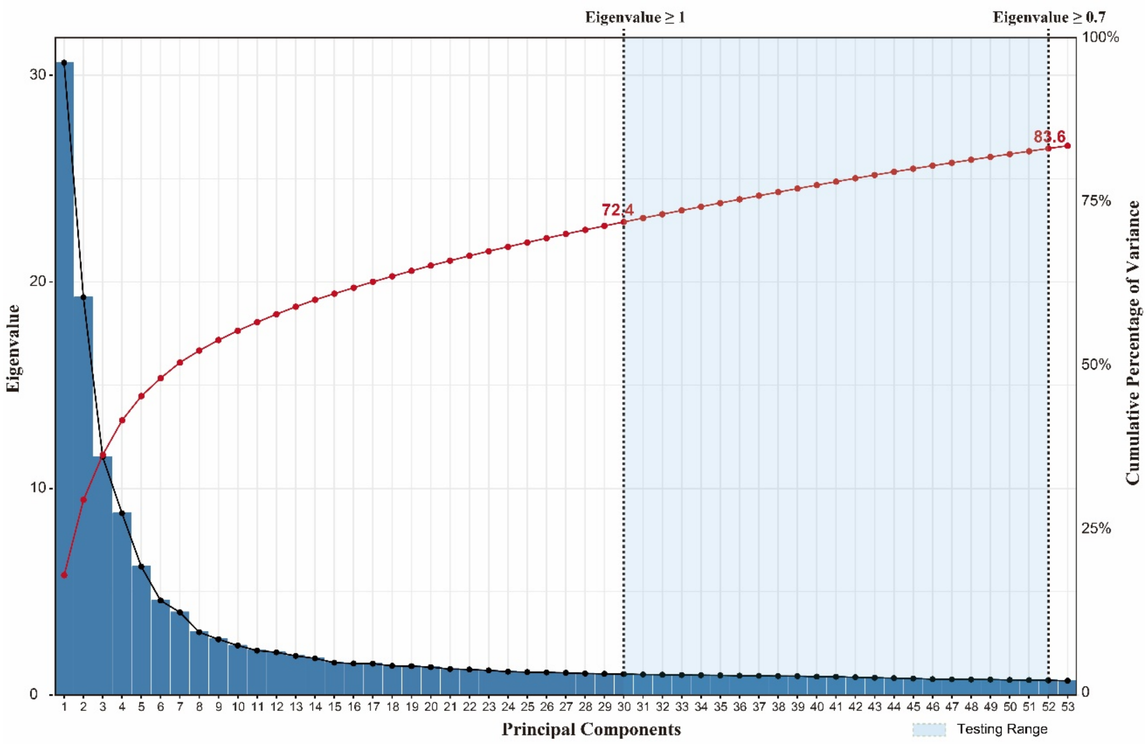


***Figure 4.2 Scree Plot: Eigenvalue vs. Percentage of Explained Variances***

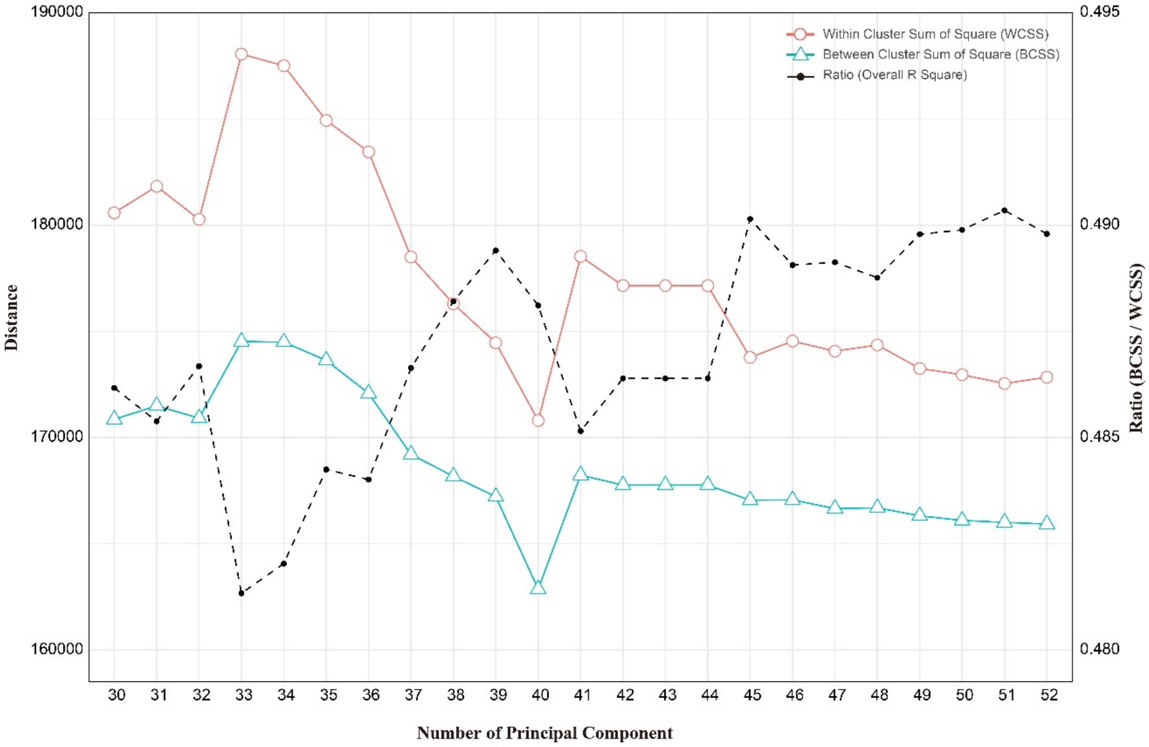


*Figure 4.3 BCSS and WCSS Result by Iteration test by Principal Components. Ratio = BCSS / WCSS*

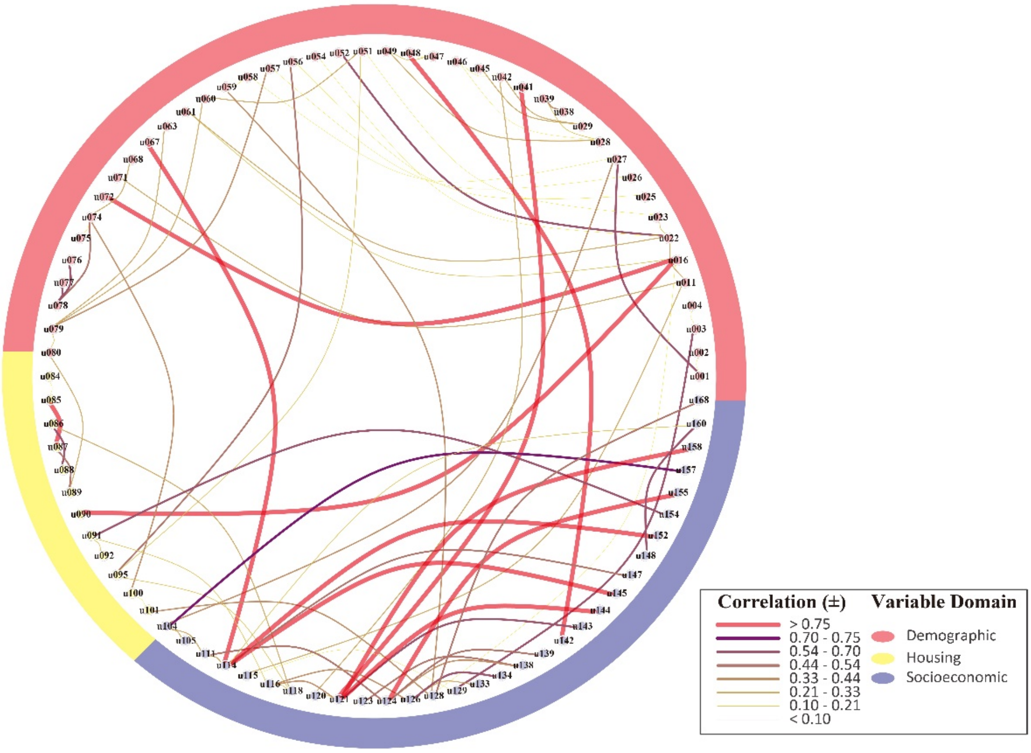


*Figure 4.4 Minimum Spanning Tree of the Census Variables after the PCA-Based Filter.*

The thickness of the curve indicates the absolute value of the person correlation coefficient. The value above ±0.75 is highlighted by red thicker line, which therefore will be removed in the next phases.

Detail about the type and frequency of variables retained for each iteration are presented in *Table 4.1*. The variables have been divided into three different domains of: demographic, socioeconomic and housing. Overall the iterations, variables in the socioeconomic domain were retained less often, indicating greater redundancy. In contrast, the housing domain was reasonably stable, and for all iterations comprised between 45 and 60 percent of the overall variables within this domain.

| PCs | Retained Variables | Demographic (D) | Socioeconomic (S) | Housing (H) | Ratio = BCSS/WCSS | D % by Total | S % by Total | H % by Total |
|---|---|---|---|---|---|---|---|---|
| 30 | 90 | 50<br>*55.6* | 28<br>*31.1* | 12<br>*13.3* | 0.4862 | 62.5 | 41.8 | 60.0 |
| 31 | 91 | 50<br>*55.6* | 29<br>*32.2* | 12<br>*13.3* | 0.4854 | 62.5 | 43.3 | 60.0 |
| 32 | 90 | 49<br>*54.4* | 29<br>*32.2* | 12<br>*13.3* | 0.4867 | 61.3 | 43.3 | 60.0 |
| 33 | 88 | 49<br>*54.4* | 27<br>*30* | 12<br>*13.3* | 0.4813 | 61.3 | 40.3 | 60.0 |
| 34 | 87 | 49<br>*54.4* | 26<br>*28.9* | 12<br>*13.3* | 0.4820 | 61.3 | 38.8 | 60.0 |
| 35 | 85 | 48<br>*53.3* | 25<br>*27.8* | 12<br>*13.3* | 0.4842 | 60.0 | 37.3 | 60.0 |
| 36 | 84 | 47<br>*52.2* | 25<br>*27.8* | 12<br>*13.3* | 0.4840 | 58.8 | 37.3 | 60.0 |
| 37 | 81 | 47<br>*52.2* | 24<br>*26.7* | 10<br>*11.1* | 0.4866 | 58.8 | 35.8 | 50.0 |
| 38 | 79 | 45<br>*50* | 24<br>*26.7* | 10<br>*11.1* | 0.4882 | 56.3 | 35.8 | 50.0 |
| 39 | 78 | 43<br>*47.8* | 25<br>*27.8* | 10<br>*11.1* | 0.4894 | 53.8 | 37.3 | 50.0 |
| 40 | 75 | 43<br>*47.8* | 23<br>*25.6* | 9<br>*10* | 0.4881 | 53.8 | 34.3 | 45.0 |
| 41 | 75 | 43<br>*47.8* | 22<br>*24.4* | 10<br>*11.1* | 0.4852 | 53.8 | 32.8 | 50.0 |
| 42 | 74 | 43<br>*47.8* | 21<br>*23.3* | 10<br>*11.1* | 0.4864 | 53.8 | 31.3 | 50.0 |
| 43 | 74 | 43<br>*47.8* | 21<br>*23.3* | 10<br>*11.1* | 0.4864 | 53.8 | 31.3 | 50.0 |
| 44 | 74 | 43<br>*47.8* | 21<br>*23.3* | 10<br>*11.1* | 0.4864 | 53.8 | 31.3 | 50.0 |
| 45 | 74 | 42<br>*46.7* | 21<br>*23.3* | 11<br>*12.2* | 0.4901 | 52.5 | 31.3 | 55.0 |
| 46 | 75 | 42<br>*46.7* | 23<br>*25.6* | 10<br>*11.1* | 0.4891 | 52.5 | 34.3 | 50.0 |
| 47 | 74 | 41<br>*45.6* | 23<br>*25.6* | 10<br>*11.1* | 0.4891 | 51.3 | 34.3 | 50.0 |
| 48 | 75 | 41<br>*45.6* | 24<br>*26.7* | 10<br>*11.1* | 0.4888 | 51.3 | 35.8 | 50.0 |
| 49 | 75 | 41<br>*45.6* | 23<br>*25.6* | 11<br>*12.2* | 0.4898 | 51.3 | 34.3 | 55.0 |
| 50 | 74 | 41<br>*45.6* | 23<br>*25.6* | 10<br>*11.1* | 0.4899 | 51.3 | 34.3 | 50.0 |
| 51 | 74 | 40<br>*44.4* | 24<br>*26.7* | 10<br>*11.1* | 0.4903 | 50.0 | 35.8 | 50.0 |
| 52 | 74 | 40<br>*44.4* | 24<br>*26.7* | 10<br>*11.1* | 0.4898 | 50.0 | 35.8 | 50.0 |
| Total | 167 | 80 | 67 | 20 |  | 55.4 | 36.0 | 53.3 |

*Table 4.1 Testing Results Showing the Number and Percentage of Overall Retained Variables and by Domain*

In the optimised result (51 PCs), 74 variables in total were retained, distributed between 40 demographic, 24 socioeconomic, 10 housing. *Table 4.2* summarises this distribution relative to those inputs used to build 2011 OAC. Most significantly, the proportion of retained variables related to demographics was much larger, while the other domain proportions remained largely similar in size.

| Domain | AVS-OAC | AVS-OAC (%) | 2011 OAC | 2011 OAC (%) |
|---|---|---|---|---|
| Demographic | 40 | 54.1% | 26 | 43.3% |
| Socioeconomic | 24 | 32.4% | 26 | 43.3% |
| Housing | 10 | 13.5% | 8 | 13.3% |
| Total | 74 | - | 60 | - |

***Table 4.2 Number of Final Census Variables Retained by Domain vs 2011 OAC***

### 4.5.1 DESCRIBING THE DERIVED CLASSIFICATION

In this penultimate section, we firstly present descriptions to accompany the optimised clustering result derived through automated variable selection (Automated Variable Selection OAC – AVS-OAC). The new classification created through this process had an average cluster size of approximately 29,037 Output areas (OAs), however, varied from 11,397 (E) and 41,399 (B) OAs, which, respectively, correspond to about 4.9% and 17.8% of the total number of OAs in the UK. By contrast, 2011 OAC varies from 8,589 OAs (2: Ethnicity Central) to 35,285 OAs (6: Urbanities), so the range of our presented clusters is larger.

*Figure 4.5* maps the geographic distribution of the AVS-OAC clusters across the UK, and also respectively highlights the cluster distribution in the largest cities, namely, London, Cardiff, Edinburgh, and Belfast. The spatial distribution highlights a useful urban-rural split, and within urban areas presents a range of differentiating clusters. Additionally, and as one might expect given the methodological choices made, London is fairly poorly segmented with the majority of inner London dominated by two clusters (i.e. Cluster D and E). This effect is similar in 2011 OAC, and indeed is discussed at length elsewhere (see Singleton and Longley, 2015). One potentially negative observation of the created classification was the emergence of two clusters that represented mainly rural areas (Clusters 1 and 6). In order to explore these patterns and wider interpretability of the cluster characteristics and later comparison with 2011 OAC, index scores (i.e. x/x¯ *100) were computed for the input variables and displayed in *Figure 4.6*, with the scores ordered by domain. These scores illustrate characteristics that are over or underrepresented for each of the eight clusters relative to the national average (a score of 100). An index score of 50 is therefore half the national average, and 200 would be double. Additionally, as is common when building a geodemographic classification, such index scores were then used to ascribe a label and brief description of each of the clusters.

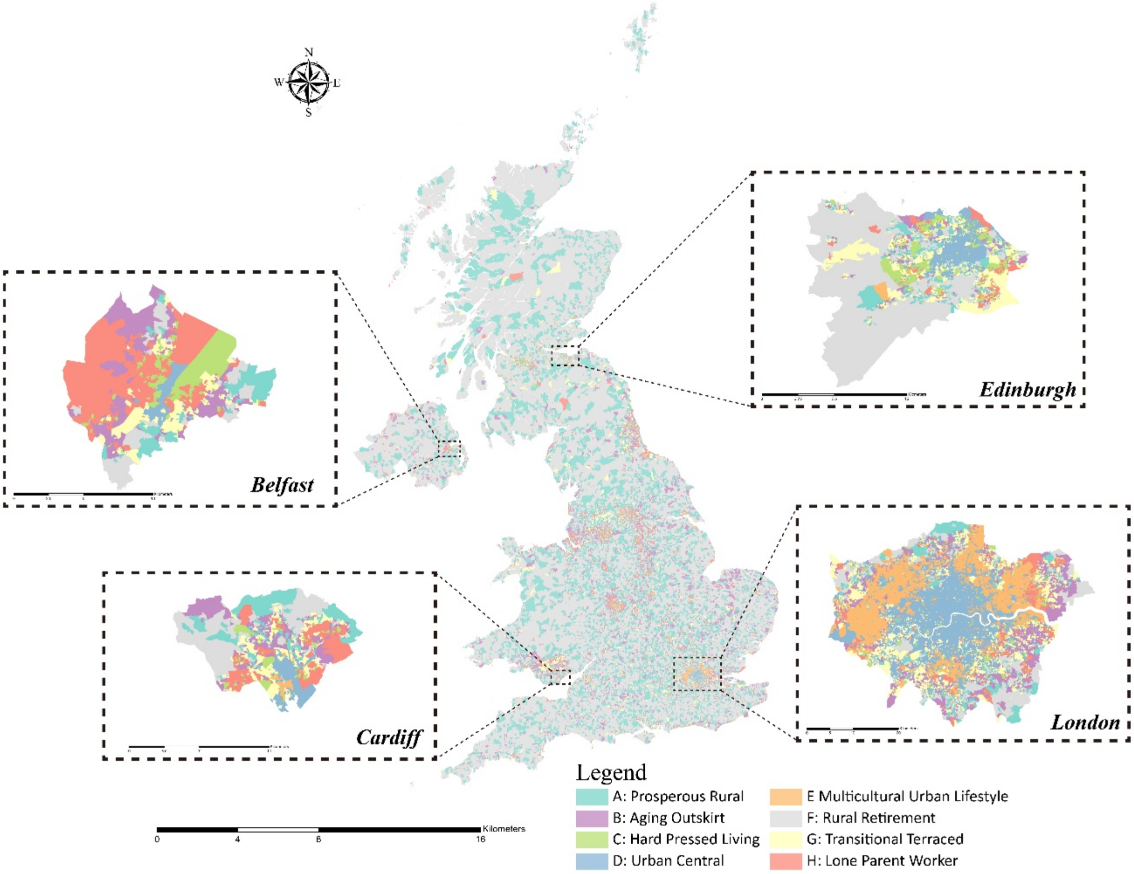
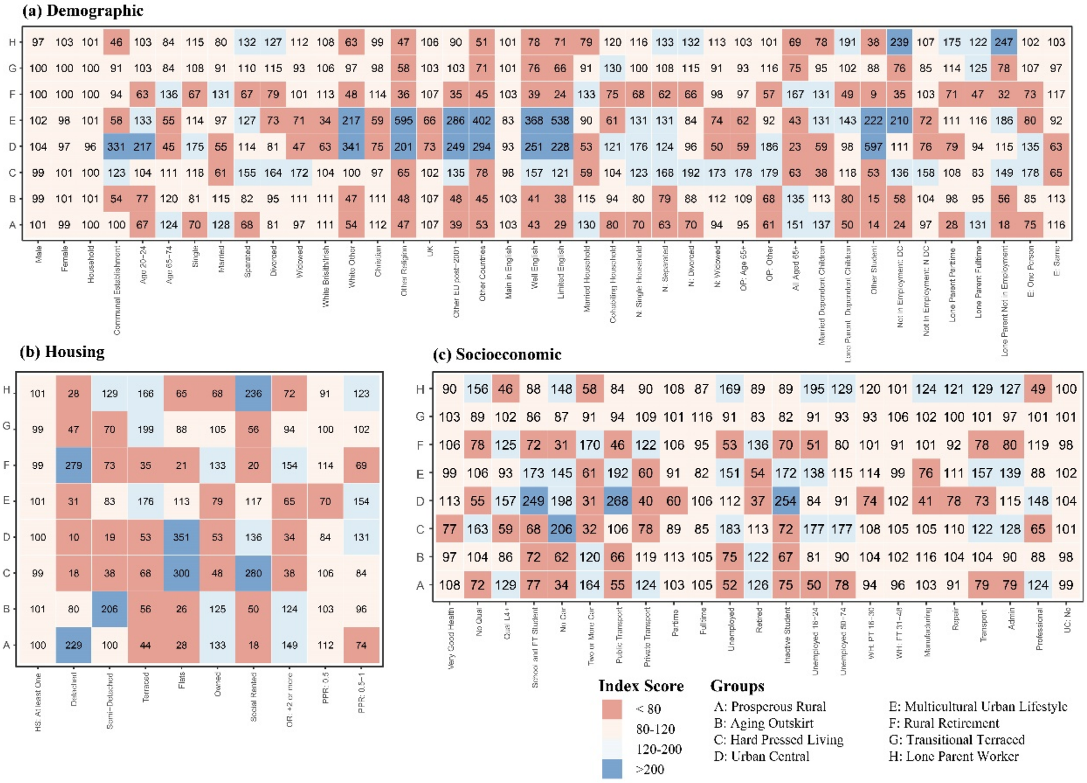


**Figure 4.3 Geographic Distribution of AVS-OACs with Highlighted Major Cities**



(a) Demographic

| | Male | Female | Household | Communal Establishment | Age 20−24 | Age 65−74 | Single | Married | Sperated | Divorced | Widowed | White Brisith/Irish | White Other | Christian | Other Religion | UK | Other EU post−2001 | Other Countries | Main in English | Well English | Limited English |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| H | 97 | 103 | 101 | 46 | 103 | 84 | 115 | 80 | 132 | 127 | 112 | 108 | 63 | 99 | 47 | 106 | 90 | 51 | 101 | 78 | 71 |
| G | 100 | 100 | 100 | 91 | 103 | 84 | 108 | 91 | 110 | 115 | 93 | 106 | 97 | 98 | 58 | 103 | 103 | 71 | 101 | 76 | 66 |
| F | 100 | 100 | 100 | 94 | 63 | 136 | 67 | 131 | 67 | 79 | 101 | 113 | 48 | 114 | 36 | 107 | 35 | 45 | 103 | 39 | 24 |
| E | 102 | 98 | 101 | 58 | 133 | 55 | 114 | 97 | 127 | 73 | 71 | 34 | 217 | 59 | 595 | 66 | 286 | 402 | 83 | 368 | 538 |
| D | 104 | 97 | 96 | 331 | 217 | 45 | 175 | 55 | 114 | 81 | 47 | 63 | 341 | 75 | 201 | 73 | 249 | 294 | 93 | 251 | 228 |
| C | 99 | 101 | 100 | 123 | 104 | 111 | 118 | 61 | 155 | 164 | 172 | 104 | 100 | 97 | 65 | 102 | 135 | 78 | 98 | 157 | 121 |
| B | 99 | 101 | 101 | 54 | 77 | 120 | 81 | 115 | 82 | 95 | 111 | 111 | 47 | 111 | 48 | 107 | 48 | 45 | 103 | 41 | 38 |
| A | 101 | 99 | 100 | 100 | 67 | 124 | 70 | 128 | 68 | 81 | 97 | 111 | 54 | 112 | 47 | 107 | 39 | 53 | 103 | 43 | 29 |

| | Married Household | Cohabiting Household | N: Single Household | N: Separated | N: Divorced | N: Widowed | OP: Age 65+ | OP: Other | All Aged 65+ | Married Dependent Children | Lone Parent Dependent Children | Other Student | Not In Employment: DC | Not In Employment: N DC | Lone Parent Parttime | Lone Parent Fulltime | Lone Parent Not in Employment | E: One Person | E: Same |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| H | 79 | 120 | 116 | 133 | 132 | 113 | 103 | 101 | 69 | 78 | 191 | 38 | 239 | 107 | 175 | 122 | 247 | 102 | 103 |
| G | 91 | 130 | 100 | 108 | 115 | 91 | 93 | 116 | 75 | 95 | 102 | 88 | 76 | 85 | 114 | 125 | 78 | 107 | 97 |
| F | 133 | 75 | 68 | 62 | 66 | 98 | 97 | 57 | 167 | 131 | 49 | 9 | 35 | 103 | 71 | 47 | 32 | 73 | 117 |
| E | 90 | 61 | 131 | 131 | 84 | 74 | 62 | 92 | 43 | 131 | 143 | 222 | 210 | 72 | 111 | 116 | 186 | 80 | 92 |
| D | 53 | 121 | 176 | 124 | 96 | 50 | 59 | 186 | 23 | 59 | 98 | 597 | 111 | 76 | 79 | 94 | 115 | 135 | 63 |
| C | 59 | 104 | 123 | 168 | 192 | 173 | 178 | 179 | 63 | 38 | 118 | 53 | 136 | 158 | 108 | 83 | 149 | 178 | 65 |
| B | 115 | 94 | 80 | 79 | 88 | 112 | 109 | 68 | 135 | 113 | 80 | 15 | 58 | 104 | 98 | 95 | 56 | 85 | 113 |
| A | 130 | 80 | 70 | 63 | 70 | 94 | 95 | 61 | 151 | 137 | 50 | 14 | 24 | 97 | 28 | 131 | 18 | 75 | 116 |

(b) Housing

| | HS: At least One | Detached | Semi−Detached | Terraced | Flats | Owned | Social Rented | OR: +2 or more | PPR: 0.5 | PPR: 0.5−1 |
|---|---|---|---|---|---|---|---|---|---|---|
| H | 101 | 28 | 129 | 166 | 65 | 68 | 236 | 72 | 91 | 123 |
| G | 99 | 47 | 70 | 199 | 88 | 105 | 56 | 94 | 100 | 102 |
| F | 99 | 279 | 73 | 35 | 21 | 133 | 20 | 154 | 114 | 69 |
| E | 101 | 31 | 83 | 176 | 113 | 79 | 117 | 65 | 70 | 154 |
| D | 100 | 10 | 19 | 53 | 351 | 53 | 136 | 34 | 84 | 131 |
| C | 99 | 18 | 38 | 68 | 300 | 48 | 280 | 38 | 106 | 84 |
| B | 101 | 80 | 206 | 56 | 26 | 125 | 50 | 124 | 103 | 96 |
| A | 100 | 229 | 100 | 44 | 28 | 133 | 18 | 149 | 112 | 74 |

(c) Socioeconomic

| | Very Good Health | No Qual | Qual L4+ | School and FT Student | No Car | Two or More Car | Public Transport | Private Transport | Partime | Fulltime | Unemployed | Retired | Inactive Student | Unemployed 16−24 | Unemployed 50−74 | WH: PT 16−30 | WH: FT 31−48 | Manufacturing | Repair | Transport | Admin | Professional | UC: No |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| H | 90 | 156 | 46 | 88 | 148 | 58 | 84 | 90 | 108 | 87 | 169 | 89 | 89 | 195 | 129 | 120 | 101 | 124 | 121 | 129 | 127 | 49 | 100 |
| G | 103 | 89 | 102 | 86 | 87 | 91 | 94 | 109 | 101 | 116 | 91 | 83 | 82 | 91 | 93 | 93 | 106 | 102 | 100 | 101 | 97 | 101 | 101 |
| F | 106 | 78 | 125 | 72 | 31 | 170 | 46 | 122 | 106 | 95 | 53 | 136 | 70 | 51 | 80 | 101 | 91 | 101 | 92 | 78 | 80 | 119 | 98 |
| E | 99 | 106 | 93 | 173 | 145 | 61 | 192 | 60 | 91 | 82 | 151 | 54 | 172 | 138 | 115 | 114 | 99 | 76 | 111 | 157 | 139 | 88 | 102 |
| D | 113 | 55 | 157 | 249 | 198 | 31 | 268 | 40 | 60 | 106 | 112 | 37 | 254 | 84 | 91 | 74 | 102 | 41 | 78 | 73 | 115 | 148 | 104 |
| C | 77 | 163 | 59 | 68 | 206 | 32 | 106 | 78 | 89 | 85 | 183 | 113 | 72 | 177 | 177 | 108 | 105 | 105 | 110 | 122 | 128 | 65 | 101 |
| B | 97 | 104 | 86 | 72 | 62 | 120 | 66 | 119 | 113 | 105 | 75 | 122 | 67 | 81 | 90 | 104 | 102 | 116 | 104 | 104 | 90 | 88 | 98 |
| A | 108 | 72 | 129 | 77 | 34 | 164 | 55 | 124 | 103 | 105 | 52 | 126 | 75 | 50 | 78 | 94 | 96 | 103 | 91 | 79 | 79 | 124 | 99 |

*Figure 4.4 AVS-OAC Results (Index scores) grouped by variable domains*

## 4.5.2 CLASSIFICATION PERFORMANCE AND COMPARISON TO 2011 OAC

In this final section, we first evaluate AVS-OAC performance internally to explore cluster robustness, and then make some external comparisons with 2011 OAC; to establish those broad similarities or differences that emerge through the application of this alternative methodology, and examine the impact this has on the overall discriminatory power.

