UNIVERSITY OF LEICESTER

DOCTORAL THESIS

Learning Digital Geographies through Geographical Artificial Intelligence

Author:
Pengyuan Liu

Supervisor:
Dr. Stefano De Sabbata
Prof. Yudong Zhang

A thesis submitted in fulfillment of the requirements for the degree of Doctor of Philosophy

in the

School of Geography, Geology and Environment

June 28, 2021

Declaration of Authorship

I, Pengyuan Liu, declare that this thesis titled, "Learning Digital Geographies through Geographical Artificial Intelligence" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: Pengyuan Liu

Date: 31/10/2020

UNIVERSITY OF LEICESTER

Abstract

College of Science and Engineering
School of Geography, Geology and Environment

Doctor of Philosophy

Learning Digital Geographies through Geographical Artificial Intelligence

by Pengyuan Liu

As the distinction between online and physical spaces rapidly degrades, digital platforms have become an integral component of how people's everyday experiences are mediated. User-generated content (UGC) shared on such platforms provides insights into how users want to represent their everyday lives, which augments and reinforces our understanding of local communities through time and layers dynamic information across and over the geographic space.

Inspired by the development of the newly arisen scientific disciplines within geography: *geographical artificial intelligence* (GeoAI), this thesis adopts deep learning approaches on graph representations of human dynamics illustrated through geotagged UGC to explore how place representations are augmented and reinforced through users' spatial experiences by classifying their multimedia activities and identifying the spatial clusters of UGC at the urban scale. Having the place representations described through UGC, this thesis explores how these representations can be used in conjunction with various official spatial statistics to understand and predict the dynamic changes of the socio-economic characteristics of places.

The principal contributions of this thesis are: (1) to provide frameworks with higher classification and prediction accuracy but requiring fewer sample data; thus, contributing to an advanced framework to summarise spatial characteristics of places; (2) to show that multimedia content provides rich information regarding places, the use of space, and people's experience of the landscape; thus, benefiting a better understanding of place representations; (3) to illustrate that the spatial patterns of UGC can be adopted as a valuable proxy to understand urban development and neighbourhood change; (4) to reinforce the concept that *Spatial is Special*. Spatial processes are commonly spatially autocorrelated. The mainstream of machine learning methods do not explicitly incorporate the spatial or spatio-temporal component to address such a speciality of spatial data. This thesis highlights the importance of explicitly incorporating spatial or spatio-temporal components in geographical analysis models.

Acknowledgements

My first and most sincere gratitude goes to my first supervisor Dr. Stefano De Sabbata, for giving me the great opportunity to pursue my PhD at the School of Geography, Geology and Environment, University of Leicester. Stefano deserves most of the credit for the completion of my doctoral studies, for his unique contribution to my research and education, for all our thoughtful discussions, and for our friendship.

I also owe a great thanks to my reviewers Dr. Nicholas Tate, Dr. Leandro Minku (University of Birmingham) and Dr. Huiyu Zhou for their critical evaluation, comments and suggestions they gave during the three annual review exercises. I would like to say a big thank you to Dr. Tumasch Reichenbacher (University of Zurich) and Dr. Andrea Ballatore (Birkbeck, University of London) for the useful feedback, comments, support and observations given in the course of writing this thesis. I would particularly like to thank Professor Alexander Kurz (Chapman University) for the initial training he offered me in understanding mathematics in machine learning. I would furthermore like to thank Professor Yudong Zhang for his helpful advice for my future career development.

I also wish to thank all my colleagues and friends in my office in the School of Geography, Geology and Environment: Mohammed Shuaibu Jamiu Ozigis, Genna Tyrell, Chris Martin, Natalia Gonzalez-Michaels and others too numerous to mention for their support throughout my period of study.

Last, but by no means least, I acknowledge all my family. I owe my parents Jian Liu and Zhen Yang everything that I might never to able to pay back. Thanks to their support and sacrifices, I could start my study in the UK and pursue my PhD. I would particularly like to thank John Graves and Dianne Graves and their family for their kind friendship. They have done so many great things for me during my stay in Leicester and supported me through all the difficult time I had in all these years. I would also like to appreciate my friends: Xiaoyang Yuan, Ye Yuan, Arandeep Lehal, Sebastian Ghetu, Bo Peng, Dr. Can Chen, Shiya Wang and Zhongcheng Qiu for their support and friendship.

Contents

List of Figures

List of Tables

List of Abbreviations

ANN	**A**rtificial **N**eural **N**etworks
API	**A**pplication **P**rogramming **I**nterface
CNN	**C**onvolutional **N**eural **N**etworks
DNN	**D**ense **N**eural **Network**
GCN	**G**raph **C**onvolutional **N**etwork
GeoAI	**Geo**graphical **A**rtificial **I**ntelligence
GNN	**G**raph **N**eural **N**etwork
GIScience	**G**eographic **I**nformation **Secience**
GIS	**G**eographic **I**nformation **S**ystem
GIR	**G**eographic **I**nformation **R**etrieval
GPS	**G**lobal **P**ositioning **S**ystems
ICT	**I**nformation and **C**ommunication **T**echnology
IoD	**I**ndices **o**f **D**eprivation
IoT	**I**nternet **o**f **T**things
JSON	**J**ava**S**cript **O**bject **N**otation
KG	**K**nowledge **G**raph
LOAC	**L**ondon **O**utput **A**rea **C**lassification
LSTM	**L**ong **S**hort-**t**erm **M**emory
LSOA	**L**ower Layer **S**uper **O**utput **A**rea
MAUP	**M**odifiable **A**real **U**nit **P**roblem
MSOA	**M**iddle **L**ayer **S**uper **O**utput **A**rea
MST	**M**inimum **S**panning **T**ree
NeSS	**Ne**ighbourhood **S**tatistic**s**
OS	**O**rdinance *S*urvey
OSM	**O**pen**S**treet**M**ap
ONS	**O**ffice for **N**ational **S**tatistics
OA	**O**utput **A**reas
OAC2011	**O**utput **A**reas **C**lassification
PCA	**P**rincipal **C**omponent **A**lysis
POI	**P**oint **O**f **I**nterest
RNN	**R**ecurrent **N**eural **N**etworks
SBD	**S**patial **B**ig **D**ata
SOA	**S**uper **O**utput **A**rea
SVM	**S**upport **V**ector **M**achine
UGC	**U**ser- **G**enerated **C**ontent
VGI	**V**olunteered **G**eographic **I**nformation
VGAE	**V**ariational **G**raph **A**uto**E**ncoder
VTCNN	**V**isual-**T**extual Fused **CNN**

Chapter 1

Introduction

Digital platforms enable users to produce vast quantities of user-generated content (UGC) online and have become an ever-increasing presence in social practices (Elwood and Leszczynski, 2013). The information created and distributed through such platforms is now a significant source of information for scholars to understand the reproduction of urban places (Shaw and Graham, 2017). The intersections between the "code" (Dodge and Kitchin, 2004) of digital platforms and space capture the "localities" of users' everyday activities, augment spatial experiences (Elwood and Leszczynski, 2013), and shape the representations of places emerging from those platforms. Such representations further contribute to a digitally layered urban environment (Zook and Graham, 2007; Shaw and Graham, 2017).

The focus of this thesis is strongly related to the studies on the understanding of place representations described by the content production from digital platforms. Information shared on such platforms provides an insight into how users want to represent their everyday life, which consistently augments and reinforces the understanding of local communities through time, and layer dynamic information across and over geographic space (Graham et al., 2015a). Therefore, this thesis encapsulates and understands UGC as "augmentations" (Ballatore and De Sabbata, 2019) of places as "time-space configurations" (Agnew and Livingstone, 2011), and aims to investigate new possibilities available to explore the representation of places at the urban scale through geotagged UGC and deep neural networks.

This thesis adopts machine learning and deep learning approaches on graph representations of human dynamics (Mocnik, 2016) illustrated through geotagged UGC to explore how place representations are augmented and reinforced through users' spatial experiences by classifying their multimedia activities and identifying the spatial clusters of UGC at the urban scale. Having the place representations described through UGC, this thesis explores how these representations can be used in conjunction with socio-economic data. Existing research has proved the connection between place representations described through UGC and local socio-economic structures (e.g., population density, education level, and income) (Ballatore and De Sabbata, 2019). In the broader context of geographical studies, such local socio-economic structures of places are often described through various official spatial statistics. This thesis aims at combining official spatial statistics with place representations described through UGC to understand and predict the dynamic changes of the socio-economic characteristics of places.

1.1 Motivation

1.1.1 User-generated content

Thanks to the fast development of computing devices, it has become increasingly easier to connect to and interact with the internet. Beyond the relatively fixed positions of desktop and mainframes, the internet nowadays has become incredibly mobile and can be easily accessed by devices such as laptop computers, smartphones, and many other portable devices. Cellular networks, WIFI, Bluetooth and various other wireless forms of communication have allowed us to carry the internet almost everywhere we go. Global positioning systems (GPS) tells us where we are with stunning accuracy. The increasing prevalence of the intersection between the internet and location-based products allow people to use smartphones, mobile devices, and computers to build up their digital life and to leave their digital footprint on the Internet (Tsou, 2015), and have brought a tremendous revolution on the use of the World Wide Web, which is now termed as the "Geographic World Wide Web" or the "GeoWeb".

The geoweb is a geographically distributed digital network of nodes that capture, produce, and communicate data with explicit spatial components (Herring, 1994). The smartphone is an important and increasingly ubiquitous example of a geoweb node with multiple integrated location-based applications (e.g., GoogleMap, Twitter, Uber). Since 48.37% population in the world owns smartphones nowadays (Turner, 2021), and billions of users are using location-based services monthly, we are producing and collecting enormous amounts of geographically referenced data every day. The accelerated incorporation of location-based services into digital information and communication technologies in the last decade has fostered major shifts in geographical practices and the studies within geographic information science (GIScience). The information production includes phenomena that have been described as spatial crowdsourcing, volunteered geographic information (VGI), UGC, as well as big data (See et al., 2016). Despite the fact that there are differences in the definitions of these terms (will be detailed in Chapter 2), they share the same basic idea of people (usually not professionals) being involved in carrying out various activities associating with geographic information science (See et al., 2016). UGC on various digital platforms is now considered as a significant source of information to understand the reproduction of urban spaces. For instance, as one of the major components of UGC, spatial social media is described as one of the defining components of geoweb technologies (Sui and Goodchild, 2001). Information created and shared on social media platforms provides an insight into how users want to represent their everyday life, which has significant impacts towards the understanding of places based on users' activities described through their content production. Most social media integrated with location-based services allow users to know and see on a map where other users or the content they produce are physically located at a particular time. As such, data generated by individuals on various social media platforms such as Twitter and Facebook can capture the diversity of spatial content produced by users and help scholars to understand social activities and experiences in specific spaces, or regarding specific events.

Over the past decades, due to the increasing use of the internet and mobile devices, digital platforms and technologies have expanded and challenged many aspects of social practices, including traditional political, social and economic activities in society (Liaropoulos, 2013). Consequently, such phenomenon raises longstanding research questions of how physical space adapts to the rapid development of digital platforms and vice versa (Goby, 2003). Goodchild (2011) discusses the idea of formalising place in the digital world. He addresses the relationship between the informal world of human discourse and the formal world of digitally represented geography where *place* stands at the central position of those platial (Gao et al., 2013) studies within geography and GIScience. Such an academic advocation later promotes a wide range of studies towards embedding the digitalised human dynamics and their interaction with space (e.g., emotions, sentiments, place descriptions, etc.) into geographical research (e.g., place-based GIScience (Gao et al., 2013) and space–place (splatial) GIScience framework (Shaw and Sui, 2020)).

The analysis of human conceptualisations of space often involves categorisations of some kind, such summarisation and categorisation of the representative geographical phenomena of a given space inform us the understanding of socio-spatial practices in places. For instance, the clustering of similar users' activities in particular geographic areas provides insights into the abstract spatial representation of places. Place in geography is broadly defined as a series of "locales" where human's everyday activities take place (Agnew and Livingstone, 2011). Place representation can be explored through "the overall information available in a target geographic space for a given data source" (Ballatore and De Sabbata, 2019, p. 880). The studies on place representation provide a sociological understanding of place and take such an understanding as a fundamental aspect of the configurations of the space (Brantner and Rodriguez-Amat, 2016). Conceptualising users as sensors within spaces (Goodchild, 2007), the intersections between an unprecedented variety of geotagged UGC from digital platforms and spaces capture the "localities" of users' everyday activities, augment spatial experiences (Elwood and Leszczynski, 2013) and shape the representations of places. Such representations further contribute to a digitally layered urban environment (Zook and Graham, 2007; Shaw and Graham, 2017).

Understanding place representations is a central problem in GIScience (Purves et al., 2019), and UGC represents a significant source of information about places. Due to the potential of digital platforms for exploring social practices in space and the narrative of places (Abernathy, 2016), social media platforms in general, and Twitter in particular, has been at the centre of data-driven analysis in GIScience and quantitative geography for about a decade (Miller and Goodchild, 2015a). Although existing studies have advanced our abilities to understand the spatial patterns of social media, they are primarily only focused on text content. However, text UGC is not the only form of communication that users post on social media platforms. Images or photos constitute around 36% of posts on Twitter(Glenn, 2012), which renders the analysis of visual data, an interesting area to explore. Despite the growing popularity of visual content in social media, limited work has been done so far on such content within the field of GIScience. The lack of visual content analysis is a severe limitation, as image content is a key component of social media posts – especially considering the rise of image-focused platforms such as Instagram or Flickr. As "a picture is worth thousand words" (Wang and Li, 2015), visual content can also provide rich information regarding places, the use of space, and people's experiences of landscape. Earlier work on the visual content of geotagged UGC mostly focused on tags or meta-data or the text posted along with the images (e.g., Hollenstein and Purves, 2010; Gao et al., 2015; Xu et al., 2017b), therefore it heavily relies on social media users tagging their posts accurately. However, information created on social media platforms tend to be noisy which contains a considerable amount of information that are challenging to interpret even for human researchers due to a variety of reasons (e.g., linguistic errors including misspellings and grammatical mistakes in the text, or posts only have not- or less-informative images and with no provided illustrative text to explain), and images are commonly attached with multiple tags which are irrelevant to the content.

Despite the fact that the growing availability of GPS-enabled devices and social media platforms has led to an increasing interest in mining geolocated content, our understanding of the role played by social media in the social construction of place has been limited by the fact that only a small percentage of social media posts are precisely geolocated (e.g., 0.85% of tweets are geolocated according to Sloan and Morgan (2015)). Existing approaches aimed at tackling such issues focus on estimating locations of users by analysing placenames present in the text using geoparsing methods (Li et al., 2012a; Li et al., 2012b; Chang et al., 2012; Purves et al., 2018) in the text content. By including rare placenames (Flatow et al., 2015) and specific geographical words (Chong and Lim, 2017) or analysing location-based topics (Eisenstein et al., 2010; Eisenstein et al., 2011), the location of each post can be estimated based on the content. Although geoparsing of textual content is the main approach used in existing research, such approaches potentially ignore social media posts that do not include location information explicitly in the content. Lansley and Longley (2016) identified a strong association between users' activity types and the spatial distribution of users, which indicates that estimating the location of users' posts solely based on a semantic understanding of their content is a feasible but also challenging research objective to explore. Such an association can also further benefit our understanding of the places by exploring whether users' activities are clustered in certain urban spaces (Gurevich and Ghosh, 2014) and how that can be used to identify their unique socio-spatial patterns. As such, being an integral part of UGC, spatial social media show their great potential contributing to place presentations through learning users' activities. Taking visual and text content into account, this thesis aims to understand place representations through analysing multimedia UGC, and estimate users' locations based on the association between users' activity types and the spatial distribution of UGC.

As discussed above, UGC is an essential source of information that describes the perceptions of places, but it is not the only source of data used to understand places. In the broader context of geography, the understanding of places is often with the socio-economic context of local spatial infrastructures (Ballatore and De Sabbata, 2019). Such understanding is commonly described by various official spatial statistics as well as using socio-demographic classification approaches. For example, geodemographic classification based on census data is a commonly adopted approach, which has a long-standing history of being created in the UK to understand socio-demographic characteristics (e.g., Harris, 2003; Gale et al., 2016). Deprivation indices are another vital approach to interpret urban development, which has been used for a wide range of analyses, from human health research (Cox et al., 2018) to socio-economic studies (Kontopantelis et al., 2018). However, the majority of socio-demographic data are commonly collected periodically. For example, census data are collected every ten years, but local areas are dynamic and may undergo changes which might not be captured by decadal censuses (Gray et al., 2018). The socio-demographic classifications remain rigid during the development of cities (Singleton et al., 2016) and become less informative as time passes from the data collection, which can lead to potential uncertainties when adopted to understand urban spaces (Gale and Longley, 2013).

As discussed in the previous paragraphs, geographers aim at understanding how places function in terms of human activities. Such an understanding of place is taken as a fundamental aspect of the configurations of the space (Brantner and Rodriguez-Amat, 2016). Some of UGC sources can provide a homogeneous spatial coverage, while others embed some locational information which is only concerning particular locations of users or their posts. For example, contributors to Wikipedia aim at providing information to include all cities and areas in a systematic way, while geotagged social media such as Twitter posts (also called as tweets) are considered as a communication process between users, which can then be adopted to study collective spatial activities and urban dynamics (Ballatore and De Sabbata, 2018; Ballatore and De Sabbata, 2019). Ballatore and De Sabbata (2018) identified how the spatial distribution of UGC is related to population density, ethnicity, education level, and income due to the bias in participation and types of activities people share or discuss. For instance, in their study, they observed that UGC generated in Greater London exhibits a significant bias towards areas characterised by a wealthier, younger, and higher-qualified population. Despite that every city and platform has its own idiosyncrasies, such connection between the spatial distribution of UGC and local socio-economic structures indicates a possibility that the digital place representation emerging from those platforms could be used as a proxy to estimate urban socio-demographic dynamics (Reades et al., 2019), thus benefiting the understanding of place for research as well as governance.

1.1.2 GeoAI

Current developments in artificial intelligence provide new directions within the study of geography and from which a novel discipline emerged, named "GeoAI" (geographical artificial intelligence) (VoPham et al., 2018). Sitting at the junction of artificial intelligence (AI), geospatial big data, and high-performance computing, GeoAI aims to "provide a promising solution technology for data- or compute-intensive geospatial problems" (Li, 2020, p. 72), which enable "machines" to perform spatial reasoning and analysis like humans.

According to *Marshall McLuhan's Law of the Media* (McLuhan, 1975), modern media can be considered as modifiable perceptive extensions of human thoughts and experience (Sui and Goodchild, 2011). The use of AI methods in digital platforms studies plays an important role in understanding place representations. However, the majority of such studies on UGC within digital geographies and GISciences focus on text content or hashtags (Brantner and Rodriguez-Amat, 2016; Shaw, 2017). Deep learning opens new opportunities for bridging the gap in visual content studies on geotagged UGC data when studying users' spatial activities, and moving beyond the use of tags chosen by the users and low- or mid-level attributes, and combine high-level representations extracted directly from the media content. Within the discipline of computer science, some work (Xu et al., 2014; You et al., 2015; Gajarla and Gupta, 2015) has been done using convolutional neural networks to analyse users' sentiments directly from the images posted on social media posts. Attracted by the growing popularity of multimedia content on social media posts, research has been widely conducted from single-modal (e.g., text or image) analysis to multimedia content (e.g., text and image) study (Cai and Xia, 2015; Nadeem et al., 2019; Ahmad and Conci, 2019), and the multimedia UGC has been widely accepted as an import source of information to support geographical studies and spatial analysis (Newsam and Leung, 2019). A typical issue when applying AI methods to study UGC is that they often require a large amount of labelled data (namely *training data*) to help an AI algorithm learn and produce sophisticated results. As such, current research has mainly focused on supervised learning approaches with well labelled and balanced data (Abudalfa and Ahmed, 2019). However, labelling large volumes of social media posts can be a lengthy and costly procedure as it requires a significant amount of human intervention. Such approaches are only viable when a pre-defined set of topics or categories has been agreed upon by a large number of stakeholders, for instance, for monitoring scheduled events or natural disasters. Such approaches are more difficult to be adopted effectively for exploratory analysis or when monitoring unexpected events. Thus far, limited attention has been given to the study of exploratory analysis, where only vague categories or no categories at all have been predefined for a specific event or a geographical phenomenon.

Many machine learning or deep learning models that have been applied to geographical analysis are a-spatial, which do not incorporate space and spatial proximity as a factor in the model itself. For example, random forest Reades et al., 2019; Alejandro and Palafox, 2019, support vector machines (SVM) (Liuying and Sichun, 2018) and principal component analysis (PCA) (Demšar et al., 2013). Although many of those conventional machine learning approaches have achieved reasonable performance, recent research has shown that spatially-explicit models substantially outperform more general models when applied to spatial data (Yan et al., 2019; Chu et al., 2019; Mac Aodha et al., 2019). *Tobler's First Law of Geography* (Tobler, 1970, p. 236) points out that "everything is related to everything else, but near things are more related than distant things". Spatial data is special because spatial processes in a region or a given space is often spatially autocorrelated. Taking gentrification in urban studies as an example, once an area gentrifies, neighbouring areas can be affected by that gentrification process independently or in conjunction with other factors. In other words, many of the variables that are usually used to model gentrification (from population age to house prices) are frequently spatially autocorrelated – that is, similar values are found in neighbouring areas.

Tobler's First Law of Geography has a prominent role in geographical research and geospatial analysis, and it has been applied to the study of geotagged digital content (Ostermann et al., 2015), where UGC nearest to a social event or activities are more similar to each other than the more distant messages (Andrade et al., 2018). Thus, the interplay of UGC and the use of space are also spatially correlated (Ostermann et al., 2015). The use of distance to define the neighbourhood and its conceptualisation as graph representations of places and human activities has long been one of the core approaches in geographic information analysis (Dacey, 1965; O'Sullivan and Unwin, 2010; Mocnik, 2016). Recent studies in deep learning models introduce a local operation—graph convolution—into the learning process. Such a process is specialised dealing with graph-structured data in the irregular spatial domain (i.e., vector model in GIS), where the input data is represented as objects and their connections (Zhu and Liu, 2018). Thus, it holds the promise to encode the geographical and temporal proximity explicitly into the model to study users' spatial activities and to further understand the place representations.

1.2 Thesis Objectives

Everyday life in urban space is increasingly experienced through, as well as produced by, coded digital information (Graham et al., 2013b). Digital representations of places are becoming pivotal for scholars to comprehend urban life using vast quantities of data produced on various digital platforms. The "coded space" (Dodge and Kitchin, 2005) produced by the content on digital platforms shape the place representations based on the amount, quality, and type of digital information available in a geographic area (Ballatore and De Sabbata, 2019). The main research objective of this thesis is proposed as:

- **Research Objective:** *How can the use of content production of UGC inform our understanding of place representations and their socio-economic characteristics?*

As discussed in the previous section, existing studies and research mainly focus on the text content of UGC to describe places through their locales and activities. Given the increasing popularity of visual content produced online, multimedia content has the potential to provide rich information regarding places, the use of space, and people's experiences of landscape. Studies within digital geographies and GIScience on the visual content of geotagged UGC mostly focused on tags or meta-data (Hollenstein and Purves, 2010; Gao et al., 2015; Xu et al., 2017b); thus they heavily rely on users tagging their posts accurately. However, information created on social media platforms tend to be noisy, and images are commonly attached with multiple tags, some of which may be irrelevant to the content, or no tags at all. As such, to better accounting for visual UGC into the studies of digital geographies and GIScience, the first research question is proposed as:

- **RQ1:** *How can we combine information extracted from both text and images from multimedia UGC to better understand places and related activities?*

UGC platforms have become major platforms for people to communicate and exchange information regarding a wide range of topics. Despite the unequal geographies of UGC platforms (Ballatore and De Sabbata, 2018), there is a growing interest in analysing such information from a geographic perspective within the field of digital geographies (Ash et al., 2018b). However, traditional qualitative analysis often struggles with tackling large datasets, and the volume of data produced daily on UGC platforms is enormous. Thus quantitative analysis and summarisation are frequently necessary steps in digital geographies. That creates a strong association with GIScience, where data mining approaches have been applied to identify users' opinions and online trends, to study the emergence of place from space through content production (Graham et al., 2015a), or to monitor events from football to earthquakes (Frias-Martinez and Frias-Martinez, 2014; Ifrim et al., 2014; Sechelea et al., 2016; Zahra et al., 2017) and to understand the digital representations of a place (Ballatore and De Sabbata, 2019).

Designing automated learning approaches on multimedia UGC is a challenging task due to the special characteristics of the data on digital platforms. Firstly, with regard to the text content, UGC on platforms such as social media are likely to be short and conversational. The content contained in a single post is limited and noisy, which may not be sufficient to express comprehensive information. For multimedia UGC, information expressed through image and text is often complementary, where users might post a short and simple text but with images that enrich or complement the text information, and vice versa. Secondly, direct semantic summarisation from images using computational technologies involves practical issues which are rooted in the nature of images. For example, two images that are expressing similar concepts or meanings may have different viewpoints, scales, illumination conditions, etc.. Thus, interpreting vast quantities of visual content is difficult and often unfeasible in the qualitative studies of UGC in GIScience and digital geographies. The rise of deep learning in computer science shows its advances in many domains of science, business and government, and keep outperforming other traditional machine learning techniques particularly in the discipline of image processing (Krizhevsky et al., 2012; Szegedy et al., 2015; He et al., 2016). That opens up the opportunity to bridge the gap between text-based analysis and multimedia studies. However, deep neural networks have been somewhat neglected in GIScience and quantitative human geography (Harris et al., 2017). That is partially due to most deep learning approaches focusing on supervised learning, while GIScience has primarily focused on unsupervised approaches, as well as modelling and exploratory tasks. Autoencoders are an unsupervised approach within deep learning that is commonly adopted for extracting and compressing high dimensional data seems to have the potential to be applied on information extraction from multimedia content for GIScience and digital geography studies. As such, to address **RQ1**, I propose and test an autoencoder-based deep learning framework to directly learn and combine information from UGC.

UGC enables scholars to understand place representations by describing their activities and locales (Ballatore and De Sabbata, 2019). Time and geolocation are important features that UGC includes with their content. Information shared on digital platforms provides an insight into how users want to share their everyday life, which consistently augments and reinforces the assumptions of local societies through time, and layer the dynamic information across and over geographic space (Graham et al., 2015a). With the idea of conceptualising social media posts as "augmentations" (Ballatore and De Sabbata, 2019) of places as "time-space configurations" (Agnew and Livingstone, 2011), I am interested in exploring if the spatio-temporal aspects of social media posts would benefit the content analysis and further inform our understanding of digital representations of the city (Pereira et al., 2013). As such, the second research question is proposed as:

- **RQ2:** *How can spatial or spatio-temporal distributions of UGC benefit our understanding of places and their representations?*

Places in geography are not isolated but are connected in many ways (Nystuen and Dacey, 1961; Noronha and Goodchild, 1992), which could be both physical and social, using measures such as distance, adjacency, and spatial interaction (Zhu and Liu, 2018). Conceptualising UGC as “augmentations” (Ballatore and De Sabbata, 2019) of places, the digital information is connected and interacts in many ways (e.g., linking posts using "following-follower" networks, –see, Sadilek et al., 2012).

The use of distance to define the neighbourhood and its conceptualisation as graph representations of places and human activities has long been one of the core approaches in geographic information analysis (Dacey, 1965; O’Sullivan and Unwin, 2010; Mocnik, 2016). In recent years, as one of the sub-disciplines of deep learning, graph neural networks have attracted increasing interests in the field of computer science because of the great expressive power on the graph-structure data (Zhou et al., 2018; Zhu and Liu, 2018), which have provided powerful models that are potentially suitable for GIScience modelling on spatial interactions of places and understanding place representations. For the second research question, I propose and test various graph-based machine or deep learning frameworks on graph representations of human activities carried out through UGC with their geographical and temporal proximity to understand places and their related activities.