An objective when building this classification was to provide an output that would make a suitable benchmark against 2011 OAC; achieved through maintaining both a broadly similar potential attribute input pool and output cluster frequency. However, a disadvantage of constraining the number of clusters to match 2011 OAC was that two very similar rural clusters emerged: Cluster A: Prosperous Rural and Cluster F: Rural Retirement; which represented considerable redundancy. When building a geodemographic classification for operational rather than methodological evaluation purposes, there is typically a stage that will test multiple potential cluster frequencies with the objective of mitigating such issues. However, conversely, the post-analysis merging or splitting of clusters is also prevalent when building many geodemographic classifications (Harris et al., 2005). For the purposes of this illustration we decided to keep this artefact, although in an operational model such as 2011 OAC, we would expect that such issues would be resolved pre or post clustering through manual intervention after stakeholder consultation.

Correspondence between 2011 OAC supergroups and AVS-OAC clusters is highlighted in *Figure 4.7* which presents the percentage by of OAs that overlap between the two classifications for the UK extent. As might be expected given the differing inputs, the correspondence between the two classifications varies; and highlights the importance of stakeholder engagement when selecting appropriate cluster representations in operational models. For example, we can see that the AVS-OAC Cluster: "F: Rural Retirement" is composed predominantly by OA identified by 2011 OAC as within Supergroups "1. Rural Residents" and "6. Suburbanites", thus representing a blend of both rural and the connecting hinterland at the periphery of urban areas. The AVS-OAC Cluster "D. Urban Central" combines many OA that are identified by the 2011 OAC Supergroups "2. Cosmopolitans", "3. Ethnicity Central", but not some other predominantly urban clusters such as "7. Constrained City Dwellers", which emerged with greater correspondence to AVS OAC Cluster "C. Hard Pressed Living". Or, the AVS-OAC cluster "B: Ageing Outskirt" can be seen to correspond to a diffuse number of 2011 OAC Supergroups located in suburban areas. A similarly defuse pattern can also be identified in "G: Transitional Terraced", although just over half of those areas identified by the 2011 OAC Supergroup "5. Urbanites" also correspond with this cluster.

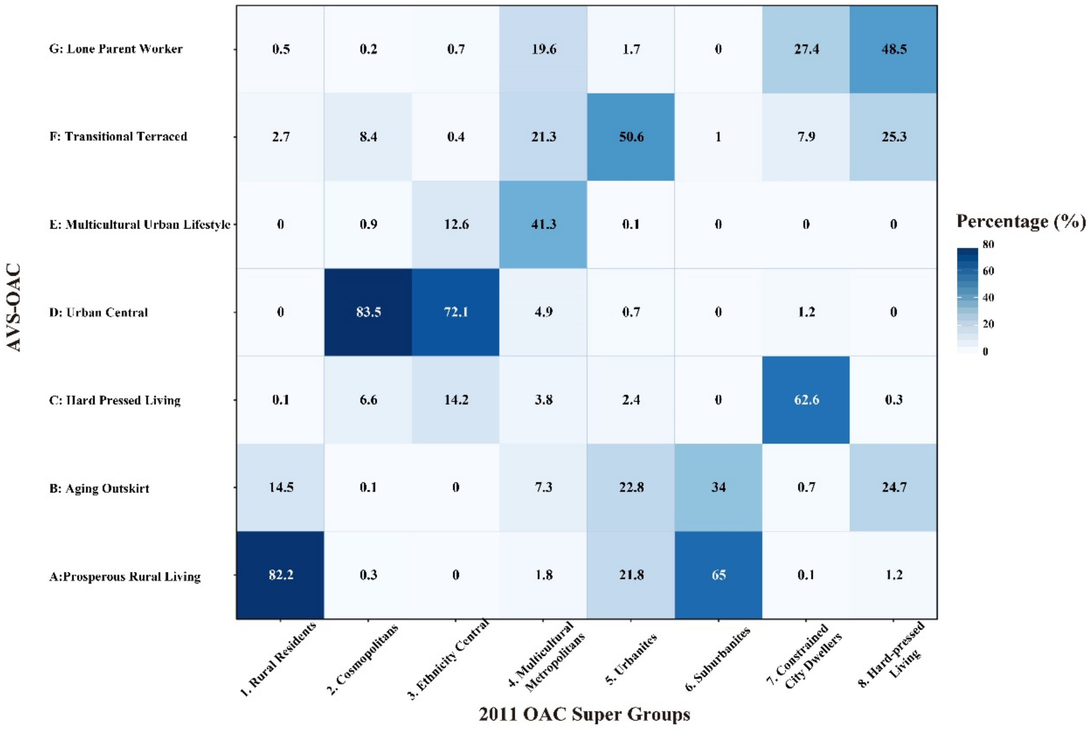


*Figure 4.5 Cross-Tabulation: OA Percentage by AVS-OAC and 2011 OAC*

As a measure of comparative clustering quality, a Total Within-cluster Sum of Squares (TWSS) statistic was calculated for each classification (i.e. AVS-OAC and 2011 OAC) by taking the sum of the squared difference between every classification input attribute within an area and the mean of the assigned cluster centroid. A higher score indicates an area where the attribute values for the OA are further from their assigned cluster mean (the centroid generated via k-means clustering), in other words, the quality of cluster assignment is poorer. Box plots in *Figures 4.8* and *4.9*, respectively delineate the TWSS by the AVS-OAC Clusters and the 2011 OAC Supergroups.

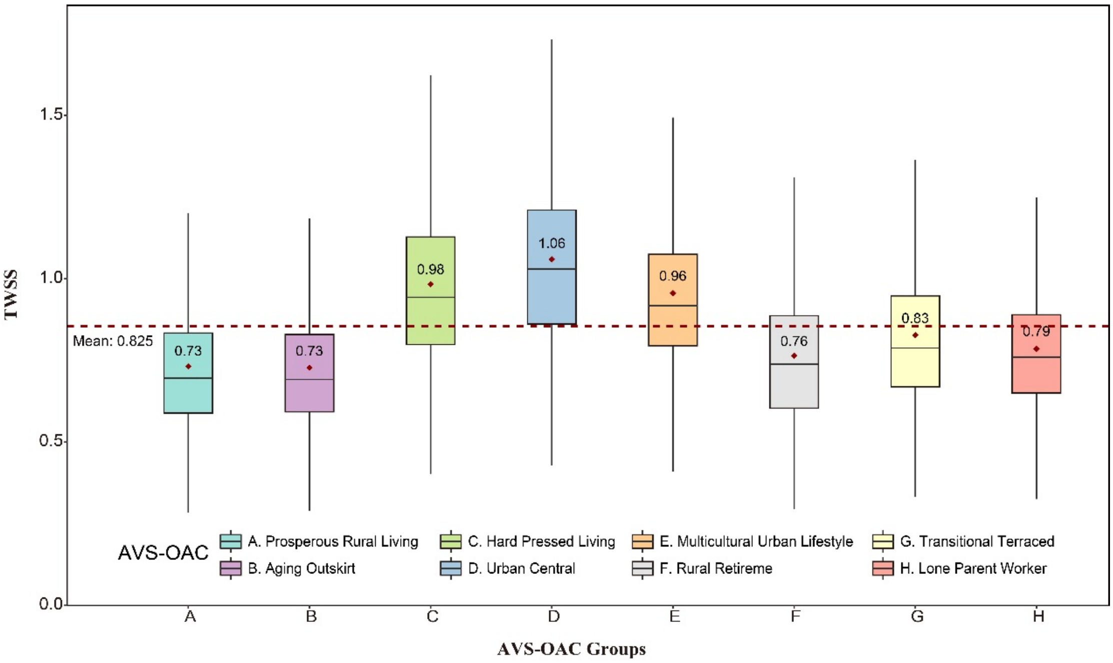


*Figure 4.6 Total Within-cluster Sum of Squares (TWSS) by the AVS-OACs.*

Mean value for each of the cluster is calculated and illustrated by the red point within the boxplot. The total mean value is presented by the dash line

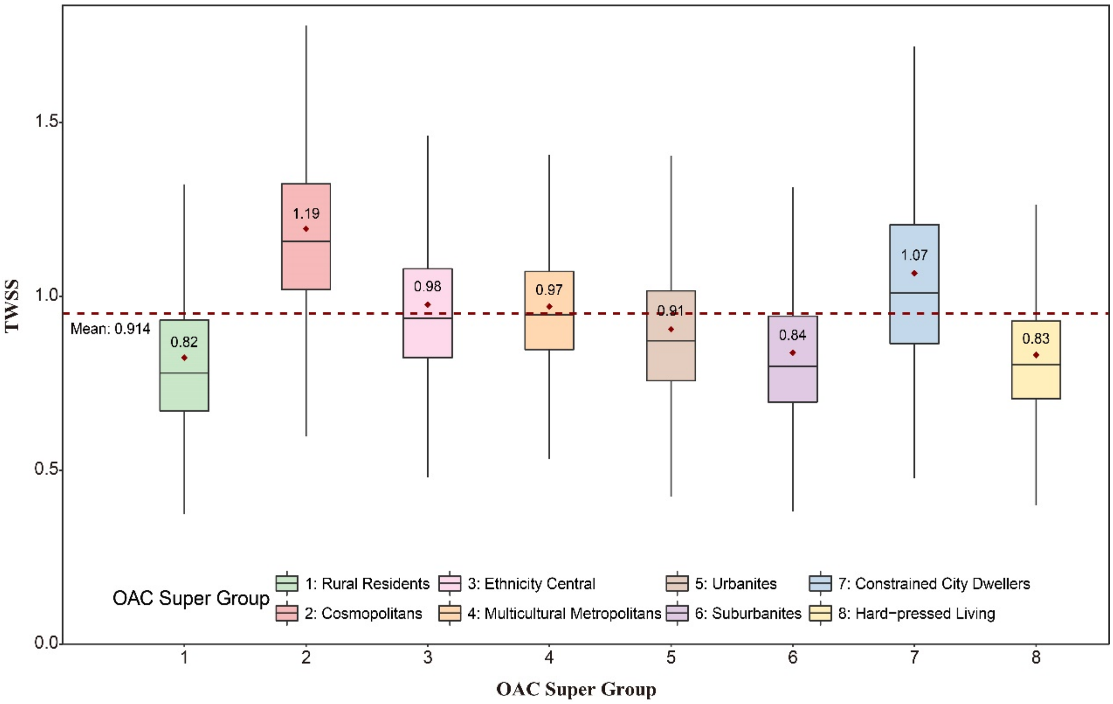


*Figure 4.7 Total Within-cluster Sum of Squares (TWSS) by the 2011 OAC.*

Mean value for each of the cluster is calculated and illustrated by the red point within the boxplot. The total mean value is presented by the dash line.

Overall, the AVS-OAC clusters have lower TWSS than 2011 OAC, statistically indicating a better fit, which is manifested by the average value (i.e. 0.825 and 0.914). Within AVS-OAC, we can see that the clusters "D: Urban Central", "C: Hard Pressed Living" and "E: Multicultural Urban Lifestyle" contain the highest TWSS value (average value, which are 1.06, 0.98, and 0.96) and the greatest variability (standard deviations, which are 0.286, 0.257, and 0.241, respectively), which might be considered the three least successful AVS-OAC clusters. These clusters are concentrated in both densely populated urban centres and transitional areas on the periphery of urban cores. In some sense, this is to be expected given the heterogeneous nature of urban centres and is an issue acute between Greater London and other parts of the UK which leads to larger variability. In particular, residents of AVS-OAC cluster "E: Multicultural Urban Lifestyle" are mainly concentrated within Greater London, which is a region known to be not well represented by 2011 OAC (Singleton & Longley, 2015).

Analysis of the geographic variability in classification performance can be expanded by mapping how well the input attributes of each OA fit their assigned cluster from both AVS-OAC and 2011 OAC, again using the TWSS statistics. The frequency of OAs that performed better by AVS-OAC relative to 2011 OAC (attributes values that are closer to their assigned cluster mean) was counted within each UK Local Authority District (LAD), and are presented in the choropleth map in *Figure 4.10*. Overall, 390 out of 404 local authority districts in the UK have greater than 50% of their constituent OAs statistically better represented by AVS-OAC relative to 2011 OAC. There are however some clear regional patterns that emerge; with particularly strong performance in Scotland and Wales where, respectively, all unitary authorities had more than 70% and 65% of OAs with better fit by AVS-OAC relative to 2011 OAC statistically. More negatively, there are some LADs that experience relatively poor performance, which are indicated by dark red in the choropleth (*Figure 4.10*): and include a cluster of boroughs alongside the River Thames within Greater London alongside some other London Boroughs. Additionally, some of the LADs located within Northern Ireland also exhibit poorer clustering performance. These instances support an argument for more consideration within an automated variable selection process of those characteristics specific to regional geographies. The need for greater regional consideration when building geodemographics is a well-established argument (Alexiou, 2016), which also points to future work outside of the scope of this paper when selecting variables automatically.

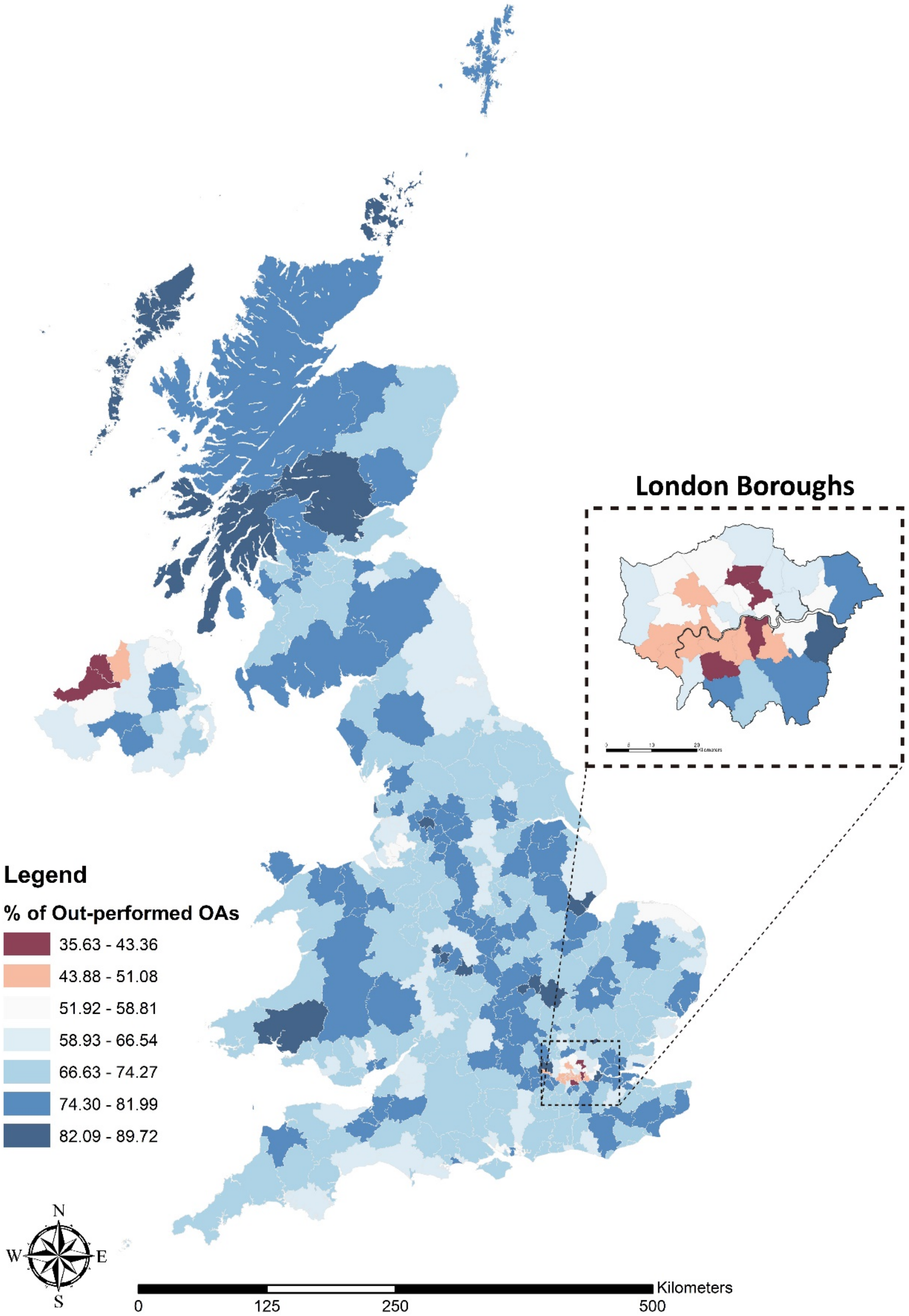


*Figure 4.8 Percentage of Out-Performed OAs by AVS-OAC by Local Authorities and London Boroughs*

## 4.6 CONCLUSIONS

The consideration of which variables are input into a cluster analysis is a common preliminary stage when building a geodemographic classification. The overarching objective is typically to achieve input parsimony, but there are multiple views on how this is best achieved, balancing theoretical considerations, practicalities of available data or attribute statistical properties and those past experiences or embedded knowledge of the classification builder(s). The primary objective of this paper was to extend such considerations by developing and testing an automated method of variable selection, and then benchmarking the presented technique within the context of building a UK national geodemographic from 2011 Census data.

The objective was to illustrate how automated variable selection could be implemented to identify inputs that produce a plausible and comparable classification. In doing so, we are not claiming that this be of equivalence to an operational model, specifically as the methodology presented here lacks user consultation; but rather provides an innovative tool that might be useful to inform variable choices. It is not difficult to envisage a build process within an operation setting where differing variable selection sets might be specified and evaluated in consultation with stakeholders.

Our heuristic process was built around Principal Component analysis that automated input variable selection, feeding these into a classification model that in our example broadly followed the 2011 OAC methodology. The application presented here was primarily data-driven for the purposes of methodological illustration; however, the technique itself is flexible and generic, and lends itself to other applications with any set of variables, thus also transferring well as a component of more theoretical expositions of geodemographic structure.

The method as implemented here was within the context of consistently available variables from the 2011 Census for the UK geographical extent. Through a five-stage variable selection procedure, 74 census variables were retained from 171 initial candidates. The clustering was constrained to mirror 2011 OAC cluster frequency, creating a final typology of eight clusters. This output was subsequently evaluated through comparison with 2011 OAC to examine both cluster similarity and relative performance. Overall, the quality of the cluster assignment is statistically better than 2011 OAC in more than 70.2% of the OAs across the UK; with particularly strong performance within Scotland and Wales.

The application of our method illustrated good comparative performance relative to 2011 OAC; however, there are several limitations that could be alleviated in future work. First, there may be potential for integrating regional and subregional evaluation when selecting variables, which might evolve into a set of heuristics that would potentially identify a more effective variable input mix. A counter view would be that this would be at the expense of computation time; and indeed, may be not resolve an inherent constraint in regional variability when building geodemographics from data pertaining to a national extent. Secondly, this automated process decouples stakeholder user input from the classification process; and as Openshaw et al. (1995) state that "there is no simple relationship between optimising a statistical measure of classification performance such as the within-cluster sum of squares and the end-users' perception of classification performance in a particular context". Such considerations could be integrated into a fuller process of classification building, which may be particularly important within the context of an operational classification, such as those built for a national statistical agency. Finally, it is also worthy of recalling that the presented method utilises PCA and there is also potential to integrate alternate and more explicitly spatial techniques, which may also enhance regionally variable performance, for example through Geographically Weighted PCA (Harris et al., 2011). Furthermore, there is also potential that additional steps could be implemented that assess an appropriate cluster frequency for a given problem, although there would be significant challenges when balancing such considerations with computational efficiency when input variable combinations were also being assessed in parallel.

This paper has presented a new methodology that optimises the selection of an initial list of candidate variables that are input into a cluster analysis used to build a geodemographic classification. The performance of this methodology is implemented within the context of the 2011 UK Census, and comparison is made with 2011 OAC. Performance was comparable to 2011 OAC over the evaluated metrics, although the shape of the classification varied, and there were also some regional differences in performance. The methodology presented here provides a generally applicable tool that integrates well with both theoretical and user embedded classification building programs over multiple international contexts, and, will likely have particular relevance for the creation of future geodemographics for the UK 2021 Census and beyond.

# 5.0 IDENTIFYING AND UNDERSTANDING ROAD-CONSTRAINED AREAS OF INTEREST (AOIS) THROUGH SPATIOTEMPORAL TAXI GPS DATA: A CASE STUDY IN NEW YORK CITY

The research presented in this chapter is an adapted version of the publication:

- Liu, Y., Singleton, A., Arribas-bel, D., and Chen, M. (2021). Identifying and understanding road-constrained areas of interest (AOIs) through spatiotemporal taxi GPS data: a case study in New York City. *Computers, Environment and Urban Systems.* 86. 101592. https://doi.org/10.1016/j.compenvurbsys.2020.101592

While the article details the study's other innovations, its main contributions to this thesis are summarised below.

1. *An enhanced methodological framework for urban mobility analysis is developed, in which space, time, and urban contexts are considered collectively.*
2. *Frameworks presented in previous articles (i.e., Chapter 3 and 4) are integrated to finalise the methodological framework proposed in this thesis.*
3. *The utility of the integrated framework has been examined by applying it to conduct an urban AOI detection task in the case study area.*

In this chapter, ***objective 2***, ***4***, ***5*** have been met.

## 5.1 ABSTRACT

Urban areas of interest (AOIs) represent areas within the urban environment featuring high levels of public interaction, with their understanding holding utility for a wide range of urban planning applications.

Within this context, our study proposes a novel space-time analytical framework and implements it to the taxi GPS data for the extent of Manhattan, NYC to identify and describe 31 road-constrained AOIs in terms of their spatiotemporal distribution and contextual characteristics. Our analysis captures many important locations, including but not limited to primary transit hubs, famous cultural venues, open spaces, and some other tourist attractions, prominent landmarks, and commercial centres. Moreover, we respectively analyse these AOIs in terms of their dynamics and contexts by performing further clustering analysis, formulating five temporal clusters delineating the dynamic

evolution of the AOIs and four contextual clusters representing their salient contextual characteristics.

## 5.2 INTRODUCTION

Urban areas of interest (AOIs) can be broadly defined as areas within an urban environment that attract people's attention, and which are often related to the generalisation of different types of urban economic activity (Hu et al., 2015; Yuan et al., 2012). AOIs are prevalently characterised by metrics describing high levels of public exposure and frequency of demand and are framed within the literature through the use of various terminologies including functionally-critical locations or urban hotspots (Cai et al., 2018; Qin et al., 2017; Zhou et al., 2015). It has been argued that a set of locations can be considered as an AOI when they involve various types of infrastructure that are of necessity for people's daily life, such as restaurants, primary workplaces, transport hubs, landmarks, entertainments, schools, and universities (Cai et al., 2018; Chen et al., 2019).

AOIs are also significant for urban transit planning, location-based services, and the management of daily travel since these areas can be utilised to assign higher priority in the allocation of public resources (Hu et al., 2015; Ma et al., 2019). Due to the wide range of applications of AOIs, successfully identifying and understanding the characteristics of such urban areas could provide a useful reference basis that benefits multiple stakeholders, including but not limited to tourism management, the identification of social functions, urban environmental study, urban vitality analysis, traffic planning, and public transit management (Alfeo et al., 2018; Chen et al., 2020; Kim, 2018; Ni et al., 2019; van der Zee et al., 2020; Zhou et al., 2019).