As discussed in the previous section, I stressed the fact that the role played by social media in our understanding of places has been limited by the fact that only a small percentage of social media posts are precisely geolocated. Existing research on location estimation tasks mainly focuses on analysing placenames with text-based methods, which potentially ignores a large chunk of data with no spatial information explicitly in the text content. Lansley and Longley (2016) has identified a strong association between users’ activity types and the spatial clusters of their content, I am interested in exploring if the location of the content can be estimated based on spatio-temporal topological structures given the users' activity types, and further understand places through the distribution of different activities. The third research question is proposed as follows:

- **RQ3:** *Can the users' activity type of social media posts reflect the location of the content and further benefit the understanding of place?*

The content production on digital platforms is grounded in the geography of users and their spatial infrastructures (Yardi and Boyd, 2010; Ballatore and De Sabbata, 2019). A sense of place is conceived as a collection of symbolic live patterns, attitudes with a spatial setting held by an individual or group (Stedman, 2002). In other words, similar social practices of users are often clustered within a geographic area. Existing methods focus on using rich information provided by text UGC to estimate the locations of social media posts (Li et al., 2012b; Li et al., 2012a; Chang et al., 2012) and understand the place. However, as discussed above, such methods ignore a large amount of data that are not explicitly geolocated in the text. The development of graph-based neural network allows to embed the features of nodes and can be an option to be adopted on estimating the locations of multimedia UGC. To address the third research question, I propose and test a framework using a variational graph autoencoder to estimate the locations of social media posts, taking into account their qualitative coded content and spatial topological structure.

As discussed in the previous section, place representation often associates with the socio-economic context of local spatial infrastructures (Ballatore and De Sabbata, 2019). This creates a strong connection to various official spatial statistics due to the fact that the socio-economic characteristics of places are often described through such statistics. Socio-demographic classification with such statistics is widely adopted to understand places at different scales. In the past decade, various approaches and indices have been adopted to understand urban development, such as geodemographic classification or indices of deprivation, and many existing research focus on socio-demographic representations at the urban as well as a national scale. However, the data for creating such socio-demographic classifications are mostly collected periodically, while UGC from digital platforms can be collected more frequently and continuously. As discussed in the previous section, Ballatore and De Sabbata (2019) illustrate how the spatial distribution of UGC is related to population density, ethnicity, education level, and income. Despite that every city and platform has its idiosyncrasies, such connection between the spatial distribution of UGC and local socio-economic structures indicate a possibility that the digital place representation emerging from those platforms could be used as a proxy to estimate urban socio-demographic dynamics (Reades et al., 2019), thus benefiting the understanding of the place from official governance perspectives. As such, the fourth research question is proposed as:

- **RQ4:** *How can the distribution of UGC benefit the modelling of urban socio-demographic change and inform our understanding of places?*

As mentioned above, places in geography are not isolated but are connected in many ways (Nystuen and Dacey, 1961; Noronha and Goodchild, 1992), which could be both physical and social, using measures such as distance, adjacency, and spatial interaction (Zhu and Liu, 2018). One place may associate with multiple defined representations described by different data sources and ways of connections and might be strongly correlated to each other, i.e., the strong correlation between geodemographic classification and financial deprivation (Dedman et al., 2006). Utilising a knowledge graph to capture the characteristics of places described by different data and the complex spatial connections among places is a feasible solution to illustrate or estimate urban dynamics in an automated way, and demonstrate how the place representation described through UGC can be used as a proxy to provide insights into urban changes. To this end, to address **RQ4**, I test a knowledge graph deep neural network approach to model urban socio-demographic changes.

1.3 Thesis Structure

The remainder of the thesis is organised as follows (see Figure 1.1), detailed introduction about the data will be provided in Chapter 3, 4, 5 and 6. Chapter 2 introduces the historical development of digital platforms and big data, and how they are associated with geographical analysis, as well as reviewing existing research and methodologies focusing on UGC data in the field of digital geographies. Chapter 3 presents an overall introduction on the data used in the thesis, as well as introducing my proposed framework and the mathematical background of all methodologies. Chapters 4, 5 and 6 are three analysis chapters. Each chapter introduces a framework developed based on the methodologies presented in Chapter 3 with case studies. In Chapter 4, I will introduce my proposed graph-based semi-supervised framework to understand places by investigating users' spatial activities using images, text and spatial or spatio-temporal information of their social media posts with case studies using geotagged Twitter data that have precise longitude and latitude pairs. In Chapter 5, I will introduce my proposed framework to estimate the geolocations of social media posts using activity types and their spatial topological structures. It will present case studies using geotagged Twitter data that have precise longitude and latitude pairs as well as data that have no precise geo-coordinates but with the attached geo-bounding box in their meta-data. In Chapter 6, I will introduce a spatial knowledge graph-based framework to predict socio-demographic changes at the urban scale, taking into account London Output Area Classification, UK Indices of Multiple Deprivation, and the distribution of geotagged social media data and Wikipedia articles in London. Chapter 7 provides a summary discussion of the results obtained from the three analysis chapters (Chapters 4, 5 and 6) with a view of harmonising all the findings and evaluating how these compare with the existing literature in addressing the research gaps observed. It also provides conclusions on the thesis's highlights, the implications of the findings and their contributions to broader knowledge, and several limitations encountered in the course of the research and the objectives of possible further research.

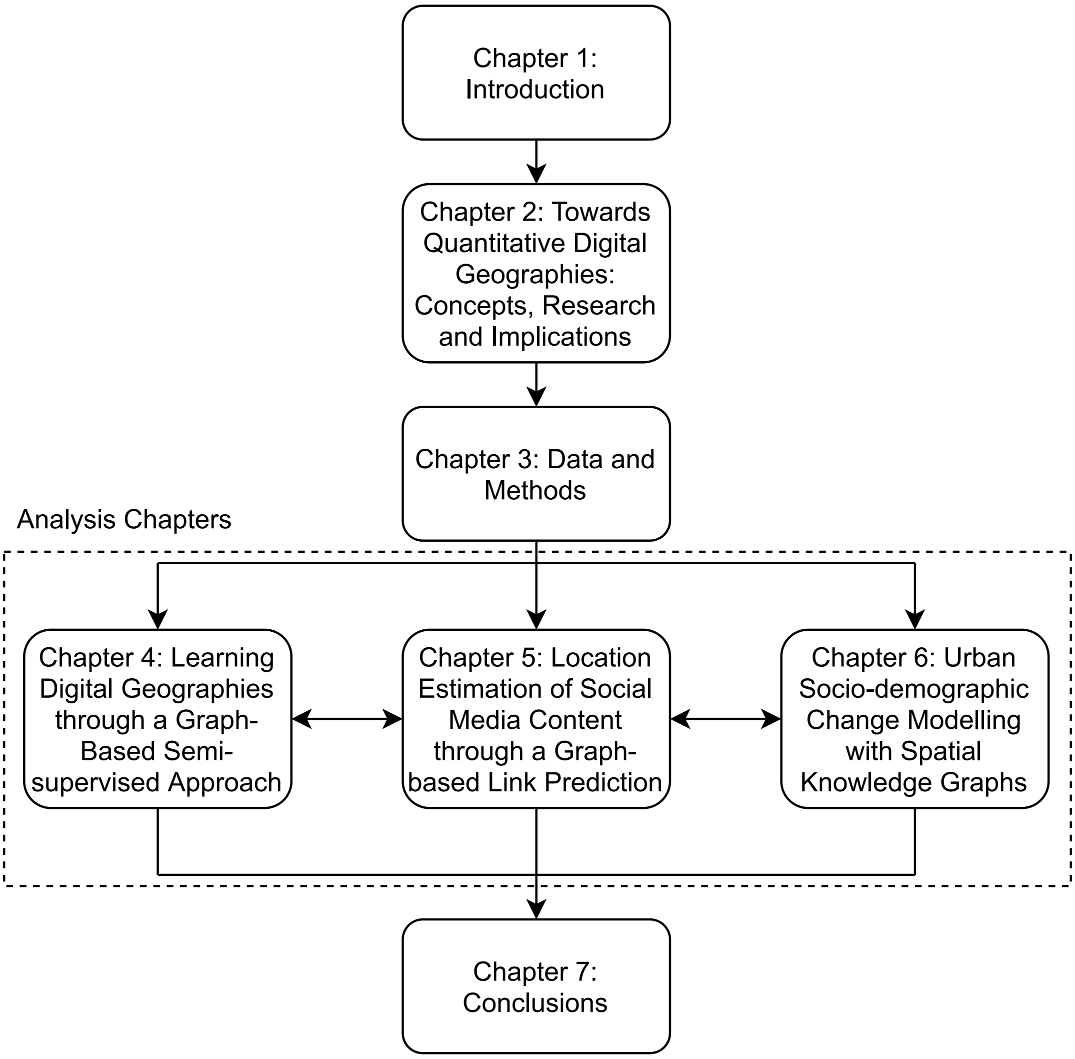

Chapter 1: Introduction
Chapter 2: Towards Quantitative Digital Geographies: Concepts, Research and Implications
Chapter 3: Data and Methods
Analysis Chapters
Chapter 4: Learning Digital Geographies through a Graph-Based Semi-supervised Approach
Chapter 5: Location Estimation of Social Media Content through a Graph-based Link Prediction
Chapter 6: Urban Socio-demographic Change Modelling with Spatial Knowledge Graphs
Chapter 7: Conclusions

FIGURE 1.1: A graphical illustration of the various chapters in the thesis and the connection between them

Chapter 2

Towards Quantitative Digital Geographies: Concepts, Research and Implications

This thesis is rooted in the discipline of digital geographies and GIScience with the conceptualisation seeing digital platforms as "code", and thus, representing the place as a "coded space". To better understand such "coded space", I use big data analytics and deep learning methodologies to explore how place representations are augmented and reinforced through content production on digital platforms and to identify the spatial clusters of UGC at the urban scale. Therefore, the background of this thesis is based on three main research areas: geographic information science (GIScience), digital geographies, and deep learning (DL). British-American geographer Michael Goodchild defined the term GIScience as the science behind the geographic information systems (GIS) (Goodchild, 1992). Later, David Mark published a more comprehensive definition for GIScience as "the development and use of theories, methods, technology, and data for understanding geographic processes, relationships, and patterns. The basic research field that seeks to redefine geographic concepts and their use in the context of geographic information systems" (Mark, 2003, p. 2). Digital geographies have emerged from the scientific awareness that knowledge is constructed, partial, situated and positioned within particular contexts (Hubbard et al., 2002). Digital platforms enable users to produce vast quantities of UGC online and have become an ever-increasing presence in social practices (Elwood and Leszczynski, 2013). Being an integral part of UGC, geotagged UGC is enriched with information describing places and their locales and activities, which draw increasing interests from scholars in many social science disciplines concerning social process embedded within a spatial context (Goodchild and Janelle, 2004). Traditional qualitative analysis in digital geographies often struggles with tackling large datasets, and the volume of data produced daily on UGC platforms is enormous. Thus quantitative analysis and summarisation are frequently necessary steps in digital geographies. That creates a strong association with GIScience, where data mining approaches have been applied to identify users' opinions and online trends, to study the emergence of place from space through content production (Graham et al., 2015a) and to understand the digital representations of a place (Ballatore and De Sabbata, 2019).

Deep learning is a class of techniques within Machine Learning technology. It "allows computational models composed of multiple processing layers to learn representations of data with multiple levels of abstraction" (LeCun et al., 2015, p. 436) and these methods have led to tremendous improvements on a large variety of domains such as visual object recognition, object detection and speech recognition. In recent years, the use of computer-based techniques for spatial data analysis has grown into an important scientific field, combining techniques from GIScience and emerging areas such as natural language processing, neurocomputing and heuristic search. Fotheringham (1997) defines geocomputation as quantitative spatial analysis where the computer plays a vital role in such analysis. Data are considered as "the driving force" (Miller and Goodchild, 2015b, p. 451) behind analysis rather than merely being a way of calibration, validation and test. Data-driven techniques are capable of handling not only large quantities of data but also a wide variety of data spreading at high speed in the world (Miller and Goodchild, 2015b). However, even though such data-driven science seeks to be exhaustive and automatically discover insights without proposing a hypothesis (Steadman, 2013), the use of data-driven methodologies must be carefully employed due to concomitant complex ethical issues (Zook et al., 2017) and other concerns about big data which will be discussed in the following section. With dramatic improvements on the computational capabilities of computers and the rise of deep learning techniques, this work is trying to explore the potential of employing deep learning methods in digital geographies discipline to handle large datasets to provide an in-depth insight of place representations associated with UGC from digital platforms.

Due to the nature of this thesis that combines a diverse set of fields, I will provide a brief introduction to each section as follows. The concepts of space and place have varied throughout recent history and have become central notions in geography. In Section 2.1, I will introduce space and place from a historical perspective of the development, and how they are evolved in "the midst of a digital turn "(Ash et al., 2018b, p. 25), whereby digital devices and software packages have become indispensable to geographic practice and scholarship across sub-disciplines.

2.1 Space and Place

The concepts of space and place have varied throughout the recent history of geography and have become central notions in disciplines like human geography. Traditionally, space was considered as a geographical container or geometric system that holds the object under study. Space is then nothing more than an abstracting instrument with no social connections for a human being (Tuan, 1979). By allowing users to lay an abstract space (a two-dimensional Euclidean container with an x- and y-axis geometry) over the Earth's surface, the use of the concept of space became particularly useful within geographical studies because space can be used to describe the exact location of every object contained in a specific area. The early adoption of space can be seen in the regional geography (Hartshorne, 1939) in the first half of the 20th century, which tasked the space with describing the location of all objects within it. Objects within such a space are completely independent of each other; the only connection among the objects are the geographical distances between each other.

In the second half of the 20th century, geographers started to critique the early conceptualisation of space as limiting and hampering the actual understanding of the social as well as spatial processes within a geographical area. Consequently, the creative thinking of space emerged. The traits of thinking of the space moved beyond considering it as an abstract geographic container which is used to describe all independent objects within it, to associating human spatial experiences and all other things in the area with the study of the space (Poorthuis, 2015). To clearly distinguish from the earlier concept of space, human geographers like Tuan (1977) conceptualised a new term in geography named as *place*. Contrary to space, place is more than just a location and can be described as a location created by human experiences, emotions and thoughts. In other words, *place* exists of *space* that is filled with meanings and objectives by human experiences situated in a particular geographic area (Tuan, 1979). Places "are specific time-space configurations made up of the intersection of many encounters between 'actants' (people and things)" (Agnew and Livingstone, 2011, p. 325), and can be viewed as a series of locales where everyday activities happen (Agnew and Livingstone, 2011). Social and cultural processes consistently participate in the process of creating, shaping and destroying places. In that sense, place plays an essential role in human cognition, social practises, and knowledge representation (Ballatore, 2016).

Thus, analysing how places are perceived and represented is crucial to interpret the underlying social and spatial practices involved with enriched human activities such as political, social and economic activities in space. The analysis of human conceptualisations of the space often involve categorisations of some kind. Such summarisation and categorisation processes of representative geographical phenomena inform us of the understanding of socio-spatial practices in the places of a given space. Thus, understanding the representation of place is a central problem in geographical studies (Purves et al., 2019). Place representation has a strong connection with information science and information systems (Purves et al., 2019), and it often refers to the overall information available in a target geographic area for a given dataset (Ballatore and De Sabbata, 2018). According to Graham et al. (2015a, p. 88), "information has always had geography. It is from somewhere; about somewhere; it evolves and is transformed somewhere; it is mediated by networks, infrastructures, and technologies: all of which exist in physical, material places". In the "pre-digital age" (Graham et al., 2015a, p. 89), studies towards the understanding of places focus on the individual level of the perceptions on a given space. For example, Lynch (1960) introduces an approach of mental mapping. His study draws images of certain aspects of the urban spaces from participators' mental memory. Taking it yet another step forward, Peter and Rodney (1974) use the same approach but with more observers, thus the understanding of places can be drawn through the collective perceptions from the observers. In the past decade, thanks to the development of digital devices and world wide web, human society has witnessed a radical change in the availability of information, such phenomena are termed as "information revolution" (Floridi, 2014, p. 87) or "data revolution" (Kitchin, 2014, p. 2). Although there remain concerns regarding the bias of human participation of the Internet due to uneven geographies of the ICT access (Graham et al., 2014b), Fuchs (2008) and Shirky (2010) highlight the ways that the digitally mediated participation allows citizens to play a more pivotal role in shaping the content and augmentations that play key roles in their lives. People nowadays are more prone to participate in the construction of knowledge and culture about places, regardless of their actual geographic locations (Lessig, 2003). Due to the vast amount of information produced on different digital devices and web daily, such information and communications technology-based (ICT-based) platforms function not only as a way of distributing, generating, monitoring, and controlling data exchange and flow across a range of the internet infrastructures (e.g., cable, WiFi, 5G network), but also as a host of associated social, economic, and political practices (Graham et al., 2015a).

Consequently, we can understand such ICT-based platforms as a geographical concept of *codes* (Dodge and Kitchin, 2005). Dodge and Kitchin (2005, p. 197) define a *code* as " an instruction or rule that has a single outcome determined by binary logic (yes or no)", and the combination of such individual logic rules produces *codes* which are able to conduct complex functions. The geographies of the conceptualised digital social practice are defined as a form of *cyber* or *virtual space* (Crang et al., 1999; Fisher and Unwin, 2001) in the studies of geography which are involved with digital information. Such cyber or virtual space has been termed as artificial reality, interactivity, and conceptual and metaphorical spaces with co-presence, low-cognitive mapping, and egalitarian and global communications (Kellerman, 2016). In other words, those are spaces that are digitally surveyed and regulated by information production from the ICT-based software and platforms. In addition to the term *cyberspace*, such digitised space are also defined as "coded space" (Dodge and Kitchin, 2005, p. 197). The geographies of such digitally mediated coded space (the amount and distribution among space and platforms of digital content about places) often relate to data at various geographic scales (e.g., streets, cities, countries) reflecting the percentage of Internet use consumed by mobile devices and other digital communication media (Dodge, 1998); thus, space is continuously reproduced through the digital information. Such coded space includes the use of check-in activities, geotagged content production and access of other forms of location-based services, and the conceptualised spatial experience described through the content production from individual users who are continuously using the Internet, and their geographic use of the websites or digital communications. Benefiting from the rise of computing technologies and the world wide web, researching digitally-augmented experience from individual users has become an interesting research objective within academia to understand how *code* operates. Gathering and analysing crowdsourced mass activities have become feasible and dynamically re-/shaped our understanding of the digital representations of places.

2.2 World Wide Web

The first generation of the world wide web is known as the web 1.0, which is the "read-only web" according to Berners-Lee (1998). In other words, the early web allowed us to search for information and read it. There was very little in the way of user interaction or content generation. With the increasing desire of users who want to participate in online activities and the rapid development of computer technologies, we moved from web 1.0 to web 2.0. The term "Web 2.0" was coined in 1999 by DiNucci (1999, p. 32) and later popularised by O'Reilly and Dougherty (2004) at the O'Reilly Media Web 2.0 Conference in late 2004. Since the emergence of web 2.0, digital platforms have started to strongly encourage user-generated content (UGC) in the form of text, video, and photo postings along with comments, tags, and ratings. Online users can communicate and participate in diverse activities with friends, colleagues and families much more conveniently and efficiently by using their mobile devices. As such, web 2.0 draws a significant number of users, and digital platforms consequently have become an ever-increasing presence in social practices, becoming part of political, social and economic activities in the society (Liaropoulos, 2013). Therefore, digital platforms are considered as an essential source for the research studying social science, including qualitative as well as quantitative fields such as computational social science that investigates questions (Lazer et al., 2009) using quantitative techniques (e.g., computational statistics and machine learning) and so-called big data (will be introduced in Section 2.5) for data mining and simulation modelling (Cioffi-Revilla, 2010).

2.2.1 Geoweb

Spatial information and geoprocessing techniques are now directly linked to many areas, such as commerce (Papamichail and Papamichail, 2007; Zhang et al., 2010), transportation (Ding, 1998), emergency response (Scholten et al., 2008), health care (Noon and Hankins, 2001) and many other domains that leverage spatial data to achieve a more comprehensive understanding of geographic areas. The integration of web technologies and GPS-enabled services has also created a strong association with the discipline of geography.

Online mapping was first introduced by Palo Alto Research Center (formerly Xerox PAPC) (Putz, 1994) in 1993 after the web 1.0 invented with the capability to show the world on a map, zooming at preferred scales and controlling the visibility of geographic features. Over the years, geoweb has witnessed a rapid increase in the development of delivery mechanisms for geographic information on the web (Haklay et al., 2008). The term geoweb has been broadly defined as "a distributed digital network of geolocated nodes that capture, produce, and communicate data that include an explicitly spatial component" (Abernathy, 2016, p. 11). As people nowadays can easily locate themselves with mobile GPS and ease getting access to the network, "we live in a geoweb" (Abernathy, 2016, p. 2). New technologies and ever fast-growing multilevel datasets that include geographic information have drawn social scientists' interests in spatial analysis within geography (Logan, 2012).

The concept of "GIS as media" (Sui and Goodchild, 2001, p. 387) creates a strong connection between geoweb and big data (this will be discussed in Section 2.5). With an increasing number of users using digital platforms which provide location-based services to geotag their locations, more studies and research from social science perspective are highlighting the importance of the collection and analysis of these massive, cross-referenced data about citizens and their activities (Crampton et al., 2013), and consequently contributing to a better understanding of the reproduction of urban spaces (Shaw and Graham, 2017).

2.3 User-generated Content

User-generated content (UGC) originates from people who contribute data, information, or media in a useful or entertaining way (Krumm et al., 2008). It could be any form of content, such as images, videos, text, and audio, which are posted by users on online platforms such as blogs and social media. In the "pre-digital age" (Graham et al., 2015a, p. 89), when the development of the digital devices and the Internet was still at their early stage, traditional "gatekeepers" such as newspaper editors, publishers, and news shows needed to approve all content and information before it was published. Thanks to the rapidly developed web technologies on Web 2.0 platforms, they have increased the prevalence of UGC, and flattened of traditional media hierarchies. The advent of UGC marked a shift among media organisations from creating online content to providing facilities for amateurs to publish their own content (Berthon et al., 2015).

Of the many UGC websites online, one of the most renowned is Wikipedia, which is the largest multilingual free encyclopedia written by users collaboratively. It is one of the top ten most visited websites worldwide and hosts 5.8 million articles in English, edited by about 130,000 monthly active editors (Wales, 2014). From a geographical perspective, in 2013, about 730,000 articles in English were associated with a location information (Ballatore and De Sabbata, 2019). Graham et al.

(2015a) point out that most editing distributed in the Global North, also for articles about places in the Global South, Wikipedia has a deep bias towards contributions from Western European and North American editors. Due to the popularity of Wikipedia, and the massive information produced on it monthly, the geographies of the information from the platform draws a wide range of scientific research towards to how Wikipedia shape our perception of the places (Graham et al., 2015b; Jenkins et al., 2016; Ballatore and De Sabbata, 2019) and its bias (Callahan and Herring, 2011; Graells-Garrido et al., 2015) as well its uneven geographies of the information (Graham et al., 2014b; Graham et al., 2015a).

Information produced on Wikipedia is certainly not the only source of UGC with geotagged information. In the rest of this section, I will introduce two important sources of UGC, social media and volunteered geographic information (VGI), that have been heavily involved in geographical studies. Note that a complete review of the literature around social media and VGI is beyond the scope of this dissertation, the introduction of them in this section includes scientific studies in relevant to understanding place representations within the fields of GIScience and digital geographies.

2.3.1 Volunteered Geographic Information

The rise of Web 2.0 represents an essential change in the way that the Web is perceived, and its products are developed, introducing an age in which ordinary users can freely share content online. In the past decades, the rapidly growing crowdsourcing techniques have been aggregated into GIScience democratisation projects (Butler, 2006), leading to Volunteered Geographic Information (VGI) initiatives, along with other geographic crowdsourcing and crowdsensing products (Pinheiro and Davis, 2018).

VGI was defined by Goodchild (2007, p. 212) as "the widespread engagement of large numbers of private citizens, often with little in the way of formal qualifications, in the creation of geographic information, a function that for centuries has been reserved to official agencies". VGI is a special subset of UGC tied to the proliferation of GPS technologies and the emergence of Web 2.0 technologies (e.g., blogs, social media, and wikis). The combination of digital platforms for sharing spatial data such as Wikimapia (2006) and OpenStreetMap (2004) (OSM) and cheap and accurate geotagging devices enable people to more actively participate in the data collection and distribution and geo-visualisations (Zook and Breen, 2017) regardless of their geographic background (Capineri, 2016b). One of the typical examples is OSM, an open-source, volunteer-generated global map that aspires to be the "Wikipedia of Maps" (Fox, 2012). Instead of scientific observations, the OSM project produces a large quantity of VGI in forms of nodes, lines, and features to a base map contributed by volunteers worldwide. As OSM operates under a Creative Commons license which ensures the data is free to use with attribution, the OSM has become a widely used resource supporting scientific research or commercial products. For example, after the Haitian earthquake, OSM became the most notable VGI to respond to crisis (Zook et al., 2010).

One of the key characteristics of VGI is that the creation of geotagged data can be derived from users' everyday activities (Zook and Breen, 2017). Such a characteristic expands VGI into the geoweb (Elwood, 2010). It has become a standard practice for users to create map mashups that adopt third-party data benefiting from the opening of the Google Maps API. Much of the data are crowdsourced and often focused on the social practices of the everyday activities, to create a whole new context for the information (Zook and Breen, 2017). Goodchild (2007) highlighted such a trend and noted that a critical novelty of VGI might be "informing about local activities in various geographic locations that go unnoticed by the world's media" (Goodchild, 2007, p. 220). As illustrated in Section 2.1, *place* can be viewed as a series of locales where everyday activities happen (Agnew and Livingstone, 2011), the conceptualisation of VGI platforms as a host producing social practices of human participants creates a strong connection towards understanding places within the discipline of geography. The individual-level content of VGI can be seen as a form of qualitative geographical information, and it provides a powerful source of information on human spatial experiences for places with a precision which was unattainable in the past using traditional time-constrained investigations (e.g., surveys, interviews, etc.) or official data (e.g., census) (Capineri, 2016a). Datasets and projects that constitute part of VGI can be seen as the result of a social creation process (Mayer et al., 2020), in which usually a large number of human participants are involved. Thus, every person holds a biased representation of their physical spaces, including geographical and geometrical features as well as information about objects (e.g., buildings, parks and streets) and their functions. VGI platforms integrate these representations from individual perceptions into a single shared conceptualisation represented in the data (Mayer et al., 2020). One of the typical VGI data sources for studying users' spatial experiences and place representations is social media data which will be discussed in the next subsection.

Given that VGI participants may vary greatly in expertise and often collect data without established protocols or standards, there is often considerable concern about the quality and usability of the data, which may influence our understanding of the places using VGI content. Early on, Flanagin and Metzger (2008) realised that it is essential to identify and develop methods and techniques to adequately evaluate VGI quality, and Goodchild (2008) highlighted the challenge to redefine the assessment of spatial accuracy in the VGI era. Although a complete review of VGI quality assessment is beyond the scope of this thesis, it is still worth noticing that there have been significant efforts from academia to deal with quality issues of VGI data. More comprehensive literature overviews of the latest developments in VGI quality assessment are presented in Barron et al. (2014) and Arsanjani et al. (2015) and Senaratne et al. (2016). Of the topics selected by the authors for future research, they emphasise the objectives of intrinsic data quality assessment, conflation methods which combine crowdsourced VGI and other data sources, and the development of credibility, reputation, and trust methodologies for crowdsourced geographic information.

Thanks to the increasing amount of data produced nowadays, studies that involved with VGI inevitably associate with quantitative data analytics. The central characteristics of such data with significant volumes, wide range of variety, and high speed of velocity often associate with the term of *big data* which will be detailed in Section 2.5.

2.3.2 Social Media

The earliest ancestor of the social media platforms that exist today is most likely USENET which was developed by two graduate students, Tom Truscott and Jim Ellis from Duke University in 1979 (Lueg and Fisher, 2012). USENET was a conglomeration of separate servers operated by various companies that store and forward messages to one another in forms of articles and threads. Users could subscribe to newsgroups depending on the topic and post articles and respond to them, forming a thread. This technique made it possible for one person to reach many other users and also spread their own voices. Following the development of USENET and the Internet, there was an explosion of different web-based services created for people to distribute content. In 1999, the platform Napster was created, apart from text content, it also allowed users to share videos and music files within the community. In 2004, social media portals like MySpace, LinkedIn and Facebook were founded. In 2005, the video platform YouTube was created, and one year later, another social media platform providing micro-blogging and social networking service called Twitter entered the market. Thanks to the development of digital devices and the Internet, social media nowadays are widely considered as interactive Web 2.0 Internet-based applications that facilitate the creation or sharing of information, ideas, sentiments and other forms of expression via virtual communities and networks (Kietzmann et al., 2011).

The term *social* highlights that the data from social media platforms represent many aspects of human life, they capture the habitual and relational interactions between family and friends or even strangers who they never met in person. Such nature of the data being collected has expanded into new realms of human geography and sociability (Poorthuis, 2015), and is a defining characteristic of social media data.

Social media data have other two core characteristics which are defined as *online* and *geotagged*. Although much of digital social media data is "online", the practical collection of "online" data is complicated, and the data collected can be categorised into different types based on the ways of accessing the data (Poorthuis, 2015). The first type is fully accessible data on the Internet (*public data*), and such data can be understood as a membership list or search results on the web that can be copied. The second type of data has controlled access on the Internet (*semi-public data*), and they can be accessed via an application programming interface (API) (e.g., Twitter's API), to obtain data that are parts of a social database. The third type of data is with controlled access and can only be collected through social means (*private data*) (asking a provider for a copy of data), such data are not, or at least very rarely, shared (e.g., Facebook transactions, mobile phone records) publicly. Each type of social media (public, semi-public, private) also implies different ethical concerns (Metcalf and Crawford, 2016) (detailed discussions will be provided later in this section). In the scope of this thesis, my case studies associate with public and semi-public data, as many researchers do, but it is still important to highlight that other researchers are not limiting themselves into these types of data.

The second term *geotagged* often refers to the *geotagging* which is an act that associates a given piece of social media data with a particular location on the earth's surface (e.g. a geotagged tweet or Foursquare check-in). A large part of geotagged data has been geotagged with longitude and latitude coordinates; in this thesis, I use the terminology *geolocated* social media to define such content. These information or data points are collected with a variety of technologies and approaches, including GPS receivers and/or WiFi and cell locative technologies with different levels of accuracy. However, longitude and latitude coordinate pairs are not the only form of geotagged data. As mentioned in Chapter 1, only a small fraction of tweets are geolocated with coordinates. This brings a significant challenge for researchers to understand geographic phenomena as recorded via online platforms. Existing approaches aimed at tackling such issues focus on estimating locations of users using the modellings of placenames with geoparsing methods (Li et al., 2012b; Li et al., 2012a; Chang et al., 2012; Purves et al., 2018) in the text content. By including rare placenames (Flatow et al., 2015) and specific geographical words (Chong and

Lim, 2017) or analysing location-based topics (Eisenstein et al., 2010; Eisenstein et al., 2011), the location of each post can be estimated based on the content. Thus, such methods have greatly enriched the data towards analysing geographic phenomenons.

The spatial and social structures of local communities in a city lead to certain collective human activities patterns (Steiger et al., 2016) clustered within different geographic areas. With the concept that users from social media as sensors of places (Goodchild, 2007), social media data can help to "sense" this type of information from urban environments. Thus, social media data provide a unique insight to places with the abundant information of sentiment as well as relationships between individuals, groups, and the physical environment (Roche, 2016). GIScience research thereby focuses on questions regarding how corresponding spatio-temporal patterns from social media networks and heterogeneous data streams can be explored, extracted, validated and aggregated. In turn, such information enables us to analyse daily spatial processes and to gain knowledge about places, especially with respect to collective human dynamics (Steiger et al., 2016). Using social media data to understand the interplay between human activities and the use of space, Li et al. (2013) explored spatio-spatial distribution of Twitter posts (tweets) and Flickr photos in California, showing that the distribution of photos is more clustered in natural parks and that Twitter posts tend to originate from areas with educated, high-income people. Hahmann et al. (2014) investigated the spatial relationship between points of interest from OSM and geolocated tweets, showing correlations at the local scale for certain topics (e.g., "railway station", "restaurant", and "supermarket") and not for others ("pub", "bakery"). Gao et al. (2017) introduced a data-synthesis-driven method using heterogeneous social media sources for detecting and extracting vague cognitive regions as one type of places, and they compared the results with a conventional human-participants study (e.g., survey). In particular, the authors assessed the spatial cognitive regions of "Northern California" and "Southern California" and found a firm correlation existed between the data-synthesis-driven method and the empirical-survey method. Although social media-derived results are often accused of being biased towards user demographics of social media platforms (Gao et al., 2017; Ballatore and De Sabbata, 2018; Ballatore and De Sabbata, 2019) and limited to the user-reachable locations, existing studies have demonstrated that social media data at least partially are able to reflect people's spatial experiences, opinions and interests in places. Thus, the collective sensing approach could benefit the understanding of places (Blaschke et al., 2018).

However, despite the fact that studies which are using social media in the context of geographical research questions in GIScience have proliferated, such studies, in particular with research that involve the mapping of social media, have also been criticised by scholars for promoting a sort of "speedy pseudopositivism" associated with a neoliberalizing "new quantitative revolution" (Wyly, 2014, p. 26). In studies of geotagged social media, researchers often over-privileged the single pair of latitude and longitude coordinates that are attached to each individual data (Shelton, 2017), and "ignoring the multiplicity of ways that space is implicated in the creation of such data" (Crampton et al., 2013, p. 132) by disregarding the socio-spatial practices embedded in the data. For example, as Crampton et al. (2013) illustrated, information that is geotagged to a particular location may not necessarily have been produced in that location, relate to that location, or exclude reference to any other geographic localities. Geotagged content often exhibits a variety of spatial referents apart from the hidden latitude and longitude coordinates attached to it. Thus, as pointed out by

Shelton (2017), research which ignores the implicit social and spatial processes embedded in this kind of data, such as many mainstream academically-oriented social media mapping projects conducted by non-social scientists, can lead to a range of decontextualised, problematic assertions (Sui, 2008; Crampton, 2011; Wilson, 2015).

Cheshire et al. (2019) identified four uncertainties of spatial analysis using social media data (e.g., Twitter), geodemographic, utilisation, semantic, and spatial. The geodemographic uncertainty relates to the self-selection of users. It is understood as a demographic bias of social media studies, where most users tend to be young, urban, affluent, and often with English as their first language. It is difficult to identify the geodemographic associations of any particular individual systematically and to understand how these biases manifest themselves spatially, as well as enabling evaluations against other datasets. Utilisation uncertainty relates to the bias of how the content production varies over space, time, and in response to different conditions and events. Since a small number of users produce a large amount of content, analyses of even vast collections of social media data may only portrait very small sections of society. Also, the spatio-temporal variation in social media usage is another crucial element of utilisation uncertainty. Temporal trends of users' activities indicate that users prefer to tweet during leisure hours; spatially, users' activities tend to be clustered at the centre of the urban areas and most activities are prone to be produced at sporting or entertainment venues. Semantic uncertainty relates to the semantic understanding of the content. Without sufficient cultural or personal context, interpreting the semantic meaning is challenging. Spatial uncertainty is similar to the issue pointed out by Crampton et al. (2013) concerning the lack of accuracy and precision measurement of social media content. Thus, an understanding of true spatial uncertainty is difficult to achieve.

Another important issue with regard to social media studies is their implication of ethical concerns. The first concern is about whether social media should be considered as *public* data which are free to access and analyse or *private* data which require extra data safety measurements and protections (e.g., General Data Protection Regulation in the European Union) (Townsend and Wallace, 2016). Despite the fact that users on each social media platform have all agreed to a set of terms and conditions which often contain clauses on how one's data may be accessed by third parties (including researchers), Boyd and Crawford (2012, p. 672) pointed out that "it is problematic for researchers to justify their actions as ethical simply because the data are accessible", and the process of ethnics evaluation can not be ignored even if the data are seemingly public. The second concern is about the *informed consent*. Unlike traditional research approaches which often involve consent forms prepared for the participants, social media-based research presents problems concerning the informed consent of participants. It is common that a social media user's data is accessed and analysed without informed consent having first been sought. "Participants" in such research are rarely aware of their participation to single studies. The third concern is about *anonymity*. In traditional research, it is straightforward to anonymise data so that the participants can not be identified. However, when working with social media data, data anonymity is more complex. It is difficult to anonymise individual data (such as tweets) when these are reproduced in publications and during presentations (Narayanan and Shmatikov, 2009). The fourth concern is the potential of harm. It is not always clear to the researchers whether or not the data they have accessed, collected, analysed or reused can be retraced in its original online context, or what the repercussions of such retracing might be (Townsend and Wallace, 2016). Markham and Buchanan (2012) suggested that researchers should be more aware of the increasing risk of harm to their participants, or the increased vulnerability of individuals or groups online. The concerns of ethical issues within social media analysis and big data studies will be further discussed in Section 2.5.

Nowadays, we immerse ourselves with increasingly abundant data. Rather than merely dismissing or rejecting analytics of social media data, it is still essential to study and develop data-driven approaches that enable social scientists to draw valuable insights from the data which are "situated and reflexive" (Kitchin, 2014, p. 1). Crampton et al. (2013, p. 130) proposed five extensions with regard to social media analysis, which they named them as "beyond geotags", to the typical practice of mapping geolocated data: (1) going beyond social media that is explicitly geographic; (2) going beyond spatialities of the 'here and now'; (3) going beyond the proximate; (4) going beyond the human to data produced by bots and automated systems, and (5) going beyond the geoweb itself, by leveraging these sources against ancillary data, such as news reports and census data. They argued that the study of social media practices should go beyond simple visualisations of content using latitude and longitude coordinates. They highlighted the significance of the temporal dimension of the data for deeper insights into the spatial and social process of the geographic phenomena. Their case study on Twitter regarding the widely reported riots following the University of Kentucky men's basketball team's victory in the 2012 NCAA championship reveals the promise of analysis that is not limited to the explicitly geographic dimensions of activity but includes a relational dimension, such as social network analysis. They highlighted the fact that social media content is not produced solely by human users, but also full of automated content producers like Twitter spam robots, which may draw uncertainties on the quantitative studies. Finally, they also highlighted the importance of including non-user-generated data, such as governmental or proprietary corporate data sources, as a supplement in social media research. Their paper has set forth a series of research directions for UGC studies, and the idea of "going beyond geotags" is a fundamental research objective rooted in the nature of this thesis.

2.4 Digital Geographies

Traditionally, extensive surveys and long periods of observation were required to collect an adequate amount of data to investigate social practices and study the associated urban representations. The rise of geoweb and technology innovation have brought geography in the midst of a digital turn (Ash et al., 2018b). Digital devices such as smartphones, satellites and digital cameras have brought great convenience to individual users who are constantly using the location-based services during our work, travel, production and leisure, thus having become indispensable to human life. As such, research interests from academia and industry have been heavily focused on associating their studies with digital platforms and social practices (Crampton et al., 2013), as well as data-driven geographies (Lazer et al., 2009). Popular social media platforms like Twitter, Facebook, Google+, LinkedIn and Instagram generate enormous amounts of content that is voluntarily shared on social networks by their users. Nowadays, people regularly use online digital platforms due to their convenience, efficiency, and significant broadcasting power for sharing information. Many aspects of human life, including how people identify and socialise with the communities, express their voice, and consume trending content and entertainment, are now highly mediated through those digital platforms (Ash et al., 2018b). People often reveal their social practices or their intent to carry out the social activities within their online communications or posts, which can now be easily accessed using a combination of methodologies such as information retrieval, natural language processing and existed social network analysis tools. Automatic retrieval of UGC removes many of the constraints associated with traditional methods which heavily relied on domain-specific expertise such as, collection time, accurate geolocation marks. As we immerse ourselves increasing deeply in a world of abundant data, much of which is geotagged, it is obvious to state that the digital phenomena have radically transformed every aspect of human life (Ash et al., 2018a). As a research field, "digital geographies" can be understood as "a turn towards the digital as object and subject of inquiry in geography, and as a simultaneous inflection of geographical scholarship by digital phenomena, is more meaningful in that it allows us to think about how the digital reshapes many geographies, mediates the production of geographic knowledge, re-configures research relationships, and itself has many geographies" (Ash et al., 2018a, p. 7).

Qualitative research that relies on data obtained by the researchers from first-hand observations, interviews, questionnaires, focus groups, participant-observation and recordings to understand how people experience place and space has been an essential approach in the discipline of geography. Since the proliferation of digital technologies, they are widely considered as the standard media of knowledge generation and analysis in digital geographies research (Ash et al., 2018b) and aid the further development of qualitative research. For example, transcriptions have been managed and analysed using qualitative software (e.g., Quirkos (2014), Kwalitan (2018), etc.) (Hinchliffe et al., 1997); social interactions can be observed in online forums using internet ethnographies (Hine, 2008); participatory research is being conducted using digital cameras and video recorders. With the help of new technologies and tools to collect and observe the data, it is increasing acknowledged that qualitative information that "born digital" as an indication of "geographies produced by the digital" (Ash et al., 2018b, p. 29). Many research projects used qualitative, resource-intensive approaches, such as studies towards regionally-specific expressions of religion (Zook and Graham, 2010; Shelton et al., 2012; Wall and Kirdnark, 2012), gendered nature of the participation on OSM (Leszczynski and Elwood, 2015; Gardner et al., 2020), language use (Graham et al., 2014a), as well as exploring how places are represented and understood differently by different people (Watkins, 2012; Graham et al., 2013b; Power et al., 2013) and more general questions regarding how and where events are discussed online (Graham et al., 2013a).

However, qualitative analysis often struggles with tackling large datasets whereas the volume of data produced daily on digital platforms is enormous. Thus, quantitative analysis and summarisation are frequently necessary steps in digital geographies. That creates a strong association with GIScience, where data mining approaches have been applied to identify users' opinions and online trends, to study the emergence of place from space through content production (Graham et al., 2015a), or to monitor events from football to earthquakes (Frias-Martinez and Frias-Martinez, 2014; Ifrim et al., 2014; Sechelea et al., 2016; Zahra et al., 2017) and to understand the digital representations of a place (Ballatore and De Sabbata, 2019). There is an increasing appetite amongst scholars for more collaborative and interdisciplinary working across the quantitative and qualitative realms to understand the potential complementary value of synchronised methodological approaches on UGC studies (Sui and DeLyser, 2012), such propositions of mixed methodologies set forth future research directions for studies within digital geographies and will be later discussed in Section 2.9.

2.5 Big Data

On June 23rd, 2008, Chris Anderson, former editor in chief of Wired magazine, published a provocative and thought-provoking article: "The end of theory: the data deluge makes the scientific method obsolete" (Anderson, 2008, p. 7). Anderson was referring to the idea that computers, algorithms, and big data can potentially generate more insightful, useful, or accurate results than specialists with domain-specific knowledge who traditionally crafted carefully targeted hypotheses and research strategies. In the era of petabyte of data as well as super-computing, sophisticated algorithms and statistical tools are at the centre of the stage to investigate into a massive amount of data to find information that could be turned into knowledge. Similarly, Prensky (2009, p. 4) indicates that "scientists no longer have to make educated guesses, construct hypotheses and models, and test them with data-based experiments and examples. Instead, they can mine the complete set of data for patterns that reveal effects, producing scientific conclusions without further experimentation". Although such arguments attract many criticisms from a wide range of disciplines and perspectives (will be detailed later in this section), they have also brought an explosion in the production of big data and the development of new epistemologies, due to the potential of capturing a whole domain and providing complete insights into the data (Kitchin, 2014).

Big data is sometimes defined as those datasets that are in huge volumes that can not be fit in a single Excel spreadsheet or stored on a single machine (Jacobs, 2009). However, the volume is not the only characteristic of big data. Massive datasets have been long produced by industry, government and academia, but given the costs and difficulties of managing such datasets, it is often practical to use sampling methods to generate summary datasets that support rapid queries (Cormode and Duffield, 2014). However, sampling methods may result in the constraint on the scope of the data, limitations on data temporality and minimization of the data size (Miller, 2010). In contrast, big data is characterised by being continuously generated in a flexible and scalable production with an exhaustive and fine-grained scope (Kitchin, 2014). Big data has been variously defined and customised within a wide range of disciplines. A comprehensive concept of big data is provided by Kitchin (2013), who draws on an extensive engagement with literature (Boyd and Crawford, 2012; Dodge and Kitchin, 2005; Laney, 2001; Marz and Warren, 2015; Mayer-Schonberger and Cukier, 2013; Zikopoulos, Eaton, et al., 2011) and suggests eight core characteristics of big data:

- huge in **volume**, consisting of terabytes or petabytes of data;
- high in **velocity**, being created in or near real-time;
- diverse in **variety**, being structured and unstructured in nature;
- exhaustive in scope, striving to capture entire populations or systems;
- fine-grained in resolution, aiming to be as detailed as possible, and uniquely indexical in identification;
- relational in nature, containing common fields that enable the conjoining of different data sets;
- flexible, holding the traits of extensionality (can add new fields easily) and scalability (can expand in size rapidly).

It has been estimated that up to 80% of big data is "spatial" with locational components attached to the data (Leszczynski and Crampton, 2016). Traditionally, spatial information is collected or generated through experts with domain-specific knowledge (e.g., census surveyors). Those datasets are usually small in volume, and the patterns within the datasets can be easily analysed through visual and statistical interpretations on the maps (Jiang and Shekhar, 2017). Nowadays, with the advanced development in remote sensors, GPS-enabled applications and the popularity of mobile devices, as well as cheap data storage and computational technologies, data are produced from a wide range of disciplines from commercial business to scientific research and engineering. Such geotagged data whose volume, velocity, and variety exceed the capability of current common spatial computing platforms are defined as *spatial big data* (SBD) (Jiang, 2016).

Compared to traditional "smaller" spatial data, SBD can make a difference in several aspects. At the macro level, SBD provides broad spatial coverage of geographic phenomena, enabling scientists to conduct large scale (global or continental) data analysis. For example, scientists can investigate the uneven geographies of access to contemporary modes of communication and uneven geographies of participation and representation at the global level based on various digital platforms (Graham et al., 2015a); the adequate information from mobility data (e.g., data retrieved from smart cards) can be adopted as a proxy for measuring urban diversity and vitality in relation to the spatial dynamics and the presence of people in the cities (Sulis et al., 2018), or supporting urban demographic predictions (Zhang et al., 2019c). At the micro level, SBD can provide high resolution with significant details, making it possible to support "precise" decision-making. As an example, geolocated social media data together with high-resolution hyper-spectral imagery or various sources of VGI data can be adopted to support disaster management and response for at certain regions. Existing research illustrates that such UGC content can be critical for sending alerts, identifying critical needs, and focusing response (Landwehr et al., 2016; Landwehr et al., 2016). Therefore, SBD is playing an increasingly important role in our physical world as well as the society.

In the article "The end of theory: the datadeluge makes the scientific method obsolete", Chris Anderson claimed that the numbers could speak for themselves with enough data, which portrays the the use of big data as being exhaustive. However, there are significant concerns about employing big data methodologies within social science. In against to Anderson's article, Kitchin (2014) suggested several critiques of big data and set forth directions for developing better situated, reflexive and contextually nuanced methodologies in the "era of big data" (González-Bailón, 2013, p. 147). Firstly, although big data may seek to be exhaustive, the data for both scientific research and commercial applications is both a representation and a sample, shaped by the usage of technologies and platforms, and it is subject to sampling bias. Thus, data are not simply natural and essential elements that collected and abstracted from the world in a neutral and objective manner. The use of data without further investigation may result in misleading conclusions of the "real-life" phenomena and can even produce harm. Datasets can be used to shape individuals' lives or stigmatise on certain groups (Danyllo et al., 2013; Barocas and Selbst, 2016; Crawford and Schultz, 2014), thus, yield discriminatory outcomes. For example, due to the categorisation based on postal codes, African-Americans in United States cities have less access to Amazon Prime same-day delivery service (Ingold and Soper, 2016).

Secondly, the process of data interpretation is framed. In other words, data are processed and interpreted within a particular scientific approach, even though the data analysis process is automatic. It indicates that regardless of how big the volume of data we have, such analytics cannot be free from human bias because "data are examined through a particular lens that influences how they are interpreted" (Kitchin, 2014, p. 5). A similar idea is proposed by Crampton et al. (2013), and they suggested that due to the knowledge is partially selected and shaped by our views, data used can be limited in their "explanatory value " (Crampton et al., 2013) regardless of how "big" datasets are. Especially when using social media data, conclusions which are drawn from the data can be naive in seeking to represent a whole society (Boyd and Crawford, 2012). For example, Ballatore and De Sabbata (2019) illustrate how the spatial distribution of UGC is related to population density, education level, and income, but different platforms exhibit a significant bias towards areas characterised by its own users' profiles and content production. As such, the generalisation of one single platform does not hold for the full understanding of the place representation.

Thirdly, the analytic methodologies within big data can be reductionist and functionalist. Some researchers developed big data analytics which could be adopted to model the social and spatial processes within cities (e.g., Bettencourt et al., 2007). They claimed that by exploring the "rules" of those process, big data could underpin the function and formation of the cities. However, researchers argue that such big data modelling approaches ignore impacts from culture, politics, policy, governance and capital, and wilfully neglects domain-specific knowledge from social science. (Graham, 2012; Kitchin, 2014).

As mentioned in Section 2.3.2, the increasingly abundant data lead to the growing awareness of ethical issues when conducting 'human-subjects' research where human experience, sentiments, opinions, social connections and behaviours are at the centre of those studies. Metcalf and Crawford (2016) examined several contentious cases of research harm in data science and proposed that it is crucial for the researchers who use big data approaches to be aware of that even data may seem public but can cause unintentional breaches of privacy and harms. They suggested that researchers should be more responsible for accessing, collecting, storing and analysing the data in an ethical and responsible way. Zook et al. (2017) proposed "ten rules" to address the complex ethical issues that will inevitably arise during "human-subjects" studies. Their works help researchers to recognise the human participants and complex systems contained within the collected data and encourage the integration of ethical questions as part of a standard workflow. They suggested that responsible big data research will ensure the research output is sound, accurate, and maximises the good while minimising harm, and can help researchers better understand society and our world.

As discussed above, a broader understanding of geographical analysis that is involved with geotagged data is required when conducting such sociological research based on big data. Instead of seeing digital platforms or other geoweb applications as a simple collection of latitude-longitude data attached to other bits of information, they should be seen as "a socially produced space that blurs the oft-reproduced binary of virtual and material spaces" (Crampton et al., 2013, p. 132). Further potential factors that may influence the results of the online platforms need to be considered, such as confidence in the accuracy of the physical location of geotagged UGC, temporal variation, the connection among users, recognition of "robots" (non-human sources), as well as a potential combination with ancillary datasets to maximise the utility of the data. While certain concerns still remain about social network analysis, research towards digital platforms associated with spatial analysis in digital geographies which are carefully designed with account for the bias in the data and results has already begun to reveal fascinating and valuable insights into societies and social practices (Abernathy, 2016).

2.6 Geographic Information Retrieval

The rapid growth of geotagged UGC in the form of collaboratively created content (primarily in the form of text) presents new opportunities for scholars to explore users' online spatial experience, the use of space, and people's experiences of the landscape. Thus, it provides us with a unique insight into how users' online spatial everyday activities carried out by digitally mediated information shape the representations of places. However, geographic content in UGC is not always explicit in the form of geotags or attached longitude-latitude coordinates pairs, but might be implicitly expressed through the content (e.g., in the form of placenames in the text, according to Aloteibi and Sanderson (2014)). Given the high volume of UGC produced daily, *geographic information retrieval* (GIR) is defined as a computational way to understand unstructured textual content to detect and resolve references to a location where the content is originated (Purves et al., 2018), and being widely studied within academia as well as commercial sectors.

During the past decade, the rapid development of information retrieval effectiveness has driven web search engines (e.g., Google) to new quality levels, and web search has become a standard and often preferred means of information finding and access. As a scientific research field, the term *information retrieval* is defined as "finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers)" (Cambridge, 2009, p. 1). As discussed in previous sections, significant amounts of data available nowadays contain spatial references to places on the Earth's surface. Traditionally, such information has been held as structured data and was tackled by methodologies from GIS (Purves et al., 2018). However, increasing amounts of data in the form of unstructured text (e.g., tweets) and images (e.g., geolocated photos on Geograph) are available for indexing and retrieval also contain spatial references. Such phenomena have promoted a growing interest in exploring how to augment conventional information retrieval approaches interacting with geographic information. GIR are a "spatially-aware search systems and support user's geographical information needs" (Purves et al., 2018, p. 1).

One of the primary tasks in GIR is to identify the candidate locative references in the text, and such a task is defined as *geoparsing* (Purves et al., 2018). There exist three types of geoparsing approaches of identifying candidate references: (1) approaches based on list lookup; (2) knowledge or rule-based methods; and (3) machine learning approaches (Leidner and Lieberman, 2011). The list lookup approach is widely considered as the simplest or the baseline approach identifying entities by looking up the previously generated lists (Purves et al., 2018). For example, Mikheev et al. (1999) illustrates that the performance using a simple list lookup for locations could achieve over 90% precision and recall of 75-85% with 5,000 locations collected from the Central Intelligence Agency World Fact Book (CIA, 2011) and evaluated on Message Understanding Conference-7 data (Chinchor, 1998). Although such approaches can achieve high performance, they are lack of the scalability because the methods heavily rely on the quality and size of the previously generated lists, and can not identify new entities not found in the lists. A more effective and sophisticated approach is to make use of the surrounding context; that is, developing rules that can capture more complex matching entities expressed within a grammar (Purves et al., 2018). Taking the sentence of *"the university is located next to the Victoria Park, north* *of Howard Road"* as an example. To locate where the *university* is, rules can be developed into a way that not only captures the entities with capitalisation (*Victoria Park* and *Howard Road*) in the text which are clearly related to the location of the university, but also the locational information that indicates the relationships between the university and the entities (*next to* and *north of*). The matching of text is performed by defining *regular expressions* that encapsulate the rules. In older geoparsing systems, such rules are generally hand-crafted, while nowadays, most modern approaches adopt machine learning or deep learning methods to induce rules automatically from previously annotated training samples using *features* that capture contextual information (Purves et al., 2018). For example, Wang et al. (2020b) introduce a Neuro-net ToPonym Recognition model targeting the language irregularities associated with social media text to recognise locations. Their proposed model extends a general bidirectional recurrent neural network (deep learning related literature review will be introduced later in this chapter) with a number of features designed to address the task of location recognition in social media messages. The model tested on GeoCorpora (Wallgrün et al., 2018) achieves 80% precision and 78% F-score, which are superior to state-of-the-art models presented in their paper. Despite machine learning or deep learning approaches have achieved reasonable performances, Purves et al. (2018) point out two issues of concern with such approaches. The first issue is about training data. This is about the issue of how much data is sufficient for a machine learning or deep learning approach to train a classifier, which can be used reliably on unseen text. The second issue regards to the generalisation of the resulting classifiers. This concerns whether the induced classifier can generalise across new unseen texts or whether they only operate successfully on the example datasets.