A traditional approach to investigate AOIs is primarily dependent on data derived from questionnaire-based methods such as field surveys or travel diaries. However, these approaches are labour-intensive, time-consuming, and error-prone, thus limiting their usefulness and applicability for large geographic areas (Yuan & Raubal, 2012). Following the rapid development and widespread use of location-based technology, large volumes of spatiotemporal data have been being collected either actively or passively, opening up new opportunities to map out and understand urban dynamics and reveal in-depth relationships between the urban fabric and the human mobility (Arribas-Bel, 2014; Qin et al., 2017). Numerous previous studies have implemented data mining techniques on heterogeneous data sources to identify urban AOIs, for instance, check-in data from social media, location data from mobile phones, and point of interest (POI) data from commercial location search engines (Chen et al., 2019; Hu et al., 2015; Kuo et al., 2018; Üsküplü et al., 2020; Xu et al., 2019; Yang et al., 2016).

Moreover, as a vital component of urban public transit, taxi trip data from GPS-enabled taxis have also been repurposed to define AOIs in many academic studies. For example, Garcia et al. (2018) utilised the origin-destination (OD) matrix extracted from 69 million records of taxi trips in NYC to identify popular taxi drop-off locations. Keler et al. (2020) investigated commuter-specific destination hotspots located in NYC by using Boro-taxi drop-off GPS points. Qin et al. (2017) applied a spatiotemporal clustering method on the taxi GPS points extracted from taxi trajectory data to detect urban hotspot areas in Wuhan. Cai et al. (2018) explored urban hotspots and computed their attractiveness index score through utilising one-week of taxi GPS trajectory data collected from 6599 taxis in Kunming.

According to the related studies (see Cai et al., 2018; Chen et al., 2019; Hu et al., 2015; Kuo et al., 2018), a typical bottom-up AOI detection framework can be summarised as comprising the following three phases:

1. **the hotspot detection phase**: identifying point clouds (i.e. the AOI prototype) through a density-based clustering method such as DBSCAN;
2. **the boundary-defining phase**: constructing closed polygons to define the AOI boundary;
3. **the analysis phase**: clarifying and exploring the characteristics of AOIs.

However, there are several aspects of these phases that require further consideration and improvement. Firstly, the hotspot detection phase is often limited to attributes in 2D planar space, which overwhelmingly concentrate on answering the question of ‘where’ but somewhat ignore the dynamic variation from the temporal aspect of AOIs. Given the fact that not every area in the urban environment is continuously recognised as a hotspot that attracts people's interest across all time periods, the omission of the temporal dimension may impose challenges in distinguishing different AOIs. For instance, office buildings and transport nodes (e.g. railway stations and airports) are defined as urban AOIs since they are both characterised by overall high traffic volume. However, the overall high traffic volume in the former AOI is more likely to be limited to two peak-time periods of commuting (i.e. morning and evening peak), whereas the latter AOI has a large traffic volume all day except at closing time.

A second research gap relates to those methods used in the boundary-defining phase. It is common to use a set of closed polygons to represent AOI geometrically, since using polygons can “provide simple and accessible representations for areas compared with clustered points” (Hu et al., 2015, 241). Many studies defined the border of an AOI by enclosing identified hotspots through convex hull or bounding box algorithms (L. Cai et al., 2018; Hollenstein & Purves, 2010). Although such methods are computationally efficient and convenient to apply, those polygons constructed through convex hulls are very likely to cover superfluous empty areas (Akdag et al., 2014). Other studies utilised the concave hull algorithms to define AOI boundaries, such as chi-shape algorithm (Hu et al., 2015) or alpha-shape (Chen et al., 2019; Kuo et al., 2018). However, concave hull algorithms are highly susceptible to parameter selection (e.g., λ in chi-shape and α in alpha shapes), which is embodied in small changes in parameter settings can make a significant difference in the shape of the calculated polygon (Chen et al., 2019). Since there is no authoritative guidance on how to obtain the optimal parameters, parameter selection is relatively subjective and can affect the quality of the results returned. Additionally, the feasibility of using polygons to represent AOIs remains to be discussed further, as such geometry only takes the impacts of human activities at AOIs into consideration, while the reshaping influences of urban structure on AOIs are neglected (Ma et al., 2019).

The third research gap relates to the analysis phase of the three-phase framework. After AOIs are identified, most existing studies mainly concentrate on their spatial distribution and morphology, but seldom do they explore those latent attributes, in terms of dynamic and contextual aspect, affecting the configuration and characterisation of an AOI. Such circumstance emerges more commonly in studies using traffic data (e.g., taxi GPS) as inputs since there is usually no extra information facilitating further in-depth analysis other than spatiotemporal coordinates.

The unique contribution of this study is the proposal of an enhanced three-phase analytical framework that improves on the aforementioned workflow within the context of a taxi GPS dataset collected for the case study area, i.e. New York City. These methodological enhancements aim to provide new substantive insight into the spatiotemporal dynamics and contextual characteristics of urban AOIs within the New York City and specifically for the Manhattan area. Firstly, we present urban AOIs as both a spatial and temporal phenomenon, implementing the ST-DBSCAN algorithm to detect spatiotemporal taxi trip hotspots. Secondly, in the process of defining the boundary of AOIs, the detected hotspots are linked to road geometry rather than enclosing them with polygons, formulating road-constrained AOIs. Finally, after the construction of AOIs, we utilise the H-K-mean clustering algorithm to conduct an in-depth analysis of these areas in respect of their spatiotemporal dynamics. Additionally, we extract several contextual variables from external open data sources and investigate the salient multidimensional characteristics of the identified AOIs through a geodemographic analysis.

The remainder of this paper proceeds as follows. Section 5.3 presents an overview of the case study area and the data used in this study, accompanying with a brief description introducing the main points of the data pre-processing and sampling. Section 5.4 provides a detailed explanation of the proposed three-phase analytical framework, ranging from essential theoretical context and algorithm introduction to detailed parameter settings and variable selection. Section 5.5 and its sub-sections respectively depict the results generated from each phase of the proposed framework, which is then followed by a summary of the work and a discussion of future directions in the context of known limitations.

## 5.3 DATA AND EXPLORATORY ANALYSIS

New York City (NYC) is the selected case study area. It is the most densely populated city within the US, with an estimated 8.4 million population distributed over a land area of approximately 784 km$^2$ (US Census Bureau, 2019). NYC is situated in the south-east of the state of New York on the US eastern seaboard, including five boroughs: Brooklyn, Queens, Manhattan, Bronx, and Staten Island. Across this area, the New York City Taxi and Limousine Commission (TLC), founded in 1971, is the agency responsible for licensing and regulating all segments of the taxi-related industry, primarily involving Medallion taxis (Yellow taxis), Street Hail Liveries (Green taxis), and For-Hive Vehicles (FHVs). In 2018, there were more than 300,000 TLC licensed vehicles servicing across the boroughs of NYC (TLC, 2018).

Data used in this study were extracted from the TLC database, involving taxi trip records jointly generated by both Yellow taxis and Green taxis in the whole year of 2015. The primary reason we used this 2015 dataset is that it is the latest and most accessible taxi trip data containing detailed GPS coordinates delineating individual taxi travels. Due to privacy issues, since the latter half of the year 2016, the TLC has replaced the provision of original taxi GPS coordinates by aggregating them into designated Taxi Zones, accordingly causing difficulties in analysing them through a density-based algorithm. It should also be mentioned that, although FHVs are occupying more and more proportions of taxi trips over the recent years (see TLC, 2018), trip record data from FHVs were excluded from this study since FHVs only began submitting trip records in Taxi Zone format after April 2015.

Data cleaning eliminated taxi trip records that were erroneous or out of bounds, such as GPS coordinates located outside of the study area or too far away from the nearest road network (⩾50 m); drop-off times that were earlier than pickup times; and unrealistic passenger counts. After the cleaning process, 150,134,156 taxi trip records were retained of the approximately 160 million original trips. *Figure 5.1* is a hexagon-binning map showing the spatial distribution of the retained taxi trip points. The majority of the NYC taxi trips (approximately 84% of the total taxi GPS points) are found within the Manhattan area, and as such, we subset the data to only focus on this area. Taking computational capacity into consideration, 1% of the samples (i.e. 1,190,646 taxi trips; 2,381,292 pickup and drop-off points), randomly selected from the pre-processed dataset, were subsequently inputted to the follow-up analysis.

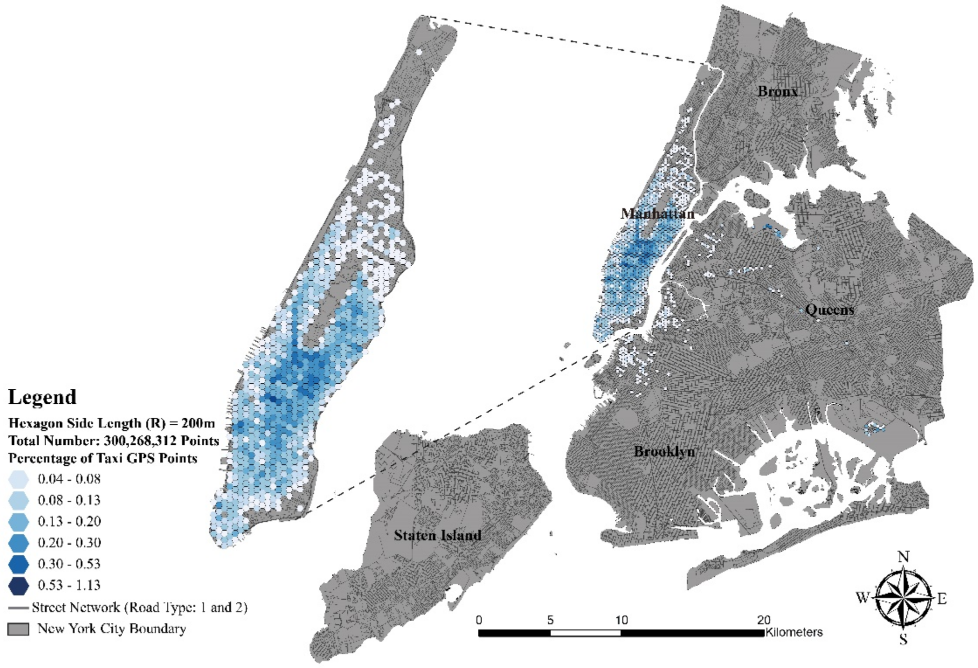


***Figure 5.1 Spatial distribution of hexagon-binning for pre-processed taxi GPS points in NYC, 2015***

The choice of a 1% random sample mirrors previous studies aiming to represent general human mobility patterns (González et al., 2008). However, to ensure the validity/stability of findings, multiple 1% random samples of the source taxi GPS data were iteratively selected and tested within our framework to examine the stability of the results. Specifically, we conducted an experiment in which 1% samples of the taxi GPS data were randomly selected multiple times, formulating several testing datasets. Then we examined the output results generated by inputting each of the testing datasets into the first two phases of our framework (introduced in *Section 5.4*). On the basis of this iterative experiment, we only retained the AOIs that can be identified every single run, assuring their stability and representativeness, and utilised them to carry out further investigations (i.e. the third phase).

## 5.4 METHODOLOGY FRAMEWORK

*Figure 5.2* presents a conceptual diagram illustrating an overview of the methodological framework proposed in this study. The framework consists of three phases, generally mirroring the conventional workflow mentioned in *Section 5.2*, but containing a methodological enhancement in each phase. Firstly, in the hotspot detection phase, we apply the ST-DBSCAN algorithm to the pre-processed taxi GPS data to detect the spatiotemporal hotspots of the taxi trips located in the case study area. The second phase is boundary-defining, which is responsible for converting the detected taxi hotspots into road-constrained AOIs through the K-Nearest Neighbour (KNN) algorithm that aggerates point clusters to their nearest road segments. The last phase of the framework is the analysis phase, which is comprised by two layers, i.e. dynamic layer and contextual layer, concentrating on extracting knowledge about the dynamic features and the contextual characteristics of the identified AOIs through clustering analysis that is carried out by using hierarchical k-means (H-K-means) algorithm. The remaining subsections respectively describe each phase of our proposed framework in more depth.

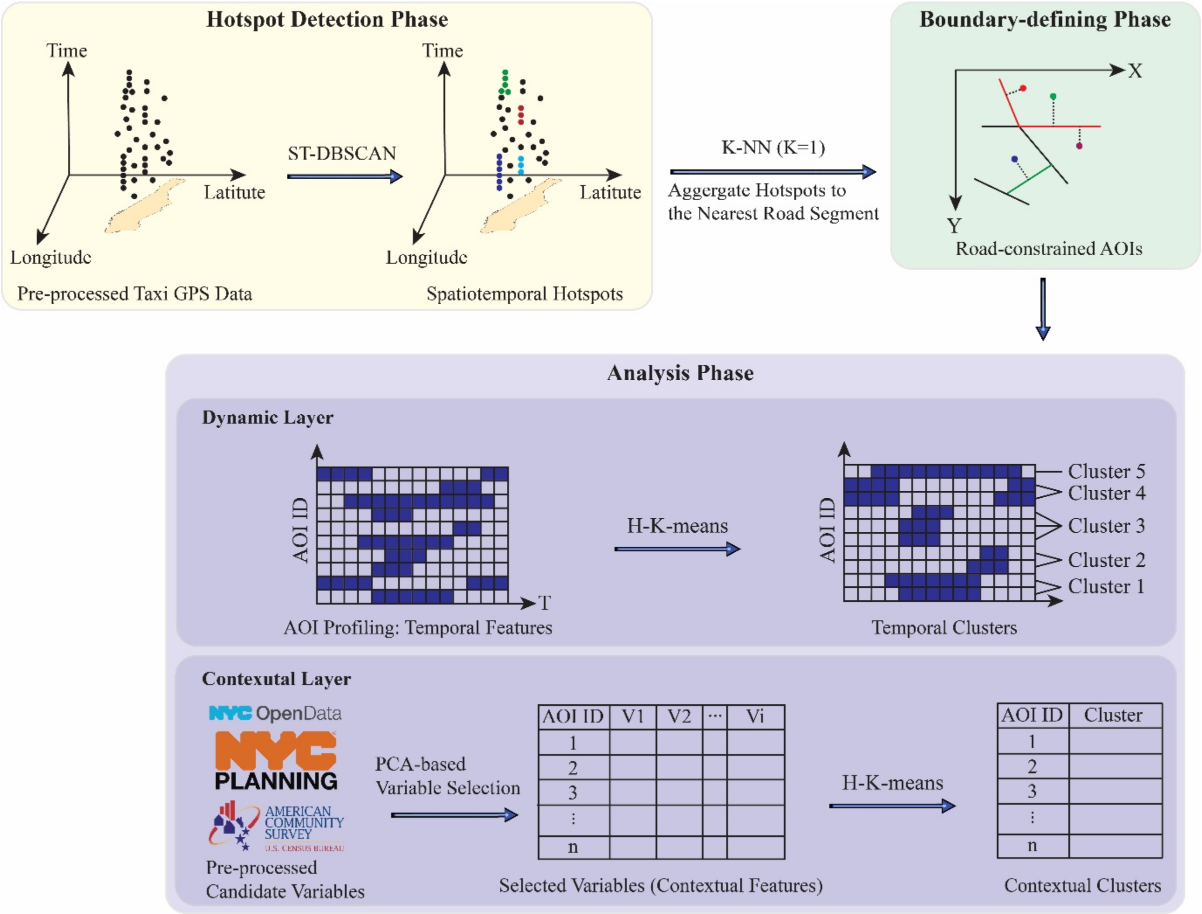


*Figure 5.2 Conceptual Diagram of the Proposed Analytical Framework*

## 5.4.1 THE HOTSPOT DETECTION PHASE

DBSCAN (density-based spatial clustering for applications with noise) is a commonly applied density-based clustering algorithm for hotspot detection (Ester et al., 1996), which is configured by two parameters: Epsilon (Eps), the search radius based on a user-defined distance measure, and MinPts, the minimum points within the Eps radius. These parameters jointly determine a minimum density threshold. Point clusters are constructed at locations in which the point density exceeds the specified threshold.

Given the advantages in distinguishing between outliers and clustered points through a relatively simple parameter setting, DBSCAN and its extensional algorithms have been widely employed by many studies to detect hotspots from large-scale geo-referenced data. For instance, Xu et al. (2019) applied DBSCAN to POI data extracted from the Baidu map API to identify the spatial agglomeration of POI-forming functional regions within Wuhan. Zhang et al. (2016) applied Grid and Kd-tree DBSCAN (GD-DBSCAN) algorithms on taxi pickup locations to identify taxi demand hotspots in Shanghai. Chen et al. (2019) implemented Hierarchical-DBSCAN (HDBSCAN) to geotagged photo data from Flickr to capture the dynamic characteristics of urban AOIs in the inner London area.

Due to the nature of DBSCAN, i.e. using only one distance (Eps) to measure similarity, DBSCAN and most of the abovementioned DBSCAN-based algorithms merely consider spatial attributes in the process of detecting hotspots, resulting in the omission of temporal attributes (Birant & Kut, 2007). However, the urban environment is a complex and constantly changing system, involving various components with multifaceted relationships and interactions (Batty, 2013b). Such complexities can be reflected in the changeable type, intensity and distribution of urban resources at different times and locations, referring to both urban dynamics and human mobility (Song et al., 2019). From the perspective of urban AOI, not all areas of the urban environment can be recognised as a hotspot over all time periods (Chen et al., 2019; Hu et al., 2015). We argue here that in many other studies that exclude a temporal dimension, this leads to the capture of only a partial representation of urban AOIs, hence, limiting our understanding of urban functions and their underlying spatiotemporal dynamics.

In order to consider spatial and temporal dimensions simultaneously, ST-DBSCAN (Spatial-temporal Density-Based Spatial Clustering of Applications with Noise), a modified extension of the traditional DBSCAN designed to analyse spatiotemporal data (Birant & Kut, 2007; Shi & Pun-Cheng, 2019), was employed to detect taxi hotspots. Generally, the primary convenience of ST-DBSCAN is that it can identify spatiotemporal clusters with arbitrary shape and noise points (Cheng et al., 2014). More specifically, according to Birant & Kut (2007), ST-DBSCAN surpasses normal DBSCAN in terms of the three following advantages: firstly, it provides cluster discoverability according to the non-spatial, spatial, and temporal values of objects; secondly, it can effectively detect noise points even when various cluster densities exist; thirdly, it improves clustering quality even if clusters are adjacent to each other. Numerous studies have highlighted the utility of ST-DBSCAN for handling complex spatiotemporal data and the application to many areas of research (see Chen et al., 2020; Iliopoulou et al., 2020; Shen & Cheng, 2016).

In common with other DBSCAN-based algorithms, ST-DBSCAN also requires predefined parameters before application. According to Birant & Kut (2007), MinPts can be determined by a heuristic method (*Equation 5.1*).

$$MinPts \approx \ln(n) \tag{5.1}$$

***Equation 5.1 n indicates the total number of observations. In this study, the observations are the 2381292 taxi GPS points located in the Manhattan area, NYC. MinPts is accordingly equal to 15.***

To define Eps (i.e. Eps1), a k-distance graph (*Figure 5.3*) delineates ascendingly sorted distances to the k-nearest neighbours for each object (where k = MinPts). An appropriate Eps value can be selected from the “first valley” of the graph (Birant & Kut, 2007, 214), where there is “an obvious and abrupt change” (Shi & Pun-Cheng, 2019, 7). For this case, we selected 70 metres as the Eps value based on this heuristic method.

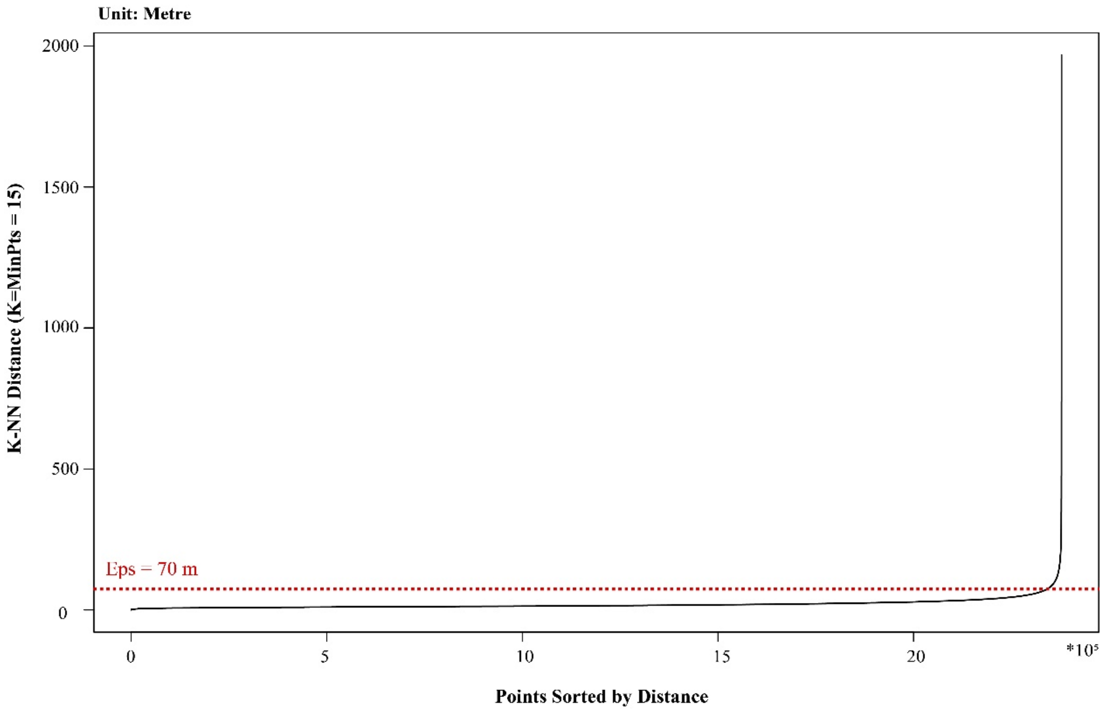


***Figure 5.3 KNN distance graph (K=15) used to determine Eps (Eps = 70m).***

In addition to MinPt and Eps, Birant & Kut (2007) introduced a second epsilon parameter, i.e. Eps2, to define the search radius for the temporal dimension. Similar to Eps1 mentioned above, a larger value for Eps2 results in broader clusters, while a smaller value generates narrower clusters, delineating a finer temporal resolution. Here we set Eps2 equal to 0.25, representing a 15-minute search radius. The primary reason for choosing this temporal resolution was approximately referenced by the average taxi trip time (i.e. 14.8 minutes), along with the consideration of a convenient result display and interpretation.

### 5.4.2 THE BOUNDARY-DEFINING PHASE

As discussed in *Section 5.2*, it is typical in the delineation of urban AOI use an enclosed polygon to define the boundary of the identified point clusters (i.e. hotspots) to formulate AOIs. However, there are growing appeals for alternative representations. Firstly, despite convex-hull and concave-hull algorithms being commonly used in many related studies, both have drawn criticism. The former is sometimes challenged for creating redundant empty areas (Akdag et al., 2014), whilst the latter is susceptible to the choice of parameters, thus involving high subjectivity (Cai et al., 2018; Chen et al., 2019; Hu et al., 2015). Secondly, it can be argued that defining AOI's boundary using enclosed polygons fails to appropriately account for the potential impacts of urban morphology on shaping AOIs since they "only considered the distribution characteristics of data that capture human activities" (Ma et al., 2019, 2). Thirdly, because of the uncertainties caused by inevitable measurement error of GPS, offset between the observed location and the actual location may be a feature of the data inputs: although vehicle GPS location should align with the road network (Yang & Gidófalvi, 2018). Taking such concerns into account, we argue that the road network is a more organic carrier of the detected point clusters, which therefore can be employed to define the boundary of urban AOIs, particularly for an application utilising taxi data since they bound these patterns of mobility (Ma et al., 2019; Yuan et al., 2012).

After projecting the detected taxi trip hotspots onto the 2D plane containing the road network, a KNN algorithm was adopted to aggregate these points to their nearest road segment, formulating road-constrained AOIs. It should be mentioned that, if the road segments are topologically connected, they are considered as one AOI, ensuring that there are no overlapping AOIs.

### 5.4.3 THE ANALYSIS PHASE

After AOIs are identified, most existing studies mainly concentrate on their spatial distribution or temporal evolution pattern, but seldom explore the latent attributes affecting the configuration of AOIs. Such circumstance emerges more commonly in studies using traffic data (e.g. taxi GPS) as inputs since there is no adequate information facilitating further analysis other than spatiotemporal coordinates (i.e. longitude, latitude, and time).

This phase consists of two layers, i.e. dynamic layer and contextual layer, which are designed to extract useful information about the detected AOIs through further clustering analysis from both dynamic and contextual perspectives. The clustering results generated through each layer will be presented and discussed in *Section 5.5*.

## THE DYNAMIC LAYER

Since the spatiotemporal hotspots were aggregated to street segments to form the road-constrained AOIs, each AOI can be regarded as proportionally containing at least one or more point clusters over a temporal sequence. Such temporal sequences depict various dynamic patterns exhibited by AOIs. Some AOIs, for instance, only appear at a particular time of day, while others have greater longevity.

In this context, the hierarchical k-means (H-K-means) clustering algorithm was adopted to classify AOIs into groups based on the similarities in their dynamic pattern. H-K-means provides a hybrid of both hierarchical clustering and k-means clustering and comprises three steps: first agglomerative hierarchical clustering is implemented to the data to create a k number of clusters; secondly, the centroids (i.e. the mean value) are calculated for each cluster; finally, these computed centroids are used as the centroid initialisation for the k-means algorithm (Arai & Ridho Barakbah, 2007; Chen et al., 2005).

The optimal number of clusters (k) is determined by Gap Statistics, introduced by Tibshirani et al.(2001) (*Equation 5.2*), which compares the total within-cluster variation for different values of k with their expected values under “an appropriate null reference distribution of the data” (p.412).

$$\boldsymbol{Gap_n(k) = \ E_n^*\{lo\,g(W_k)\} - log(W_k)} \qquad \boldsymbol{(5.2)}$$

***Equation 5.2 $E_n^*$ denotes the expectation under a sample size n from the reference distribution. $W_k$ is the pooled within-cluster sum of squares around the cluster means. The estimation of the optimal clusters k will be the value that maximises $Gap_n(k)$.***

The clustering results could portray the picture of ‘urban pulse’ answering questions, such as where AOIs are and when they emerge and disappear.

## THE CONTEXTUAL LAYER

As mentioned previously, due to a lack of further detail on journey purpose, it is insufficient to solely use taxi GPS data to understand the characteristics of identified urban AOIs, for example, to explore what specific features of these AOIs attract taxi passengers and further affect their travel behaviour. In order to gain greater insight into the identified AOIs and improve their interpretability, it is helpful to import supplementary data capturing some contextual attributes that potentially influence individual’s travel behaviour, as well as to apply the corresponding analytical method to extract meaningful information about the salient characteristics of urban context from these datasets (Liu & Cheng, 2020). In this study, we utilised a geodemographic classification methodology to extract salient contextual characteristics exhibited by each identified AOIs.

Geodemographic classification is an analytical framework that provides categorical summaries of multidimensional socioeconomic, demographic and built environment characteristics for small geographic areas (Singleton et al., 2017). The detailed processes to build a geodemographic classification and the advantages of such classification are well documented (see Alexiou, 2016; Harris et al., 2005; Leventhal, 2016; Singleton et al., 2017). Geodemographic classification has an expansive and international lineage, with utility for both private and public sectors applications and for various geographic extents (Gale et al., 2016; Singleton & Longley, 2015; Singleton & Spielman, 2014). The implementation of geodemographic classification for this study can be regarded as a bespoke application designed to differentiate urban AOIs in Manhattan, NYC.

Numerous studies have investigated the linkage between urban context and travel behaviour over the past decades (Cervero & Kockelman, 1997; Dieleman et al., 2002; Ewing & Cervero, 2010; Ma et al., 2014; Pan et al., 2009). For instance, Ewing & Cervero (2010) found that an individual's travel mode choice can be influenced by the demographic and socioeconomic characteristics of the household as well as the built environment characteristics of the surrounding area, which provided additional ‘D’ variables to the well-established ‘three Ds’ principle (i.e. density, diversity, and design) introduced by Cervero & Kockelman (1997). More recently, Liu et al. (2020) presented a study containing a systematic literature review over 29 contemporary studies related to the impacts of the urban context on people's travel behaviour. They pointed out that although most of the studies still aligned with the 'D' variables, some of the variables they used have beyond the scope of the traditional 'D' variables, implying broader or context-specific considerations. They further categorised those variables into four domains, namely, Land Use and Built Environment (LB), Location and Accessibility (LA), Socioeconomic and Demographic (SD), and Transit-related (T), guiding the variable selection for their research about creating a contextual transit-oriented development (TOD) typology for NYC.

Given the overlapping research context and case study area, we acknowledged the systematic literature review conducted by Liu et al. (2020) and utilised their presented four variable-domains as a reference to guide our initial variable selection. With extra consideration of the availability and consistency of the data (note that the 2015 taxi data were used in this study), 52 candidate variables were initially selected (*Table 5.1*), which were extracted from the following four open data sources, i.e. American Community Survey (ACS), NYC Open Data, Smart Location Database (SLD), and NYC Planning.

Inevitably, such a large number of candidate variables and the resulting high dimensionality we argue would lead to harmful effects in the following cluster analysis. Numerous studies have discussed the negative impact caused by the high dimensionality on the clustering performance, which is also known as 'the dimensional curse', including dramatically increasing the demand for computational power and storage capacity, lowering the efficiency of the clustering algorithm, impairing the output interpretability (Guyon & Elisseeff, 2003; Renjith et al., 2020; Weber et al., 1998). Apart from the potential threats from high dimensionality, multicollinearity between the candidate variables is also problematic (Sambandam, 2003). The existence of variable pairs with high correlation is harmful to the clustering performance since such dimensions are effectively assigned more weight during the clustering process (Harris et al., 2005; Sambandam, 2003).

In order to alleviate the adverse impacts of high dimensionality and multicollinearity, we employed a principal component analysis (PCA)-based variable selection framework, proposed by Liu et al. (2019), to "select the smallest possible subset of variables that can represent the main variance within a universe of potential inputs being considered" (Liu et al., 2019, 253). PCA is a feature transformation methods, which has a long history of being applied across multiple disciplines to accomplish dimensionality reduction (Ma et al., 2019; Malhi & Gao, 2004; Webber, 1975). Through linear transformation, PCA finds a set of orthogonal space to maximise the variance in each coordinate axis (Abdi & Williams, 2010), to project high-dimensional data onto a low-dimensional representation, while preserving the original data features as much as possible (Ma et al., 2019). The variable-selection framework proposed by Liu et al. (2019) consists of multiple stages, that not only select variables according to the average contribution of the input variables to the principal components (PCs) but also filters variables based on their correlation between each other. The minimum spanning tree (MST) was integrated into the framework to filter out variable pairs with relatively high correlation (correlation coefficient ≥ ± 0.75). Additionally, their framework also considers such impacts on overall clustering quality, which provides additional utility for this study. A full description of the PCA-based variable selection framework, its properties, parameter settings, and relative strengths and weaknesses is beyond the scope of this section however presented by Liu et al. (2019).

| Data Sources | Code | Domain | Variables Title | Description | Checklist |
|---|---|---|---|---|---|
| ACS | B01001 | SD | Age: 0- 4 | % of population aged between 0 and 4 | |
| | | SD | Age: 5 - 14 | % of population aged between 5 and 14 | * |
| | | SD | Age: 15 - 19 | % of population aged between 15 and 19 | |
| | | SD | Age: 20 - 24 | % of population aged between 20 and 24 | |
| | | SD | Age: 25 - 44 | % of population aged between 25 and 44 | * |
| | | SD | Age: 45 - 64 | % of population aged between 45 and 64 | |
| | | SD | Age: 65 & above | % of population aged 65 and above | * |
| | B08303 | LA | TTtW: < 5 | % of workers whose travel time to work is less than 5 minutes | |
| | | LA | TTtW: 5 - 14 | % of workers whose travel time to work is between 5 and 14 minutes | * |
| | | LA | TTtW: 15 - 29 | % of workers whose travel time to work is between 15 and 29 minutes | |
| | | LA | TTtW: 30 - 44 | % of workers whose travel time to work is between 30 and 44 minutes | * |
| | | LA | TTtW: 45 - 59 | % of workers whose travel time to work is between 45 and 59 minutes | * |
| | | LA | TTtW: > 60 | % of workers whose travel time to work is longer than 60 minutes | |
| | B15003 | SD | EA: No school | % of population have no qualifications | * |
| | | SD | EA: Elementary school | % of population attained kindergarten to 5th grade | |
| | | SD | EA: Middle school | % of population attained 6th to 8th grade | |
| | | SD | EA: High school | % of population attained 9th to 12th grade | * |
| | | SD | EA: College / Bachelor | % of population attained college or bachelor's degree | |
| | | SD | EA: Master / Doctorate | % of population attained master or doctorate degree | * |
| | B19013 | SD | Median Income | Household median income in the past 12 months | |
| | B24010 | SD | OT: M.B.S.A. | % of workers in management, business, science, and art occupations | * |
| | | SD | OT: S. | % of workers in service occupations | * |
| | | SD | OT: S.O. | % of workers in sales and office occupations | |
| | | SD | OT: N.C.M. | % of workers in natural resources, construction, and maintenance occupations | |
| | | SD | OT: P.T.M. | % of workers in production, transportation, and material moving occupations | |
| | B01003 | LB | Population Density | Number of populations by area (km$^2$) | |
| SLD | D4a | LA | D4a | Distance from the population-weighted centroid to the nearest transit stop (meters) | |
| | D1c | LA | Job Density | Gross employment density (jobs/acre) | * |
| | D4d | T | Transit Frequency | Aggregate frequency of transit service per square mile | * |
| NYCOD | CSCL | LB | Intersection Density | Number of street intersections by road length | * |
| | STC | LB | Tree Density | Number of street trees by road length | |
| | Bicycle | T | Bike Facilities | Number of Citi-bike, bicycle routes and parking shelters by road length | * |
| | Bus | T | Bus Facilities | Number of bus stops by road length | * |
| NYCP | MapPLUTO | LB | LU: R | % of building & poi categorised as residential use | * |
| | | LB | LU: C | % of building & poi categorised as commercial use | * |
| | | LB | LU: TU | % of building & poi categorised as transport and utility | * |
| | | LB | LU: PSCI | % of building & poi categorised as public service and institution | * |
| | | LB | LU: OSR | % of building & poi categorised as open space and recreation | |
| | | LB | LU: V | % of building & poi categorised as vacant | |
| | | LB | LU: Mixed | % of building & poi categorised as mixed-use | * |
| | | LB | FAR | Floor area ratio (gross floor area/area of plot) | * |
| | | LB | Landmark Density | Number of landmarks by road length | * |
| | | LB | BT: Detached | % of building unit categorised as detached | |
| | | LB | BT: Attached | % of building unit categorised as attached | * |

| | | | | |
|---|---|---|---|---|
| | LB | BT: Semi-Attached | % of building unit categorised as semi-attached | |
| | LB | BT: Apartment | % of building unit categorised as apartment | |
| | LB | YB: 2010 / Later | % of building built in 2010 or later | * |
| | LB | YB: 2000 – 2009 | % of building built between 2000 and 2009 | * |
| | LB | YB: 1980 – 1999 | % of building built between 1989 and 1999 | |
| | LB | YB: 1960 – 1979 | % of building built between 1960 and 1979 | |
| | LB | YB: 1940 – 1959 | % of building built between 1940 and 1959 | |
| | LB | YB: 1939 / Earlier | % of building built in 1939 or later | * |

***Table 5.1 Initial 52 candidate variables and selected variables from the PCA-based variable selection framework proposed by Liu et al. (2019).***

Many of the variables related to specific points of interest, and as such were aggregated into the road-constrained AOIs using the KNN algorithm (K=1) that was applied in the boundary-defining phase. Values of some variables, such as Floor Area Ratio (FAR), were averaged during the aggregation process, whereas others (e.g. many of the ACS variables) were aggregated based up their intersection with the AOI. The last column of *Table 5.1* shows the checklist indicating the contextual variables that were selected after the application of the PCA-based selection method. 27 out of 52 candidate variables were included as inputs.

After the selected variables were assembled for each AOI, the Box-Cox transformation (Box & Cox, 1964) (*Equation 5.3*) was employed to convert abnormally distributed variables to approximate normality. Furthermore, since the variables are measured on different scales, z-scores (*Equation 5.4*) were applied as a method of standardisation.

$$\boldsymbol{x_i'} = \begin{cases} \dfrac{\boldsymbol{x_i^{\lambda} - 1}}{\boldsymbol{\lambda}} & , \quad \boldsymbol{if\ \lambda \neq 0}; \\ \boldsymbol{\log x_i} & , \quad \boldsymbol{if\ \lambda = 0}. \end{cases} \tag{5.3}$$

***Equation 5.3 $x_i'$ is the transformed value; λ ranges from -5 to 5, which can be estimated using the profile likelihood function to achieve 'optimal value'.***

$$\boldsymbol{z_i} = \frac{\boldsymbol{x_i - \mu}}{\boldsymbol{\sigma}} \tag{5.4}$$

***Equation 5.4 $z_i$ is the standardised value, $x_i$is an original value, $\mu$ is the mean of $x_i$, and $\sigma$ is the standard deviation from the mean.***

The variables were subsequently clustered through H-K-means, and the Gap Statistics mentioned in *Section 5.4.3*. were utilised once again to define the optimal number of clusters. The clustering result provides summary measures of the urban context, revealing the salient characteristics distinguishing AOIs from other urban areas. Furthermore, in order to improve the interpretability of revealed clusters, it is typical to assign shorthand names and written "pen portraits" descriptions for each of the clusters within the built geodemographic classification (Alexiou, 2016; Harris et al., 2005).

## 5.5 RESULTS

### 5.5.1 IDENTIFIED AOIS IN MANHATTAN, NYC

*Figure 5.4* presents the spatial distribution of the 31 identified urban AOIs. These areas are featured by major transportation hubs, such as the West 39th Street Ferry Terminal (AOI 18), Pennsylvania Station (AOI 15), and Grand Central Station (AOI 16); famous cultural venues, such as the Lincoln Centre for the Performing Arts (AOI 26), the Whitney Museum of American Art (AOI 8), and the Metropolitan Museum of Art (AOI 30); open spaces, such as Central Park (AOI 24) and Union Square (AOI 6); and some other tourist attractions, prominent landmarks, and commercial centres, such as Columbus Circle (AOI 25), the Empire State Building (AOI 13), the Rockefeller Centre (AOI 20), and the One World Trade Centre (AOI 1).

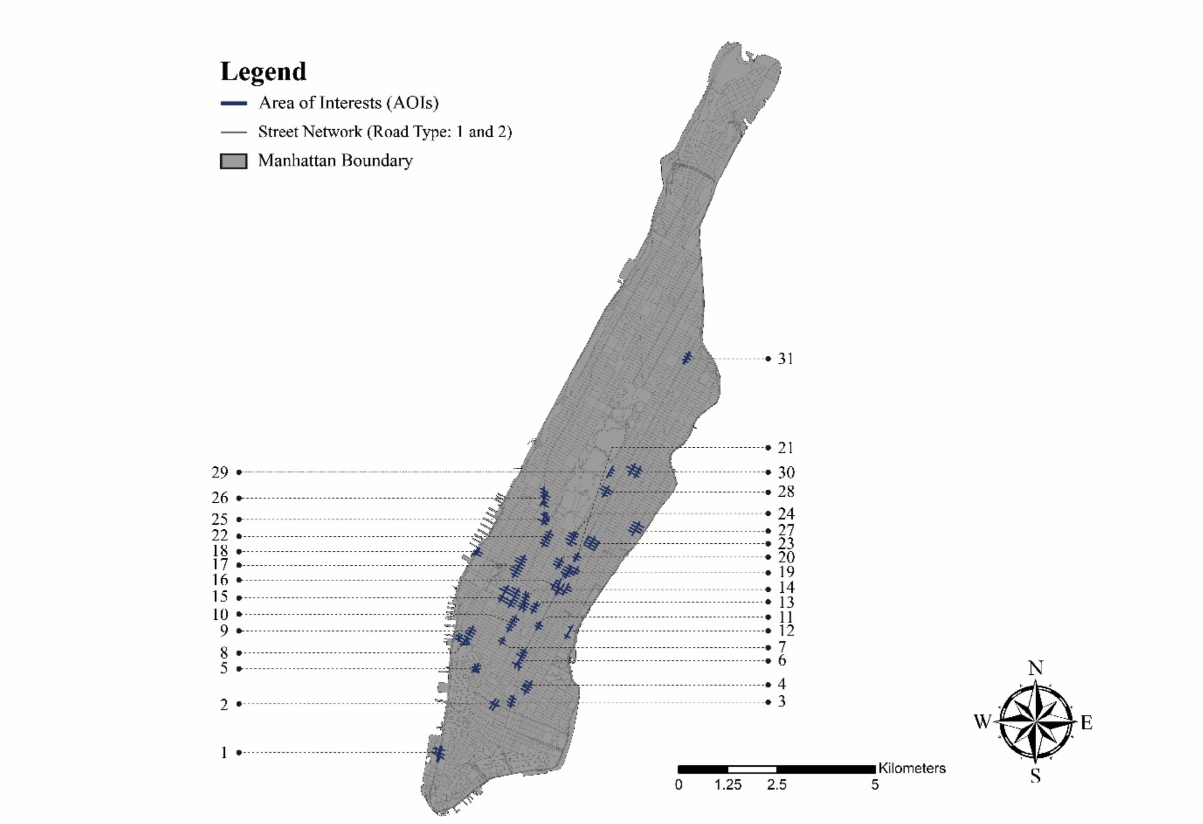


***Figure 5.4 Geographic Distribution of 31 Identified AOIs in NYC.***

### 5.5.2 DYNAMIC FEATURES OF AOIS

As discussed earlier, an advantage of the ST-DBSCAN algorithm is that in addition to the spatial attributes of the urban AOI, the temporal characteristics are also preserved, enabling further exploration of their dynamic evolution throughout the day. As such, the 31 identified urban AOIs were further classified into five temporal clusters representing different types of dynamic patterns. *Figure 5.5* contains a sorted heatmap presenting the temporal distribution of the clustering results, followed by a map showing their spatial distribution (*Figure 5.6*). Based on such patterns, furthermore, shorthand names and descriptive profiles were generated for each AOI cluster.

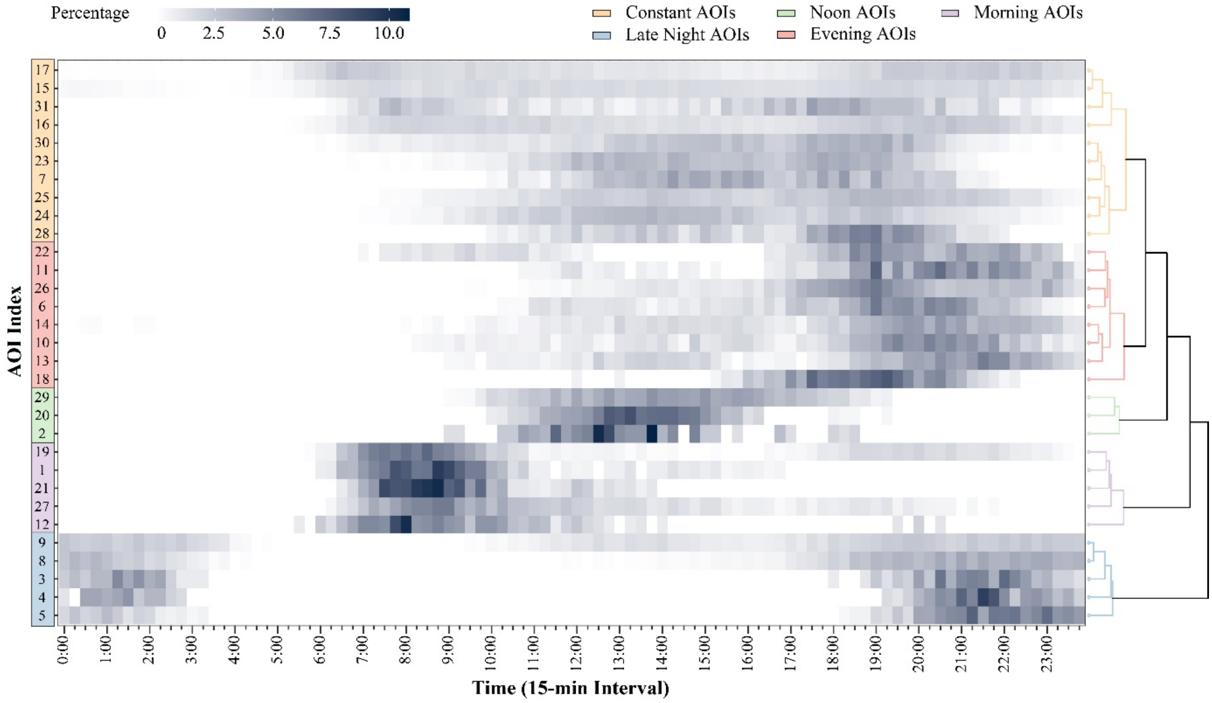


***Figure 5.5 The Temporal Distribution of AOIs (by 15-Minute Interval).***

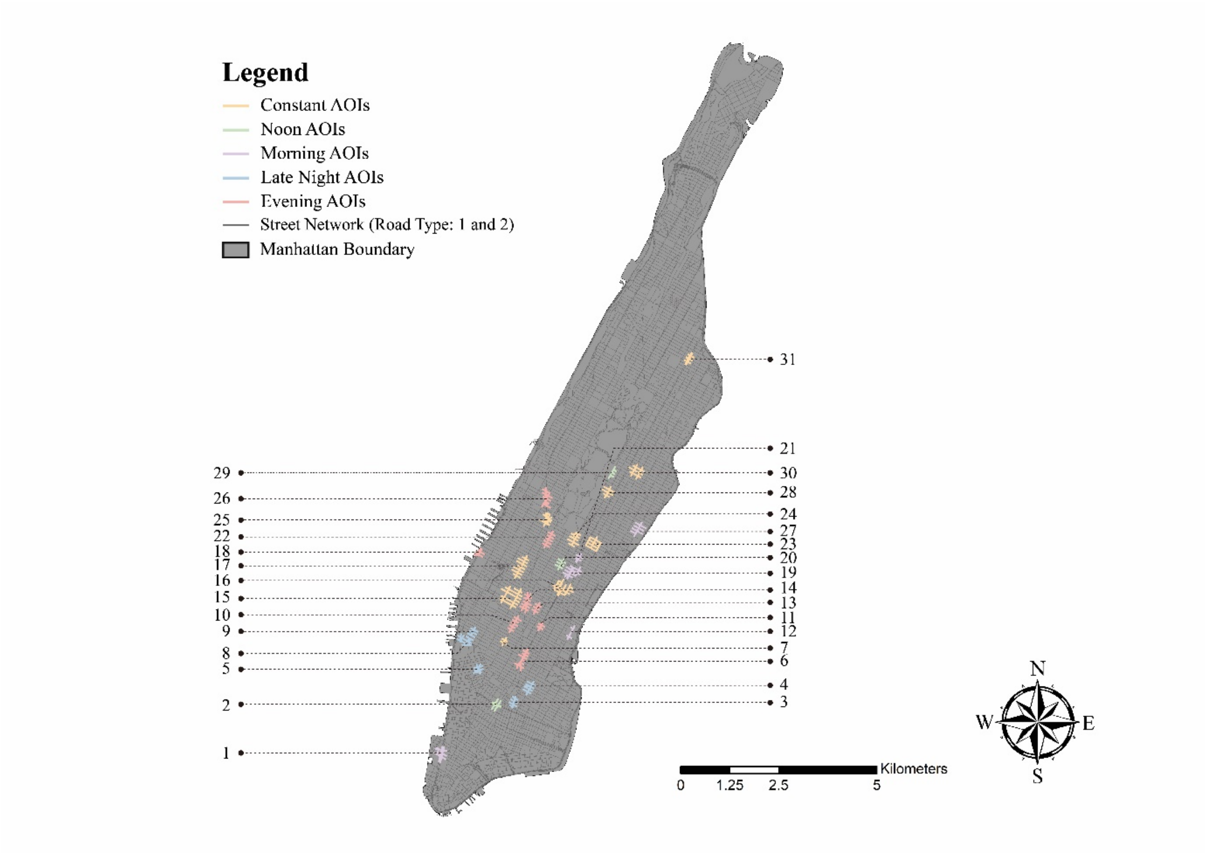


*Figure 5.6 Geographic Distribution of Five Temporal Clusters.*

## CONSTANT AOIS

Most AOIs classified in this group are located in Midtown of Manhattan, covering various major transit hubs (e.g. Pennsylvania Station, AOI 15), and integrated commercial, retail centres (e.g. Rockefeller Centre, AOI 20). AOIs from this cluster are continuously exposed to a high volume of taxi activity lasting approximately the whole day, and as such is one of the most stable AOIs in Manhattan.

## NOON AOIS

AOIs of this group distribute evenly across Manhattan from north to south, with no specific agglomerations. These AOIs record gradually increased taxi flow at around 9:30, a peak at high noon, and a reduction after 17:30, which could be affected by business opening hours.

## MORNING AOIS

Experiencing high taxi travel demand between 6:00 and 10:30 in the morning, AOIs in this group are primarily identified in areas proximal to major commercial centres (e.g. One World Trade Centre, AOI 1) or public institutions, such as hospitals and medical institutions (e.g. Weill Cornell Medical Centre, AOI 27), which could indicate a typical morning peak commuting pattern.

#### LATE NIGHT AOIS

AOIs from this group are mainly identified in south Manhattan. AOIs begin to emerge after 17:30 and continuously attract taxi travels until 3:00 in the early morning of the next day, which might either suggests a recreational pattern reflecting the nightlife in Manhattan or residential-oriented pattern, or combination of both.

#### EVENING AOIS

AOIs of this group are diffuse over Manhattan from Midtown (Union Square, AOI 6) to the Upper West Side (Lincoln Square, AOI 26). These AOIs emerge at around 17:00, peak at around 21:30, and entirely disappear before midnight, indicating an off-peak recreational-oriented travel pattern.

### 5.5.3 THE CONTEXTUAL FEATURE OF AOIS

*Figure 5.7* presents a map illustrating the spatial distribution of the geodemographic classification that was generated from applying H-K-means to the 27 variables retained by the PCA variable selection. The identified 31 AOIs were classified into four clusters, i.e. Major transit hubs, High-rise integrated commercial, Residential heritage mix, and Public institution mix, delineating four different salient multidimensional characteristics extracted from the contextual variables.

Index scores (i.e. x/x̄ *100) were computed for the retained variables and were displayed within each cluster in Figure 8. These scores reflect the (over-) underrepresentation of a target attribute compared to the average value (i.e. a score of 100). An index score of 50 would be quivalent to a rate that is half the average, and 200 would be double. Using both the map and scores, descriptive profiles were generated.