Thanks to the emerging research and development of geoparsing systems, a wide range of studies that utilise the systems to solve practical issues such as text-based location estimation have also been proposed within the scientific discipline. In particular, in studies on UGC, although digital platforms have profound impacts on understanding geographical social practices, the lack of geolocated data has been a longstanding issue within the spatial analysis of digital platforms. The research theme of location estimation and prediction opens up the opportunity to complement the issue of lack of data, as well as researching into human mobility patterns. Chen et al. (2013) proposed a location estimation framework targeting on social media data. The framework firstly sorts the tweets of each user in a chronology order, and employs a latent dirichlet allocation-based topic modelling approach to discover the personal interest distribution of each user. Then, it proposes a *function-interests mapping* method to construct the hidden relationship between users' interests to the functions of real physical places. Such functions of real physical places are defined by the functions of each tweet's closest *Point of Interest* (POI). Finally, given the historical locations of users and the *function-interests mapping*, they estimate the locations of the social media posts using a Bayesian model to predict the current location from the history travel records of the users. Their framework shows a 70% accuracy tested on Weibo (a well-known Chinese social media platform), considering if the distance between the estimated location and the actual location of posts are within 1-kilometre distance. Their studies presented in this paper demonstrates that a user's location is strongly related to his or her interest. Moncla et al. (2014a) and their substantial work Moncla et al. (2014b) developed an unsupervised algorithm that employs clustering algorithms to estimate a spatial footprint of toponyms which are not found in gazetteers. They evaluate their approach with a corpus of hiking description in three different languages. However, Skoumas et al.

(2016) argued that such proposed assumptions in the hiking text consider no uncertainties exist. In other words, their works assumed that the text description from a human is precisely accurate, which is potentially an over strong assumption in real-world applications. Skoumas et al. (2016) introduced a text-based location estimation framework that adopts information extraction methods to identify toponyms and spatial relations in the text content, and they formulated a quantitative approach based on distance and orientation features to represent the spatial relations. Probability density functions for spatial relations are defined through a greedy expectation maximisation-based algorithm. These PDFs then were used to estimate unknown object locations and achieve high-quality location estimation results as evidenced by a range of real-world datasets constructed using travel blogs. They also argued that their framework is robust regarding handling the uncertainties derived from crowdsourced textual data.

Location estimation can be analysed through not only text content, but also other properties attached to UGC, which can also be retrieved through GIR methods. For example, research has focused on location estimation based on users' social interactions with other people. Backstrom et al. (2010) examined the interaction between geographical information and social relationship through a maximum likelihood approach to estimate a user's location by knowing the geographic information of the user's friends. Davis Jr et al. (2011) predicted users' locations using a voting mechanism with three adjusting parameters based on a Twitter following-follower network and demonstrated that the size of social interactions between users plays a vital role in the location estimation process. That is, the large friends networks users have, the better information can be provided for location estimation. Similarly, Rout et al. (2013) applied a Support Vector Machine (SVM) to classify users' locations based on features that are extracted from a follower-based network on Twitter. Kong et al. (2014) proposed a framework named SPOT, which infers users' locations by measuring social closeness. Moreover, peoples' activities can also be employed as useful information for the location estimation tasks. For example, Ye et al. (2013) exploited the check-in category information to model the underlying user movement pattern using a mixed hidden Markov model to predict the category of user activity and its most likely locations given the estimated category.

As discussed in the previous paragraphs, Purves et al. (2018) have pointed out that the most modern approaches for geoparsing are through machine learning or deep learning models. As a matter of fact, the innovation of the newly raised machine learning and deep learning technologies also provide new directions of study within other geographical research disciplines, which will be detailed in the next section.

2.7 Artificial Intelligence in GIScience

Artificial Intelligence (AI) is a term frequently applied to the project of developing machine learning or deep learning algorithms and systems endowed with the intellectual processes characteristic of humans, such as the ability to reason, discover meaning, generalise, or learn from past experience. Machine learning is a sub-field of computer science that gives computers the capability to learn from experiences without being explicitly programmed (Samuel, 1959). Machine learning techniques are designed with the principle to explore the nature of data with well-designed algorithms and models to process sample inputs for making predictions or classifications. Machine learning algorithms have empowered many aspects of modern society, from recommendations on e-commercial websites (e.g., Ding et al., 2002;Yang et al., 2004) to medical diagnosis (e.g., Kononenko et al., 1997). However, conventional machine learning techniques lack abilities to process data in their raw format (LeCun et al., 2015). In past decades, developing a machine learning system required careful engineering and considerable domain expertise to transform the raw data (such as the pixel values of an image) into a suitable internal representation for further classification or clustering tasks. As a sub-class of techniques of machine learning, deep learning methods are composed of simple but non-linear modules, and every module transforms the data representation from a lower level into a higher level with more abstract representations by applying a back-propagation algorithm, to discover complicated and subtle structures within large and high-dimensional datasets (LeCun et al., 2015). Deep learning has shown its advances in many domains of science, business and policy decisions, and kept outperforming other machine learning techniques within disciplines such as image recognition (Krizhevsky et al., 2012; Szegedy et al., 2015; He et al., 2016), speech recognition (Mikolov et al., 2011; Hinton et al., 2012) and natural language processing (Collobert et al., 2011; Sutskever et al., 2014).

2.7.1 Artificial Neural Network

Artificial neural networks (ANN) are a set of algorithms based on the idea of artificial neurons, loosely inspired by the biological neural networks that constitute human brains, which are designed to recognise patterns. Considered the first generation of neural networks, *perceptrons* (Rosenblatt, 1958) are simply computational models of a single neuron. As shown in Figure 2.1, a neuron is a place where computation happens. A neuron combines input from the data with a set of weights, which either amplify or dampen that input, thereby assigning significance to inputs concerning the task the algorithm is trying to learn.

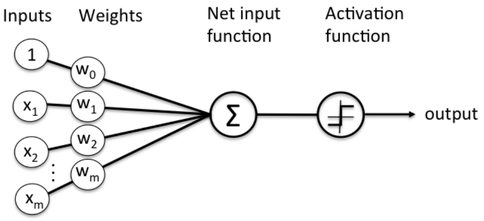


FIGURE 2.1: Schematic of Rosenblatt's perceptron (Image by BERGHOUT Tarek, via MathWorks, BSD-3-Clause).

Deep learning often refers to "stacked neural networks"; that is, neural networks composed of multiple processing layers (commonly more than three layers, including the input layer and output layer). As can be seen in Figure 2.2, a layer is a row of those neurons which computationally process the input fed through the network. Each layer's output is simultaneously the subsequent layer's input, starting from an initial input layer receiving the data.

In each iteration of the learning process of the network, the inputs are fed into the neuron, processed, and result in an output. The error which is back-propagated is usually the difference between the input and the output data. The network updates

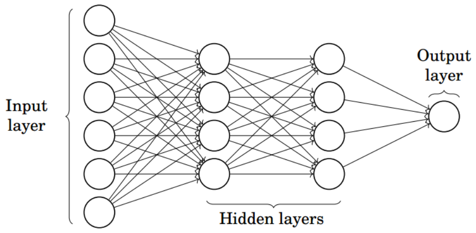


FIGURE 2.2: Schematic of a multi-layer perceptron (Image by Michael Nielsen, via Neural Networks and Deep Learning online free book, Determination Press, CC BY-SA 3.0).

the weights in the network accordingly for performance optimisation in the next iteration. Detailed introductions will be provided in Chapter 3.

2.7.2 Convolutional Neural Network

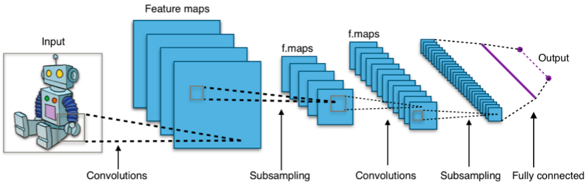


FIGURE 2.3: A traditional CNN design. (Image by Aphex34, via Wikimedia Commons, CC BY-SA 4.0)

Convolutional neural networks (CNNs) a type of ANNs which is primarily used for image processing tasks, but can also be adopted for other types of input, such as audio or text. A typical use case for CNNs is where the network is fed with images, and it classifies the categories of the input data. The performance of these deep neural networks has already exceeded human performance in object recognition tasks. The top-5 classification error rate (a fraction of test images for which the correct label is not among the five labels considered most probable by the model) performed by human annotators on the large scale ImageNet dataset[1] has been reported to be 5.1% (Russakovsky et al., 2015), whereas a state-of-the-art CNN (He et al., 2016) achieves a top-5 error rate of 3.57%. CNNs consist of filters or kernels or neurons that have learnable weights or parameters and biases. As shown in Figure 2.3, each filter takes some inputs, performs convolution and optionally follows it with a non-linearity (Uçar et al., 2017). The spatial relationship between pixels is preserved by adopting convolution using small squares of the input image. The input image is convoluted by employing a set of learnable neurons, and the convolutional layer produces an activation map as a layer output. Such output is further fed into the following convolutional layers. The pooling layer is another essential element in CNNs, which significantly reduces the dimensionality of each activation map but continues to preserve the most crucial information. A commonly found pooling technique is max pooling, as shown in Figure 2.3, which calculates the maximum value in each patch of each feature map. Using pooling layers is aimed at achieving better generalisation, faster convergence, robust to translation and distortion, and they are usually placed between convolutional layers.

[1] http://www.image-net.org/

Within the discipline of geography, the remote sensing community has extensively used CNNs for scene classification (natural and urban) (e.g., Zhang et al., 2017; Zhang et al., 2019a; Zhou et al., 2020), change detection (e.g., Wang et al., 2018a; Wang et al., 2018b; Seydi et al., 2020), and other image analysis tasks in recent years. Some CNNs have also been adopted to support research in GIScience, for example, studies on green space and urban built environment (e.g., Wang et al., 2019a; Wang et al., 2020a), traffic prediction and analysis (e.g., Zhang et al., 2019b; Zheng et al., 2019; Zhang et al., 2019d), geodemographics (De Sabbata and Liu, 2019), etc.. More examples of the research that adopt CNNs will be provided in Section 2.7.6.

2.7.3 Recurrent Neural Networks

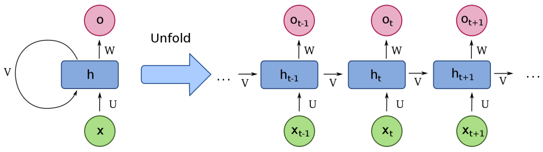


FIGURE 2.4: A RNN and the unfolding in time of the computation involved in its forward computation. (Image by Ixnay, via Wikimedia Commons, CC BY-SA 4.0)

Recurrent neural networks (RNNs), of which long short-term memory (LSTM) networks are the most potent and well-known subset, are a type of ANNs designed to recognise patterns in sequences of data, such as time-series analysis (e.g., Gers et al., 2002), stock markets prediction (e.g., Chen et al., 2015) and natural language processing (e.g., Nowak et al., 2017). One of the outstanding characteristics of RNNs is that they take time and sequence into account, they have a temporal dimension. RNNs process an input sequence one element at a time, maintaining in their hidden units a "state vector" that implicitly contains information about the history of all the past elements of the sequence. In a more intuitive way to understand such type of networks, RNNs function similar to the human way of processing data, they combine two sources of input, the present and the recent past, to determine how they respond to new data. One of the typical use of RNNs is shown in Figure 2.4, the artificial neurons get inputs from other neurons at previous time steps, a RNN can map an input sequence with elements x_t into an output sequence with elements o_t, with each o_t depending on all the previous x'_t (for $t' \leq$t).

RNNs have been widely adopted in geospatial domains, in particular dealing with time-series data to achieve real-time prediction and analyses (Reichstein et al., 2019; Li, 2020). For example, Ma et al. (2015) proposed an LSTM-based network for travel speed prediction. Their empirical results on data from Beijing indicate that LSTMs have the ability to capture long-term dependencies over the time-series of the traffic data. Xu et al. (2017a) developed a framework that combines LSTMs with mixture density networks (MDNs) to predict taxi demand in New York, United States. In their approach, the city is previously divided into smaller areas, and then the LSTM-based model is used to jointly predict the taxi demand for the given time period (20-60 minutes) in all the areas. Yu et al. (2017) introduced a LSTM-based framework for forecasting peak-hour traffic and identified unique characteristics of the traffic data. In their research, they identified that feeding the time stamps such as *time of day* and *day of week* as input to the model will significantly improve the model performance for the accurate peak-hour forecasting, and, LSTM can learn the traffic patterns taking into account the historical average traffic as well as the interruption caused by accidents.

2.7.4 Graph Theory and Graph Convolutional Network

Graph theory is the studies on graphs. In mathematics, graph structures are used to model pairwise relations between objects. A graph is made up of vertices or nodes which are connected by edges. An adjacency matrix of a graph is a square matrix to represent the graph. The elements of the matrix indicate whether pairs of vertices are adjacent or not in the graph. Depending on whether the edges of a graph have directions, the graphs can be classified into two types, directed graphs and undirected graphs. In an undirected graph, edges have no directions and indicate *two-way* relationships between nodes. Hence, the graph can be traversed in either direction. The adjacency matrix of an undirected graph is symmetric. On the other hand, a directed graph is a set of vertices connected by edges, with each node having a direction associated with it.

The use of various measures to define the neighbourhood and its conceptualisation as a graph network has long been one of the core approaches in geographical information analysis (O'Sullivan and Unwin, 2010). Spatial weights are mathematical structures used to represent such spatial relationships. A spatial weight $w_{i,j}$ represents a geographical relationship between locations i and j. Many spatial analysis methods, such as spatial autocorrelation statistics, and regionalisation algorithms, rely on spatial weights. These relationships can be formulated with several criteria, including contiguity and geospatial distance. A commonly-used type of weight is the *Queen* contiguity weight, which reflects adjacency relationships as a binary indicator variable (0 or 1) expressing whether or not a polygon shares an edge or a vertex with another polygon. These weights are symmetric, in which when a polygon a neighbours to a polygon b, both $w_{a,b} = 1$ and $w_{ba} = 1$. Another widely-adopted spatial weight is the *Rook* weights. *Rook* weights are also a type of contiguity weight, but consider polygons as neighbouring only when they share an edge. The *Rook* neighbours of a polygon may be different from its *Queen* neighbours, depending on how the observation and its surrounding polygons are configured. There are also many distance-based approaches to determine spatial weights, such as KNN distance, kernel weights, distance thresholds, etc.. Despite the differences, all the distance-based approaches are defining neighbours between polygons if there are within a given distance defined by a form of distance measurement. Such a spatial weight matrix can be seen as a form of graph adjacency matrix, and the spatial relationships can be represented as graphs.

In this thesis, undirected graphs are used to construct spatial and spatio-temporal graphs connecting social media posts using distance-based approaches; directed graphs are adopted to construct spatial knowledge graphs. The implementation details will be introduced in Chapters 4, 5, and 6.

Graph convolution, in general, is defined as a filter moving over the nodes of the graph, with the adjacency matrix determining the area captured by the filter. An intuitive understanding of Graph Convolutional Network (GCN) is for each node in the graph, and the graph convolution process will aggregate the information from its connected neighbours. By propagating through the hidden layers, GCN is able to produce useful feature representations of nodes in the graph, thus benefits further downstream tasks, such as classification, link prediction or the generation of graph embeddings (Wu et al., 2020b). GCN is a generalisation of CNNs to deal with graph structured data in the irregular spatial domain. The convolutional filter in GCNs can be extended to be localised in the spectral domain of the objects' features (Defferrard et al., 2016; Henaff et al., 2015), which are suitable for modelling the complex spatial patterns in geographical data that generally contain both Euclidean spatial information and non-Euclidean feature information (Liu et al., 2015).

2.7.5 Knowledge Graph

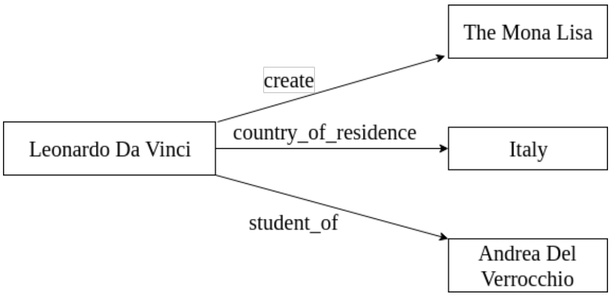


FIGURE 2.5: Knowledge graph example.

The term *Knowledge Graph* (KG) denotes a collection of labelled and interconnected descriptions of entities (namely triples). These entities can be real-world objects, events, situations or abstract concepts, where their descriptions have a formal structure that enables both human expertise and computer programs to process them efficiently and unambiguously. According to Nickel et al. (2015), the majority of KGs are constructed in a curated (e.g., WordNet), collaborative (e.g., Wikidata, Freebase), or auto semi-structured (e.g., YAGO Hoffart et al. (2013)) fashion rather than an automated unstructured approach (Mai et al., 2020). KG is commonly organised as a set of concepts, relations, and facts, which are associated by two kinds of types entity, relation, entity and entity, attribute, attribute value (Zhang et al., 2008). For example, Figure 2.5 shows a simple example of the knowledge graph, entities such as "Italy", or "Leonardo Da Vinci", are represented as nodes in the graph, and relationships such as "country_of_residence", are represented as edges. Entity descriptions contribute to one another, forming a network, where each entity depicts a part of the description of the network, related to it.

Taking geographic information into account, KG has played a vital play in answering geographic research questions from various perspectives, such as geographic knowledge graph completion (Qiu et al., 2019), geographic ontology alignment (Zhu et al., 2016), geographic question answering (Mai et al., 2019; Mai et al., 2020), etc.. At present, most geographic knowledge graphs are organised as universal knowledge graphs, such as the common sense geographic knowledge base (CSGKB) (Zhang et al., 2008) and CrowdGeoKG (Chen et al., 2017). Instead of traditional gazetteers, CSGKB employs a data structure that connects the notions of geographic features, geographic locations, spatial relationships and administrators for GIR tasks. CrowdGeoKG applies a crowdsourced geographic knowledge graph that derives various types of spatial entities (*OSMNode*,*OSMWay* and *OSMRelation*) from OpenStreetMap and enriches them with geo-entities (e.g., administrative regions) that are extracted from Wikidata with richer knowledge contributed by volunteers. Knowledge graphs with geographic information are more complicated than general graphs. Due to the sparsity of information in knowledge graphs (e.g., missing triples), entities, relations, and attributes cannot easily and directly answer many geographic queries or questions without spatial or non-spatial reasoning (Mai et al., 2020).

However, how to encode geographic knowledge (i.e., locations) into a knowledge graph and methodologies remains as a domain-specific challenge, and research in this area is still in its first steps. Trisedya et al. (2019) encoded geographic coordinates as a sequence of characters (string) and used a compositional function to encode these coordinate strings for geographic entities alignment adopting TransE. To integrating spatial distance relations between geographic entities, Mai et al. (2019) and Qiu et al. (2019) borrowed the translation assumption from TransE. In TransE, if a triple $< h, r, t >$ exists, the entity embeddings h, t should be connected by the relational vector r, i.e., $h + r \approx t$. For example, two facts $< China, Capital, Beijing >$ and $< UK, Capital, London >$ will enjoy a relation that $China - UK \approx Beijing - London$ in the embedding space. Using such a translation assumption, TransE predicts the existence of a triple by measuring the distance between the head entity and the tail entity after a translation enforced by the corresponding relation. Both Mai et al. (2019) and Qiu et al. (2019) implemented such an assumption, and further introduced sampling procedures to incorporate the distance between two entities into the knowledge graph. The geographic distance between two entities in the graph determines the frequency of sampling of a triple; thus, a triple has higher sampling frequency closer in both geographic space and embedding space. Despite the novelty of their work, the estimated entity similarities are based on some form of distance measures among entities with the designed data conversion process and ignore their absolute positions or relative directions. That is, their frameworks do not explicitly incorporate the absolute geo-locations and spatial distance between the entities into the knowledge graphs, while the data conversion process can be unnecessarily expensive and leads to information loss. To address such an issue, Mai et al. (2020) proposed a novel framework to directly encode entity locations into a high-dimensional vector space, which conserves more abundant spatial information than distance measures. Their work was the first KG embedding model that consolidates location encoding into the model architecture instead of relying on some forms of distance measure among entities.

The study of knowledge graphs in the task of geographic question answering forms the basis of an interesting research discipline of GeoAI. The concepts of geographical knowledge graph will be adopted in this thesis in Chapter 6.

2.7.6 GeoAI

The scientific field of geographical artificial intelligence (GeoAI) (Li, 2020) has been recently formed from combining innovations in spatial science with the rapid growth of methods in AI (VoPham et al., 2018) and lies in its applications using machine learning or deep learning techniques to address real-world problems. Section 2.6 has provided some discussions about how a recurrent network-based deep learning method is applied in geoparsing on social media messages (Wang et al., 2020b), which can also be seen as an application in GeoAI. Apart from geoparsing, GeoAI applications and systems are also widely researched within other geographical studies. For example, GeoAI opens up opportunities and applications in health and healthcare, where location plays a pivotal role in both population and individual health. Boulos et al. (2019) surveyed the GeoAI applications in health and healthcare and summarised the benefits of those applications in disciplines within the domains of public health, precision medicine, and IoT (Internet of Things)-powered "smart healthy cities and regions".

Within the context of GeoAI, few studies have a focus on the analysis of UGC and digital platforms. For instance, Chen et al. (2017) developed a convolutional neural network-based framework to extract text related to traffic information from Sina Weibo, and indicate that the rich information embedded in online social media data can help improve traffic prediction by using deep learning framework. Huang and Carley (2017) proposed a convolutional neural network framework to predict tweets locations, and the results of their framework tremendously outperform conventional machine learning approaches (STACKING (Han et al., 2014)). Their research can benefit more tweets to be accurately located by country, and city, of origin. They demonstrate at the country level, the more tweets that come from the country, the better the prediction their model can provide. However, inferring locations at city level remains a challenging task as the results achieved in the paper are mixed, for about half the tweets it is difficult to infer the locations. Huang et al. (2018) introduced an end-to-end, fully supervised framework to report geo-located flood events using Twitter posts. They adopted two convolutional neural network architectures to extract representations from texts and images, and combine both representations for filtering out flood-related tweets from a massive tweets pool. Their method is reported having around 80% accuracy, which can significantly improve the traditional selection process on Twitter data regarding disaster management, which is time- and labour-consuming. Zhu and Liu (2018) proposed a graph convolutional neural network-based approach to model spatial patterns with check-in data from a social media platform (Sina Weibo), and suggested their framework can achieve satisfying results in the prediction of intra-urban POI check-in patterns and can be modified to be applied to other geographical applications such as spatial interpolation, site selection and event detection.

To advance GeoAI research needs high-quality geospatial datasets. Many deep neural networks need to be trained on a large set of well-labelled training data. It has long been recognised in the field of deep learning that a trained model is only as good as the quality of the training data. Therefore, data are "no longer the only resources to be mined by computational tools but are becoming part of the tools" (Janowicz et al., 2020, p. 630). However, sampling data for the training process may share the same concerns towards big data studies, as mentioned in Section 2.5. Data selected for training and addressing research questions may be subject to sampling bias and can not be free from human discrimination. The increasingly abundant data also leads to awareness of ethical issues when training deep learning models to understand human activities, sentiments and experiences, which sometimes can lead to severe issues in society (e.g., inequality, unfair democratic elections, etc.) (O'neil, 2016). Bolukbasi et al. (2016) demonstrate an interesting example using word embedding, a popular framework to represent text data as vectors used in many machine learning and natural language processing tasks. Their studies identified that even word embeddings trained on Google News articles exhibit female and male gender stereotypes to a disturbing extent. For example, doctors are male and nurses are female, women are sensitive and men are successful, etc.

Moreover, due to the multi-layer nonlinear structures of deep neural networks, deep learning has been often criticised for being non-transparent and their predictions not traceable by humans (Buhrmester et al., 2019). In other words, deep learning has been widely recognised as a "black box". The nature of being "black box" has limited human's understanding of the learning and inference process of deep neural networks. Without explicitly knowing how models process the data inherently, it is often questioned how much can we trust the output of deep learning models. As an example of such trust issues of automated decision-making by the algorithms, the European Union's General Data Protection Regulation has restricted the use of AI and automated decision-making algorithms to access people's sensitive information (e.g., age, sex, ancestry, name or place of residence, etc.). If a result affects individuals, they should be able to demand explanations of the algorithmic decision made about them (Goodman and Flaxman, 2017). Further discussions on "deep learning as a black box" issues will be addressed in Chapter 7.

2.8 Quantitative Urban Geography

The constant emergence and development of cities and urban regions bring significant changes to their socio-demographic composition, which have been widely studied in the field of GIScience. Various approaches and indices have been adopted to understand urban development. In this section, for the scope of this thesis, I will briefly provide introductions to two different official spatial statistics: *census data* and *UK Index of Multiple Deprivation*, and how they are incorporated into quantitative studies to understand urban dynamics.

One approach in analysing urban dynamics is to observe change at the level of individual neighbourhoods (Modai-Snir and Ham, 2018). Despite the risk that some neighbourhood processes at a granular level cannot be observed through quantitative data (Barton, 2016), there remains the challenge of defining a neighbourhood in the first place (Reades et al., 2019). According to Knaap et al. (2019), there is no precise definition of "neighbourhood in either spatial extent or social composition". For the scope of this thesis, I take the definition by Galster (2001, p. 2112) as the starting point to define the term *neighbourhood* in the context of my study: "the bundle of spatially-based attributes associated with clusters of residences, sometimes in conjunction with other land uses". As discussed by Reades et al. (2019, p. 923), such a definition "does not establish neighbourhoods as discrete, bounded entities as it does not directly provide the size of the neighbourhood, but it provides a basis for defining neighbourhoods on different spatial scales through the 'bundling' of attributes". Following such a definition, neighbourhoods, in the context of this thesis (Chapter 6 in particular) are the spatial units defined by Office for National Statistics (ONS) (i.e., output areas, lower layer super output area) which underpin the the operationalisation of the 2011 Output Area Classifications and 2015 & 2019 English Index of Multiple Deprivation in the UK (Gale, 2014; Gale et al., 2016).

2.8.1 Census Data and Geodemographic Classification

Census data is a unique source that is collected periodically (e.g., every ten years in the UK), which detailed socio-demographic statistics that underpin national policy-making with population estimates and projections to help funding and plans. For example, census data which show the population that work in different occupations and industries can be used for designing new jobs and training policies or supporting investment decisions; information regarding ethnic groups can help with evaluating equal opportunities policies.

A geodemographic classification is defined as a process grouping geographical neighbourhoods, or small areas, in terms of their socio-economic characteristics. Such a process is generally achieved by applying a clustering algorithm (e.g., *k-means* (Hartigan and Wong, 1979)) on a dataset of composite socio-demographic variables collected from static data such as population census, which are not updated regularly. Observing changes in the geodemographic classification of areas over time is a useful approach to analysing the spatial development of the socio-economic structure of towns and cities (Brown, 1991). The combination of various open geodemographic indicators enhances not only our understanding of the dynamics of urban areas but also the assessment of policies from local and national governments. For instance, Longley (2005) suggested that analysing the dynamics of geodemographics will assist decision-makers in understanding the geographies of public service consumption. Batey and Brown (2007) proposed an assessment tool to examine policy initiatives focused on urban development. They suggested that geodemographic classification is a flexible tool to provide useful information for urban planners to assess the quality of urban policy initiatives. Although there are issues and concerns within academia limiting the broader utility of geodemographic classifications, such as the omission of classification uncertainty estimates, the 'black box' nature of the methods for scientific replication, etc. (see, Fisher and Tate, 2015; Longley, 2007), the Output Area Classification (OAC2011) of the 2011 census data created by Gale et al. (2016) has become an essential tool for researchers to understand socio-demographic patterns at the urban, regional and national scales in the UK. By utilising census data derived from open geodemographics, Liu and Cheng (2018) enhanced the interpretation of the transportation patterns within cities, and further illustrated its potential usefulness in public transport planning.