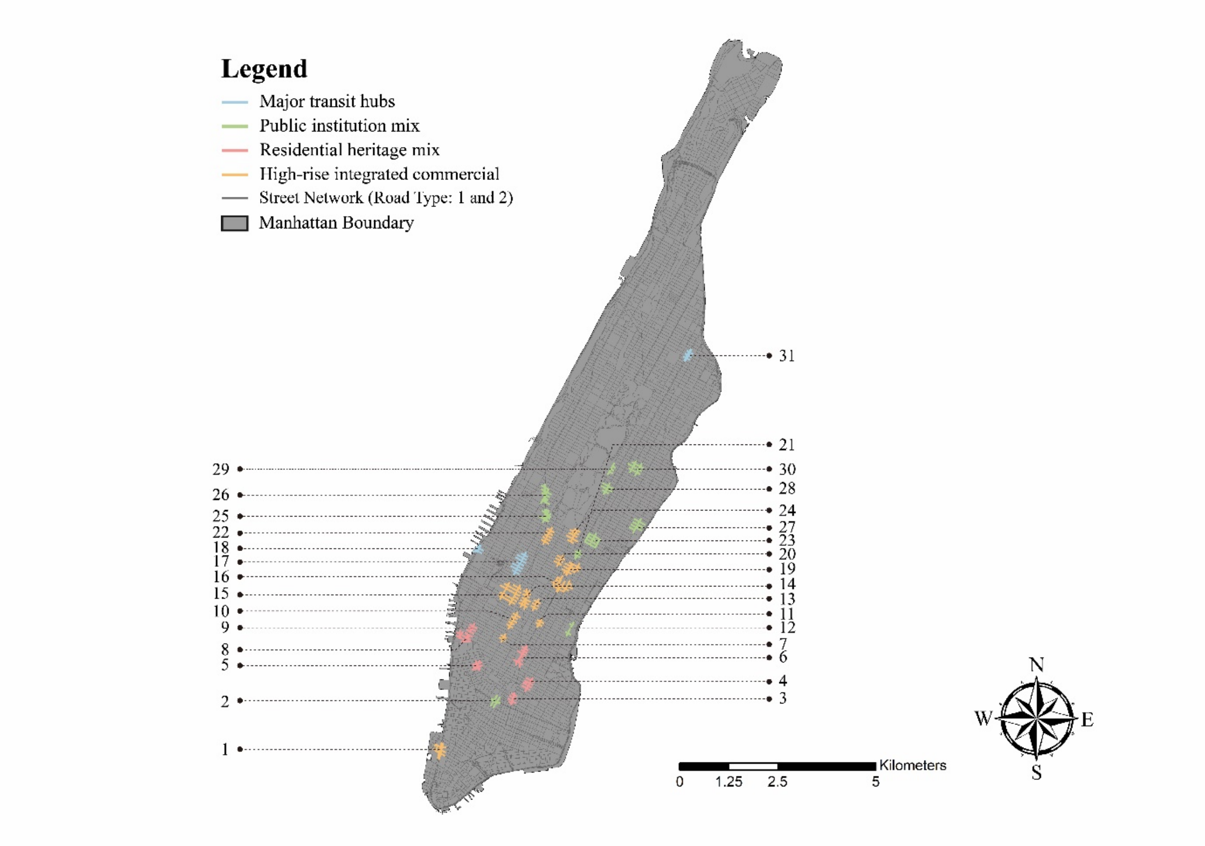
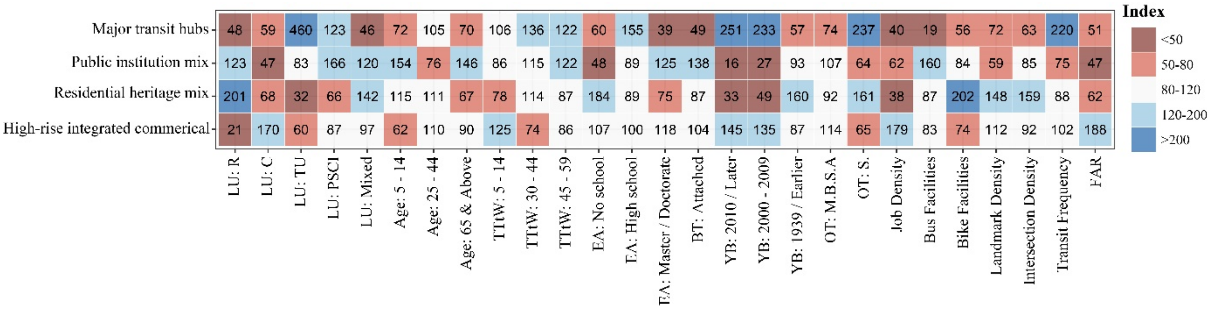


*Figure 5.7 Geographic Distribution of Four Contextual Clusters.*



| | LU: R | LU: C | LU: TU | LU: PSCI | LU: Mixed | Age: 5 - 14 | Age: 25 - 44 | Age: 65 & Above | TTtW: 5 - 14 | TTtW: 30 - 44 | TTtW: 45 - 59 | EA: No school | EA: High school | EA: Master / Doctorate | BT: Attached | YB: 2010 / Later | YB: 2000 - 2009 | YB: 1939 / Earlier | OT: M.B.S.A | OT: S. | Job Density | Bus Facilities | Bike Facilities | Landmark Density | Intersection Density | Transit Frequency | FAR |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Major transit hubs | 48 | 59 | 460 | 123 | 46 | 72 | 105 | 70 | 106 | 136 | 122 | 60 | 155 | 39 | 49 | 251 | 233 | 57 | 74 | 237 | 40 | 19 | 56 | 72 | 63 | 220 | 51 |
| Public institution mix | 123 | 47 | 83 | 166 | 120 | 154 | 76 | 146 | 86 | 115 | 122 | 48 | 89 | 125 | 138 | 16 | 27 | 93 | 107 | 64 | 62 | 160 | 84 | 59 | 85 | 75 | 47 |
| Residential heritage mix | 201 | 68 | 32 | 66 | 142 | 115 | 111 | 67 | 78 | 114 | 87 | 184 | 89 | 75 | 87 | 33 | 49 | 160 | 92 | 161 | 38 | 87 | 202 | 148 | 159 | 88 | 62 |
| High-rise integrated commerical | 21 | 170 | 60 | 87 | 97 | 62 | 110 | 90 | 125 | 74 | 86 | 107 | 100 | 118 | 104 | 145 | 135 | 87 | 114 | 65 | 179 | 83 | 74 | 112 | 92 | 102 | 188 |

*Figure 5.8 Index Scores by Four Contextual Clusters.*

## MAJOR TRANSIT HUBS

AOIs of this cluster cover primary public transit nodes in Manhattan, predominantly manifested by the high level of transit frequency and the surrounding transport-oriented buildings and facilities. These nodes facilitate inter-/intra city flows, including a ferry terminal (AOI 18), railway station (AOI 31), and an interstate bus terminal (AOI 17).

#### HIGH-RISE INTEGRATED COMMERCIAL

Commercial-use skyscrapers are very likely to be located in proximity to AOIs from this group since the average floor area ratio is dramatically higher than the average, exemplified by the high-rise office buildings near the One World Trade Centre (AOI 1). These areas are likely to be the leading employment destinations in Manhattan due to the short travel-to-work time and the high level of the job density.

#### RESIDENTIAL HERITAGE MIX

AOIs of this cluster mainly agglomerate in Midtown Manhattan. Areas approximating to these AOIs are likely to contain many old buildings built earlier than 1939 and have had been primarily utilised for residential purposes, while the mixed-use buildings and facilities are also much in evidence (e.g. multipurpose areas near the Pennsylvania Station, AOI 15). Landmark destinations within these AOIs are significantly higher than the regional average, which may be attractive for tourists and travellers.

#### PUBLIC INSTITUTION MIX

These AOIs are prevalently located in Upper Manhattan, although they can be found across Manhattan. Buildings or facilities located near this type of AOIs are likely to be used for many purposes, including residential usages, retailing markets, culture venues, public services (e.g. hospitals), and research or educational institutions.

### 5.5.4 INTEGRATED SPATIOTEMPORAL DYNAMICS AND CONTEXT

The main objective of this study was to understand how AOIs are represented both from contextual and spatiotemporal perspectives. Accordingly, the intersection of the temporal and contextual classifications was analysed through cross-tabulation, and the result presented in *Figure 5.9*. The heatmap illustrates the frequency and proportion of AOIs categorised at the intersection of the two typologies. The result indicates a general correspondence between the two classifications with some emerging differences.

As the major gateways of NYC and interchange platforms facilitating multimodal inter-/intra-city journeys, two out of three AOIs from the ‘Major transit hubs’ unsurprisingly correspond to the ‘Constant AOIs’ featuring consistent exposure to high volumes of taxi traffic throughout the day. It should be noticed that although the areas near the ferry terminal (i.e. AOI 18) are also classified as ‘Major transit hubs’, these areas are only recognised as an AOI after 16.30 (i.e. Evening AOIs), which might indicate a typical evening return peak use.

The intersection also reveals regular commuting patterns. Nearly 60% of those AOIs classified as ‘High-rise integrated commercial’ are respectively occupied by ‘Morning AOIs’ and ‘Evening AOIs’, manifesting a typical bimodal commuting pattern. However, there is also correspondence between the AOIs categorised as ‘Residential heritage mix’ and ‘Late Night AOIs’, suggesting a residential-oriented function.

Moreover, characterised by mixed and compact land use, AOIs from the ‘High-rise integrated commercial’ and ‘Public institution mix’ categories present various temporal usage patterns, which with more defuse representation over the four temporal clusters, with the exception of ‘Late Night AOIs’. Such a pattern reflects a wide variety of essential roles in people’s daily life, which could satisfy multiple demands, including entertainment, public services, commuting, shopping, tourism and other aspects.

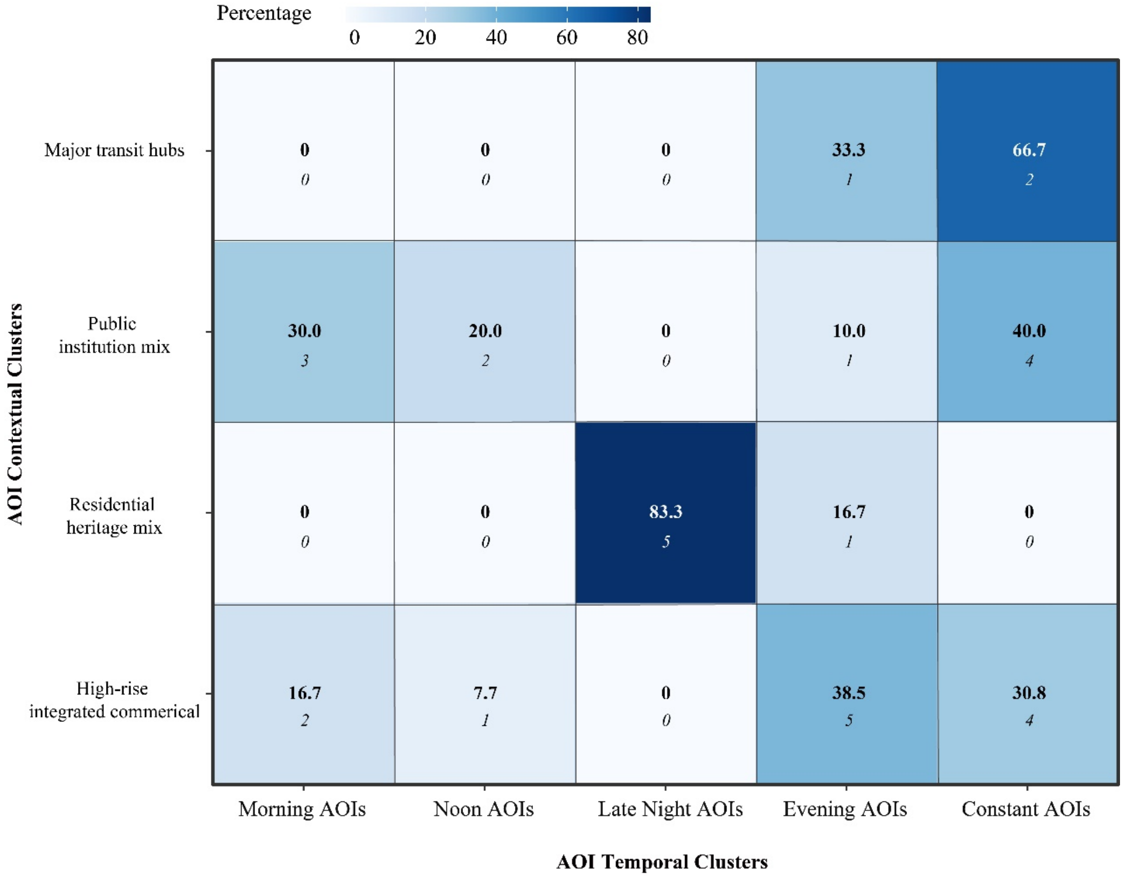


***Figure 5.9 Cross-Tabulation: AOI Frequency and Percentage by Contextual Clusters and Temporal Clusters in Manhattan.***

Italic number shows the actual number of AOIs; Bold number shows the percentage.

## 5.6 DISCUSSION AND CONCLUSIONS

The measurement and ascription of urban AOIs are of continued interest within the field of urban mobility studies. The wide availability of large-scale spatiotemporal data has enabled a variety of new methods of identifying and understanding urban AOIs through the application of density-based cluster analysis, which can generally be conceptualised into a framework comprising three phases: hotspot detection, boundary-defining, and analysis. We identified how such frameworks as those currently implemented contain several limitations across each phase. Firstly, due to the nature of the traditional DBSCAN algorithm, many of the existing studies overwhelmingly concentrated on the spatial aspect of the AOI, while a more integrated view combining spatial and temporal dimensions was somewhat overlooked. Secondly, using enclosed polygon to define the boundary of AOI from those identified hotspot clusters may not form the most appropriate units for analysis given that they lack the attributes of the underlying urban morphology that may inform the identified patterns. Finally, after AOIs are identified, most existing studies neglect the characterisation of those latent attributes affecting the formation of AOIs.

Within this context, our study proposed a new analytical framework that is guided by a conventional three-phase workflow, yet addressed the abovementioned research. The ST-DBSCAN algorithm was employed as the core of the first phase to detect spatiotemporal hotspots. In the second phase, the road network was used to define the boundary of urban AOI; and finally, the dynamic features and contextual features of urban AOI were exposed and investigated. The proposed framework was applied to a taxi GPS dataset extracted from the selected case study area, New York City.

Our enhanced framework identified 31 unique AOIs across the spatial extent of Manhattan. Most of the AOI locations were highly correlated to famous places, such as landmarks, culture venues, open spaces, commercial centres, and transit stations. The spatiotemporal dynamics of the extracted AOIs were considered through further cluster analysis conducted using the H-K-means algorithm. The 31 detected AOIs were classified into five unique clusters (i.e. Temporal Clusters), respectively, representing different types of spatiotemporal activity. Furthermore, the contextual features of AOIs were considered by importing 52 candidate variables from supplementary open data portals. A PCA-based variable selection framework proposed by Liu et al. (2019) was employed to filter out redundant variables, which eventually retained 27 variables that identified five salient AOI clusters (i.e. Contextual Clusters). These clusters were named, described, and mapped. Through cross-tabulating the abovementioned two types of AOI clusters, a high degree of correspondence was found, reflecting the interrelation between the context and dynamics of AOIs.

The utility of defining road-constrained AOIs alongside their dynamic and contextual characteristics we envisage will benefit multiple stakeholders. For urban planners and policymakers, they are more likely to identify urban areas with greater priority and issue more context-based policies, assisting in allocating limited urban resources more effectively. For transport agencies and operators, enhanced spatiotemporal information about the urban AOIs could help to mitigate traffic congestions and provide timely adjustment to the provision of public transport. For taxi drivers, enhanced knowledge of trip hotspots will assist in making more purposeful route selections to maximise the potential for passenger demand. For tourists and travellers, the identified urban AOIs might be utilised as an informative city guide; and for retailers and business managers, our results could assist them with site selection and targeted advertising.

One limitation of this study relates to the parameter selection of ST-DBSCAN. The method used in this study to define MinPt and Eps is primarily based on the heuristic method suggested by the Birant & Kut (2007), which requires further justification in terms of practical application. In another context, Chen et al. (2019) suggested using 1% of the observations to define the MinPt in their study on the detection of urban AOIs in London. In our case, however, if 1% of the observations were employed to define the parameter, the algorithm would fail to identify any clusters since the Minpt is too large (i.e. MinPt is more than 2000). As we discussed previously, there are no standard rules guiding the parameter selection, meaning that the parameter setting may be adjusted according to the actual conditions. As such, we envisage further work looking at optimised methods for parameter selection. Nonetheless, despite such caveat, this paper has presented an innovative methodological framework to identify and understand urban AOIs in terms of both context and dynamics, and will likely be a useful framework for applications within other urban contexts.

The presented approach is extendable in many ways. One direction of future work that would be favourable to the quality of value of the outcomes is the integration with the other emerging datasets. Since the landscape of the traditional taxi market has been changing by the rapid rise of ‘ride-hailing’ businesses such as Uber and Lyft, a growing number of taxi travellers replace their traditional on-street-hailing with more convenient app-hailing (NYDOT, 2018; Willis & Tranos, 2021). In this context, it is possible to either compare the spatiotemporal differences between the urban AOIs formed by the traditional taxi GPS data and those formed by the app-based for-hire vehicle data; or integrate them together to deepen our understandings about the urban AOIs more comprehensively within the context of the current taxi market. Furthermore, with more public transit datasets are becoming publicly available, it is possible to identify and compare AOIs through using data from other travel modes, which might demonstrate manifold differences of interest between multimodal travellers, e.g. active mobility and motorised road users (Keler et al., 2020).

# 6.0 CONCLUSION AND FUTURE WORK

The final chapter provides a summary of the thesis contributions. First, in Section 6.1, a chapter-by-chapter synopsis of this thesis is presented, in which the main outcomes and contributions of each chapter are re-emphasised. Second, in Section 6.2, the limitations of this thesis are discussed, followed by prospects for future research directions.

## 6.1 THESIS SUMMARY AND CONTRIBUTIONS

The thesis was introduced in Chapter 1 and provided the general context of the research alongside hypothesised outcomes aligned to research objectives. The challenges of urbanisation were discussed, including a specific focus on urban mobility. The potential applications of new forms of urban data were discussed within the context of enabling urban analysts to engage in evidence-based urban planning activities. Much of this discussion was framed within the context of advancements of ICTs and the increased availability of urban (big) data; positing that such changes enable the improved analysis of urban mobility problems from spatial, temporal and contextual dimensions. The principal goal of this thesis has been to develop a knowledge-discovery framework that effectively integrated the dynamic and multidimensional contexts of urban mobility. To achieve this, five major research objectives were set (listed below), which have been accomplished through the development of this thesis.

- *Objective 1*: to summarise and improve the typical workflow of building a geodemographic classification from existing literature
- *Objective 2*: to identify and select variables that are commonly used in both traditional and recent studies to build the classification through a comprehensive literature review
- *Objective 3*: to handle the adverse effects caused by high dimensionality in the dataset by using the dimensionality reduction method
- *Objective 4*: to extract urban and human mobility patterns from multi-sourced urban data while concurrently considering dynamic and contextual urban contexts
- *Objective 5*: to apply the developed framework in the target case study area to manifest its utilities and contributions to the existing literature.

Chapter 2 achieved *Objective 1* by conducting a literature review of the relevant research. First, the research background of this thesis was elaborated in further detail. This included a discussion of how the "Big Data deluge" and its manifestation within urban settings could be understood through a framework based around the intentionality of collection. Urban data were then classified into two major categories (i.e., *purposeful data* and *organic data*), with the advantages and drawbacks clarified for both. It was argued that although *organic data* benefit urban mobility research due to their large-scale coverage and fine spatiotemporal granularity, *purposeful data* still have their own advantages in capturing detailed contextual information. These advantages can be utilised to overcome the drawback of solely using *organic data* by adding contextual enrichment. Urban studies, and especially urban mobility studies, were argued to benefit from incorporating both organic and purposeful urban data as these collectively build upon each others' strengths to offset their weakness. Second, the chapter reviewed the Data, Information, Knowledge, and Wisdom (DIKW) paradigm and its application for contemporary urban analytics. The concept of urban contexts was introduced amid the discussion of urban analytics and was followed by a review of its utility for urban and human mobility research. The latter half of the literature review focused specifically on the methodology of geodemographics and how geodemographic classification can be created. This was necessary given that geodemographics provided both the theoretical and methodological foundation for the thesis, which significantly influenced the construction of the proposed methodology framework developed and implemented by this thesis.

Chapter 3 fulfilled *Objectives 2*, *4* and *5*. The prototype analytical framework was outlined in this chapter, which integrated dynamic and multidimensional contexts distilled from the multi-sources of urban data to provide a comprehensive understanding of the urban environment and urban mobility. The proposed framework was primarily based on the typical workflow of building an application-specific geodemographic classification; however, it made several improvements represented by the variable enrichment from the dynamic and contextual perspectives. In particular, a systematic literature review (SLR) with the aim of identifying those variables that are commonly used in contemporary studies affecting urban mobility and differentiating urban transit areas. Through the SLR, a variable specification was gathered; following four domains, namely: land use and built environment, location and accessibility, socioeconomic and demographic, and transit-related. Such findings significantly expanded upon variable selection methods utilised within the existing literature that were overwhelmingly based on the ‘Ds’. The SLR served as the theoretical framework that guided the initial variable selection for building any mobility-related classifications.

Chapter 4 achieved *Objective 3* by developing a PCA-based variable selection framework to mitigate the adverse effects on cluster performance of high dimensionality input data when building geodemographic classification. This automated five-stage framework employed traditional PCA in a computationally intensive setting to select a subset of appropriate candidate variables from the initial inputs for creating any geodemographic classification. However, within the proposed framework, the variable selection was not solely based on the traditional PCA results; additional factors were considered during and after the clustering process. For instance, after PCA filtering, the highly correlated variable pairs were further examined by the minimum spanning tree (MST) method, in which variables with the fewest branches in the MST were eliminated from the candidate pool; furthermore, variables that adversely affect the final clustering quality (measured by the ratio between BCSS and WCSS) were also filtered out. The performance of this variable selection framework was then evaluated by comparing it to the benchmark geodemographic classification (i.e., 2011 OAC). The several statistical indicators demonstrate that through the implementation of the proposed methodology, the quality of the cluster assignment was improved relative to the 2011 OAC.

Chapter 5 fulfilled *Objectives 2*, *4*, and *5*. The two frameworks developed in Chapter 3 and 4 were integrated with this chapter. This was then employed to address a typical urban mobility task, namely, identifying and characterising urban areas of interest (AOIs) through taxi GPS data. This analytical framework of urban AOI discovery was primarily derived from a three-phase workflow that was summarised from the existing studies, but containing methodological enhancements in each phase. First, instead of using the frequently used DBSCAN approach, which only considers the spatial distribution of the GPS points, the ST-DBSCAN algorithm was implemented on the taxi dataset to identify spatiotemporal hotspots. Such implementation considers spatial and temporal distributions simultaneously, identifying taxi travel hotspot clusters in a space-time cube. Second, the road networks were innovatively utilised as the carrier of the detected spatiotemporal hotspots, therefore configuring the road-constrained AOIs (i.e., the aggregation between road network with spatiotemporal hotspots). By contrast, traditional studies usually employed a closed polygon approach to defining AOIs’ boundary, and it could therefore be argued that they fail to consider the reshaping effects of urban morphology. Third, the proposed frameworks introduced in Chapter 3 and 4 were employed in the AOI analysis phase. These formulated the dynamic and contextual layers, which provided a comprehensive analysis of urban AOIs in terms of spatiotemporal dynamics and urban contexts. The results demonstrate that the contextual enrichment significantly improved the interpretability of the spatiotemporal AOIs identified in the case study area, thereby facilitating a comprehensive understanding of the urban environment and mobility patterns.

In summary, the work conducted in this thesis has presented a new methodological framework that enables the analysis of urban mobility through multiple sources of urban data; and considering both the dynamic and multidimensional aspects of urban contexts. The proposed framework has demonstrated its utility in the case study areas by resolving real-world mobility-problems (e.g., creating context-based TOD classification and urban AOI detection). Moreover, since the methodological framework established in this thesis was applied to open data obtained from the case study areas to exemplify its utility, it will likely be a helpful framework for applications within other international urban contexts.

## 6.2 LIMITATIONS AND FUTURE OUTLOOK

Although the thesis has presented many novel contributions to the current research and literature, several criticisms may be levelled at some limitations that were encountered during the development of the proposed methodological framework. These limitations have been outlined in detail within those papers making up the thesis (specifically, see Section 3.7, 4.6, and 5.6). These arguments are reiterated here alongside suggestions of future works based on them.

First, as stated in Section 2.1, although the rapid development of ICTs and the increased availability of urban data have made it possible to simultaneously analyse urban mobility from spatial, temporal, and contextual perspectives, computational challenges continue to limit such work since there is always a lag between the ability to understand data and the ability to produce and collect them (Kong et al., 2020). This computational bottleneck has been manifested by the compromises made in this research. For instance, in Chapter 3, one of the main reasons why the weekends' turnstile data were excluded from the analysis is that the passengers' weekend travel contained too much noise (i.e. random trips or outliers). Such noisy data could not be handled well using traditional distance-based clustering methods, and therefore the weekend travel patterns were not able to be revealed effectively. Moreover, the ST-DBSCAN algorithm employed in Chapter 5 is very memory-intensive, directly leading to the decision to merely take 1% random samples from the whole year taxi GPS dataset. Given the limited computational power, how to effectively and efficiently process urban (big) data is one of the future directions not only for this study but also for other similar urban studies seeking to derive actionable knowledge from raw datasets. As such, one of the potential future directions of this work may be to address the problem of balancing the speed-accuracy and the privacy-utility tradeoffs during the data aggregation (Asikis & Pournaras, 2020; Yun et al., 2019). Future studies could ascertain how to aggregate data to achieve better computing performance without losing too much detailed information. For example, in the study outlined in Chapter 5, it is possible to aggregate the individual taxi GPS points to the designated areas, such as TAZ (see Chen et al., 2019; Zhang et al., 2016), and then adopt a grid- or area-based ST-DBSCAN to explore their spatiotemporal patterns. Alternatively, it would be helpful to transpose the first two phases of the framework by first snapping GPS points to the road network segment, also referred to as the map-matching process (Yang & Gidófalvi, 2018), and then using a network-based ST-DBSCAN algorithm to detect taxi aggregation hotspots or so-called hot routes. While problems related to the ecological fallacy or MAUP might occur due to the data aggregation, these hypothetical attempts could significantly improve the performance of the data mining and machine learning algorithms while maintaining a reasonable level of spatiotemporal granularity.