2.8.2 UK Index of Multiple Deprivation

According to the official document of the government report published in 2019 by McLennan et al. (2019, p. 9), "the Index of Multiple Deprivation(IMD) is the official measure of relative deprivation in England and is part of a suite of outputs that form the Indices of Deprivation (IoD). It follows an established methodological framework in broadly defining deprivation to encompass a wide range of an individual's living conditions. People may be considered to be living in poverty if they lack the financial resources to meet their needs, whereas people can be regarded as deprived if they lack any kind of resources, not just income".

Deprivation indices have been widely considered as another vital approach to modelling urban development (Sloggett and Joshi, 1998) and the evolution of the cities. Dickson and Young (1985) suggested that the study of the spatial distribution of deprivation indices at the regional level can reshape regional development policies; hence it is necessary for local government to conduct much detailed analysis and experiments before decision making. Pacione (1989) illustrated the significance of having structural-level "people policies" and local-level "place-policies" on urban planning of places within Scottish cities with high-level poverty and deprivation through an urban deprivation index. Talbot (1991) suggested that indicators of urban deprivation are useful for identifying and catering for underprivileged areas, such as health care planning within urban areas. The 2015 & 2019 English Indices of Deprivation for Lower-layer Super Output Areas (LSOAs) across England were published by Smith et al. (2015) and McLennan et al. (2019), which have been used for a wide range of analysis, from health (Cox et al., 2018) to socio-economic studies (Kontopantelis et al., 2018).

2.8.3 Urban Dynamics

For the first time in human history, the majority of population in the world lives in cities (Ritchie and Roser, 2018). Metropolitan areas have become the immediate sphere that is described through human existence and experiences, and such a sphere is changing rapidly (Schneider-Sliwa, 2001). Low-income countries are subject to distinct urban population growth and consequently are increasingly challenging to manage. A typical example is China. In the late 1970s, China launched its economic reforms, since then China has witnessed unprecedented economic growth and urbanisation process. The velocity and complexity of China's transition and urbanisation, in terms of the economic, social, and environmental improvements, have and still are dramatically changing human experiences for the people who live inside the country (Wei and Ye, 2014). Since 2012, more than half of the population in China have been living in cities. Such a rapid increase in urban population is putting forward severe challenges for housing, food, jobs, social services, and environmental sustainability. On the other hand, highly industrialised or developed countries are likewise undergoing a historic change in their cities or metropolitan areas. Facing the global trends in economics, society, and politics, urban regions in those highly developed countries require new and extensive economic and local policies to meet increasing intra-urban competition for investments and taxpayers (Smelser, Baltes, et al., 2001). Therefore, the socio-spatial structure of cities and metropolitan areas changes over time. Research targeting at urban evolution and dynamics are necessary "to monitor trends of urban development, to provide basic information for the optimisation of local strengths, to design urban development concepts which adequately provide for local needs and demands, and for urban planning which equally respects collective decisions, increased competition, modern urban structures, and the individual 'feeling' of the city" (Schneider-Sliwa, 2001, p. 16008).

Understanding the human dimension and spatial dynamics of cities from an aggregate and demographic perspective using a collection of human-focused spatially-referenced data has been widely facilitated by researchers within the field of quantitative urban geography (Manley and Dennett, 2019). Traditionally, many of those data would have been collected as official spatial statistics (e.g., census, IMD) or empirical observations and surveys, which has contributed to our understanding of static demographic profiles of the city or creating snapshots of everyday human activities and their interactions with urban areas. However, because the official spatial statistics are commonly collected periodically while the socio-spatial structure of cities and metropolitan areas are continuously changing, many studies have attempted to investigate new forms of data to gain insights into the patterns and processes exhibited by humans in cities in at a much finer temporal resolution. Sulis et al. (2018) proposed a computational approach using smart cards used in public transport to measure the spatio-temporal variations of urban vitality and diversity in the city of London concerning the presence of people in different areas of London and the spatial dynamics. They demonstrated that smart cards could provide a high spatial and temporal resolution to observe meaningful urban dynamics in relation to human activities. The smart card data were also adopted by Zhang et al. (2019c) to explore the relationships between the city's inhabitants' travel patterns and their socio-demographic profiles (e.g., age, working status). Their work explicitly explained why specific travel patterns are presented in the city, which is useful for city planners and transport operators to forecast travel demands and provide personalised transportation services. Manley and Dennett (2019) used a combination of mobile phone transactional data and a fine-grained building-use dataset to derive useful information about urban spatial dynamics. The study focused on the analysis of the city's inhabitants and their activities from the mobile phone transactional data and their relations to the physical features of the city and the mobile phone usage in the urban context.

Recently, the use of machine learning in topics of interest to urban studies has proliferated. One of the typical examples is to use neural network-based statistical modelling. Arribas-Bel et al. (2011) adapted a self-organising map algorithm which taking advantage of its properties as a data-reduction as well as a clustering technique to addresses the issue of urban sprawl in Europe from a multidimensional point of view. Their study identifies the most sprawled areas and characterising them in terms of population size. In the paper, they categorise and extract the most relevant six dimensions from the literature that are divided into two main categories: urban morphology, which includes as variables the scattering of urban development, the connectivity of the area, and the availability of open space; and internal composition, which focuses on how the socio-spatial structure is constructed in the area. These are then calculated for a sample of the major European cities that uses several sources to obtain the best possible dataset to measure urban sprawl. Being one of the early research that adopted and adapted machine learning approaches to study urban dynamics of the urban sprawl, their study pointed out research questions and hypotheses within the discipline (e.g., how to study the importance of temporal element in the urban dynamics modelling) which are still interesting to many researchers (Zhang et al., 2013; Hu and Zhang, 2020). Another emerging field within urban studies is to use machine learning models on visual content to explore urban dynamics. Naik et al. (2017) developed a computer vision method to measure changes in the physical appearances of neighbourhoods from street-level imagery. The method is developed in a relatively naive and straightforward way that it adopts a support vector regression algorithm and takes two mid-level image features GIST descriptors and texton maps (specific terminologies in computer vision, they can be understood as vector features encoding the shapes and textures present in an image) as input to calculate the perception of safety (so-called "Streetscore" in their paper). They then correlate the measured changes with neighbourhood characteristics to determine which characteristics predict neighbourhood improvement. Instead of street-level images, remote sensing images are also can be adapted to study the urban deprivation changes. Arribas-Bel et al. (2017) provided evidence on the usefulness of very high spatial resolution (VHR) imagery in gathering socio-economic information in urban settlements. They used land-cover, spectral, structure and texture features extracted from a Google Earth image of Liverpool (UK) to evaluate their potential to predict Living Environment Deprivation at a small statistical area level (lower layer super output area). Their study proves that Random Forest is the best model (compared to other models in the paper) in predicting the deprivation level of the neighbourhoods.

Currently, after the "digital turn "(Ash et al., 2018b), the possibility of collecting qualitative and social evidence with new data, such as UGC, has generated broad interest in using it to better understand social synergies in the city context; as well as motivating innovations in the development of citizen-centric approaches (Acedo et al., 2018). The citizen-centric approaches base themselves on the human–space interactions (Roche, 2016), which are mainly dependent on our capability to understand the use of space and the corresponding place representations. Such innovations of citizen-centric approaches have created a strong connection to the study of digital geographies, which has the key objective to explore the interaction between human and space through the online content production and the resulting production of digitally coded urban space. With the help of rapid development of ICT-based big data analytical research and tools, citizen-centric approaches conceptualise citizen as senors (Goodchild, 2007) that produce an enormous amount of geographical data with or without consent (See et al., 2016), and can evolve into a more cooperative and sophisticate process to aggregate and measure real sensing in the human–urban interaction (Acedo et al., 2018) and the corresponding urban dynamics.

2.9 Towards Quantitative Digital Geographies

Spatial distribution of UGC is accepted by scholars as a valuable resource to advance research on specific urban aspects (Anselin and Williams, 2016; Arribas-Bel et al., 2015). The representation and interpretation of data retrieved from social media provide a means by which to assess different urban dynamics and create sociodemographics of the cities. Ballatore and De Sabbata (2018, 2019) studied the geographic distribution of Twitter, OSM objects, Foursquare venues and Wikipedia articles in London (UK) (Ballatore and De Sabbata, 2018) and Los Angeles (California, US) (Ballatore and De Sabbata, 2019). Exploratory spatial analysis and regression-based models indicated that the four UGC platforms present distinct geographies of place representations. They illustrated how population density, ethnicity, education, and income are related to a high density of social media and VGI content in both cities, although each platform has its own peculiarities and not all findings are mirrored in the two cities.

The studies mentioned above demonstrate that the content production from digital platforms is grounded in the geography of their users and their digital infrastructures. The amount, quality, and type of digital information available in a geographic area consistently shape and reshape place representations.

Understanding social practices in the context of specific events through the analysis of the content production of UGC has also been regarded as an essential research objective. Bruns (2012) analysed hashtag (a type of metadata tag which enables users to apply dynamic, user-generated tagging that helps other users easily find messages with a specific theme or content) conversations in a Twitter message. The author extracts public Twitter activity data around specific hashtags, and for processing these data in order to analyse and visualise the *reply* (an act that allows senders to direct public messages even to users whom they do not already follow) networks existing between users as a static network, and to highlight the dynamic structure of reply conversations over time. Similar research focus on hashtag activity to social practices in the context of specific events were also conducted by Bruns and Burgess (2011) and Wohn and Na (2011). These studies constitute a comprehensive understanding of the Twitter dynamics and of the broad range of social interactions that produced from the platform, providing in-depth categorisations of the events and of their Twitter-based characteristics. Although the above mentioned three papers are not specifically within the study discipline of digital geographies, it is still worth to mention here to provide a broader picture how can social media activities and real-world events mutually impacted. The similar research objective is brought into geography; thus, geographers can link Twitter activities to the local events in the physical space to explore how they are associated, as well as detecting and monitoring unusual events (e.g., disasters). Andrienko et al. (2013) described a visual analysis approach for examining the frequently tweeted words and their spatio-temporal patterns. They first adopt a spatio-temporal *term usage cluster analysis* to identify general patterns and topic terms to explore what people tweet about, where, when, and how often. Then, the authors provide an in-depth analysis by categorising the content keywords of the tweets and conduct visual analysis on the distribution of keywords by identifying spatial clusters at the urban scale. Agarwal et al. (2018) analysed the tweets spatially regarding a political event, the *UK-EU referendum*. They performed a *geo-spatial sentiment analysis* on tweets content as well as hashtags based on the location of the events compared to the distribution of the geospatial tweets for that particular event on a global level. Both studies show that the use of geotagged social media posted by ordinary citizens is a valuable source providing insights about people and the space where they live in, as well as their social connections (e.g., sentiments, opinions and activities) to the particular events locally and globally.

2.10 Summary and Outlook

This chapter in each of the nine sections – *Space and Place*, *The World Wide Web*, *User Generated Content*, *Digital Geographies*, *Big Data*, *Geographic Information Retrieval*, *Deep Learning in GIScience*, *Quantitative Urban Geography* and *Towards Quantitative Digital Geographies* – attend to the myriad ways that introducing different research concepts which later will be expanded in each chapter of this thesis. The representation of place is a central problem in GIScience (Purves et al., 2019) but how we understand places is shifting because of the rapid development of the World Wide Web and digital platforms. Traditional qualitative analysis within digital geographies often struggles with tackling large datasets, and the volume of data produced daily on digital platforms is enormous. Thus, quantitative analysis and summarisation are frequently necessary steps in digital geographies. That creates a strong association with GIScience, where data mining approaches have been applied to identify users' opinions and online trends, to study the emergence of place from space through content production (Graham et al., 2015a). However, existing research on the visual content of geotagged UGC mostly focused on tags or meta-data (e.g., Hollenstein and Purves, 2010; Gao et al., 2015; Xu et al., 2017b), while they heavily rely on users tagging their posts accurately. However, information created on social media platforms tend to be noisy, and images are commonly attached with multiple tags, in which some of them may be irrelevant to the content, or no tags at all. The rise of deep learning algorithms and techniques provides scholars useful tools to analyse visual content of UGC explicitly, which inspires the study of this thesis to incorporate multi-media content towards understanding place representations with more abundant information (images and text) from digital platforms.

Moreover, the use of distance to define the neighbourhood and its conceptualisation as graph representations of places and human activities has long been one of the core approaches in geographic information analysis (Dacey, 1965; O'Sullivan and Unwin, 2010; Mocnik, 2016). In recent years, as one of the sub-disciplines of deep learning, graph neural networks have attracted increasing interests in the field of computer science because of the great expressive power on the graph-structure data (Zhou et al., 2018; Zhu and Liu, 2018), which have provided powerful models that are potentially suitable for GIScience modelling on spatial interactions of places and understanding place representations. As such, I adopt the use of graph conceptualisations of the spatial interactions of UGC and explore the use of graph neural networks to understand places representations and their socio-economic characteristics better.

In Chapter 4, I will introduce my proposed graph-based semi-supervised deep learning (combination of CNN, RNN and GCN) framework to classify users' activities using the images, text and spatial or spatio-temporal information of their social media posts. In Chapter 5, I will introduce my proposed framework (a variant of GCN) to estimate the geo-locations of social media posts using activity types and their spatial topological structures. Ballatore and De Sabbata (2019) identified how the spatial distribution of UGC is related to population density, ethnicity, education level, and income. Traditionally, such socio-economic characters are commonly illustrated through official spatial statistics. The connection between the spatial distribution of UGC and local socio-economic structures indicate a potential possibility that the digital place representation emerging from those platforms could be used as a proxy to estimate socio-demographic dynamics, thus benefiting the understanding of the place from the official governance perspectives. In Chapter 6, I will introduce a spatial knowledge graph-based framework to predict socio-demographic changes at the urban scale, taking into account London Output Area Classification, UK Indices of Deprivation, and distributions of geotagged social media data and Wikipedia articles in London.

Chapter 3

Data and Methods

This chapter introduces the data adopted for case studies and mathematical details of each deep learning approach employed in this thesis.

3.1 Data

The overview of the data adopted in this thesis is shown in Figure 3.1, and this section will introduce the datasets in the following subsections.

3.1.1 Twitter Data

This thesis conducts case studies on Twitter, which is a micro-blogging platform for providing individuals with ways of connecting with family and friends, and it has over the time become a significant form of social communication on a global scale. Twitter generates a vast amount of messages (which are also named as tweets) per day with extremely heterogeneous contents ranging from primary sports events coverage to natural disasters. Since late 2009, Twitter has allowed each tweet to be geotagged with a specific latitude and longitude. Users can become citizen-sensors (Goodchild, 2007) which capture information from where they are and spread real-time geotagged data. As mentioned in Chapter 1, 0.85% of the Twitter feed output is geotagged with coordinates (Sloan and Morgan, 2015), which is equal to roughly 4 million tweets a day, produced by a population only marginally different to the overall platform population (Ballatore and De Sabbata, 2018). The massive amount of collectable geotagged information on Twitter opens the opportunities for scholars to view online trends, collect human perspectives and understand digitally-mediated place representations through the users' activities carried out by their content production. Despite the fact that in June 2019, Twitter decided to remove the ability for users to add precise location information in their text-based posts (TwitterSupport, 2019), users can still share precise their locations through Twitter's updated camera or geotag their positions in the form of a hierarchy-level of *Twitter Place* (will be introduced later in this subsection). Twitter remains as one of the major platforms providing VGI data to support geographical studies.

The dataset in this thesis consists of all geotagged tweets posted within UK between January 8th, 2018 and December 31st, 2018, through an application programming interfaces provided by Twitter, Inc. named *Twitter APIs*.

According to the official documents published on Twitter's website (Twitter, 2021a): APIs allow users to request and deliver information. This is done by allowing a software application to call an address (also known as an *endpoint*) that corresponds with a specific type of information a company provides (endpoints are generally unique like phone numbers). Twitter allows access to parts of its service via APIs to allow developers to build software that can integrate with Twitter, such as a solution that

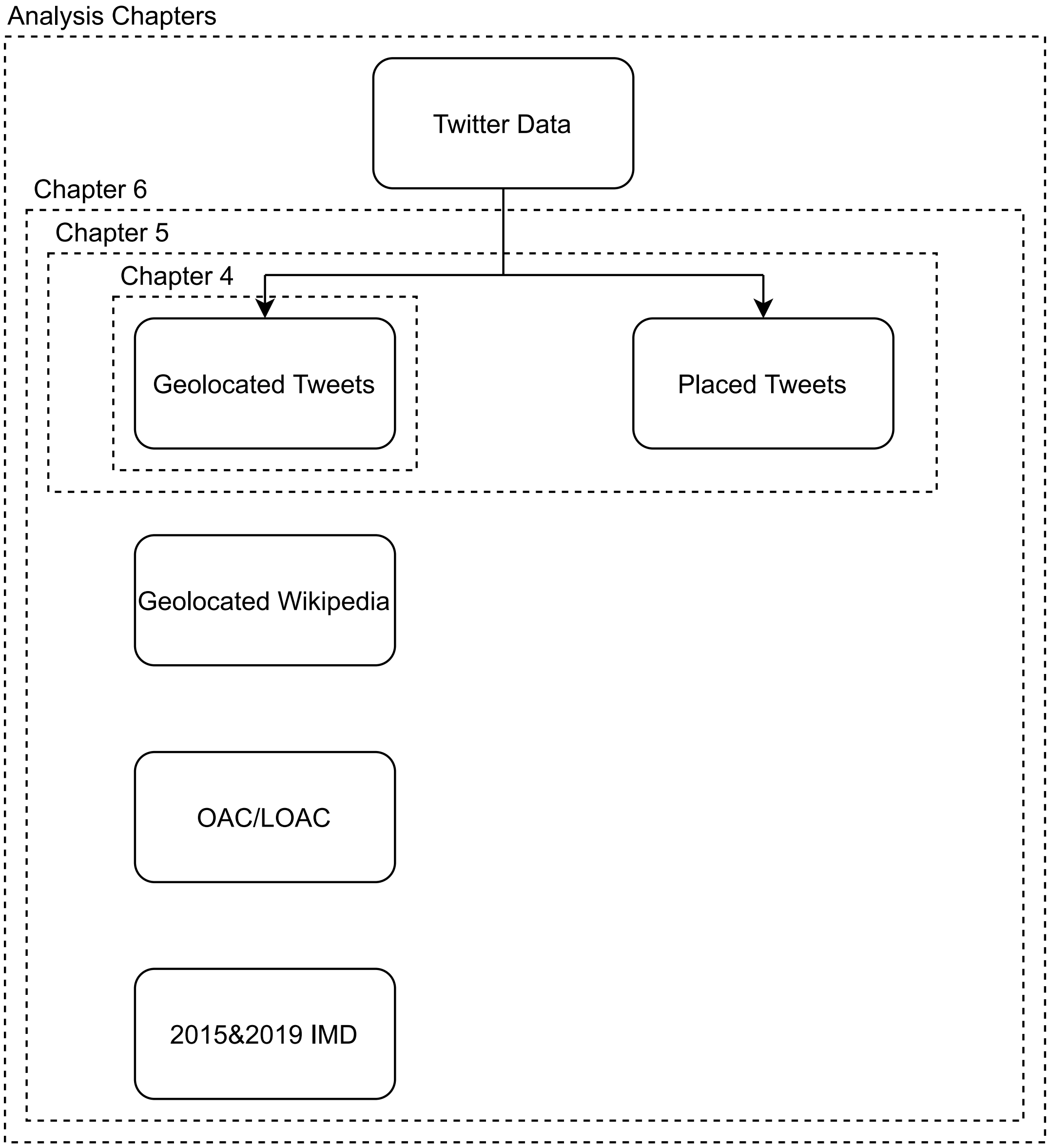


FIGURE 3.1: A graphical illustration of different data in the case studies presented in each analysis chapter.

helps a company respond to customer feedback on Twitter. Thanks to APIs' flexibility and ease of use, Twitter also has supported significant amount of scientific research towards such a social media platform. The Twitter APIs return tweets a data structure encoded using *JavaScript Object Notation* (JSON). JSON is a data format based on key-value pairs, with named attributes and associated values. These attributes and their state are used to describe objects. Twitter serves each tweet as JSON, which encapsulate core attributes that describe the object. Each Tweet has an author, a message, a unique ID, a timestamp of when it was posted, and sometimes geo-metadata shared by the user. Each Tweet also generates *entity* objects, which include arrays of tweet contents such as hashtags, mentions, media (images, animated GIFs or videos), and links (metadata such as the fully unwound URL and the webpage's title and description).

In particular, regarding the geo-metadata of the tweets, tweets can be "geotagged" with a location through the Twitter user-interface or when posting a Tweet using the API. A tweet's location can be a "point" location with a geo-coordinates pair, or a *Twitter Place* with a "bounding box" that describes a larger area ranging from a venue to an entire region where the tweet is generated. There are two "root-level" JSON objects used to describe the location associated with a tweet: *coordinates* and *place* as shown below:

LISTING 3.1: Two "root-level" JSON objects of a tweet's geo-metadata.

```
{
 "coordinates": {},
  "place": {}
 }
```

The *place* object is always present when a tweet is geotagged, while the *coordinate* object is only present non-null when the tweet is geotagged with an exact location. If an exact location is provided, the *coordinate* object will provide a [*long*, *lat*] array with the geo-coordinates, and a *Twitter Place* that corresponds to that location will be assigned. When a user decides to geotag a location to a tweet, they are presented with a list of candidate *Twitter Places* which associate an object "*place_type*" presented within the "*place*" object. The granularity of *place_type* must be one of the five types: *poi* (point of interests, additionally sourced by Foursquare (Twitter, 2021b)), *neighbourhood*, *city*, *admin* or *country*.

In this thesis, I used a dataset of geotagged tweets across the United Kingdom between January 8th, 2018 and December 31st, 2018 using the official Twitter APIs. Although there was a disruption in the data collection between September 27th and October 3rd (five days are missing: 28-09-2018, 29-09-2018, 30-09-2018, 01-10-2018, and 02-10-2018), and some minor disruptions in a few days throughout the year (in each case a few minutes to a couple of hours are missing), the dataset is consisted of 5,870,022 geotagged tweets. All the data are maintained in encrypted form in the Enhanced Security Research Drive of the University of Leicester and will be deleted after five years from collection. For the data that have been sampled for processing as in Chapters 4 and 5 (see introduction about the sampling process in the corresponding chapters), any post containing sensitive data has been excluded from the study. In order to ensure privacy (Metcalf and Crawford, 2016), no post is included in the dissertation, and ethical fabrication (Webb et al., 2017) is used to illustrate some examples

It is worth mentioning that due to the popularity of Twitter, there exists a significant amount of non-human postings (e.g., company account for advertising purpose, or weather forecast stations for reporting weather every hour) on the platform.

Therefore, I designed a 3-step process to filter non-human bots. Firstly, I looked at all the users with more than 3,650 tweets (that is more than 10 tweets per day), and exclude those users in the dataset if their accounts are non-human which only produce advertisements or any other contents that are automatically generated. Secondly, I aggregated tweets into each lower layer super output area (LSOA) and select those areas that have more than 1825 tweets to count the number of tweets per user per LSOA. Then, I analysed each of those users who has more than 730 tweets (2 tweets per day) and exclude the users who are bots. It ends up with 4,565,424 tweets left as the Twitter dataset used in this thesis. It is important to notice that despite I designed a 3-step process to exclude as many bots as possible in the dataset, it is still impossible to exclude every bot. The development of more complex Twitter bot detection methods is beyond the scope of this thesis, and it is indeed one of the limitations of this thesis which will be discussed in Chapter 7 under the section of *Uncertainty*. With such a dataset, each case study samples different parts of data from this dataset, and the details will be introduced in the following analysis chapters (Chapter 4, 5 and 6).

As tweets can be geotagged in forms of coordinates or *Twitter Places*, to clarify different types of tweets which will be investigated in this thesis, I use two terminologies to define the tweets:

- **geolocated tweets**: tweets that have longitude and latitude.
- **placed tweets**: as introduced in the previous subsection, a tweet's location can be a "point" location with a coordinates pair or a *Twitter Place* with a "bounding box" that describes a larger area ranging from a venue to an entire region where the tweet is attached to. I defined such tweets with no coordinates but bounding boxes as *placed tweets*.

3.1.2 Wikipedia Data

The second VGI source in this thesis consists of 173,117 geolocated Wikipedia articles located in England with all language editions collected in 2017 using the databases available on Wikimedia Toolforge (Wikimedia, 2003).

The geolocated of Wikipedia articles include features such as monuments, notable buildings, parks, and head-quarters of organisations. Wikipedia only allows for geotags in the form of points, and even large geographical entities are geotagged to a point. For example, the article about the University of Leicester is associated with a point that has precise latitude and longitude (N 52° 36′ 57.8196", W 1° 7′ 45.3108"). The decision about where to locate entities depends on the combination of the platform's guidelines and the editors' arbitrary choices. As a result, the same entity can be tagged in different locations in different language editions. For instance, at the time of writing, Manchester in the English Wikipedia is geotagged on the Albert Square, whereas the Chinese Wikipedia geotags Manchester at the Gartside Garden which is approximately 1.1 miles away from the Albert Square. Such inconsistencies of geolocations of Wikipedia articles might lead to uncertainties when analysing place representations of the digital platforms. Although this is beyond the scope of this thesis, it is also another limitation of this thesis which will be discussed in Chapter 7 under the section of *Uncertainty*.

3.1.3 Neighbourhood Statistics Geography

In the United Kingdom, the *Office for National Statistics* (ONS) maintains a series of geo-codes (namely ONS codes) to describe a range of geographical areas of the UK, to tabulate official statistical data such as the census. These codes include output areas, super output areas referring to the *Government Statistical Service* of which ONS is part.

Output areas (OAs) were initially produced to support the publication of 2001 Census outputs. They were designed based on postcode blocks after the census data were available, to standardise population size, geographical shape, and social homogeneity (in terms of dwelling types and housing tenure) (Tait, 2012). The OA is the lowest geographical level at which census estimates are provided, and each OA contains at least 40 households and 100 persons with the target size being 125 households (ONS, 2016). The OAs created in 2001 were maintained as much as possible for supporting the publication of the 2011 Census (less than 3% were changed according to ONS (2011)).

Super Output Areas (SOAs) are an assortment of geographical areas developed to facilitate the calculation of the UK Indices of Multiple Deprivation 2004 and subsequently for a range of additional *Neighbourhood Statistics (NeSS)*. The creation of SOAs aimed to generate a collection of areas with consistent size, whose boundaries would be stable and potentially remain no changes, suitable for the publications of statistical data such as the Indices of Deprivation. Each SOA is an aggregation of adjacent OAs with similar social characteristics (Tait, 2012). Each lower layer super output areas (LSOA) has a minimum population of 1000 with a mean size of 1500 and generally consist of between four and six contiguous OAs. Similarly, contiguous LSOAs in groups are organised into Middle Layer Super Output Areas (MSOAs).