Second, it is inevitable to end up with a dataset with high dimensionality when integrating dynamic and multidimensional contexts to analyse the complex urban environment and its mobility. This will typically result in the ‘curse of dimensionality’, which brings uncertainties and adversely affecting the clustering results (see Section 4.2). Although this thesis sought to utilise various approaches (e.g., SOM and the PCA-based variable selection methods) to mitigate such adverse effects, the issues caused by high dimensionality cannot be addressed entirely. Given that most unsupervised classifiers employed in this thesis, such as k-means, hierarchical clustering, and H-K-means, are based on distance measurement (i.e., using distance to measure the similarity between attributes), such limitation will be further amplified. Moreover, the high-dimensionality exists not only in the dataset assembled by contextual variables but also in one configured by temporal variables. The size of the time intervals used to aggregate human mobility, also known as the temporal window, determines the dimensionality of the dynamic aspect. For instance, due to the inherent temporal resolution of the subway turnstile data used in Chapter 3, the passenger flow counts were aggregated into a four-hour temporal window, meaning that a day comprises 12 variables (i.e., six for entry and six for exit). In terms of the dataset with a finer temporal resolution, such as the taxi GPS data used in Chapter 5, the temporal interval size was set as 15 minutes (i.e., a day comprises 96 variables). Moreover, since some mobility data contain real-time changes, such temporal resolution may need to be more fine-grained. For instance, more than half of the bike trips duration in the NYC CitiBike bike-sharing system last less than 10 minutes (Sokoloff, 2018). To perform a similar mobility-related analysis, the size of the temporal window aggregating the bike trips would be five minutes or even less, thereby resulting in an exponential expansion in dimensionality. This explosive growth in dimensionality will undoubtedly add more uncertainties to the clustering result, degenerating the level of result interpretability. Alongside reducing dimensionality by carrying out complex variable selection work, another promising approach is to conduct the clustering task by using more sophisticated data mining and machine learning algorithms that do not predominantly rely on distance measures. For instance, Liu and Cheng (2020) innovatively employed a text-mining algorithm in the urban mobility study. They applied the Latent Dirichlet Allocation (LDA) algorithm, a generative model-based clustering technique, on the smart card dataset to classify London underground passengers characterising similar travel behaviour. Similarly, in order to identify the spatiotemporal functions of metro stations in Shanghai, Wang et al. (2017) proposed a Doc2vec-based semantic framework (IS2Fun) for characterising the semantic distribution of subway stations based on human mobility patterns and POIs. Moreover, as the neural network techniques have matured, it is possible to utilise graph-based data clustering methods to mitigate the challenges caused by high dimensionality (see Liu & Barahona, 2020).

Finally, while this thesis provided an innovative framework for incorporating dynamic and multidimensional contexts into urban mobility studies, it may be limited in aspects of the dynamic nature of the urban environment. As discussed in Chapter 2, although certain urban environment components, such as the neighbourhood’s socioeconomic and demographic background, are classified as a relatively static category, they do inherently change-over-time. For instance, the disparity of research about the daytime and nighttime population has been well-documented and has a lengthy history of research (Akkerman & Shimoura, 2012; Moss & Qing, 2012; Schmitt, 1956; Sleeter & Wood, 2006). This daytime-nighttime differential is predominantly influenced by the heterogeneity in people’s living and working locations, resulting in different sociodemographic characteristics at different times of the day but in the same spatial location. Furthermore, it should be noted that even the ‘most static’ urban environment components, such as urban land use and POIs, may alter their main functionality at different times of the day. Some mixed-use urban areas are predominately used for one certain type of function, such as commercial use, during the day, and then switch to another type, such as residential use, during the evening and nighttime hours. This time-varying shifting is more evident in modern metropolitan areas, such as NYC and London, where mixed land use and compact urban development are more prominent. With the expanded diversity and availability of urban data, it would be advantageous in future urban mobility research to convert these ‘static’ contextual attributes into dynamic time-varying attributes. For example, future research could combine both daytime and nighttime data to analyse the dynamic change of people’s sociodemographic characteristics in relation to their mobility. Moreover, it would be beneficial in future works to link the multi-sourced urban data together. For example, future studies could establish a linkage between the conventional survey-based land use data, such as POIs, and the emerging crowdsourcing data, such as OSM data, to obtain the POIs’ opening time and thus formulate a time-varying POI dataset. Future works would be hugely benefited by such linked data.

However, despite the abovementioned limitations, this thesis has made significant contributions to the existing research in urban analytics through developing an enhanced methodological framework that is based on conventional geodemographic analysis to investigate the urban environment and urban mobility more comprehensively. The developed framework has presented a typical DIKW workflow that enables the analysis of urban mobility through multi-sourced urban data; and systematically integrates both the dynamic and multidimensional aspects of urban contexts, contributing to a better understanding of urban mobility from space, time and urban context aspects. The proposed framework has demonstrated its utility in the case study areas by resolving real-world mobility problems. Furthermore, since the framework was designed to use open data to exemplify its adaptability, it will likely be a valuable framework for implementations within other urban contexts on a global scale.

Moreover, beyond the scope of urban study, the methodological framework developed in this thesis could be potentially applied to various mobility-related fields, facilitating evidence-based and effective decision-making. For instance, in public transit planning, the developed framework could be employed to monitor and assess the effectiveness of transit policies introduced by the public transit agency. In my previous research, namely, Liu & Cheng (2020), a similar analytical framework was applied to the Oyster Card data in London to assess the overall performance of the TfL’s Night Tube campaign, to some extent manifesting the utility of this analytical framework (see Liu & Cheng, 2020). Comparing to that work, instead of using the existing general-purpose geodemographic classification (i.e., 2011 OAC) to contextualise human mobility pattern, the analytical framework developed in this thesis advocates creating an application-specific classification to conduct contextual enrichment, which could further improve the result interpretability and provide a target-specific application.

Furthermore, this framework could be utilised to analyse the ongoing COVID-19 pandemic in the public health field. The knowledge gained can be used as the evidence basis guiding the governmental implementation of non-pharmaceutical interventions (NPIs), such as regional or national lockdowns, with the goal of containing the viral transmission effectively. Many recent studies (see Grantz et al., 2020; Yabe et al., 2020) have proved a strong positive correlation between human mobility and coronavirus transmission. Therefore, analysing people's mobility patterns can help decision-makers monitor and even predict the outbreak hotspots in both space and time. For example, in my recent research (see Cheng et al., 2021), a part of the analytical framework[14] developed in this thesis was implemented to conduct COVID-19 transmission analysis based upon individual patients' trajectory data collected in China. In that research, we revealed the spatiotemporal patterns of patients' mobility and the transmission stages of COVID-19 from Wuhan to the rest of China at finer spatial and temporal scales. In addition, with higher data availability of contextual information, it is feasible to undertake further contextual enrichment to evaluate and profile the severity or vulnerability for each hotspot area and the households who live there, suggesting the areas that will be most at risk from pandemic waves (Daras et al., 2021). Based on the findings, policymakers could make further modifications (e.g., whether easing or strengthening the restrictions) for existing NPIs.

[14] Only the spatiotemporal paradigm (see Figure 1.1) part was used due to the lack of personal contextual information

# BIBLIOGRAPHY

Abdi, H., & Williams, L. J. (2010). Principal component analysis. In *Wiley Interdisciplinary Reviews: Computational Statistics*. https://doi.org/10.1002/wics.101

Abreu, M., & Öner, Ö. (2020). Disentangling the Brexit vote: The role of economic, social and cultural contexts in explaining the UK's EU referendum vote. *Environment and Planning A: Economy and Space*, *52*(7), 1434–1456. https://doi.org/10.1177/0308518X20910752

Aditya Shastry, K., & Sanjay, H. A. (2020). *Data Analysis and Prediction Using Big Data Analytics in Agriculture* (pp. 201–224). Springer, Singapore. https://doi.org/10.1007/978-981-15-0663-5_10

Adnan, M. (2011). *Towards Real-Time Geodemographic Information Systems: Design, Analysis and Evaluation*. University College London.

Akdag, F., Eick, C. F., & Chen, G. (2014). *Creating Polygon Models for Spatial Clusters* (pp. 493–499). Springer, Cham. https://doi.org/10.1007/978-3-319-08326-1_50

Akkerman, A., & Shimoura, S. (2012). Discrete choice in commuter space: Small area analysis of diurnal population change in the Tokyo Metropolitan Region. *Computers, Environment and Urban Systems*. https://doi.org/10.1016/j.compenvurbsys.2012.03.001

Alberti, M., McPhearson, T., & Gonzalez, A. (2018). Embracing Urban Complexity. In *Urban Planet*. https://doi.org/10.1017/9781316647554.004

Alexiou, A. (2016). *Putting "Geo" into Geodemographics: Evaluating the performance of national classification systems within regional contexts*. https://livrepository.liverpool.ac.uk/3007463/

Alexiou, A., Riddlesden, D., & Singleton, A. (2020). The Geography of Online Retail Behaviour. In Paul Longley, J. Cheshire, & A. Singleton (Eds.), *Consumer Data Research* (pp. 97–109). UCL Press. https://discovery.ucl.ac.uk/id/eprint/10046615/1/Consumer-Data-Research.pdf

Alexiou, A., & Singleton, A. (2015). Geodemographic Analysis. In A. Singleton & C. Brunsdon (Eds.), *Geocomputation: a practial primer* (pp. 137–152). SAGE Publication Ltd.

Alexiou, A., Singleton, A., & Longley, P. A. (2016). A classification of multidimensional open data for urban morphology. *Built Environment*. https://doi.org/10.2148/benv.42.3.382

Alfeo, A. L., Cimino, M. G. C. A., Egidi, S., Lepri, B., & Vaglini, G. (2018). A Stigmergy-Based Analysis of City Hotspots to Discover Trends and Anomalies in Urban Transportation Usage. *IEEE Transactions on Intelligent Transportation Systems*. https://doi.org/10.1109/TITS.2018.2817558

Andrienko, G., Andrienko, N., Boldrini, C., Caldarelli, G., Cintia, P., Cresci, S., Facchini, A., Giannotti, F., Gionis, A., Guidotti, R., Mathioudakis, M., Muntean, C. I., Pappalardo, L., Pedreschi, D., Pournaras, E., Pratesi, F., Tesconi, M., & Trasarti, R. (2020). (So) Big Data and the transformation of the city. *International Journal of Data Science and Analytics*. https://doi.org/10.1007/s41060-020-00207-3

Angel, S., Parent, J., Civco, D. L., Blei, A., & Potere, D. (2011). The dimensions of global urban expansion: Estimates and projections for all countries, 2000-2050. *Progress in Planning*, *75*(2), 53–107. https://doi.org/10.1016/j.progress.2011.04.001

Arai, K., & Ridho Barakbah, A. (2007). Hierarchical K-means: an algorithm for centroids initialization for K-means. In *Rep. Fac. Sci. Engrg. Reports of the Faculty of Science and Engineering*.

Arribas-Bel, D. (2014). Accidental, open and everywhere: Emerging data sources for the understanding of cities. *Applied Geography*. https://doi.org/10.1016/j.apgeog.2013.09.012

Arribas-Bel, D., & Schmidt, C. R. (2013). Self-organizing maps and the US urban spatial structure. *Environment and Planning B: Planning and Design*. https://doi.org/10.1068/b37014

Asikis, T., & Pournaras, E. (2020). Optimization of privacy-utility trade-offs under informational self-determination. *Future Generation Computer Systems*. https://doi.org/10.1016/j.future.2018.07.018

Atkinson-Palombo, C., & Kuby, M. J. (2011). The geography of advance transit-oriented development in metropolitan Phoenix, Arizona, 2000-2007. *Journal of Transport Geography*. https://doi.org/10.1016/j.jtrangeo.2010.03.014

Austin, M., Belzer, D., Benedict, A., Esling, P., Haas, P., Miknaitis, G., Wampler, E., Wood, J., Young, L., & Zimbabwe, S. (2010). Performance-Based Transit-Oriented Development Typology Guidebook. *Report*.

Bação, F., & Lobo, V. (2010). *Introduction to Kohonen's Self-Organizing Maps*. https://www.semanticscholar.org/paper/Kohonen-'-s-Self-Organizing-Maps-Bação-Lobo/779d1280f296d887c502c5bfeac6e7fdb7a7b0ab?p2df

Badr, H. S., Du, H., Marshall, M., Dong, E., Squire, M. M., & Gardner, L. M. (2020). Association between mobility patterns and COVID-19 transmission in the USA: a mathematical modelling study. *The Lancet. Infectious Diseases*, *20*(11), 1247–1254. https://doi.org/10.1016/S1473-3099(20)30553-3

Bajracharya, A., & Shrestha, S. (2017). Analyzing Influence Of Socio-Demographic Factors On Travel Behavior Of Employees A Case Study Of Kathmandu Metropolitan City Nepal. *International Journal of Scientific & Technology Research*.

Bara, M. W., Ahmad, N. B., Modu, M. M., & Ali, H. A. (2018). Self-organizing map clustering method for the analysis of e-learning activities. *Proceedings of Majan International Conference: Promoting Entrepreneurship and Technological Skills: National Needs, Global Trends, MIC 2018*. https://doi.org/10.1109/MINTC.2018.8363155

Bassett, K., & Short, J. (1980). *Housing and Residential Structure: Alternative Approaches*. Routledge.

Bates, A. (2015). *Classification for Output Areas Classification*. *April*. www.nationalarchives.gov.uk/doc/open-government-licence/

Batey, P., & Brown, P. (1995). From human ecology to customer targeting: the evolution of geodemographics. In P Longley & G. Clarke (Eds.), *GIS for Business and Service Planning* (pp. 150–166). https://books.google.co.uk/books?hl=en&lr=&id=w7F9mAnrGj8C&oi=fnd&pg=IA1&ots=tsneEOWc6f&sig=7WlvcEqPhsKozSlzBSILjAmKO9s&redir_esc=y#v=onepage&q&f=false

Batty, M. (2013a). Big data, smart cities and city planning. *Dialogues in Human Geography*. https://doi.org/10.1177/2043820613513390

Batty, M. (2013b). *The new science of cities*. MIT Press. https://mitpress.mit.edu/books/new-science-cities

Batty, M. (2019). Urban analytics defined. *Environment and Planning B: Urban Analytics and City Science*, *46*(3), 403–405. https://doi.org/10.1177/2399808319839494

Bernard, A., Bell, M., & Charles-Edwards, E. (2014). Life-course transitions and the age profile of internal migration. *Population and Development Review*. https://doi.org/10.1111/j.1728-4457.2014.00671.x

Betis, G., Larios, V. M., Petri, D., Wu, X., Deacon, A., & Hayar, A. (2018). The ieee smart cities initiative - Accelerating the smartification process for the 21st century cities [point of view]. *Proceedings of the IEEE*. https://doi.org/10.1109/JPROC.2018.2814239

Bhattacharjee, S., & Goetz, A. R. (2016). The rail transit system and land use change in the Denver metro region. *Journal of Transport Geography*. https://doi.org/10.1016/j.jtrangeo.2016.02.004

Bibri, S. E. (2021a). Data-driven smart sustainable cities of the future: An evidence synthesis approach to a comprehensive state-of-the-art literature review. *Sustainable Futures*. https://doi.org/10.1016/j.sftr.2021.100047

Bibri, S. E. (2021b). Data-driven smart sustainable cities of the future: urban computing and intelligence for strategic, short-term, and joined-up planning. *Computational Urban Science*. https://doi.org/10.1007/s43762-021-00008-9

Bibri, S. E. (2021c). The core academic and scientific disciplines underlying data-driven smart sustainable urbanism: an interdisciplinary and transdisciplinary framework. *Computational Urban Science*. https://doi.org/10.1007/s43762-021-00001-2

Bibri, S. E., & Krogstie, J. (2020). The emerging data–driven Smart City and its innovative applied solutions for sustainability: the cases of London and Barcelona. *Energy Informatics*. https://doi.org/10.1186/s42162-020-00108-6

Bibri, S. E., & Krogstie, J. (2018). The Big Data deluge for Transforming the Knowledge of Smart Sustainable Cities. *Proceedings of the 3rd International Conference on Smart City Applications - SCA '18*, 1–10. https://doi.org/10.1145/3286606.3286788

Birant, D., & Kut, A. (2007). ST-DBSCAN: An algorithm for clustering spatial-temporal data. *Data and Knowledge Engineering*. https://doi.org/10.1016/j.datak.2006.01.013

Box, G. E. P., & Cox, D. R. (1964). An Analysis of Transformations. *Journal of the Royal Statistical Society: Series B (Methodological)*. https://doi.org/10.1111/j.2517-6161.1964.tb00553.x

Cai, J., Huang, B., & Song, Y. (2017). Using multi-source geospatial big data to identify the structure of polycentric cities. *Remote Sensing of Environment*. https://doi.org/10.1016/j.rse.2017.06.039

Cai, L., Jiang, F., Zhou, W., & Li, K. (2018). Design and Application of an Attractiveness Index for Urban Hotspots Based on GPS Trajectory Data. *IEEE Access*. https://doi.org/10.1109/ACCESS.2018.2869434

Cattuto, C., van den Broeck, W., Barrat, A., Colizza, V., Pinton, J. F., & Vespignani, A. (2010). Dynamics of person-to-person interactions from distributed RFID sensor networks. *PLoS ONE*. https://doi.org/10.1371/journal.pone.0011596

Cervero, R., & Kockelman, K. (1997). Travel demand and the 3Ds: Density, diversity, and design. *Transportation Research Part D: Transport and Environment*. https://doi.org/10.1016/S1361-9209(97)00009-6

Chang, V. (2021). An ethical framework for big data and smart cities. *Technological Forecasting and Social Change*, *165*(December 2020), 120559. https://doi.org/10.1016/j.techfore.2020.120559

Charlton, M., Openshaw, S., & Wymer, C. (1985). Some New Classifications of Census Enumeration Districts in Britain: A Poor Man's ACORN. *Journal of Economic and Social Measurement*, *13*(1), 69–96. http://mural.maynoothuniversity.ie/6111/

Chavez-Baeza, C., & Sheinbaum-Pardo, C. (2014). Sustainable passenger road transport scenarios to reduce fuel consumption, air pollutants and GHG (greenhouse gas) emissions in the Mexico City Metropolitan Area. *Energy*, *66*(2), 624–634. https://doi.org/10.1016/j.energy.2013.12.047

Chen, B., Tai, P. C., Harrison, R., & Pan, Y. (2005). Novel hybrid hierarchical-K-means clustering method (H-K-means) for microarray analysis. *2005 IEEE Computational Systems Bioinformatics Conference, Workshops and Poster Abstracts*. https://doi.org/10.1109/CSBW.2005.98

Chen, C., Chen, J., & Barry, J. (2009). Diurnal pattern of transit ridership: a case study of the New York City subway system. *Journal of Transport Geography*. https://doi.org/10.1016/j.jtrangeo.2008.09.002

Chen, C., Ma, J., Susilo, Y., Liu, Y., & Wang, M. (2016). The promises of big data and small data for travel behavior (aka human mobility) analysis. In *Transportation Research Part C: Emerging Technologies*. https://doi.org/10.1016/j.trc.2016.04.005

Chen, Meixu, Arribas-Bel, D., & Singleton, A. (2019). Understanding the dynamics of urban areas of interest through volunteered geographic information. *Journal of Geographical Systems*. https://doi.org/10.1007/s10109-018-0284-3

Chen, Meixu, Arribas-Bel, D., & Singleton, A. (2020). Quantifying the characteristics of the local urban environment through geotagged flickr photographs and image recognition. *ISPRS International Journal of Geo-Information*. https://doi.org/10.3390/ijgi9040264

Chen, Min, Mao, S., & Liu, Y. (2014). Big data: A survey. *Mobile Networks and Applications*. https://doi.org/10.1007/s11036-013-0489-0

Chen, T., Bowers, K., Cheng, T., Zhang, Y., & Chen, P. (2020). Exploring the homogeneity of theft offenders in spatio-temporal crime hotspots. *Crime Science*, *9*(1), 9. https://doi.org/10.1186/s40163-020-00115-8

Chen, Y., Zhang, Z., & Liang, T. (2019). Assessing urban travel patterns: An analysis of traffic analysis zone-based mobility patterns. *Sustainability (Switzerland)*. https://doi.org/10.3390/su11195452

Cheng, T., Haworth, J., Anbaroglu, B., Tanaksaranond, G., & Wang, J. (2014). Spatiotemporal data mining. In *Handbook of Regional Science*. https://doi.org/10.1007/978-3-642-23430-9_68

Cheng, T., Lu, T., Liu, Y., Gao, X., & Zhang, X. (2021). Revealing spatiotemporal transmission patterns and stages of COVID-19 in China using individual patients' trajectory data. *Computational Urban Science*. https://doi.org/10.1007/s43762-021-00009-8

Chorus, P., & Bertolini, L. (2011). An application of the node place model to explore the spatial development dynamics. *The Journal of Transport and Land Use*.

Cockings, S., Martin, D., & Harfoot, A. (2020). Developing a National Geodemographic Classification of Workplace Zones. *Applied Spatial Analysis and Policy*, *13*(4), 959–983. https://doi.org/10.1007/s12061-020-09337-4

CTOD. (2013). *Transit-Oriented Development Typology Strategy for Allegheny County*. http://ctod.org/pittsburgh/201302pittsburgh-tod-book-web.pdf

Cullen, I. G. (1972). Space, Time and the Disruption of Behaviour in Cities. *Environment and Planning A: Economy and Space*. https://doi.org/10.1068/a040459

Dalton, C. M., & Thatcher, J. (2015). Inflated granularity: Spatial "Big Data" and geodemographics. *Big Data and Society*, *2*(2), 1–15. https://doi.org/10.1177/2053951715601144

Daras, K., Alexiou, A., Rose, T. C., Buchan, I., Taylor-Robinson, D., & Barr, B. (2021). How does vulnerability to COVID-19 vary between communities in England? Developing a Small Area Vulnerability Index (SAVI). *Journal of Epidemiology and Community Health*. https://doi.org/10.1136/jech-2020-215227

Daras, K., Green, M. A., Davies, A., Barr, B., & Singleton, A. (2019). Open data on health-related neighbourhood features in Great Britain. *Scientific Data*, *6*(1), 1–10. https://doi.org/10.1038/s41597-019-0114-6

Das, G., Chattopadhyay, M., & Gupta, S. (2016). A comparison of self-organising maps and principal components analysis. *International Journal of Market Research*. https://doi.org/10.2501/IJMR-2016-039 de Palma, A., & Lindsey, R. (2001). Transportation: Supply and Congestion. *International Encyclopedia of the Social & Behavioral Sciences*, 15882–15888. https://doi.org/10.1016/b0-08-043076-7/02318-4

Debenham, J. (2002). *Understanding Geodemographic Classification: Creating The Building Blocks For An Extension*. http://eprints.whiterose.ac.uk/5014/

Demartines, P., & Blayo, F. (1992). Kohonen self-organizing map: is the normalization necessary? *Complex Systems*, *6*, 105–123. http://wpmedia.wolfram.com/uploads/sites/13/2018/02/06-2-2.pdf

Department for Communities and Local Government. (2010). *The English Indices of Deprivation 2010 Stoke-on-Trent - Summary* (Issue March). http://webapps.stoke.gov.uk/uploadedfiles/Indices of Deprivation 2010 - Summary.pdf

Dieleman, F. M., Dijst, M., & Burghouwt, G. (2002). Urban form and travel behaviour: Micro-level household attributes and residential context. *Urban Studies*. https://doi.org/10.1080/00420980220112801

Dirgahayani, P., & Choerunnisa, D. N. (2018). Development of Methodology to Evaluate TOD Feasibility in Built-up Environment (Case Study: Jakarta and Bandung, Indonesia). *IOP Conference Series: Earth and Environmental Science*. https://doi.org/10.1088/1755-1315/158/1/012019

Domínguez-Mujica, J., González-Pérez, J., & Parreño-Castellano, J. (2011). Tourism and human mobility in Spanish Archipelagos. *Annals of Tourism Research*. https://doi.org/10.1016/j.annals.2010.11.016

Dupont, L., Morel, L., & Guidat, C. (2015). Innovative public-private partnership to support Smart City: the case of “Chaire REVES.” *Journal of Strategy and Management*. https://doi.org/10.1108/JSMA-03-2015-0027

Eräranta, S., & Staffans, A. (2015). From situation awareness to smart city planning and decision making. *CUPUM 2015 - 14th International Conference on Computers in Urban Planning and Urban Management*.

Ester, M., Kriegel, H., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. *KDD’96*, 226–231. https://doi.org/10.5555/3001460.3001507

Ettema, D., Friman, M., Gärling, T., & Olsson, L. E. (2016). Travel Mode Use, Travel Mode Shift and Subjective Well-Being: Overview of Theories, Empirical Findings and Policy Implications. In D. Wang & S. He (Eds.), *Mobility, Sociability and Well-being of Urban Living* (pp. 129–150). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-662-48184-4_7

Everitt, B., Landau, S., Leese, M., & Stahl, D. (2011). *Cluster Analysis* (5th ed.). Wiley.

Ewing, R., & Cervero, R. (2001). Travel and the built environment: A synthesis. *Transportation Research Record*. https://doi.org/10.3141/1780-10

Ewing, R., & Cervero, R. (2010). Travel and the built environment. *Journal of the American Planning Association*. https://doi.org/10.1080/01944361003766766

Financial Network Analytics. (2012). *Correlation Networks*.

Fricke, M. (2019). The knowledge pyramid: The dikw hierarchyt. In *Knowledge Organization*. https://doi.org/10.5771/0943-7444-2019-1-33

Gale, C., Singleton, A., Bates, A., & Longley, P. (2016). Creating the 2011 area classification for output areas (2011 OAC). *Journal of Spatial Information Science*, *12*(2016), 1–27. https://doi.org/10.5311/JOSIS.2016.12.232

Gantz, B. J., Reinsel, D., & Shadows, B. D. (2012). Big Data , Bigger Digital Shadow s , and Biggest Grow th in the Far East Executive Summary: A Universe of Opportunities and Challenges. *Idc*, *2007*(December 2012), 1–16.

Gao, J., Ettema, D., Helbich, M., & Kamphuis, C. B. M. (2019). Travel mode attitudes, urban context, and demographics: do they interact differently for bicycle commuting and cycling for other purposes? *Transportation*, *46*(6), 2441–2463. https://doi.org/10.1007/s11116-019-10005-x

Garcia, J. C., Avendaño, A., & Vaca, C. (2018). *Where to go in Brooklyn: NYC Mobility Patterns from Taxi Rides* (pp. 203–212). Springer, Cham. https://doi.org/10.1007/978-3-319-77703-0_20

Gershenson, C. (2016). *Improving Urban Mobility by Understanding its Complexity*. http://arxiv.org/abs/1603.04267

González, M. C., Hidalgo, C. A., & Barabási, A. L. (2008). Understanding individual human mobility patterns. *Nature*. https://doi.org/10.1038/nature06958

Gordon, P., Kumar, A., & Richardson, H. W. (1989). The influence of metropolitan spatial structure on commuting time. *Journal of Urban Economics*. https://doi.org/10.1016/0094-1190(89)90013-2

Grantz, K. H., Meredith, H. R., Cummings, D. A. T., Metcalf, C. J. E., Grenfell, B. T., Giles, J. R., Mehta, S., Solomon, S., Labrique, A., Kishore, N., Buckee, C. O., & Wesolowski, A. (2020). The use of mobile phone data to inform analysis of COVID-19 pandemic epidemiology. *Nature Communications*. https://doi.org/10.1038/s41467-020-18190-5

Guo, J., Nakamura, F., Li, Q., & Zhou, Y. (2018). Efficiency Assessment of Transit-Oriented Development by Data Envelopment Analysis: Case Study on the Den-en Toshi Line in Japan. *Journal of Advanced Transportation*. https://doi.org/10.1155/2018/6701484

Gurteen, D. (1998). Knowledge, Creativity and Innovation. *Journal of Knowledge Management*. https://doi.org/10.1108/13673279810800744

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. In *Journal of Machine Learning Research*. https://doi.org/10.1162/153244303322753616

Haken, H., & Portugali, J. (2021). *Cities as Hybrid Complex Systems*. https://doi.org/10.1007/978-3-030-63457-5_2

Hao, J., Zhu, J., & Zhong, R. (2015). The rise of big data on urban studies and planning practices in China: Review and open research issues. *Journal of Urban Management*. https://doi.org/10.1016/j.jum.2015.11.002

Harris, P., Brunsdon, C., & Charlton, M. (2011). Geographically weighted principal components analysis. *International Journal of Geographical Information Science*. https://doi.org/10.1080/13658816.2011.554838

Harris, R., Sleight, P., & Webber, R. (2005). *Geodemographics, GIS and Neighbourhood Targeting*. John Wiley & Sons Ltd. https://www.wiley.com/en-gb/Geodemographics%2C+GIS+and+Neighbourhood+Targeting-p-9780470864135

Harvey, A. S., & Taylor, M. E. (2000). Activity settings and travel behaviour: A social contact perspective. *Transportation*. https://doi.org/10.1023/a:1005207320044

Hennock, E. P. (1991). Concepts of poverty in the British social surveys from Charles Booth to Arthur Bowley. In M. Bulmer, K. Bales, & K. Kish Sklar (Eds.), *The Social Survey in Historical Perspective*. Cambridge University Press (CUP). https://scholar.google.com/scholar_lookup?title=Concepts of Poverty in the British Social Surveys from Charles Booth to Arthur Bowley&author=EP. Hennock&publication_year=1991

Hickman, R., & Banister, D. (2014). *Transport, Climate Change and the City* (1st ed.). Routledge. https://www.routledge.com/Transport-Climate-Change-and-theCity-1st-Edition/Hickman-Banister/p/book/9780415660020

Higgins, C. D., & Kanaroglou, P. S. (2016). A latent class method for classifying and evaluating the performance of station area transit-oriented development in the Toronto region. *Journal of Transport Geography*. https://doi.org/10.1016/j.jtrangeo.2016.02.012

Hincks, S., Kingston, R., Webb, B., & Wong, C. (2018). A new geodemographic classification of commuting flows for England and Wales. *International Journal of Geographical Information Science*, *32*(4), 663–684. https://doi.org/10.1080/13658816.2017.1407416

Hollenstein, L., & Purves, R. S. (2010). Exploring place through user-generated content: Using Flickr tags to describe city cores. *Journal of Spatial Information Science*. https://doi.org/10.5311/JOSIS.2010.1.3

Hsieh, H.-P., Lin, S.-D., & Zheng, Y. (2015). Inferring Air Quality for Station Location Recommendation Based on Urban Big Data. *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 437–446. https://doi.org/10.1145/2783258.2783344

Hu, Y., Gao, S., Janowicz, K., Yu, B., Li, W., & Prasad, S. (2015). Extracting and understanding urban areas of interest using geotagged photos. *Computers, Environment and Urban Systems*. https://doi.org/10.1016/j.compenvurbsys.2015.09.001

Huang, R., Grigolon, A., Madureira, M., & Brussel, M. (2018). Measuring transit-oriented development (TOD) network complementarity based on tod node typology. *Journal of Transport and Land Use*. https://doi.org/10.5198/jtlu.2018.1110

Huff, J. O., & Hanson, S. (1986). Repetition and Variability in Urban Travel. *Geographical Analysis*. https://doi.org/10.1111/j.1538-4632.1986.tb00085.x

Hynes, M. (2017). At a Crossroads: Investigating Automobility and Its Implications for Local Urban Transport Policy Design. *Urban Science*, *1*(2), 14. https://doi.org/10.3390/urbansci1020014

Iliopoulou, C. A., Milioti, C. P., Vlahogianni, E. I., & Kepaptsoglou, K. L. (2020). Identifying spatio-temporal patterns of bus bunching in urban networks. *Journal of Intelligent Transportation Systems: Technology, Planning, and Operations*. https://doi.org/10.1080/15472450.2020.1722949

Ivan, I., Boruta, T., & Horák, J. (2012). Evaluation of railway surrounding areas: The case of Ostrava city. *WIT Transactions on the Built Environment*. https://doi.org/10.2495/UT120131

Jain, D. K., Dubey, S. B., Choubey, R. K., Sinhal, A., Arjaria, S. K., Jain, A., & Wang, H. (2018). An approach for hyperspectral image classification by optimizing SVM using self organizing map. *Journal of Computational Science*. https://doi.org/10.1016/j.jocs.2017.07.016

Jiang, S., Ferreira, J., & Gonzalez, M. C. (2012). Discovering urban spatial-temporal structure from human activity patterns. *Proceedings of the ACM SIGKDD International Workshop on Urban Computing - UrbComp '12*, 95. https://doi.org/10.1145/2346496.2346512

Jolliffe, I. T. (1972). Discarding Variables in a Principal Component Analysis. I: Artificial Data. *Applied Statistics*. https://doi.org/10.2307/2346488

Jolliffe, I. T. (2002). *Principal Component Analysis* (2nd ed.). Springer.

Jun, M. J., Choi, K., Jeong, J. E., Kwon, K. H., & Kim, H. J. (2015). Land use characteristics of subway catchment areas and their influence on subway ridership in Seoul. *Journal of Transport Geography*. https://doi.org/10.1016/j.jtrangeo.2015.08.002

Kaiser, H. F. (1960). The Application of Electronic Computers to Factor Analysis. *Educational and Psychological Measurement*. https://doi.org/10.1177/001316446002000116

Kaisler, S., Armour, F., Espinosa, J. A., & Money, W. (2013). Big Data: Issues and Challenges Moving Forward. *2013 46th Hawaii International Conference on System Sciences*, 995–1004. https://doi.org/10.1109/HICSS.2013.645

Kamruzzaman, M., Baker, D., Washington, S., & Turrell, G. (2014). Advance transit oriented development typology: Case study in brisbane, australia. *Journal of Transport Geography*. https://doi.org/10.1016/j.jtrangeo.2013.11.002

Kandt, J., & Batty, M. (2021). Smart cities, big data and urban policy: Towards urban analytics for the long run. *Cities*. https://doi.org/10.1016/j.cities.2020.102992

Kang, C., Ma, X., Tong, D., & Liu, Y. (2012). Intra-urban human mobility patterns: An urban morphology perspective. *Physica A: Statistical Mechanics and Its Applications*. https://doi.org/10.1016/j.physa.2011.11.005

Kassambara, A. (2017). *Hierarchical K-Means Clustering: Optimize Clusters*. https://www.datanovia.com/en/lessons/hierarchical-k-means-clustering-optimize-clusters/

Kattiyapornpong, U., & Miller, K. E. (2009). Socio-demographic constraints to travel behavior. *International Journal of Culture, Tourism and Hospitality Research*, *3*(1), 81–94. https://doi.org/10.1108/17506180910940360

Keler, A., Krisp, J. M., & Ding, L. (2020). Extracting commuter-specific destination hotspots from trip destination data–comparing the boro taxi service with Citi Bike in NYC. *Geo-Spatial Information Science*. https://doi.org/10.1080/10095020.2019.1621008

Khan, M. A. U. D., Uddin, M. F., & Gupta, N. (2014). Seven V's of Big Data understanding Big Data to extract value. *Proceedings of the 2014 Zone 1 Conference of the American Society for Engineering Education - "Engineering Education: Industry Involvement and Interdisciplinary Trends", ASEE Zone 1 2014*. https://doi.org/10.1109/ASEEZone1.2014.6820689

Kim, M. K., Kim, S. P., Heo, J., & Sohn, H. G. (2017). Ridership patterns at subway stations of Seoul capital area and characteristics of station influence area. *KSCE Journal of Civil Engineering*. https://doi.org/10.1007/s12205-016-1099-8

Kim, M. K., Kim, S., & Sohn, H. G. (2018). Relationship between Spatio-Temporal travel patterns derived from smart-card data and local environmental characteristics of Seoul, Korea. *Sustainability (Switzerland)*. https://doi.org/10.3390/su10030787

Kim, Y. L. (2018). Seoul's Wi-Fi hotspots: Wi-Fi access points as an indicator of urban vitality. *Computers, Environment and Urban Systems*. https://doi.org/10.1016/j.compenvurbsys.2018.06.004

Kinsella, S. (2007). *Indices of Multiple Deprivation : 2000 , 2004 and 2007 Table 2 : Child Poverty Ranking ( IDACI ) and main IMD Ranks for Wirral wards in 2000. December.*

Kitchin, R. (2013). Big data and human geography: Opportunities, challenges and risks. *Dialogues in Human Geography*. https://doi.org/10.1177/2043820613513388

Kitchin, R. (2014a). Big Data, new epistemologies and paradigm shifts. *Big Data and Society*. https://doi.org/10.1177/2053951714528481

Kitchin, R. (2014b). The Data Revolution: Big Data, Open Data, Data Infrastructures & Their Consequences. In *The Data Revolution: Big Data, Open Data, Data Infrastructures & Their Consequences*. https://doi.org/10.4135/9781473909472

Kohonen, T. (1998). The self-organizing map. *Neurocomputing*. https://doi.org/10.1016/S0925-2312(98)00030-7

Kong, L., Liu, Z., & Wu, J. (2020). A systematic review of big data-based urban sustainability research: State-of-the-science and future directions. In *Journal of Cleaner Production*. https://doi.org/10.1016/j.jclepro.2020.123142

Kourtit, K., Elmlund, P., & Nijkamp, P. (2020). The urban data deluge: challenges for smart urban planning in the third data revolution. *International Journal of Urban Sciences*. https://doi.org/10.1080/12265934.2020.1755353

Krefis, A., Augustin, M., Schlünzen, K., Oßenbrügge, J., & Augustin, J. (2018). How Does the Urban Environment Affect Health and Well-Being? A Systematic Review. *Urban Science*, *2*(1), 21. https://doi.org/10.3390/urbansci2010021

Kumar, H., Singh, M. K., Gupta, M. P., & Madaan, J. (2020). Moving towards smart cities: Solutions that lead to the Smart City Transformation Framework. *Technological Forecasting and Social Change*. https://doi.org/10.1016/j.techfore.2018.04.024

Kuo, C. L., Chan, T. C., Fan, I. C., & Zipf, A. (2018). Efficient method for POI/ROI discovery using Flickr geotagged photos. *ISPRS International Journal of Geo-Information*. https://doi.org/10.3390/ijgi7030121

Laney, D. (2001). 3D Data Management: Controlling Data Volume, Velocity, and Variety. *Application Delivery Strategies*.

Lathia, N., Quercia, D., & Crowcroft, J. (2012). *The Hidden Image of the City: Sensing Community Well-Being from Urban Mobility* (pp. 91–98). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-31205-2_6

Lee, S., Yi, C., & Hong, S. P. (2013). Urban structural hierarchy and the relationship between the ridership of the Seoul Metropolitan Subway and the land-use pattern of the station areas. *Cities*. https://doi.org/10.1016/j.cities.2013.06.010

Leventhal, B. (2013). Census-taking in the United Kingdom: 2011 and beyond. *Journal of Direct, Data and Digital Marketing Practice*. https://doi.org/10.1057/dddmp.2013.4

Leventhal, B. (2016). *Geodemographics for Marketers: Using Location Analysis for Research and Marketing*. Kogan Page.

Lew, A., & McKercher, B. (2006). Modeling tourist movements: A local destination analysis. *Annals of Tourism Research*. https://doi.org/10.1016/j.annals.2005.12.002

Liu, Y., & Cheng, T. (2020). Understanding public transit patterns with open geodemographics to facilitate public transport planning. *Transportmetrica A: Transport Science*. https://doi.org/10.1080/23249935.2018.1493549

Liu, Y., Singleton, A., & Arribas-Bel, D. (2019). A Principal Component Analysis (PCA)-based framework for automated variable selection in geodemographic classification. *Geo-Spatial Information Science*, *22*(4), 251–264. https://doi.org/10.1080/10095020.2019.1621549

Liu, Y., Singleton, A., & Arribas-Bel, D. (2020). Considering context and dynamics: A classification of transit-orientated development for New York City. *Journal of Transport Geography*. https://doi.org/10.1016/j.jtrangeo.2020.102711

Liu, Z., & Barahona, M. (2020). Graph-based data clustering via multiscale community detection. *Applied Network Science*. https://doi.org/10.1007/s41109-019-0248-7

Long, Y., & Liu, L. (2016). Transformations of urban studies and planning in the big/open data era: a review. In *International Journal of Image and Data Fusion*. https://doi.org/10.1080/19479832.2016.1215355

Longley, P., Ashby, D., Webber, R., & Li, C. (2006). Geodemographic classifications, the digital divide and understanding customer take-up of new technologies. *BT Technology Journal*, *24*(3), 67–74. https://doi.org/10.1007/s10550-006-0077-y

Longley, Paul. (2005). Geographical information systems: A renaissance of geodemographics for public service delivery. *Progress in Human Geography*. https://doi.org/10.1191/0309132505ph528pr

Longley, Paul, & Goodchild, M. (2008). The Use of Geodemographics to Improve Public Service Delivery. In J. Hartley, C. Donaldson, C. Skelcher, & M. Wallace (Eds.), *Managing to Improve Public Services* (pp. 176–194). Cambridge University Press (CUP).

Longley, Paul, Goodchild, M., Maguire, D., & Rhind, D. (2015). *Geographic Information Science and Systems* (4th ed.). John Wiley & Sons, Ltd. https://www.wiley.com/en-gb/Geographic+Information+Science+and+Systems%2C+4th+Edition-p-9781119031307

Love, R. L., & Chapin, F. S. (1976). Human Activity Patterns in the City: Things People Do in Time and Space. *Contemporary Sociology*. https://doi.org/10.2307/2064142

Lyu, G., Bertolini, L., & Pfeffer, K. (2016). Developing a TOD typology for Beijing metro station areas. *Journal of Transport Geography*. https://doi.org/10.1016/j.jtrangeo.2016.07.002

Ma, A. T. H., Chow, A. S. Y., Cheung, L. T. O., Lee, K. M. Y., & Liu, S. (2018). Impacts of tourists' sociodemographic characteristics on the travel motivation and satisfaction: The case of protected areas in South China. *Sustainability (Switzerland)*. https://doi.org/10.3390/su10103388

Ma, H., Meng, Y., Xing, H., & Li, C. (2019). Investigating road-constrained spatial distributions and semantic attractiveness for area of interest. *Sustainability (Switzerland)*. https://doi.org/10.3390/su11174624

Ma, J., Mitchell, G., & Heppenstall, A. (2014). Daily travel behaviour in Beijing, China.An analysis of workers' trip chains, and the role of socio-demographics and urban form. *Habitat International*. https://doi.org/10.1016/j.habitatint.2014.04.008

Maat, K., van Wee, B., & Stead, D. (2005). Land use and travel behaviour: Expected effects from the perspective of utility theory and activity-based theories. *Environment and Planning B: Planning and Design*. https://doi.org/10.1068/b31106

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*.

Mahrsi, M. K. El, Côme, E., Baro, J., & Oukhellou, L. (2014). Understanding Passenger Patterns in Public Transit Through Smart Card and Socioeconomic Data. *The 3rd International Workshop on Urban Computing (UrbComp 2014)*.

Malhi, A., & Gao, R. X. (2004). PCA-based feature selection scheme for machine defect classification. *IEEE Transactions on Instrumentation and Measurement*. https://doi.org/10.1109/TIM.2004.834070

Martin, K. E. (2015). Ethical issues in the big data industry. *MIS Quarterly Executive*, *14*(2), 67–85. https://doi.org/10.4324/9780429286797-20

Mason, G. A., & Jacobson, R. D. (2006). *Fuzzy Geographically Weighted Clustering*. *1998*, 1–7.

Mclennan, A. D., Noble, S., Noble, M., Plunkett, E., Wright, G., & Gutacker, N. (2019). *IoD2019_Technical_Report*. *September*.

McPhearson, T., Haase, D., Kabisch, N., & Gren, Å. (2016). Advancing understanding of the complex nature of urban systems. In *Ecological Indicators*. https://doi.org/10.1016/j.ecolind.2016.03.054

McPhearson, T., Parnell, S., Simon, D., Gaffney, O., Elmqvist, T., Bai, X., Roberts, D., & Revi, A. (2016). Scientists must have a say in the future of cities. In *Nature*. https://doi.org/10.1038/538165a

Meng, X., Zhang, K., Pang, K., & Xiang, X. (2020). Characterization of spatio-temporal distribution of vehicle emissions using web-based real-time traffic data. *Science of the Total Environment*. https://doi.org/10.1016/j.scitotenv.2019.136227

Michel, A., & Ribardière, A. (2017). Identifying urban resources to read socio-spatial inequalities. *EchoGéo*, *39*, 0–6. https://doi.org/10.4000/echogeo.14943

Miljkovic, D. (2017). Brief review of self-organizing maps. *2017 40th International Convention on Information and Communication Technology, Electronics and Microelectronics, MIPRO 2017 - Proceedings*. https://doi.org/10.23919/MIPRO.2017.7973581

Monajem, S., & Ekram Nosratian, F. (2015). The evaluation of the spatial integration of station areas via the node place model; an application to subway station areas in Tehran. *Transportation Research Part D: Transport and Environment*. https://doi.org/10.1016/j.trd.2015.07.009

Moorthy, J., Lahiri, R., Biswas, N., Sanyal, D., Ranjan, J., Nanath, K., & Ghosh, P. (2015). Big Data: Prospects and Challenges. *Vikalpa*. https://doi.org/10.1177/0256090915575450

Moss, M., & Qing, C. (2012). *The Dynamic Population of Manhattan*. http://citeseerx.ist.psu.edu/viewdoc/summary;jsessionid=8F1D344B6C1E50AE2608CE9D3F23D297?doi=10.1.1.392.6121

Mouratidis, K. (2018). Built environment and social well-being: How does urban form affect social life and personal relationships? *Cities*, *74*(November 2017), 7–20. https://doi.org/10.1016/j.cities.2017.10.020

MTA. (2016). *Introduction to Subway Ridership*. http://web.mta.info/nyct/facts/ridership/

Murphy, S., & Smith, M. (2014). Geodemographic model variable selection spacial data mining of the 2011 Irish census. *Souvenir of the 2014 IEEE International Advance Computing Conference, IACC 2014*. https://doi.org/10.1109/IAdCC.2014.6779395

Narayanaswami, S. (2016). *Urban transportation: trends, challenges and opportunities*. http://parisinnovationreview.com/articles-en/urban-transportation-trends-challenges-and-opportunities

Nasri, A., & Zhang, L. (2014). The analysis of transit-oriented development (TOD) in Washington, D.C. and Baltimore metropolitan areas. *Transport Policy*. https://doi.org/10.1016/j.tranpol.2013.12.009

Natita, W., Wiboonsak, W., & Dusadee, S. (2016). Appropriate Learning Rate and Neighborhood Function of Self-organizing Map (SOM) for Specific Humidity Pattern Classification over Southern Thailand. *International Journal of Modeling and Optimization*. https://doi.org/10.7763/ijmo.2016.v6.504

Nativi, S., Santoro, M., Giuliani, G., & Mazzetti, P. (2020). Towards a knowledge base to support global change policy goals. *International Journal of Digital Earth*, *13*(2), 188–216. https://doi.org/10.1080/17538947.2018.1559367

Ni, X., Huang, H., Meng, Y., Zhou, S., & Su, B. (2019). An urban road-traffic commuting dynamics study based on hotspot clustering and a new proposed urban commuting electrostatics model. *ISPRS International Journal of Geo-Information*. https://doi.org/10.3390/ijgi8040190

Noble, M., Wright, G., Smith, G., & Dibben, C. (2006). Measuring multiple deprivation at the small-area level. *Environment and Planning A*. https://doi.org/10.1068/a37168

Nurulin, Y., Skvortsova, I., Tukkel, I., & Torkkeli, M. (2019). Role of knowledge in management of innovation. *Resources*. https://doi.org/10.3390/resources8020087

NYDOT. (2018). NYC Mobility Report. *White Paper*, *June*, 7–8. http://www.nyc.gov/html/dot/downloads/pdf/mobility-report-2018-print.pdf

O'Dowd, L. (2003). Ecological fallacy. In R. Miller & J. Brewer (Eds.), *The A-Z of Social Research* (pp. 84–85). SAGE Publication Ltd. http://ndl.ethernet.edu.et/bitstream/123456789/25472/1/4.pdf.pdf

Office for National Statistics. (2015). *2011 Census General Report*. https://www.ons.gov.uk/census/2011census/howourcensusworks/howdidwedoin2011/2011censusgeneralreport

Office for National Statistics. (2016). *Quality and methods*. https://www.ons.gov.uk/census/2011census/2011censusdata/2011censususerguide/qualityandmethods

Oliver, L. N., Schuurman, N., & Hall, A. W. (2007). Comparing circular and network buffers to examine the influence of land use on walking for leisure and errands. *International Journal of Health Geographics*. https://doi.org/10.1186/1476-072X-6-41

Openshaw, S., Blake, M., & Wymer, C. (1995). Using neurocomputing methods to classify Britain's residential areas. *Innovations in GIS 2*.

Openshaw, S., & Wymer, C. (1995). Classifying and Regionalising Census Data. In S. Openshaw (Ed.), *Census User's Handbook* (pp. 353–361). Geoinformation International.

Orford, S., Dorling, D., Mitchell, R., Shaw, M., & Smith, G. D. (2002). Life and death of the people of London: A historical GIS of Charles Booth's inquiry. *Health and Place*. https://doi.org/10.1016/S1353-8292(01)00033-8

Pacheco, E. (2015). *Unsupervised Learning with R*. Packt Publishing.

Pan, H., Shen, Q., & Zhang, M. (2009). Influence of urban form on travel behaviour in four neighbourhoods of Shanghai. *Urban Studies*. https://doi.org/10.1177/0042098008099355

Panimalar, A., Shree, V., & Kathrine, V. (2017). The 17 V's of Big Data. *International Research Journal of Engineering and Technology(IRJET)*, *4*(9), 3–6. https://irjet.net/archives/V4/i9/IRJET-V4I957.pdf

Papa, E., Carpentieri, G., & Angiello, G. (2018). A TOD classification of metro stations: An application in Naples. In *Green Energy and Technology*. https://doi.org/10.1007/978-3-319-77682-8_17

Pasichnyi, O., Wallin, J., Levihn, F., Shahrokni, H., & Kordas, O. (2019). Energy performance certificates — New opportunities for data-enabled urban energy policy instruments? *Energy Policy*. https://doi.org/10.1016/j.enpol.2018.11.051

Pelletier, M. P., Trépanier, M., & Morency, C. (2011). Smart card data use in public transit: A literature review. *Transportation Research Part C: Emerging Technologies*. https://doi.org/10.1016/j.trc.2010.12.003

Pollack, S., Gartsman, A., Benedict, A., & Wood, J. (2014). Rating the Performance of Station Areas for Effective and Equitable Transit Oriented Development. *Transportation Research Board 2014 Compendium of Papers*.

Qin, K., Zhou, Q., Wu, T., & Xu, Y. Q. (2017). Hotspots detection from trajectory data based on spatiotemporal data field clustering. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences - ISPRS Archives*. https://doi.org/10.5194/isprs-archives-XLII-2-W7-1319-2017

Rani, K. S., Kumari, M., Singh, V. B., & Sharma, M. (2019). Deep Learning with Big Data: An Emerging Trend. *2019 19th International Conference on Computational Science and Its Applications (ICCSA)*, 93–101. https://doi.org/10.1109/ICCSA.2019.00005

Ravenstein, E. G. (1885). The Laws of Migration. *Journal of the Statistical Society of London*. https://doi.org/10.2307/2979181

Reibel, M. (2011). Classification approaches in neighborhood research: Introduction and review. In *Urban Geography*. https://doi.org/10.2747/0272-3638.32.3.305

Reinsel, D., Gantz, J., & Rydning, J. (2018). The Digitization of the World - From Edge to Core. *Framingham: International Data Corporation*, *November*, US44413318. https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf

Renjith, S., Sreekumar, A., & Jathavedan, M. (2020). *Pragmatic Evaluation of the Impact of Dimensionality Reduction in the Performance of Clustering Algorithms* (pp. 499–512). Springer, Singapore. https://doi.org/10.1007/978-981-15-5558-9_45

Riddlesden, D., & Singleton, A. D. (2014). Broadband speed equity: A new digital divide? *Applied Geography*. https://doi.org/10.1016/j.apgeog.2014.04.008

Robinson, G. (1998). *Methods and Techniques in Human Geography*. John Wiley & Sons, Ltd.