3.1.4 Geodemographic Classifications

Geodemographic classification based on census data is a commonly adopted approach within GIScience and has a longstanding history of being created in the UK to understand socio-demographic characteristics. The *Output area classification* (OAC2011) for the UK and the *London output area classification* (LOAC) of 2011 census data created by Gale et al. (2016) and Longley and Singleton (2014) have become important tools for researchers to understand socio-demographic patterns at both the national scale and the urban scale in the UK. For the scope of this thesis, 8 super-groups in England (Gale, 2014) and 8 super-groups within the Greater London (Longley and Singleton, 2014) as shown in Table 3.1 and Table 3.2 will later be adopted in Chapter 6.

3.1.5 English Indices of Deprivation

English Indices of Deprivation (IMDs) are long-established datasets within England to classify the relative deprivation (a measure of poverty) of 32,844 Lower-layer Super Output Areas (LSOAs). Multiple components of deprivation of a place are weighted with different strengths and compiled into a single score of deprivation. The scores of places are divided into ten equal groups (or deciles) according to their deprivation ranks. The use of IMDs in social analysis recognises that fact that deprivation in a place has many interacting components, and aims to measure deprivation from many perspectives rather than adopting a single number describing the concept of deprivation. IMDs are widely considered as an improvement over simpler measures of deprivation, for instance, low average household disposable

TABLE 3.1: List of 8 super-groups from OAC2011 in England.

OAC2011 Super-groups	Descriptions
Rural Residents	The population of this super-group live in rural areas that are far less densely populated compared with elsewhere in the country.
Cosmopolitans	The majority of the population in this super-group live in densely populated urban areas. They are more likely to live in flats and communal establishments, and private renting is more prevalent than nationally.
Ethnicity Central	The population of this group is predominately located in the denser central areas of London, with other inner urban areas across the UK having smaller concentrations. All non-white ethnic groups have a higher representation than the UK average especially people of mixed ethnicity or who are Black, with an above average number of residents born in other EU countries.
Multicultural Metropolitans	The population of this super-group is concentrated in larger urban conurbations in the transitional areas between urban centres and suburbia.
Urbanites	The population of this group are most likely to be located in urban areas in southern England and in less dense concentrations in large urban areas elsewhere in the UK.
Suburbanites	The population of this super-group is most likely to be located on the outskirts of urban areas.
Constrained City Dwellers	This super-group has a lower proportion of people aged 5 to 14 and a higher level aged 65 and over than nationally.
Hard-Pressed Living	The population of this group is most likely to be found in urban surroundings, predominately in northern England and southern Wales. Households are more likely to have non-dependent children and are more likely to live in semi-detached or terraced properties, and to socially rent.

TABLE 3.2: List of 8 super-groups from LOAC in London.

LOAC Super-groups	Descriptions
Intermediate Lifestyles	Employment levels are average for London, and are split between full and part-time working in a range of intermediate occupations (clerical, sales, service).
High Density and High Rise Flats	Concentrations of this super-group are found in densely populated areas of flats.
Urban Elites	This super-group comprises young professionals working in the science, technology, finance and insurance sectors.
City Vibe	There are many young, single professionals in this super-group, living in Zone 2 of the London travel network.
London Life-Cycle	Predominantly White in ethnic composition in this super-group. Residents are highly qualified, employment rates are high, and employment is concentrated in the technical, scientific, finance, insurance and real estate industries.
Settled Asians	Residents of this super-group identify themselves with their Asian origins, although many are second or subsequent generation British residents.
Aging City Fringe	Many of the residents in this super-group are over 45, and many are above state pensionable age.
Multi-Ethnic Suburbs	Residents of this super-group are drawn from a wide range of non-White ethnic groups and White groups are less represented than average for London.

income (Saunders, 2004). IMDs capture variables such as the advantage of access to a good school and the disadvantage of exposure to high levels of air pollution to provide a complete picture regarding living conditions per household. Taking the latest English Indices of Deprivation 2019 (McLennan et al., 2019) as an example, seven domains of deprivation are considered and weighted as follows: *Income* (22.5%), *Employment* (22.5%), *Education* (13.5%), *Health* (13.5%), *Crime* (9.3%), *Barriers to Housing and Services* (9.3%) and *Living Environment* (9.3%). Each of these domains has multiple components. For example, according to McLennan et al. (2019), the *Barriers to Housing and Services* considers seven components including levels of household overcrowding, homelessness, housing affordability, and the distance by road to four types of key amenity (post office, primary school, supermarket, and GP surgery).

In this thesis, I adopt data from 2015 IMD deciles and 2019 IMD deciles. The deprivation deciles within the UK in both 2015 and 2019 are between 1 and 10. The

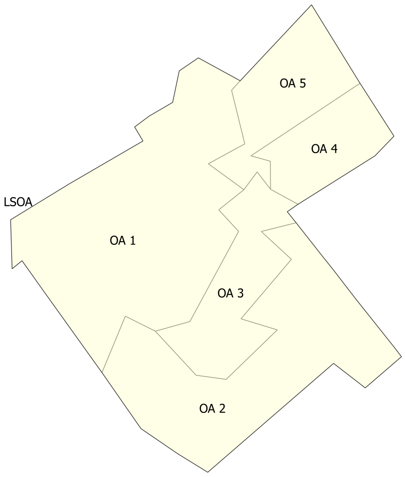


FIGURE 3.2: A demonstration of the area unit difference between LSOA and OA. The LSOA shown in the image contains 5 OAs (denoted by OA 1-5). Map boundaries source: Office for National Statistics licensed under the Open Government Licence v.3.0.

smaller the number is, the worse deprivation it represents.

3.1.6 Scales

The output area classification and English Indices of Multiple Deprivation will be adopted and combined in Chapter 6. However, it is important to notice that these two datasets are organised at different area units — output area classification was calculated on OA-level neighbourhoods, while English Indices of Deprivation (IMDs) were calculated on LSOA-level neighbourhoods. The spatial extent of LSOAs is larger than the OAs, for example, as shown in Figure 3.2. In Chapter 6, I will introduce that the deprivation decile of an OA in this thesis is simply defined as the decile of the LSOA that contains it. I am aware that this is in part problematic due to uncertainties and local variation — aggregated value calculated at one area unit (LSOA) might not necessarily apply equally to all parts of the area (OA). Detailing the implications of this issue is beyond the scope of this dissertation, and it will be considered part of the known unknowns of the framework presented in Chapter 6. The issue will be discussed further in the concluding chapter and future studies

3.2 Neural Networks

In Chapter 2, I have provided a brief and high-level introduction on the types of neural networks (ANN, CNN, RNN and GCN) that are adopted in this thesis. To better understand what deep learning techniques are and to have a more comprehensive introduction on the methodologies and frameworks which are adopted in the following chapters, it is helpful to start with some basic concepts of neural network models regarding their the technical components.

3.2.1 Linear Classifiers and Loss Function

To help ease the introduction in this section, I will explain the concepts using image classification as an example. Assuming there is a training dataset of a type of data such as images $x_i \in R^D$, and each image is associated with a label y_i, where $i =$

$1, 2...N$ and $y_i = 1, 2...K$. That is, there are N examples (each with a dimensionality D) and K distinct categories. The objective is to define the score function $f : R^D \to R^K$ that maps the raw image pixels to class scores. The simplest linear mapping function is widely defined as:

$$y_i = f(x_i, W, b) = Wx_i + b \tag{3.1}$$

where I assume the image x_i has all of its pixels flattened out to a single column vector of shape $[D \times 1]$. The matrix W (of size $[K \times D]$), and the vector b (of size $[K \times 1]$) are the parameters of the function. The parameters in W are often called the *weights*, and b is called the *bias vector* because it influences the output scores, but without interacting with the actual data x_i. It is important to highlight that the input of this linear mapping function (x_i, y_i) is given and fixed, but the weight W and the bias vector b are adjustable. Consequently, the objective of the function is to set those parameters so that the predicted class scores are consistent with the ground truth labels in the training data.

Before moving on, it is worth to mention a common simplifying trick to representing the two parameters W, b as one. Such a trick combines the two sets of parameters into a single matrix that holds both of them by extending the vector x_i with one additional dimension that always holds the constant one which is a default bias dimension. With the extra dimension, the new linear mapping function will simplify to a matrix multiply:

$$f(x_i, W, b) = Wx_i \tag{3.2}$$

When a researcher uses an algorithmic classifier to classify such a dataset that I mentioned above, there will naturally exist mismatch between the output labels produced by the classifier and the ground truth labels. Thus, a *loss function* (or sometimes also referred to as the *cost function* or the *objective*) is defined to measure such inconsistencies between the output labels and the ground truth labels. Intuitively, the loss will be high if the classifier poorly classifies the training data, and it will be low if the classifier performs well. Taking one of the most popular loss function Multi-class Support Vector Machine (SVM) loss (Wang and Xue, 2014) as an example, the SVM loss is designed so that the SVM would "enable" the correct class for each image to a have a score higher than the incorrect classes by fixed margin $\triangle$. Given the pixels of image x_i and the label y_i which specifies the index of the correct class, the score function (1.1) takes the pixels and computes the vector $f(x_i, W, b)$ of class scores (abbreviate to s). Thus, the SVM loss the i-th sample x_i which has j-th class label is formalised as:

$$\mathcal{L}_i = \sum_{j \neq y_i} max(0, s_j - s_{y_i} + \triangle) \tag{3.3}$$

the threshold at zero $max(0, -)$ is often called the *hinge loss*. In reality, the issue of overfitting happens when model learns the signal as well as noise in the training data and would not perform well on new data on which model was not trained on. To ameliorate such an issue, it is common to add a *regularization penalty* $R(W)$ (e.g., L^1 regularization and L^2 regularization) in the loss function. Notice that $R(W)$ is not a function of the data; it is only based on the weights W. Therefore, the full Multi-class SVM loss becomes:

$$\mathcal{L} = \frac{1}{N} \sum_i L_i + \lambda R(W) \tag{3.4}$$

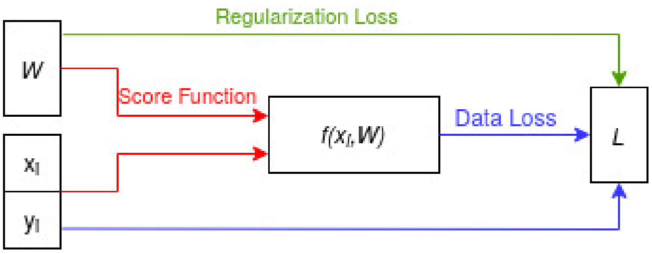


FIGURE 3.3: summary of the information flow for *classifiers*, *loss functions* and *optimization*.

where N is the number of training examples, and λ is a hyperparameter. The loss function lets us quantify the quality of any particular set of weights W. To find W that minimises the loss function, I will introduce another important component in the next section *Optimization*.

3.2.2 Optimisation

The loss function is used to quantify the quality of any particular set of weights W, and the goal of optimisation is to find such a set of W which can minimise the loss function. Intuitively, optimisation is to find a direction in the weight-space (a mathematical conceptualised space of possible parameter values) that would improve the weight vector W and produce a lower loss. Such a process can be done by calculating the gradient of the loss function. The mathematical expression for such a calculation is:

$$\frac{df(x)}{dx} = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h} \tag{3.5}$$

The procedure of repeatedly evaluating the gradient and then performing a parameter update is called *Gradient Descent* which is currently by far the most common and established way of optimising neural network loss functions. There are many different methods (e.g., Stochastic Gradient Descent methods, Adaptive Gradient methods, etc.) to perform *Gradient Descent*; their mathematical details are beyond the scope of this thesis. For a more detailed tutorial of standard methods for machine learning (not just deep learning), see Curtis and Scheinberg (2017) and Bottou et al. (2018).

Thus far, I have introduced three key components of a neural network model. Figure 3.3 summarises the information flow for *classifiers*, *loss functions* and *optimization* and how they interact with each other. The dataset of pairs of (x, y) is given and fixed. The weights start as random values and can change. During the forward pass, the classifier computes class scores, stored in vector f. The loss function contains two components: The data loss computes the compatibility between the scores f and the labels y. The regularisation loss is only a function of the weights. During Gradient Descent, it computes the gradient on the weights (and optionally on data if necessary) and use them to perform a parameter update during Gradient Descent.

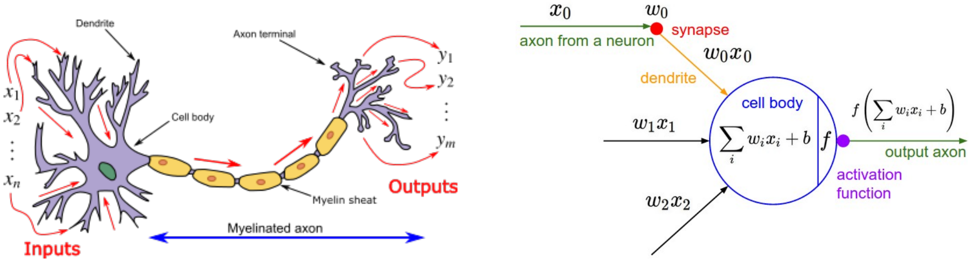
3.2.3 Back-propagation and Neural Networks

Back-propagation is a recursive application of the chain rule that computes the gradients of a given expression $f(x)$, where x is a vector of inputs and the objective is to compute the gradient of f at x ($\nabla f(x)$). Instead of providing complicated mathematical background of the process, an intuitive understanding of back-propagation is that given a classifier and a loss function, the back-propagation calculates the gradient of the loss function concerning the classifier's weights. It is a crucial process in a multi-layer neural network, the "backwards" part of the name stems from the fact that calculation of the gradient proceeds backwards through the network. The gradient on the weights of the final layer is calculated first, and the gradient on the weights of the first layer is calculated last. Partial computations of the gradient from one layer are reused in the computation of the gradient for the previous layer. This backwards flow of the error (refers to the mismatch between the model's outputs and ground truth labels) allows for efficient computation of the gradient at each layer.

Before moving on, I will provide a general introduction to the neural networks without using brain analogies. Recall the proposed image classification task and the linear classifier that I introduced previously as $f = Wx$, where W was a weight matrix (of size $[K \times D]$) and x was an input column vector containing all pixel data of the image (with shape $[D \times 1]$). An example of a 2-layer neural network instead can be defined as $f = W_2 \max(0, W_1 x)$, where W_1 could be a $[N \times D]$ matrix transforming the image into a N-dimensional intermediate vector, and $W_1 x$ can be seen as the first layer of this 2-layer neural network. The function $max(0, -)$ is a non-linearity activation function that is applied element-wise. The activation function defines the output of the first layer given an input or set of input. In other words, it decides whether the learned feature of a layer will be passed on to the next layer and helps to block information which the network considers as trivial. Finally, the matrix W_2 would then be of size $[K \times N]$, so that the neural network again get K numbers out to interpret the classification scores. The parameters W_2, W_1 are learned with *Gradient Descent*, and their gradients are derived with back-propagation. Similarly, a 3-layer neural network could analogously be defined as: $f = W_3 \max(0, W_2 \max(0, W_1 x))$ with W_1, W_2 and W_3 are parameters to be learned.

As mentioned in Chapter 2, the discipline of neural networks is inspired by the biological neural networks that constitute human brains. The basic computational unit of the brain is a neuron. The human nervous system has approximately 86 billion neurons, and they are connected with approximately 10^{14} - 10^{15} synapses. Figure 3.4(A) shows that each neuron receives input signals from its dendrites and produces output signals along its (single) axon, and the axon eventually branches out and connects via synapses to dendrites of other neurons. In the computational model of a neuron in Figure 3.4(B), the signal x_0 that travels along the axons interact multiplicatively with the dendrites of the other neuron ($w_0 x_0$) based on the synaptic strength at that synapse w_0. As can be seen from Figure 3.4(B), the mathematical equation in the cell body looks similar to the linear classifier that I introduced earlier. In fact, a neuron in a neural network model can be modelled as a single linear classifier.

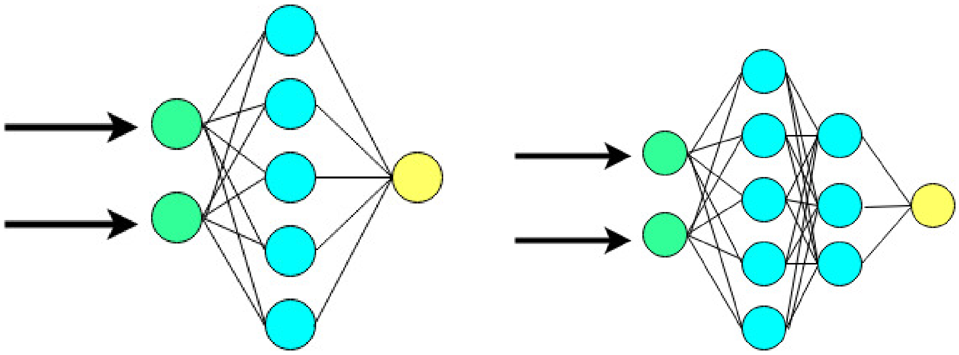
Therefore, as shown in Figure 3.5 neural networks are modelled as collections of neurons that are connected in an acyclic graph. The representation of the input data output by each layer is called the hidden representation of the data. Note that regardless which type of neural networks, the basic concepts of *neuron*, *back-propagation*, *loss function* and *optimization* are similar. With the concepts mentioned above in mind, I will introduce the frameworks and methods in the rest of this chapter.



(A) Human nervous system (image by Egm4313.s12, via Wikimedia Commons, CC BY-SA 4.0).

(B) Mathematical model.

FIGURE 3.4: A cartoon drawing of a biological neuron (A) and its mathematical model (B).



(A) A 2-layer Neural Network (one hidden layer of 5 neurons (or units) and one output layer with 1 neurons), and two inputs.

(B) A 3-layer neural network with three inputs, two hidden layers of 5 and 3 neurons each and one output layer.

FIGURE 3.5: Artificial neural networks. Notice that in both cases there are connections (synapses) between neurons across layers, but not within a layer.

3.3 Multi-modal Autoencoder

To address the research question **RQ1** mentioned in Chapter 1, I propose a multi-modal autoencoder to extract the combined representations of images and text from social media posts. In this section, I will introduce what autoencoders are and my proposed multi-modal autoencoder.

The definition of an autoencoder is a type of neural network to learn compressed data representations in an unsupervised manner (Kramer, 1991). It learns to copy its input to its output and has an internal hidden layer to describe the representations for the input. A typical neural network architecture for an autoencoder is shown in Figure 3.6. It is constituted by an encoder which maps the input into a representation, and a decoder which maps the representation to a reconstruction of the original input.

Despite the fact that autoencoders are often trained with a single layer encoder and a single layer decoder, adding more layers symmetrically into both parts has been evidenced to be able to produce better data representations compared to single-layer autoencoders (Goodfellow et al., 2016). Figure 3.7 shows an autoencoder with three fully connected layers, and autoencoders remain the flexibility to add more layers or substitute the fully connected layers to other neural architectures (e.g., long short-term memory neural network (LSTM) layers, convolutional network ((CNN))

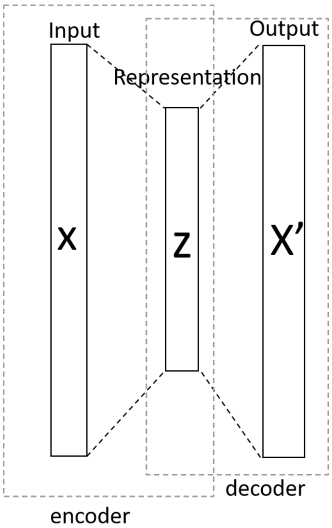


FIGURE 3.6: Schema of a basic autoencoder. **x** is the input, **x′** is the output, and **z** is the compressed representation

layers, etc.).

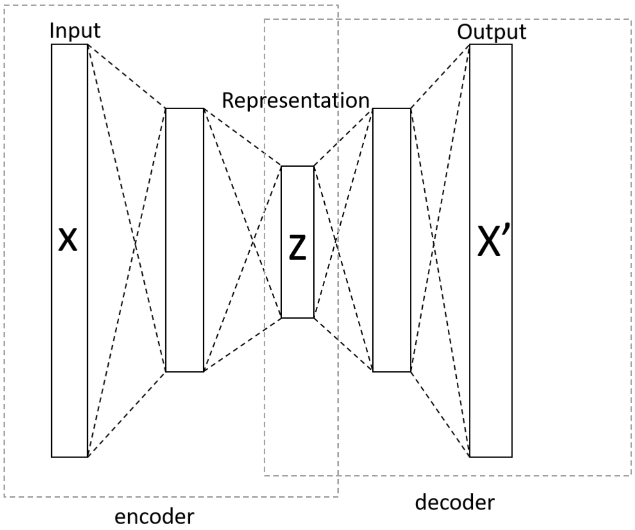


FIGURE 3.7: An autoencoder with 3 fully connected layers.

One of the primary uses of autoencoders is dimensionality reduction (Goodfellow et al., 2016). Comparing with principal component analysis (PCA) (Wold et al., 1987), a traditional dimensionality reduction approach, the main advantage of autoencoders is their non-linearity, which allow the model to learn more powerful generalisations, and reconstruct back the input with a significantly lower loss of information (Hinton and Salakhutdinov, 2006). Therefore, the studies of autoencoder have massively benefited the research and applications in a broad spectrum of domains, ranging from information retrieval (Salakhutdinov and Hinton, 2009) to image processing (Cho, 2013; Guo et al., 2017).

The success of the autoencoders to compress data into lower dimensions have attracted a growing interest within the field of common representation learning (CRL), wherein different modalities of the data are represented in a common subspace (Chandar et al., 2016).

Ngiam et al. (2011) proposed a multi-modal autoencoder (MAE) to learn a common representation by reconstructing two modalities. Given any modality, the model learns to reconstruct itself and the other modality. Chandar et al. (2016) further developed an MAE based approach named Correlational Neural Networks (Corrent) which integrated with a canonical correlation analysis (Hotelling, 1992) layer to ensure the learned representation can be highly correlated. This is particularly interesting in my study, where I assume that text content and image content of a social media post are correlated to each other. In other words, text and images are jointly expressing the activities which users are doing—as such, finding a common representation of a multi-media content of social media is essential for further analysis. My proposed approach is based on Cornet, which is able to learn joint representations by maximising correlation of two views when projected to common subspace. As shown in Appendix A, the dense layers of the original Corrent are replaced with Resnet-style convolution layers (Ledig et al., 2017) for learning image representations, and an LSTM layer for text representations learning. The objective is not only to minimise the self-construction error, but also the cross-reconstruction error from image and texts, and maximise the correlation between the hidden representations of both parts. I achieved this by minimising the objective function introduced in the original Corrnet paper (Chandar et al., 2016):

$$\mathcal{L}_{\mathcal{Z}} = \sum_{i=1}^{N}(L(z_i, g(h(z_i))) + L(z_i, g(h(x_i))) + L(z_i, g(h(y_i)))) - \lambda corr(h(X), h(Y)) \tag{3.6}$$

$$corr(h(X), h(Y)) = \frac{\sum_{i=1}^{N}(h(x_i - \overline{h(X)})(h(y_i - \overline{h(Y)}))}{\sqrt{(\sum_{i=1}^{N}(h(x_i - \overline{h(X)})^2(\sum_{i=1}^{N}(h(y_i - \overline{h(Y)})^2}} \tag{3.7}$$

considering a dataset $\mathcal{Z} = \{z_i\}_{i=1}^{N}$ where all data have inputs from two channels of media text and images X and Y. Each data z_i can be represented as $z_i = (x_i, y_i)$, where $x_i \in X$ and $y_i \in Y$. L is the squared error reconstruction error, and λ is the scaling parameter. $\overline{h(X)}$ is the mean vector for the hidden representation $h(x_i)$ of the text part input and $\overline{h(Y)}$ is the mean vector for the hidden representation $h(y_i)$ of the image part input. $h(z) = f(Wx + Vy + b)$, where W and V are two $k \times d_i$ weight matrix and b is a $k \times 1$ bias vector. Equation 3.6 is the objective function of the proposed multi-modal autoencoder. The first term is the objective function that allows learning meaningful hidden representations. The second term ensures that both images and text output from the decoder can be reconstructed using only text representations. Similarly, the third term ensures that both images and text output from the decoder can be reconstructed using only image representations. The fourth term ensures that the combined representations are highly correlated, and it is defined in Equation 3.7.

Extracting representations from social media posts using my proposed multi-modal autoencoder is a crucial fundamental stage for further analysis. The extracted representations will be adopted for a semi-supervised classification task detailed in Section 4.1.

3.4 Graph Convolutional Network

To address research question **RQ2**, I adopt a graph convolutional network (GCN) as the method to classify users' activities on social media platforms. In this section, I will introduce the concept of graph theory and the mathematical details of GCN.

Graphs are a kind of data structure which models a set of entities (represented as nodes) and their relationships (represented as edges). Within the studies of machine learning, the use of graphs has attracted a wide range of attention because of their extraordinary expressive power. The graphs can be adopted to denote for a large number of systems across various disciplines, including social science (e.g., social networks Hamilton et al. (2017)), natural science (e.g., protein-protein interaction networks Fout et al. (2017)), knowledge graph (e.g., Trouillon et al., 2016), etc..

Graph neural network (GNN) was first proposed in Scarselli et al. (2008), which extended existing neural networks for processing the data represented in graph domains. To better introduce GNN, all the notations in this subsection are following the notations in Zhou et al. (2018). The target of a GNN is to learn a state embedding $h_v \in \mathbb{R}^m$, h_v is the hidden representation of node v which contains the information of neighborhood for each node, and can be used to produce the outputs o_v such as nodes' labels; $\mathbb{R}^m$ is m-dimensional Euclidean space. In GNN, the *local transition function* f is an essential element that is shared among all nodes and updates the node state according to the input neighbourhood. The h_v and o_v are defined as:

$$h_v = f(x_v, x_{co[v]}, h_{ne[v]}, x_{ne[v]}) \tag{3.8}$$

$$o_v = g(h_v, x_v) \tag{3.9}$$

where $x_v, x_{co[v]}, h_{ne[v]}, x_{ne[v]}$ are features of node v, the features of its edges, the states, and the features of the nodes in the neighborhood of v, respectively. g is defined as a *local output function* which describes how the output is produced. With the *local transition function* and *local output function* of a node in a graph, the *global transition function* F and *global output function* G are defined as:

$$H = F(H, X) \tag{3.10}$$

$$O = G(H, X_N) \tag{3.11}$$

where H, O, X and X_N are the vectors constructed by stacking all the states, all the outputs, all the features, and all the node features, respectively. The loss function for such a GNN is defined as:

$$\mathcal{L} = \sum_{i=1}^{p} (t_i - o_i) \tag{3.12}$$

and t_i represents the ground-truth label of a target node. The learning algorithm is based on a gradient descent strategy. The specific type of graph neural network which will be introduced in the next subsection is one of the variants of GNN, which is developed based on graph convolutions to gather information from each node's neighbours and specific updaters to update nodes' hidden representations.

In this thesis, I adopt the graph structures constructed using social media posts and use it to take advantage of recent advances in graph convolutional neural networks. I consider each social media post as a node in a graph network and define the edges linking the nodes based on a geographical neighbourhood (i.e., based on the spatial distance between the location of the geotags attached to the posts) to construct an adjacency matrix A. Several definitions of geographical neighbourhoods are taken into account, which will be detailed in Section 4.3.1. I then frame the problem of classifying each tweet in such a spatial graph as a graph-based semi-supervised learning task, and a graph convolution network (GCN) (Kipf and Welling, 2016a) is adopted for efficient information propagation through the graph.

As already mentioned in Chapter 2, graph convolution, in general, is defined as a filter moving over the nodes of the graph, with the adjacency matrix determining the area captured by the filter. The graph convolution process will aggregate the information from a node's connected neighbours. By propagating through the hidden layers, GCN is able to produce useful feature representations of nodes in the graph, thus benefits the further classification task.