Rodrigue, J.-P. (2020). *The Geography of Transport Systems* (5th ed.). Routledge. https://www.routledge.com/The-Geography-of-Transport-Systems/Rodrigue/p/book/9780367364632

Rojas, R. (2015). *The Curse of Dimensionality*. https://www.inf.fu-berlin.de/inst/ag-ki/rojas_home/documents/tutorials/dimensionality.pdf

Rossetti, R. J. F. (2015). Internet of Things (IoT) and Smart Cities. *IEEE Xplore] Readings on Smart Cities --[Editorial*.

Rowley, J. (2007). The wisdom hierarchy: Representations of the DIKW hierarchy. *Journal of Information Science*. https://doi.org/10.1177/0165551506070706

Rubinstein, I. S. (2013). Big data: The end of privacy or a new beginning? *International Data Privacy Law*. https://doi.org/10.1093/idpl/ips036

Sambandam, R. (2003). Cluster Analysis Gets Complicated. In *Marketing Research*.

Sampson, R. (2012). *Great American City: Chicago and the Enduring Neighborhood Effect*. University of Chicago Press. https://press.uchicago.edu/ucp/books/book/chicago/G/bo5514383.html

Santeo, K., Nayang, N., & Ibrahim, S. (2016). Improving the tool for analyzing Malaysia's demographic change: data standardization analysis to form geo-demographics classification profiles using k-means algorithms. *Geografia : Malaysian Journal of Society and Space*.

Schäfer, P., Pflugmacher, D., Hostert, P., & Leser, U. (2018). Classifying land cover from satellite images using time series analytics. *CEUR Workshop Proceedings*.

Schmitt, R. C. (1956). Estimating Daytime Populations. *Journal of the American Planning Association*. https://doi.org/10.1080/01944365608979227

She, Z., King, D. M., & Jacobson, S. H. (2017). Analyzing the impact of public transit usage on obesity. *Preventive Medicine*, *99*, 264–268. https://doi.org/10.1016/j.ypmed.2017.03.010

Shen, J., & Cheng, T. (2016). A framework for identifying activity groups from individual space-time profiles. *International Journal of Geographical Information Science*, *30*(9), 1785–1805. https://doi.org/10.1080/13658816.2016.1139119

Shi, Z., & Pun-Cheng, L. S. C. (2019). Spatiotemporal data clustering: A survey of methods. In *ISPRS International Journal of Geo-Information*. https://doi.org/10.3390/ijgi8030112

Shirkhorshidi, A. S., Aghabozorgi, S., & Ying Wah, T. (2015). A Comparison study on similarity and dissimilarity measures in clustering continuous data. *PLoS ONE*, *10*(12), 1–20. https://doi.org/10.1371/journal.pone.0144059

Sinaga, K. P., & Yang, M. S. (2020). Unsupervised K-means clustering algorithm. *IEEE Access*, *8*, 80716–80727. https://doi.org/10.1109/ACCESS.2020.2988796

Singh, Y. J., Lukman, A., Flacke, J., Zuidgeest, M., & Van Maarseveen, M. F. A. M. (2017). Measuring TOD around transit nodes - Towards TOD policy. *Transport Policy*. https://doi.org/10.1016/j.tranpol.2017.03.013

Singleton, A. (2010). The Geodemographics of Educational Progression and their Implications for Widening Participation in Higher Education. *Environment and Planning A: Economy and Space*, *42*(11), 2560–2580. https://doi.org/10.1068/a42394

Singleton, A. (2016). Cities and Context: The Codification of Small Areas through Geodemographic Classification. In R. Kitchin & S. Perng (Eds.), *Code and the City* (pp. 215–235). Routledge.

Singleton, A., Alexiou, A., & Savani, R. (2020). Mapping the geodemographics of digital inequality in Great Britain: An integration of machine learning into small area estimation. *Computers, Environment and Urban Systems*, *82*, 101486. https://doi.org/10.1016/j.compenvurbsys.2020.101486

Singleton, A., Dolega, L., Riddlesden, D., & Longley, P. A. (2016). Measuring the spatial vulnerability of retail centres to online consumption through a framework of e-resilience. *Geoforum*, *69*, 5–18. https://doi.org/10.1016/j.geoforum.2015.11.013

Singleton, A., & Longley, P. (2009). Creating open source geodemographics: Refining a national classification of census output areas for applications in higher education. *Papers in Regional Science*. https://doi.org/10.1111/j.1435-5957.2008.00197.x

Singleton, A., & Longley, P. (2015). The internal structure of Greater London: a comparison of national and regional geodemographic models. *Geo: Geography and Environment*. https://doi.org/10.1002/geo2.7

Singleton, A., & Spielman, S. (2014). The Past, Present, and Future of Geodemographic Research in the United States and United Kingdom. *Professional Geographer*, *66*(4), 558–567. https://doi.org/10.1080/00330124.2013.848764

Singleton, A., Spielman, S., & Folch, D. (2017). *Urban Analytics*. SAGE Publication Ltd. https://uk.sagepub.com/en-gb/eur/urban-analytics/book249267

Sleeter, R., & Wood, N. (2006). Estimating daytime and nighttime population density for coastal communites in oregon. *44th Urban and Regional Information Systems Association Annual Conference, British Columbia.*

Sleight, P. (1993). *Targeting Customers: How to Use Geodemographic and Lifestyle Data in Your Business*. NTC Publications.

Sleight, P. (1997). *Targeting customers: How to use geodemographic and lifestyle data in your business* (2nd ed.). NTC Publications.

Smith, T., Noble, M., Noble, S., Wright, G., McLennan, D., & Plunkett, E. (2015). *The English Indices of Deprivation 2015 - Technical Report. Department for Communities and Local Government. London, UK.* (Issue September). https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/464485/English_Indices_of_Deprivation_2015_-_Technical-Report.pdf

Smolan, R., & Erwitt, J. (2012). *The Human Face of Big Data*. Sterling.

Sohn, K. (2013). Feature Mapping the Seoul Metro Station Areas Based on a Self-Organizing Map. *Journal of Urban Technology*. https://doi.org/10.1080/10630732.2013.855514

Sokoloff, J. (2018). *How far do people commute using Bike Sharing Systems?* https://medium.com/data-tale/how-far-do-people-travel-in-bike-sharing-systems-faf0295bc75a

Someh, I., Davern, M., Breidbach, C. F., & Shanks, G. (2019). Ethical issues in big data analytics: A stakeholder perspective. *Communications of the Association for Information Systems*. https://doi.org/10.17705/1CAIS.04434

SONG, J., & DEGUCHI, A. (2013). EVALUATION AND TYPOLOGY OF RAILWAY STATION AREAS IN A 30KM CIRCUMFERENCE SURROUNDING CENTRAL TOKYO FROM VIEW OF TRANSIT-ORIENTED DEVELOPMENT. *Journal of Architecture and Planning (Transactions of AIJ)*. https://doi.org/10.3130/aija.78.413

Song, S., Xia, T., Jin, D., Hui, P., & Li, Y. (2019). UrbanRhythm: Revealing urban dynamics hidden in mobility data. In *arXiv*.

Spielman, S., & Folch, D. (2015). Social area analysis and self-organizing maps. In C. Brunsdon & A. Singleton (Eds.), *Geocomputation: a practial primer* (pp. 152–168). SAGE Publication Ltd.

Spielman, S., & Singleton, A. (2015). Studying Neighborhoods Using Uncertain Data from the American Community Survey: A Contextual Approach. *Annals of the Association of American Geographers*. https://doi.org/10.1080/00045608.2015.1052335

Spielman, S., & Thill, J. C. (2008). Social area analysis, data mining, and GIS. *Computers, Environment and Urban Systems*. https://doi.org/10.1016/j.compenvurbsys.2007.11.004

Staricco, L., & Vitale Brovarone, E. (2018). Promoting TOD through regional planning. A comparative analysis of two European approaches. *Journal of Transport Geography*. https://doi.org/10.1016/j.jtrangeo.2017.11.011

Stevenson, M., & Gleeson, B. (2019). Complex Urban Systems: Compact Cities, Transport and Health. In *Integrating Human Health into Urban and Transport Planning* (pp. 271–285). Springer International Publishing. https://doi.org/10.1007/978-3-319-74983-9_14

Sung, H., & Choi, C. G. (2017). The link between metropolitan planning and transit-oriented development: An examination of the Rosario Plan in 1980 for Seoul, South Korea. *Land Use Policy*. https://doi.org/10.1016/j.landusepol.2017.01.045

Syam, A., Khan, A., & Reeves, D. (2012). Demographics do matter: An analysis of people's travel behaviour of different ethnic groups in Auckland. *WIT Transactions on the Built Environment*. https://doi.org/10.2495/UT120441

Taki, H. M., Maatouk, M. M. H., Qurnfulah, E. M., & Aljoufie, M. O. (2017). Planning TOD with land use and transport integration: a review. *Journal of Geoscience, Engineering, Environment, and Technology*. https://doi.org/10.24273/jgeet.2017.2.1.17

Tang, Jiliang, Alelyani, S., & Liu, H. (2014). Feature selection for classification: A review. In *Data Classification: Algorithms and Applications*. https://doi.org/10.1201/b17320

Tang, Jinjun, Liu, F., Wang, Y., & Wang, H. (2015). Uncovering urban human mobility from large scale taxi GPS data. *Physica A: Statistical Mechanics and Its Applications*, *438*, 140–153. https://doi.org/10.1016/j.physa.2015.06.032

Taxi and Limousine Commission. (2018). *2018 Fact Book*. https://www1.nyc.gov/assets/tlc/downloads/pdf/2018_tlc_factbook.pdf

Thakuriah, P., Tilahun, N. Y., & Zellner, M. (2017). *Big Data and Urban Informatics: Innovations and Challenges to Urban Planning and Knowledge Discovery* (pp. 11–45). Springer, Cham. https://doi.org/10.1007/978-3-319-40902-3_2

Thériault, M., & Des Rosiers, F. D. (2013). Modeling Urban Dynamics: Mobility, Accessibility and Real Estate Value. *Modeling Urban Dynamics: Mobility, Accessibility and Real Estate Value*. https://doi.org/10.1002/9781118558041

Thiele, T., Singleton, A., Pope, D., & Stanistreet, D. (2016). Predicting students' academic performance based on school and socio-demographic characteristics. *Studies in Higher Education*, *41*(8), 1424–1446. https://doi.org/10.1080/03075079.2014.974528

Thomas, R., Pojani, D., Lenferink, S., Bertolini, L., Stead, D., & van der Krabben, E. (2018). Is transit-oriented development (TOD) an internationally transferable policy concept? *Regional Studies*. https://doi.org/10.1080/00343404.2018.1428740

Tian, J., Azarian, M. H., & Pecht, M. (2014). Anomaly Detection Using Self-Organizing Maps-Based K -Nearest Neighbor Algorithm. *Proceedings of the European Conference of the Prognostics and Health Management*.

Tibshirani, R., Walther, G., & Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*. https://doi.org/10.1111/1467-9868.00293

Timms, D. (1971). *The Urban Mosaic: Towards a Theory of Residential Differentiation*. Cambridge University Press (CUP).

Tobler W. (1970). A computer movie simulating urban growth in the Detroit region. *Economic Geography*.

Townsend, P. (1987). Deprivation. *Journal of Social Policy*, *16*(2), 125–146. https://doi.org/10.1017/S0047279400020341

Tran Thi Hoang, G., Dupont, L., & Camargo, M. (2019). Application of Decision-Making Methods in Smart City Projects: A Systematic Literature Review. *Smart Cities*. https://doi.org/10.3390/smartcities2030027

Transport for London. (2017). *Transport Classification of Londoners (TCoL). February*.

Troy, A. (2017). Geodemographic Segmentation. In *Encyclopedia of GIS* (pp. 667–678). Springer International Publishing. https://doi.org/10.1007/978-3-319-17885-1_456

Tumwebaze, I. K., Rose, J. B., Hofstra, N., Verbyla, M. E., Okaali, D. A., Katsivelis, P., & Murphy, H. M. (2021). Bridging Science and Practice-Importance of Stakeholders in the Development of Decision Support: Lessons Learned. *Sustainability*. https://doi.org/10.3390/su13105744

Udovičić, M., Baždarić, K., Bilić-Zulle, L., & Petrovečki, M. (2007). What we need to know when calculating the coefficient of correlation? *Biochemia Medica*. https://doi.org/10.11613/bm.2007.002

UN. (2015). *Transforming our world: the 2030 Agenda for Sustainable Development*. https://sustainabledevelopment.un.org/post2015/transformingourworld

UNDESA. (2018). World Urbanization Prospects. In *Demographic Research* (Vol. 12). https://population.un.org/wup/Publications/Files/WUP2018-Report.pdf

United States Census Bureau. (2020). *Subjects Included in the Survey*. https://www.census.gov/programs-surveys/acs/guidance/subjects.html

Urban, S. (2009). Is the Neighbourhood Effect an Economic or an Immigrant Issue? A Study of the Importance of the Childhood Neighbourhood for Future Integration into the Labour Market. *Urban Studies*, *46*(3), 583–603. https://doi.org/10.1177/0042098008100996

US Census Bureau. (2019). *QuickFacts*. https://www.census.gov/quickfacts/newyorkcitynewyork

Üsküplü, T., Terzi, F., & Kartal, H. (2020). Discovering Activity Patterns in the City by Social Media Network Data: a Case Study of Istanbul. *Applied Spatial Analysis and Policy*. https://doi.org/10.1007/s12061-020-09336-5

Vale, D. S. (2015). Transit-oriented development, integration of land use and transport, and pedestrian accessibility: Combining node-place model with pedestrian shed ratio to evaluate and classify station areas in Lisbon. *Journal of Transport Geography*. https://doi.org/10.1016/j.jtrangeo.2015.04.009

van der Zee, E., Bertocchi, D., & Vanneste, D. (2020). Distribution of tourists within urban heritage destinations: a hot spot/cold spot analysis of TripAdvisor data as support for destination management. *Current Issues in Tourism*. https://doi.org/10.1080/13683500.2018.1491955

van Lierop, D., Maat, K., & El-Geneidy, A. (2017). Talking TOD: learning about transit-oriented development in the United States, Canada, and the Netherlands. *Journal of Urbanism*. https://doi.org/10.1080/17549175.2016.1192558

van Tilburg, C. (2006). Traffic and congestion in the Roman Empire. In *Traffic and Congestion in the Roman Empire*. https://doi.org/10.4324/9780203968031

Vickers, D., & Rees, P. (2007). Creating the UK National Statistics 2001 output area classification. *Journal of the Royal Statistical Society. Series A: Statistics in Society*. https://doi.org/10.1111/j.1467-985X.2007.00466.x

Wan, L., Gao, S., Wu, C., Jin, Y., Mao, M., & Yang, L. (2018). Big data and urban system model - Substitutes or complements? A case study of modelling commuting patterns in Beijing. *Computers, Environment and Urban Systems*. https://doi.org/10.1016/j.compenvurbsys.2017.10.004

Wang, H., Huang, H., Ni, X., & Zeng, W. (2019). Revealing Spatial-Temporal Characteristics and Patterns of Urban Travel: A Large-Scale Analysis and Visualization Study with Taxi GPS Data. *ISPRS International Journal of Geo-Information*, *8*(6), 257. https://doi.org/10.3390/ijgi8060257

Wang, J., Kong, X., Rahim, A., Xia, F., Tolba, A., & Al-Makhadmeh, Z. (2017a). IS2Fun: Identification of Subway Station Functions Using Massive Urban Data. *IEEE Access*. https://doi.org/10.1109/ACCESS.2017.2766237

Wang, J., Kong, X., Rahim, A., Xia, F., Tolba, A., & Al-Makhadmeh, Z. (2017b). IS2Fun: Identification of Subway Station Functions Using Massive Urban Data. *IEEE Access*, *5*, 27103–27113. https://doi.org/10.1109/ACCESS.2017.2766237

Wang, J., Kong, X., Xia, F., & Sun, L. (2019). Urban Human Mobility. *ACM SIGKDD Explorations Newsletter*, *21*(1), 1–19. https://doi.org/10.1145/3331651.3331653

Wang, S., Sun, L., Rong, J., Hao, S., & Luo, W. (2016). Transit trip distribution model considering land use differences between catchment areas. *Journal of Advanced Transportation*. https://doi.org/10.1002/atr.1431

Wang, W., Pan, L., Yuan, N., Zhang, S., & Liu, D. (2015). A comparative analysis of intra-city human mobility by taxi. *Physica A: Statistical Mechanics and Its Applications*, *420*, 134–147. https://doi.org/10.1016/j.physa.2014.10.085

Ward, J. H. (1963). Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association*. https://doi.org/10.1080/01621459.1963.10500845

Webb, R., Bai, X., Smith, M. S., Costanza, R., Griggs, D., Moglia, M., Neuman, M., Newman, P., Newton, P., Norman, B., Ryan, C., Schandl, H., Steffen, W., Tapper, N., & Thomson, G. (2018). Sustainable urban systems: Co-design and framing for transformation. *Ambio*. https://doi.org/10.1007/s13280-017-0934-6

Webber, R. (1975). Liverpool Social Area Study 1971 Data: Final Report. In *PRAG Technical Paper 14*. Planning Research Applications Group, Centre for Environmental Studies. https://catalogue.nla.gov.au/Record/746774

Webber, R. (1978). Making the Most of the Census for Strategic Analysis. *Town Planning Review*, *49*(3), 274. https://doi.org/10.3828/tpr.49.3.l56260145j64ml57

Webber, R., & Burrows, R. (2018). *The Predictive Postcode: The Geodemographic Classification of British Society*. SAGE Publication Ltd. https://uk.sagepub.com/en-gb/eur/the-predictive-postcode/book254638

Webber, R., & Craig, J. (1978). *Socio-Economic Classifications of Local Authority Areas (Studies on Medical and Population Subjects)*. Office of Population, Censuses and Surveys.

Weber, R., Schek, H. J., & Blott, S. (1998). A Similarity-Search Analysis Methods and Performance Study for in High-Dimensional Spaces. *Proceedings of the 24th VLDB Conference*, 1–8.

Wegener, M. (1995). Current and future land use models. *Land Use Model Conference, February*, 19–21. http://spiekermann-wegener.com/pub/pdf/MW_Dallas.pdf

Willis, G., & Tranos, E. (2021). Using ‘Big Data’ to understand the impacts of Uber on taxis in New York City. *Travel Behaviour and Society*, *22*, 94–107. https://doi.org/10.1016/j.tbs.2020.08.003

Winters, M., Buehler, R., & Götschi, T. (2017). Policies to Promote Active Travel: Evidence from Reviews of the Literature. In *Current environmental health reports*. https://doi.org/10.1007/s40572-017-0148-x

World Bank. (2020). *Urban Development: Overview*. https://www.worldbank.org/en/topic/urbandevelopment/overview

World Economic Forum. (2019). Global Risks Report 2019. In *Geneva Switzerland*.

Wu, Y., Wang, L., Fan, L., Yang, M., Zhang, Y., & Feng, Y. (2020). Comparison of the spatiotemporal mobility patterns among typical subgroups of the actual population with mobile phone data: A case study of Beijing. *Cities*. https://doi.org/10.1016/j.cities.2020.102670

Xia, F., Liu, L., Jedari, B., & Das, S. K. (2016). PIS: A Multi-Dimensional Routing Protocol for Socially-Aware Networking. *IEEE Transactions on Mobile Computing*, *15*(11), 2825–2836. https://doi.org/10.1109/TMC.2016.2517649

Xia, F., Wang, J., Kong, X., Wang, Z., Li, J., & Liu, C. (2018). Exploring Human Mobility Patterns in Urban Scenarios: A Trajectory Data Perspective. *IEEE Communications Magazine*. https://doi.org/10.1109/MCOM.2018.1700242

Xiong, C., Hu, S., Yang, M., Luo, W., & Zhang, L. (2020). Mobile device data reveal the dynamics in a positive relationship between human mobility and COVID-19 infections. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(44), 27087–27089. https://doi.org/10.1073/pnas.2010836117

Xu, F. F., Lin, B. Y., Lu, Q., Huang, Y., & Zhu, K. Q. (2016). Cross-region traffic prediction for China on OpenStreetMap. *Proceedings of the 9th ACM SIGSPATIAL International Workshop on Computational Transportation Science, IWCTS 2016*. https://doi.org/10.1145/3003965.3003972

Xu, W. A., Guthrie, A., Fan, Y., & Li, Y. (2017). Transit-oriented development in China: Literature review and evaluation of TOD potential across 50 Chinese cities. *Journal of Transport and Land Use*. https://doi.org/10.5198/jtlu.2017.1217

Xu, Z., Cui, G., Zhong, M., & Wang, X. (2019). Anomalous urban mobility pattern detection based on GPS trajectories and POI data. *ISPRS International Journal of Geo-Information*. https://doi.org/10.3390/ijgi8070308

Yabe, T., Tsubouchi, K., Fujiwara, N., Wada, T., Sekimoto, Y., & Ukkusuri, S. V. (2020). Non-compulsory measures sufficiently reduced human mobility in Tokyo during the COVID-19 epidemic. *Scientific Reports*. https://doi.org/10.1038/s41598-020-75033-5

Yang, C., & Gidófalvi, G. (2018). Mining and visual exploration of closed contiguous sequential patterns in trajectories. *International Journal of Geographical Information Science*, *32*(7), 1282–1304. https://doi.org/10.1080/13658816.2017.1393542

Yang, X., Zhao, Z., & Lu, S. (2016). Exploring spatial-temporal patterns of urban human mobility hotspots. *Sustainability (Switzerland)*. https://doi.org/10.3390/su8070674

Yigitcanlar, T., & Cugurullo, F. (2020). The sustainability of artificial intelligence: an urbanistic viewpoint from the lens of smart and sustainable cities. *Sustainability (Switzerland)*. https://doi.org/10.3390/su12208548

Yin, H. (2008). The self-organizing maps: Background, theories, extensions and applications. *Studies in Computational Intelligence*. https://doi.org/10.1007/978-3-540-78293-3_17

Yuan, J., Zheng, Y., & Xie, X. (2012). Discovering regions of different functions in a city using human mobility and POIs. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. https://doi.org/10.1145/2339530.2339561

Yuan, N. J., Zheng, Y., & Xie, X. (2012). Segmentation of Urban Areas Using Road Networks. *Msr-Tr-2012-65*. https://www.microsoft.com/en-us/research/wp-content/uploads/2016/02/mapsegmentation.pdf

Yuan, Y, & Raubal, M. (2012). Extracting Dynamic Urban Mobility Patterns from Mobile Phone Data. In N. Xiao, M. Kwan, M. Goodchild, & S. Shekhar (Eds.), *Geographic Information Science. GIScience 2012. Lecture Notes in Computer Science* (pp. 354–367). Springer.

Yuan, Yihong, & Raubal, M. (2012). Extracting dynamic urban mobility patterns from mobile phone data. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *7478 LNCS*, 354–367. https://doi.org/10.1007/978-3-642-33024-7_26

Yun, Y., Park, Y., Woo, C., & Lim, S. (2019). Speed Accuracy Trade-off in Pedestrian and Vehicle Detection Using Localized Big Data. *Proceedings - 2018 IEEE International Conference on Big Data, Big Data 2018*. https://doi.org/10.1109/BigData.2018.8622161

Zemp, S., Stauffacher, M., Lang, D. J., & Scholz, R. W. (2011). Classifying railway stations for strategic transport and land use planning: Context matters! *Journal of Transport Geography*. https://doi.org/10.1016/j.jtrangeo.2010.08.008

Zhang, L., Chen, C., Wang, Y., & Guan, X. (2016). Exploiting Taxi Demand Hotspots Based on Vehicular Big Data Analytics. *2016 IEEE 84th Vehicular Technology Conference (VTC-Fall)*, 1–5. https://doi.org/10.1109/VTCFall.2016.7881010

Zhang, Yan, Wang, L., Zhang, Y. Q., & Li, X. (2012). Towards a temporal network analysis of interactive WiFi users. *EPL*. https://doi.org/10.1209/0295-5075/98/68002

Zhang, Yang, Sari Aslam, N., Lai, J., & Cheng, T. (2020). You are how you travel: A multi-task learning framework for Geodemographic inference using transit smart card data. *Computers, Environment and Urban Systems*, *83*(April), 101517. https://doi.org/10.1016/j.compenvurbsys.2020.101517

Zhao, K., Tarkoma, S., Liu, S., & Vo, H. (2016). Urban human mobility data mining: An overview. *2016 IEEE International Conference on Big Data (Big Data)*, 1911–1920. https://doi.org/10.1109/BigData.2016.7840811

Zhou, T., Liu, X., Qian, Z., Chen, H., & Tao, F. (2019). Automatic identification of the social functions of areas of interest (AOIS) using the standard hour-day-spectrum approach. *ISPRS International Journal of Geo-Information*. https://doi.org/10.3390/ijgi9010007

Zhou, Y., Fang, Z., Thill, J. C., Li, Q., & Li, Y. (2015). Functionally critical locations in an urban transportation network: Identification and space-time analysis using taxi trajectories. *Computers, Environment and Urban Systems*. https://doi.org/10.1016/j.compenvurbsys.2015.03.001

Zhou, Y., Fang, Z., Zhan, Q., Huang, Y., & Fu, X. (2017). Inferring social functions available in the metro station area from passengers' staying activities in smart card data. *ISPRS International Journal of Geo-Information*. https://doi.org/10.3390/ijgi6120394

Zikopoulos, P., Eaton, C., Deroos, D., Deutsch, T., & Lapis, G. (2012). Understanding big data: analytics for enterprise class Hadoop and streaming data. In *Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data*.