In Kipf and Welling (2016a), they model a graph-based classification task as *f(X,A)*, where X is the extracted information from the multi-modal autoencoder for each post, and A is the adjacency matrix for the graph. The model is expected to produce a node-level output Z as:

$$Z = f(X, A) = softmax(H^{(L)}) \tag{3.13}$$

which satisfies the layer-wise propagation rule for GCN:

$$H^{(L+1)} = \sigma(\hat{D}^{-\frac{1}{2}}\hat{A}\hat{D}^{-\frac{1}{2}}H^{(L)}W^{(L)}) \tag{3.14}$$

with $\hat{A} = A + I_N$. I_N is the identity matrix of A and $W^{(L)}$ denotes the trainable weight matrix of the Lth layer of the neural network. $\hat{D}_{ii} = \sum_j \hat{A}_{ij}$, and $\sigma(\cdot)$ represents a non-linear activation function using $ReLu(\cdot) = max(0, \cdot)$. $H^{(L)}$ is the activation matrix for the Lth layer; for example, $H^{(0)} = X$ and $H^{(L)} = \hat{A}ReLu(H^{(L-1)})W^{(L)}$. The softmax activation in formula (3.3) is used for classifying nodes into their corresponding categories. I calculate the cross-entropy error as the loss function over all labeled nodes in the graph:

$$\mathcal{L} = -\sum_{l \in \mathcal{Y}_L} \sum_{f=1}^{F} \mathcal{Y}_{lf} \ln Z_{lf} \tag{3.15}$$

where $\mathcal{Y}_L$ is the set of nodes that have labels.

3.5 Variational Graph Autoencoder

In this section, I will introduce a variational graph autoencoder, which I adapt to estimate the locations of social media posts using a link prediction process to address research question **RQ3**.

Link prediction is defined as predicting the future or missing relationships from nodes in a complex graph based on the observed graph structure and node attributes (Martinčić-Ipšić et al., 2017). Kipf and Welling (2016b) proposed an unsupervised learning model on graph-structured data based on a variational autoencoder (Kingma and Welling, 2013; Rezende et al., 2014) named Variational Graph Autoencoder (VGAE). The model is composed of a graph convolutional network (GCN) (Kipf and Welling, 2016a) as its encoder and an inner product decoder.

The inference model of the encoder is parameterized by a two-layer GCN:

$$q(Z|X, A) = \Pi_{i=1}^{N} q(z_i|X, A), \text{ with } q(z_i|X, A) = \mathcal{N}(z_i|\mu_i, diag(\sigma_i^2)), \tag{3.16}$$

where $\mu = GCN_\mu(X, A)$ is the matrix of mean vector μ_i and $log\sigma = GCN_\sigma(X, A)$. A is introduced as an adjacency matrix of the graph; z_i is the stochastic latent variables summarized in matrix Z, and node features are summarized in matrix X. The two-layer GCN is defined as $GCN(X, A) = \hat{A}ReLu(H^{(0)})W^{(1)}$, where $H^{(0)} = \hat{A}XW^{(0)}$; $W^{(i)}$ denotes for the weight matrices. $GCN_{\mu}(X, A)$ and $GCN_{\sigma}(X, A)$ share first-layer parameters $W^{(0)}$. $ReLu(\cdot) = max(0, \cdot)$ and $\hat{A} = D^{-\frac{1}{2}}AD^{-\frac{1}{2}}$ is the symmetrically normalized adjacency matrix of A.

An inner product between latent variables is used to define the generative model of the VGAE:

$$p(A|Z) = \Pi_{i=1}^{N}\Pi_{j=1}^{N}p(A_{ij}|z_i, z_j), \text{ with } p(A_{ij} = 1|z_i, z_j) = \sigma(z_i^T, z_j), \tag{3.17}$$

where $\sigma(\cdot)$ is the logistic sigmoid function.

The learning process of the autoencoder is defined by optimizing the variational lower bound L:

$$L = \mathbb{E}_{q(Z|X,A)}[logp(A|Z)] - KL[q(Z|X, A)||p(Z)], \tag{3.18}$$

$KL[q(\cdot)||p(\cdot)]$ is the Kullback-Leibler divergence (Kullback and Leibler, 1951) between $q(\cdot)$ and $p(\cdot)$.

The results summarised in Kipf and Welling (2016b) demonstrate the ability of VGAE on link prediction tasks using different state of the art citation networks as benchmarks. In this thesis, VGAE model is adapted on geographically connected social media spatial networks and estimates the potential locations of each new post from users.

3.6 Complex Graph Embeddings of Knowledge Graph

To address the research question **RQ4**, I proposed a spatial knowledge graph to model the socio-demographic changes at the urban scale using a link prediction process. In this section, I will describe the method I adopt to perform link prediction on the knowledge graph.

In this thesis, I adopt a state-of-art method ComplEx (Trouillon et al., 2016) to predict relations for a spatial knowledge graph which later will be introduced in Chapter 6. Consider $\mathcal{R}$ and $\mathcal{E}$ the set of relations and entities in a knowledge graph. The knowledge graph aims to recover the matrices of scores $\mathbf{X}_r$ for all relations $r \in \mathcal{R}$. Given two entities s and $o \in \mathcal{E}$, the probability that the fact $r(s, o)$ exists is defined as:

$$P(Y_{rso} = 1) = \sigma(\phi(r, s, o; \theta)), \tag{3.19}$$

where $\mathbf{Y}$ is the partially observed sign matrix, ϕ is a scoring function based on a factorization of the observed relations, and θ denotes the parameters (embeddings of entities and relations) of the model.The scoring function $\phi(r, s, o; \theta)$ is further defined as:

$$\begin{aligned}
\phi(r, s, o; \theta) = Re(w_r, e_s, \bar{e}_o) = Re(\sum_{k=1}^{K} w_{rk}e_{sk}\bar{e}_{ok}) \\
=< Re(w_r), Re(e_s), Re(e_o) > \\
+ < Re(w_r), Im(e_s), Im(e_o) > \\
+ < Im(w_r), Re(e_s), Im(e_o) > \\
- < Im(w_r), Im(e_s), Re(e_o) >,
\end{aligned} \tag{3.20}$$

where e_s and e_o are embeddings of the entities s and o, and w_r is the embedding for the relation r. The objective function of this model is designed by minimising the negative log-likelihood of the logistic model with L^2 regularization on the parameters θ:

$$\mathcal{L} = \min_{\theta} \sum_{r(s,o)\in\omega} log(1 + exp(-Y_{rso}\phi(s,r,o;\theta))) + \lambda||\theta||_2^2 \tag{3.21}$$

The results summarised in the original paper of this model (Trouillon et al., 2016) demonstrate the model is powerful in link prediction tasks on knowledge graphs. In this thesis, I adopt this model to predict socio-demographic patterns of IMD deciles in the case studies in both England and Greater London.

3.7 Summary

In this chapter, I provided a brief introduction on the datasets used in this thesis, including geotagged Twitter data and Wikipedia data, OAC and LOAC, 2015 and 2019 English Indices of Multiple Deprivation. In addition, I provide an in-depth introduction to neural networks and their key components so that it benefits my further explanations of the methodologies and frameworks that I develop and adopt in this thesis. In the rest of this thesis, I will illustrate how each of the frameworks is used to answer geographic questions in different case studies (Chapter 4, 5, and 6).

Chapter 4

Classification Learning through a Graph-Based Semi-supervised Approach

Part of this work is presented in this chapter has been published as:

- **Pengyuan Liu and Stefano De Sabbata**, *2019*. Learning Digital Geographies through a Graph-Based Semi-supervised Approach. In *15th International Conference of GeoComputation*.

The extended journal version of this paper is published as:

- **Pengyuan Liu and Stefano De Sabbata**, 2021. A graph-based semi-supervised approach to classification learning in digital geographies[1]. *Computers, Environment and Urban Systems*, 86, p.101583. DOI: https://doi.org/10.1016/j.compenvurbsys.2020.101583

.

4.1 Introduction

Understanding place representation is a central problem in GIScience (Purves et al., 2019), and as discussed in Section 2.3, UGC represents a significant source of information about places. Traditionally, extensive surveys and long periods of observation were required to collect an adequate amount of data to investigate social practices and study the associated urban representations. Nowadays, people regularly use online digital platforms due to their convenience, efficiency, and significant broadcasting power for sharing information. People often reveal their social practices or their intent to carry out social activities within their online communications or posts.

Due to the potential of digital platforms for exploring social practices in space and the narrative of places (Abernathy, 2016), social media platforms in general, and Twitter in particular, have been at the centre of data-driven analysis in GIScience and quantitative geography for about a decade (Miller and Goodchild, 2015a). Although existing studies have advanced our abilities to understand the spatial patterns of social media, they are primarily only focused on text content. However, text UGC is not the only form of communication that users post on social media platforms. Digital platforms have become increasingly visual over the past decade as visual content has become more prevalent as content posted online (Gleason et al., 2019),

[1] Code to reproduce my experiments is available at: `https://github.com/PengyuanLiu1993/PhD_Thesis_Codes_PengyuanLiu/tree/master/GCN_Activities_Classification`

which renders the analysis of visual data, an interesting area to explore. Despite the growing popularity of visual content in social media, limited work has been done so far on such content within the field of GIScience. The lack of visual content analysis is a severe limitation, as image content is a key component of social media posts – especially considering the rise of image-focused platforms such as Instagram or Flickr. "A picture is worth thousand words" (Wang and Li, 2015, p. 1584), visual content can also provide rich information regarding places, the use of space, and people's experiences of landscape.

Studies within digital geographies on the visual content of geolocated UGC mostly focused on their tags or meta-data (e.g., Hollenstein and Purves, 2010; Gao et al., 2015; Xu et al., 2017b), while they heavily rely on users tagging their posts accurately. However, information created on social media platforms tend to be noisy, and images are commonly attached with multiple tags, in which some of them may be irrelevant to the content. One of the objectives of this chapter is to answer the research question **RQ1** proposed in Chapter 1:

- *How can spatial or spatio-temporal distributions of UGC benefit our understanding of places and their representations?*

UGC enables scholars to understand place representations by describing their activities and locales (Ballatore and De Sabbata, 2019). Time and geolocation are important features that UGC includes with their content. Information shared on digital platforms indicate the patterns of users' everyday life, which consistently augment and reinforce the assumptions of local societies through time, and layer the dynamic information across and over geographic space (Graham et al., 2015a). As discussed in Chapter 2.3.2, the relationship between content and space and time on digital platforms can help the studies go beyond "geotags" (Crampton et al., 2013). Based on the conceptualisation of social media posts as "augmentations" (Ballatore and De Sabbata, 2019) of places as "time-space configurations" (Agnew and Livingstone, 2011), the second objective of this Chapter is to test the hypothesis that the spatio-temporal aspects of social media posts would benefit the content analysis and further inform our understanding of digital representations of the city (Pereira et al., 2013) (***RQ2*** proposed in Chapter 1).

- *How can spatial or spatio-temporal distributions of social media posts benefit the semantic categorization of their contents?*

In this chapter, I present and test an approach to the exploratory analysis of social media content capable of classifying posts based not only on the textual component but also taking into account their visual content. The latter is a crucial contribution of my approach, as only a handful of papers account for images when conducting quantitative analyses of social media content (Gao et al., 2015; Xu et al., 2017b; Huang et al., 2018). Furthermore, with a conceptualisation of posts as “augmentations”(Graham et al., 2015a) of places, understood as “time-space configurations”(Agnew and Livingstone, 2011), I go beyond the geotag (Crampton et al., 2013) by developing graph convolutional networks that account for the relationships between each post and its spatio-temporal neighbours. To the best of my knowledge, the proposed model is the first to account for all four aspects (text and images, as well as geographical and temporal information) using a deep learning approach.

The novelties of this study are:

- this study devises a graph-based framework for understanding place representations through multimedia content (text and images, as well as geographical and temporal information) from digital platforms;

- this chapter highlights the importance of the spatial proximity and spatial-temporal patterns of UGC and how they can aid the model to better classify users' spatial activities.

4.2 Case Study

This research is interested in understanding the digital representations (Graham et al., 2013b) of the city through users' spatial activities carried out by their related geolocated social media posts. The case study includes a geographical analysis of social media multimedia content regarding a certain set of topics (e.g., posts about personal life, trending news, or entertainment) in London. In digital geographies studies, it is common to qualitatively explore a sample of a few hundred tweets, and conduct content and visual analysis by categorising the sampled tweets through a relatively small number of "codes" (i.e., labels – see, e.g., Felt, 2016; Awcock, 2018).

In this scenario, the research objective in the Chapter is to develop an approach that could learn from a small coded sample and apply it to a much larger dataset, thus aiding research in digital geographies, bridging quantitative and qualitative.

The labelled dataset could then be used for further exploration of the specific topics. Due to the fact that the manual labelling process of the Twitter data is challenging and time-consuming, for the scope of this thesis, I set myself a limit of three working days to manually label a set of tweets randomly sampled from the dataset introduced in Section 3.1, thus reaching a total of 701 tweets manually labeled as discussed below. I manually labeled the randomly sampled 701 tweets into 11 different categories: *Animals*, *Entertainment*, *Food*, *Nature*, *News*, *Personal*, *Places and attractions*, *Social*, *Sports*, *Work* and *Not informative*. The latter category includes advertisement and other content that was difficult to interpret. The category *Personal* includes content related to personal daily activities such as shopping or selfies, whereas tweets in the category *Social* are related to social activities (e.g., parities). The category *Work* mostly contains tweets related to offices environment or the description of users' work. As illustrated in Figure 4.1, the sampled dataset is unbalanced as certain categories are more represented than other, for instance, there are 166 tweets regarding *Places and attractions* while only 8 tweets in the category *Animals*.

It is important to emphasise that the number of categories of activities is a designing choice and I labelled social media only based on their text and image content rather than labelling them based on their location or geographic content explicitly. That is, the labels used in the case study are not geographical per se. The predefined classification created is relatively generic, but still very subjective and expressing the interests and understanding of the authors on the classified content. Other authors might prefer to incorporate the category *Work* into *Personal*, or clearly differentiate a diverse set of *Sports* (e.g., football or tennis). However, this is not an issue in the scope of this experiment. The objective of my proposed approach is to provide a framework to classify large volumes of social media posts that is unrealistic to process manually, based on a set of labels tailored to a specific project or task. As such, the preparation of the data fits the scenario.

Traditional classification tasks in computer science tend to use datasets created using a top-down approach with a set of well-balanced categories as benchmarks, which are optimised to test the effectiveness of new algorithms. Given the aim of this approach, I decided to use a "real-life" dataset, retrieved from Twitter directly, which is much noisier and could be difficult to categorise even for human assessors.

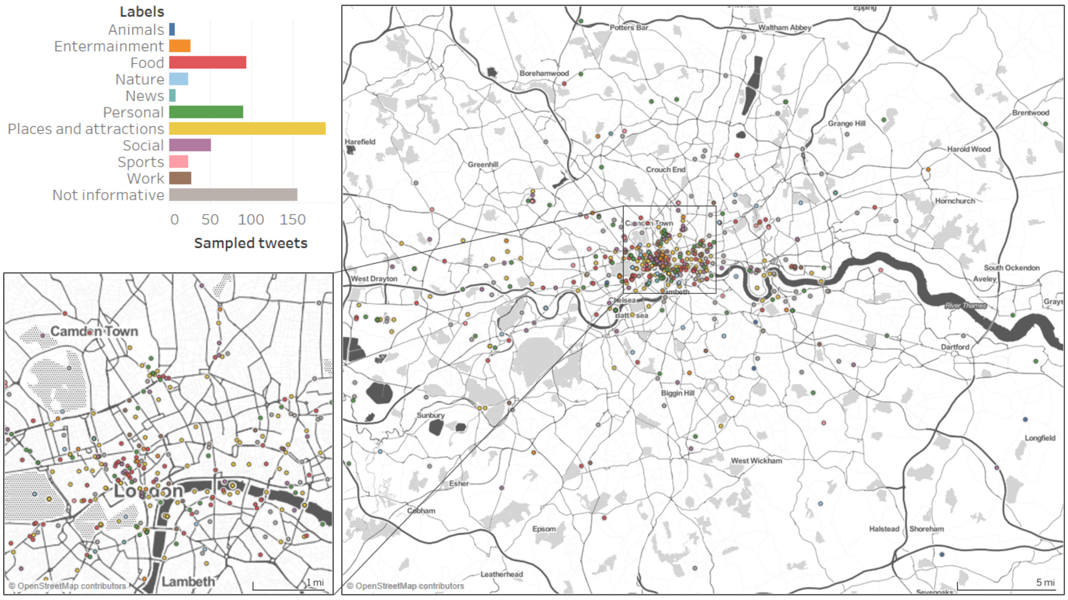


FIGURE 4.1: Distribution of labeled tweets used for training and testing. Map tiles by Stamen Design, under CC BY 3.0. Data by OpenStreetMap, under ODbL.

Even for tweets within the same category, the information tends to be much fuzzier compared to datasets used in traditional classification tasks.

As shown in Figure 4.1, the distribution of tweets in the dataset is heavily concentrated in the central area of the inner boroughs of London, while only a few tweets are located in the suburban areas of the external boroughs. The impact of this skewed geographical distribution on the creation of the spatial graph was one of the reasons that led to testing the diverse set of approaches which will be discussed in Section 4.3. Most categories seem to follow this general pattern, and while some expected clusters can be identified (e.g., *Food* in Soho, or *Nature* in Hyde Park), there seems to be no clear-cut geographic clustering of the categories among the 701 sampled tweets.

4.3 Methodology

Places in geography are not isolated but are connected in many ways (Nystuen and Dacey, 1961; Noronha and Goodchild, 1992), which could be both physical and social, using measures such as distance, adjacency, and spatial interaction (Zhu and Liu, 2018). Conceptualising social media posts as “augmentations” (Ballatore and De Sabbata, 2019) of places, the digital information is connected in many different ways (i.e., linking posts using "following-follower" networks, –see, Sadilek et al. (2012)). In recent years, graph neural networks have attracted increasing interests in the field of computer science, as one of the sub-disciplines of deep learning. The main reason behind such interest is because of the great expressive power on the graph-structure data (Zhou et al., 2018), which have provided powerful models that are potentially suitable for GIScience modelling on spatial interactions of places and understanding place representations. This section introduces different approaches to construct graphs using social media posts. The graphs are constructed not only based on the location information of each tweet but also using the temporal information to explore whether knowing the temporal element of social media posts would benefit my proposed framework to better understand users' activities.

The framework is illustrated in Figure 4.2. First, a stacked multi-modal autoencoder model is used to extract dense representations from both texts and images of tweets. Second, graph convolutional network (GCN) is applied based on the graph constructed with geo-coordinates from the social media posts to do the semi-supervised classification. That is, the relationship between features and labels is learnt throughout the process and updated based on the information that neighbours exchange with each other. As such, each node of the neural network learns locally, focusing on one social media post. The neural network node works towards understanding the relationship between content and assigned labels in a locally defined subset, taking into account that particular post and all its spatial neighbours. The knowledge acquired locally for each social media post at one layer is added to the information available for that post at the following layer. Note that the introduction for multi-modal autoencoder and GCN has been provided in Chapter 3.

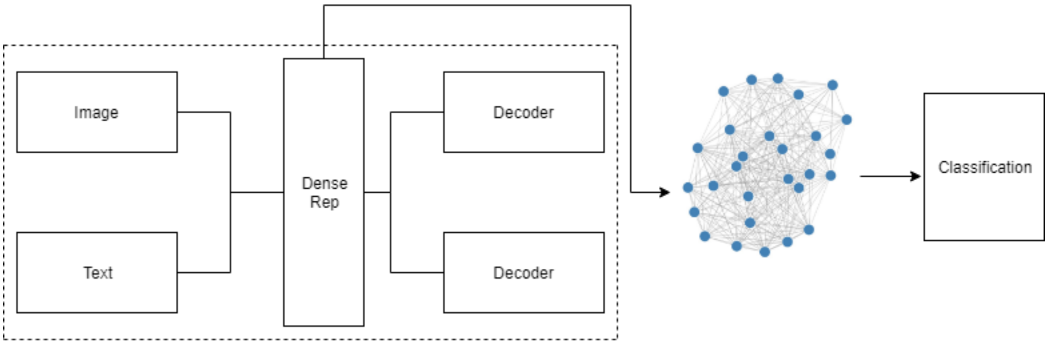


FIGURE 4.2: Methodology flowchart.

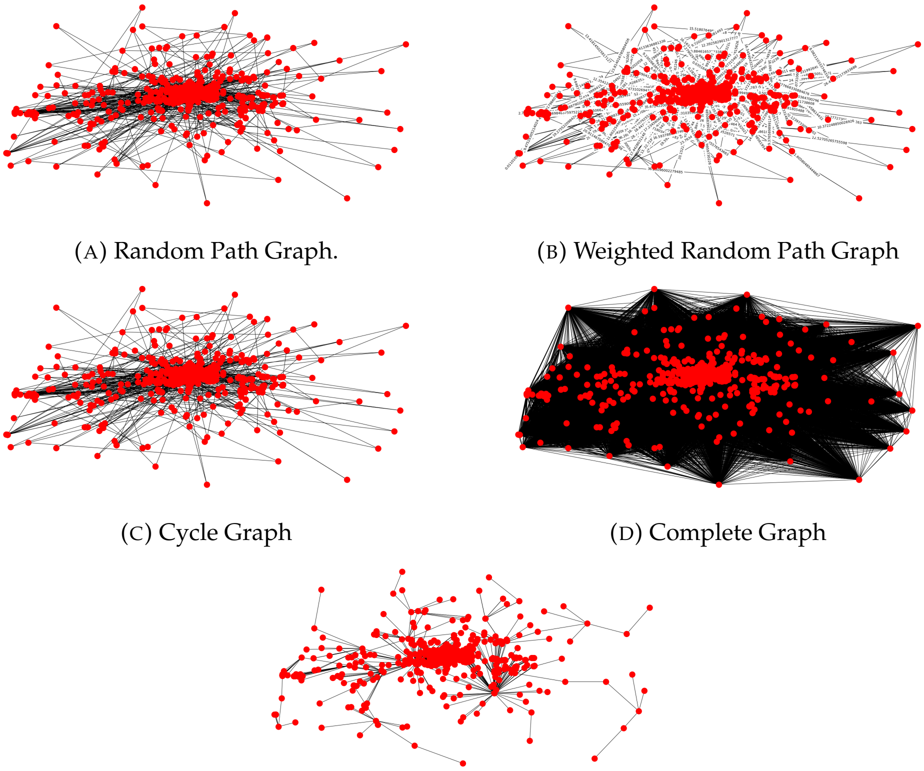
I then postulate that the local learning process described above should allow the neural network to take better advantage of spatial clusters of information as well as the temporal information. In turn, that approach should deliver better performance in understanding labels that are spatio-temporally clustered, as it is commonly the case in geolocated social media, which focus on spatio-temporal content. To the best of my knowledge, by the time the work in this chapter is published, it is the first to account for all four aspects (text, image, geolocation, and temporal information) using a deep learning approach to classify users' online activities.

4.3.1 Graph Construction

I tested a variety of graphs that were constructed using the tweets presented in the case study. I classified the graphs based on whether they account for the absolute positions of tweets and distances between the tweet pairs into three different categories: *a-spatial graphs*, *semi-spatial graphs* and *spatial graphs*.

A-spatial Graphs

A-spatial graphs do not take into account the absolute positions of the tweets and distances between the pairs of nodes in the graph. I tested three different *a-spatial graphs* in the experiments:



(A) Random Path Graph.

(B) Weighted Random Path Graph

(C) Cycle Graph

(D) Complete Graph

(E) Minimum Spanning Tree

FIGURE 4.3: Different spatial graph structures

- **Random Path Graph:** A path graph is a graph that can be drawn so that all of its vertices and edges lie on a single straight line (Gross and Yellen, 1999). I randomly assign tweets in a line so that they are linked to each other one by one, as shown in Figure 4.3(a). If two nodes are connected to each other $A_{ij} = 1$ in its adjacency matrix, otherwise $A_{ij} = 0$.

- **Random Cycle Graph:** A cycle graph is a graph containing a single cycle through all nodes shown in Figure 4.3(d). It is randomly generated in the same way as the random path graph, plus adding a link between the beginning and the end nodes.

- **Complete Graph:** A complete graph is a graph in which each pair of graph vertices is connected by an edge shown in Figure 4.3(c).

Note that although *Random Path Graph*, *Cycle Graph* and *Complete Graph* are connected in the form that each tweet is connected to another, such graphs do not take into account the absolute positions of the tweets and distances between the pairs of nodes in the graph. For example, *Random Path Graph* is constructed by connecting all nodes in the graph with a straight line, and it can start from any arbitrary node as long as all the nodes can lie on the same line by the end of the graph construction. Thus, the absolute positions of tweets are not useful in such a case. Therefore, such graphs can be seen as the nodes are connected without the spatial component, whereby the spatial locations of tweets have no impact in those graph construction processes.

Semi-spatial Graphs

Semi-spatial graphs do not take into account the absolute positions of the tweets but are constructed with the information of the distances between the pairs of nodes in the graph. Following the experiment in the previous section, I tested three different *semi-spatial graphs* in the experiments:

- **Weighted Random Path Graph:** Same structure as *Path Graph* shown in Figure 4.3(b), however, the weights for edges are defined by spatial interaction as:

$$A_{ij} = 1/(1 + distance) \tag{4.1}$$

  where *distance* denotes for the spatial distance between tweets. However, as two tweets may share same locations (i.e., the users posted them at the same locations with same coordinates), the spatial distance can be 0. Therefore, $1 + distance$ is to avoid such an error which may occur in the calculations. Because GPS-enabled smartphones are typically accurate to within a 4.9 metres range (Van Diggelen and Enge, 2015), the 1 meter added in this equation will unlikely raise further uncertainties in the calculations and results.

- **Weighted Random Circle Graph:** Same structure as the cycle graph, however, the weights for edges are defined by spatial interaction.

- **Weighted Complete Graph:** Same structure as the complete graph, but the weights for edges are defined by spatial interaction.

Spatial Graphs

Spatial graphs take into account the absolute positions of the tweets as well as the information of the distances between the pairs of nodes in the graph. I tested two *spatial graphs* listed below:

- **Minimum Spanning Tree (MST):** I first generate a series of graphs based on spatial adjacency using distances ranging from 2 kilometres to 15 kilometres. I then calculate the minimum spanning tree for each one of those graphs to further minimise the number of connections. In Figure 4.3 (e) is an example of a minimum spanning tree calculated starting from the 9 kilometres spatial adjacency. If two nodes are connected to each other $A_{ij} = 1$ in its adjacency matrix, otherwise $A_{ij} = 0$.

- **Weighted Minimum Spanning Tree (Weighted MST):** Same structure as minimum spanning tree, but the weights for edges are defined by the same spatial interaction defined in Equation (4.1).

4.3.2 Spatio-Temporal graph

The temporal component of social media post (Yang and Leskovec, 2011) is a key aspect to move beyond the simple geotag (Crampton et al., 2013). The temporal evolution of the social media trend has clear links to emerging events in the physical world (Wang et al., 2016a), which leave “data shadows” behind them (Shelton et al., 2014). Spatio-temporal analysis has been widely adopted in the study of digital geographies (Cheng and Wicks, 2014; Gomide et al., 2011; Lee et al., 2011), to identify sociospatial patterns of online events (Crampton et al., 2013; Luo et al., 2016), or to monitor and surveillance nature disasters (Wang et al., 2016b; Martín et al., 2017).

To explore the usefulness of the temporal component of social media posts in better understanding its relationship with the assigned labels, I tested two graphs based on two different spatio-temporal distances and a weighted graph with distance information on the edges.

- **Spatio-temporal neighbourhood (StN), Euclidean Distance:** To be consistent with spatial distance, I transform the time series information equivalent to the spatial distance, and I define such process as *temporal-spatial distance transformation*. That is, the temporal differences between social media posts are measured in a defined spatial distance (see more details in Section 4.4.2). I define the first spatio-temporal distance as:

  $$STDist = \sqrt{easting_{dist}^2 + northing_{dist}^2 + time\_difference_{dist}^2} \tag{4.2}$$

  where the distance is calculated using the British National Grid (Crossley, 1999); "easting" and "northing" denotes to the longitude and latitude of each post; $time\_difference_{dist}$ is the defined *temporal-spatial distance transformation*; for example, $time\_difference_{dist} = 1\ metre$ if the temporal difference between two posts is 12 hours. An example of the constructed graph using Euclidean distance is shown in Figure 4.4(a). I ran a series of experiments using different distances to equate time and space, and the results are presented in Section 4.4.

- **Spatio-temporal neighbourhood (StN), temporally-weighted Euclidean Distance:** Chang et al. (2007) defined a spatio-temporal similarity measure to compute spatio-temporal relevance between two trajectories of moving objects on road networks, which is known as spatio-temporal distance:

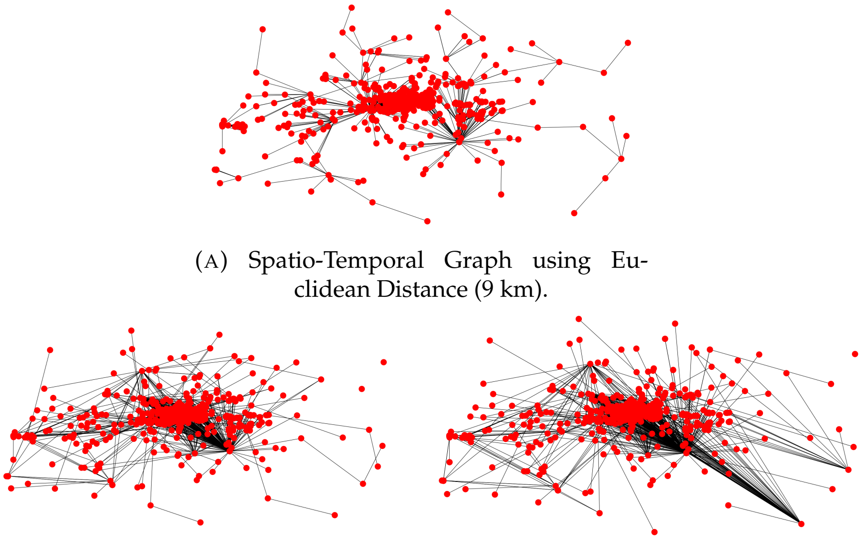
  $$STDist = (SD + \delta * TD)/2 \tag{4.3}$$

  where δ is the spatio-temporal weight; SD and TD denote to spatial distance and temporal distance, respectively. An example of using such distance can be seen in Figure 4.4(b). As each entity in my dataset represents a point in the space-time continuum, rather than a trajectory, I propose the following definition of the distance between two points into:

  $$STDist = \sqrt{SD^2/2 + (\delta * TD)^2/2} \tag{4.4}$$

  where SD is defined as $\sqrt{easting_{dist}^2 + northing_{dist}^2}$. It is a variation on the Euclidean distance, but taking into account of an additional spatial weight δ defined in formula (8) to define the impact of the temporal distance. In this paper, I keep δ as 20 same in Chang et al. (2007), and the results are presented in Section 4.4. Thus, I define such approach as *Temporal weighted Euclidean Distance*. An example of the constructed graph using temporal weighted Euclidean distance is shown in Figure 4.4(c).

- **Spatio-temporal neighbourhood (StN), distance and temporally weighted Graph:** Given the best results reported in Section 4.4 are achieved by using the graph defined by Equation (4.4), I define this graph same as the temporally weighted Euclidean Distance model, but the weights for edges are defined by



(A) Spatio-Temporal Graph using Euclidean Distance (9 km).

(B) Spatio-Temporal Graph using distance defined in Chang et al. (2007) (9km)

(C) Spatio-Temporal Graph using Temporal weighted Euclidean Distance (9 km).

FIGURE 4.4: Different spatio-temporal graph structures

spatial-temporal interaction as:

$$A_{ij} = 1/(1 + distance_{ST}) \tag{4.5}$$

where $distance_{ST}$ is the distance calculated by Equation (4.4).

4.3.3 Baseline methods

In order to test the capability of my proposed semi-supervised multimedia classification framework, I compare it with eight baselines developed from various methods focusing on text content and image content, as well as the spatial component of the tweets:

A-spatial Baselines with Text and Images

I set up baselines which employ a traditional machine learning algorithm and two neural network-based deep learning approaches to compare their performance with my proposed graph-based semi-supervised classification framework.

- **SVM**: I adopt a traditional machine learning approach Support Vector Machine (SVM) (Cortes and Vapnik, 1995) on the extracted representations from multi-modal autoencoder to classify tweets. Traditional machine learning methods such as SVM has a long-standing history being adopted for social media classification and spatial analysis within the field of geography (Guo and Chen, 2014; Qi et al., 2019). Although in recent years, deep learning methods have proved to outperform such a traditional machine learning method in various disciplines, SVM is still worth to be set up as a basic baseline in comparison with the proposed GCN framework due to its popularity within academic studies.
- **Dense Neural Network (DNN)**: I adopt a 3-layer dense neural network (DNN) on the extracted representations from multi-modal autoencoder to classify tweets.

Due to its strong ability of generalisation, DNN as one type of deep learning techniques has been widely adopted in various social media analytic studies (Ghani et al., 2019). Thus, the 3-layer DNN is chosen as another baseline.

- **Visual-textual Fused CNN (VTCNN)**: Inspired by Huang et al. (2018), I design an end-to-end deep learning framework using two stacked CNN to extract representations from images and text simultaneously, and concatenate them in the middle layer of the framework for twitter classification. Huang et al. (2018) can be seen as a direct comparison to my proposed framework, although such a method is primarily a supervised training framework which usually requires training data in a large size and to be well-labelled.

Note that the baseline VTCNN is the only end-to-end training framework among all the baselines. For other baselines introduced in this section, representation extraction (from images, text or both) and classification are two separated steps. It is also important to highlight that these three baselines do not take into account the locational information of the tweets, and they perform classification purely based on the multimedia content (images and text) of tweets. Thus, they are set up as the comparisons to my framework, which is performed on *a-spatial graphs* introduced in Chapter 4.3.1.

A-spatial Baseline with Text Only

Doc2Vec + Label Propagation: I use Doc2Vec (Le and Mikolov, 2014) to extract text representation and a traditional semi-supervised machine learning approach Label Propagation (LP) (Zhu and Ghahramani, 2002) to classify tweets. Such a method has been widely adopted on online content analysis, for example, sentiment analysis (Mishra et al., 2019; Wadawadagi and Pagi, 2020). As a semi-supervised machine learning approach, LP is used to assess the performance of the GCN framework, which is also a semi-supervised learning framework. Such a framework targets on classifying users' activities using their text content; thus, it can be used for demonstrating whether multimedia content analysis is superior to content analysis which only targeting on the text.

A-spatial Baseline with Images Only

CNN autoencoder + Label Propagation: I use a CNN autoencoder (Mao et al., 2016) which is the same structure as I adopted in the multi-modal autoencoder to extract image representation and use LP approach to classify tweets.

Spatial Baselines with Text Only

Doc2Vec + GCN (MST): I use Doc2Vec to extract text representation and GCN on a spatially constructed graph (MST) to classify tweets. This baseline is set up as a direct comparison to the previous baseline (**Doc2Vec + Label Propagation**).

LSTM autoencoder + GCN (MST): I use an LSTM autoencoder to extract text representation and GCN on a spatially constructed graph to classify tweets. Despite the fact that the previous two baselines are designed for the purpose of showing the classification results based on text, Doc2Vec is not considered as a deep learning approach to extract text representation. As mentioned in Section 3.3, my proposed multi-modal autoencoder contains an LSTM encoder extracting text representations from UGC content. Thus, I further design such a baseline as one of the comparisons to assess whether multimedia content analysis is superior to content analysis which only targeting on the text.

Spatial Baselines with Images Only

CNN autoencoder + GCN (MST): I use the CNN autoencoder to extract image representation, and GCN on a spatially constructed graph (MST) to classify tweets. This baseline and the previous baseline (**CNN autoencoder + Label Propagation**) are designed to assess whether multimedia content analysis is superior to content analysis which only targeting on images.

4.3.4 Model training

The text content of each tweet was pre-processed using tokenisation, stop words removal and case folding. The resulting text was then vectorised using Word2Vec, a publicly available word embeddings model trained with a 400 million Twitter dataset[2].

The image content of each tweet was converted to greyscale and re-sized them into 158 × 158 uniform size images. After the extraction of the encoded features using the encoder component of the autoencoder has been completed, I randomly selected a certain amount of tweets (as discussed below) from the dataset as training data for the GCN. The number of tweets in each category in the dataset is unbalanced as shown in Figure 4.1. Therefore, a completely random sample of data as a training dataset may lead to the consequence that some categories have no tweets included, especially for categories such as *Animals* which only have 8 tweets. To guarantee the model is trained on every category introduced in Section 4.2, I ensured that at least four tweets from each category were selected in the training sample. For the same reason, I evaluated the model based on both the classification accuracy and F1 score. The accuracy and F1 score are commonly used in deep learning and machine learning studies to measure the performance of the models. Accuracy is measured by a percentage of how many labels in the predefined dataset (manually labelled in Section 4.2) are correctly classified by the model; F1 score is a measure of a weighted average of the precision (the proportion of correct predictions among all predictions of a certain class) and recall (the proportion of examples of a certain class that have been predicted by the model as belonging to that class) of the model.

I trained a two-layer GCN model with 0.5 dropout rate for both layers, $L2$ regularisation factor for the first GCN layer and 8 as the number of hidden units. I trained the GCN model for a maximum of 3000 epochs (training iterations) using Adam (Kingma and Ba, 2014) with a learning rate of 0.01, and early stopping with a window size of 300, that is the model stop training if the validation loss does not decrease for 300 consecutive epochs. Trainable weights initialisation and feature vectors normalisation remain the same as in Kipf and Welling (2016a). My framework is designed in Keras (Chollet et al., 2015) with Tensorflow (Abadi et al., 2015) as backend and the training producer was performed using Nivida GPU Geforce GTX 1080 (NVIDIA et al., 2020).

Figure 4.5 illustrates the curve for the training accuracy by randomly sampling the number of tweets as the training data varies from 50 to 600, incremented by 50. The test is performed using the Random Path Graph structure, and it shows that the best number of samples is around 200. Once the size is over 200, there is a slight drop in the performance of the GCN, and the accuracy tends to be stable. The

[2] https://github.com/loretoparisi/word2vec-twitter

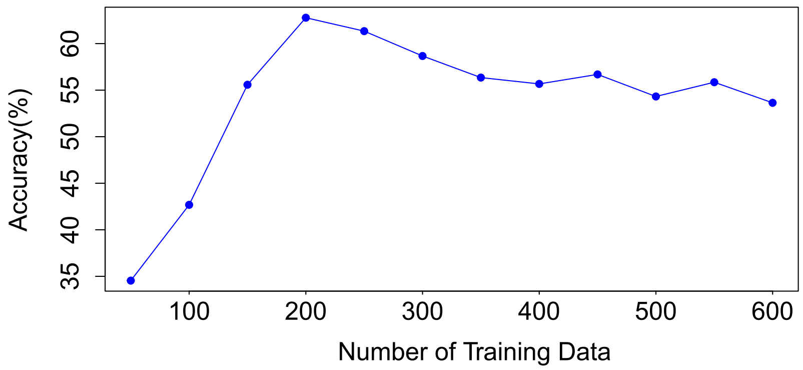


FIGURE 4.5: Variation of accuracy based on the number of training data.

reason behind why the performance drops after the number of training data is more than 200 is that the training data are becoming too unbalanced that the model could properly handle, whereas when the number of tweets fewer than 200, the model is struggling to perform the desired classifications without sufficient training samples. The result demonstrates that my framework can achieve a reasonably good result with only partially labelled data.

4.4 Results

4.4.1 Spatial graph

The experiments started with activity classification using GCN on the graph structures with no defined spatial interaction (see Chapter 4.3.1, Equation 4.1). As summarised in Table 4.1, the results reveal that my best GCN approach successfully categorises each tweet into its corresponding category based on partially labelled data with an accuracy of 72.57%. Figure 4.6 shows how the manually assigned labels compare to the model output, and it illustrates how most of the errors are due to some *Food* and most of *Nature* tweets being labelled as *Personal* by the model, many *Sports* tweets being labelled as *Places and attractions*, and most *Work* tweets being labelled as *Not informative*. Figure 4.7 is a visual comparison between corrected and incorrectly classified tweets. Both correctly and incorrectly predicted tweets are largely clustered in central London and they are not easily separable to each other. The incorrectly classified tweets seem not to follow any specific spatial patterns.

Those results are achieved on a training sample of 200 randomly selected tweets and despite a fairly imbalanced and noisy (contains a considerable amount of information that are challenging to interpret even for human researchers due to a variety of reasons as mentioned in Section 1.1) dataset. The GCN seems to perform better on a sparse – non necessarily simple, but less dense graph structure, as the best results are obtained with a graph structure constructed by creating a weighted minimum spanning tree using a 3 kilometres range, whereas the classification accuracy and F1 score on the two complete graphs are much lower compared to the other spatial graph structures. Clearly, choosing a suitable distance range for creating graph structure is essential within the framework. The results show that finding a geographic graph that has an appropriate density of connections within a reasonable distance range can significantly improve the performance of my graph-based semi-supervised framework.

Model input	Representation Extractor	Model	Accuracy	Micro-F1 Score
A-spatial with Images and Text	Multi-modal Autoencoder	SVM (no graph structure)	15.87%	9.13%
	Multi-modal Autoencoder	DNN (no graph structure)	11.20%	4.35%
	(VTCNN itself)	VTCNN (no graph structure)	16.00%	8.37%
	Multi-modal Autoencoder	GCN (Random Path Graph)	62.78%	56.87%
	Multi-modal Autoencoder	GCN (Cycle Graph)	68.63%	65.94%
	Multi-modal Autoencoder	GCN (Complete Graph)	23.75%	15.65%
A-spatial with Text	Doc2Vec	Label Propagation	18.31%	3.40%
A-spatial with Images	CNN autoencoder	Label Propagation	26.76%	4.20%
Spatial with Text	Doc2Vec	GCN (MST (3 km))	26.43%	24.32%
	LSTM autoencoder	GCN (MST (3 km))	36.66%	35.95%
Spatial with Images	CNN autoencoder	GCN (MST (3 km))	71.07%	70.51%
Semi-spatial with Images and text	Multi-modal Autoencoder	GCN (Weighted Random Path Graph)	65.34%	63.15%
	Multi-modal Autoencoder	GCN (Weighted Cycle Graph)	68.83%	67.67%
	Multi-modal Autoencoder	GCN (Weight Complete Graph)	23.66%	18.15%
Spatial with Images and text	Multi-modal Autoencoder	GCN (MST (2 km))	56.73%	51.89%
	Multi-modal Autoencoder	**GCN (MST (3 km))**	**72.57%**	**69.10%**
	Multi-modal Autoencoder	GCN (MST (5 km))	61.60%	57.83%
	Multi-modal Autoencoder	GCN (MST (8 km))	55.55%	52.24%
	Multi-modal Autoencoder	GCN (MST (10 km))	54.67%	48.67%
	Multi-modal Autoencoder	GCN (MST (15 km))	51.64%	47.25%
	Multi-modal Autoencoder	**GCN (Weighted MST (3 km))**	**73.57%**	**72.89%**
Spatio-temporal with Images and Text	Multi-modal Autoencoder	GCN (StN (Euclidean, 2 km))	67.33%	64.53%
	Multi-modal Autoencoder	GCN (StN (Euclidean, 3 km))	70.28%	68.45%
	Multi-modal Autoencoder	GCN (StN (Euclidean, 4 km))	69.58%	67.25%
	Multi-modal Autoencoder	GCN (StN (Euclidean, 5 km))	69.15%	66.73%
	Multi-modal Autoencoder	GCN (StN (as defined in Chang et al. (2007), 2 km))	63.24%	60.24%
	Multi-modal Autoencoder	GCN (StN (as defined in Chang et al. (2007), 3 km))	66.57%	63.83%
	Multi-modal Autoencoder	GCN (StN (as defined in Chang et al. (2007), 4 km))	69.89%	65.27%
	Multi-modal Autoencoder	GCN (StN (as defined in Chang et al. (2007), 5 km))	69.24%	67.51%
	Multi-modal Autoencoder	GCN (StN (temporally-weighted, 2 km))	69.58%	65.32%
	Multi-modal Autoencoder	GCN (StN (temporally-weighted, 3 km))	72.32%	69.68%
	Multi-modal Autoencoder	**GCN (StN (temporally-weighted, 4 km))**	**78.98%**	**76.72%**
	Multi-modal Autoencoder	GCN (StN (temporally-weighted, 5 km))	74.56%	71.41%
	Multi-modal Autoencoder	**GCN (StN (distance-temp.-weighted, 4 km))**	**80.08%**	**78.65%**

TABLE 4.1: Comparisons of different graph structures. (Best results achieved.)

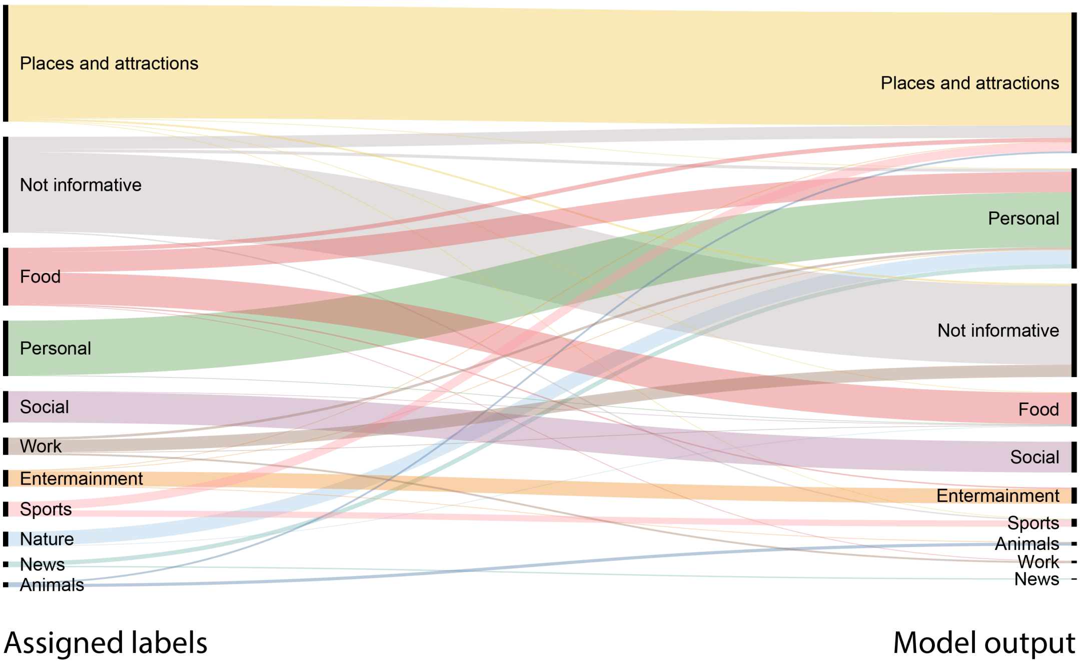


FIGURE 4.6: Comparing manually assigned labels and model (Minimum Spanning Tree, 3 km) output.

As shown in the table, GCN on the spatial graph constructed using Minimum Spanning Tree with 3 kilometres radius achieves the best results among other structures, I choose the same distance radius for constructing the weighted Minimum

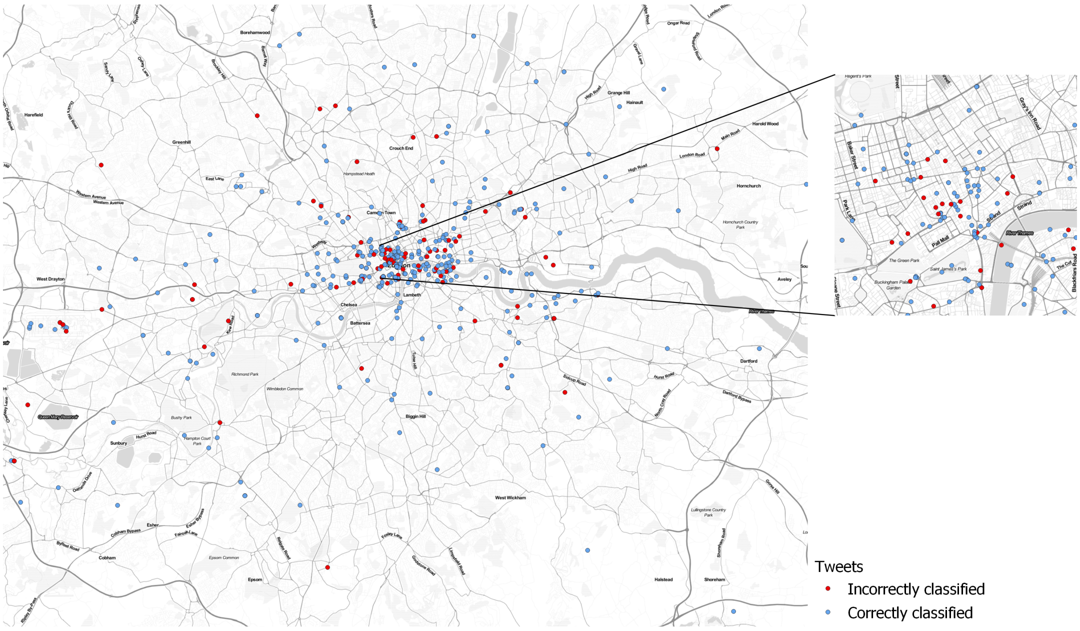


FIGURE 4.7: Visualisation between corrected and incorrectly classified tweets. Map tiles by Stamen Design, under CC BY 3.0. Data by OpenStreetMap, under ODbL.

Spanning Tree. The results show an even better accuracy of 73.57%, and it illustrates that knowing the local context (i.e., tweets posted nearby) can help my framework to better understand the content.

Based on whether the models account for the absolute positions of tweets and spatial distances between the pairs of nodes in the graph or not, I classified the graphs into three major categories as introduced in Section 4.3.1 and shown in Table 4.1: *a-spatial graphs*, *semi-spatial graphs* and *spatial graphs*. It is evident that the more abundant spatial information the model has, the higher performance the GCN can achieve. GCN on weighted Minimum Spanning Tree is clearly higher than the results achieved by GCN on the semi-spatial graphs (weighted Random Path Graph, weighted Cycle Graph and weighted Complete Graph), and the results achieved by GCN on the a-spatial graphs (Random Path Graph, Cycle Graph and Complete Graph).

It is also interesting to find that the results achieved by GCN are significantly higher than the traditional supervised learning method SVM, that performed the classification purely based on the extracted features from the stacked multi-modal autoencoder and their corresponding categories, with no geographical knowledge. As discussed above, the labelling process used in the case study does not take into account geographical locations of tweets, but only based on the textual and image content, that is the same information provided to the SVM. As shown in Table 4.1, my GCN framework outperforms this traditional supervised machine learning method.

Moreover, as evident by further experiments, my proposed framework outperforms the two deep learning methods DNN and VTCNN. It is important to highlight that these two frameworks were originally designed for supervised learning tasks with large and well-defined training data. In the context of my task, those two frameworks are inevitably overfitted during the training phrase with only 200 training data points, which is considered as a relatively small and noisy sample. However, the problem of insufficient training samples is not an issue for GCN framework.

These findings are particularly interesting from a geographical perspective. The GCN approaches on spatial graphs are clearly superior to the a-spatial and semi-spatial models, and the Minimum Spanning Tree approach, which encodes the geography of the tweets, is able to outperform the other approaches, which use random or complete graphs. As the labels used are not geographical per se and have not been assigned based on the tweet's location, this seems to indicate that the geographies of tweets can provide valuable insight into their content.

Additionally, following my experiment design in Section 4.3.3, Table 4.1 also summarises the results of the baseline methods which only adopt text content and image respectively rather than using combined representations for the classification. As GCN with spatial graph constructed using the minimum spanning tree with 3 kilometres as radius achieves reasonably well classification, I implement the same settings for GCN models in the baseline experiments. The result shows that the GCN model outperforms the traditional machine learning semi-supervised approach Label Propagation. Also, the classification solely relying on text content proves to be unreliable with comparably low accuracy. Furthermore, although the classification on image content achieves worse results compared to multimedia content, it produces a competitive classification output with relatively high accuracy and F1 score.

This is particularly interesting from a social science perspective, as it proves the evidence that visual content offers richer complementary information than what the accompanying text reveals (Borth et al., 2013), and the image content of tweets dominates human judgment at the labelling stage.

4.4.2 Spatio-Temporal graph

With the spatial and temporal information of tweets encoded in the graph, GCN achieves even better results. As mentioned in Section 4.3.2, to be consistent with the spatial distance, I transform the time series information equivalent to the spatial distance. In Table 4.1, I summarise the results of experiments on spatio-temporal graphs using $10\,meter = 12\,hours$. The choice of 12 hours can cover half day activities of users in a day, and it aims to better capture the temporal "localities" of different activities posted by the users. The topological structure using Minimum Spanning Tree based on temporal weighted Euclidean distance with a radius as 4 kilometres achieves the best results for both accuracy (78.98%) and F1 score (76.72%), which are significantly higher than the results achieved by GCN on the graphs merely with spatial information. A further performance improvement is obtained when applying the weighted minimum spanning tree using the same distance radius (80.08% accuracy and 78.65% F1 score). The performance is superior compared with the results achieved by spatial graphs discussed above.

The findings also illustrate that despite the variation of the graphs constructed using different types of spatio-temporal distance and distance radius, the results achieved prove to be rather stable with higher accuracy and F1 score compared with spatial graphs. These findings are interesting from spatio-temporal analysis perspective, as they illustrate that adding a temporal component of tweets can help the GCN model to produce a better semantic categorisation of their multimedia contents.

I also designed further experiments using different temporal-spatial distance transformations on the graphs, and explored their impact on classification accuracy. As shown in Table 4.1, the best results for graphs constructed using Spatio-Temporal

Transformations	Spatio-temporal Euclidean Distance	Temporal Weighted Euclidean Distance
1m = 12hr	60.53%	65.56%
10m = 6 hr	65.47%	72.08%
10m = 8 hr	68.23%	75.82%
10m = 12 hr	**70.28%**	**80.08%**
10m = 24 hr	68.85%	77.07%

TABLE 4.2: Comparisons between different temporal-spatial distance transformations. (Best results achieved.)

Euclidean Distance and Temporal Weighted Euclidean Distance are achieved with a radius equal to 3 and 4 kilometres. As such, I use 3 and 4 kilometres as default radius to construct graphs, respectively, for these two approaches. Table 4.2 shows the results obtained with different temporal-spatial distance transformations including 1 $meter = 12$ $hours$, 10 $meter = 6$ $hours$, 10 $meter = 8$ $hours$, 10 $meter = 12$ $hours$ and 10 $meter = 24$ $hours$. These tests allowed me to test different temporal "localities" and how they compare against spatial "localities" in capturing events and spatio-temporal patterns. The results indicate that 10 $meter = 12$ $hours$ performs best in the context of my dataset.

As shown in Table 4.1 and Table 4.2, GCN performs best when used in combination with my proposed spatio-temporal weighted distance. As I discussed in Section 4.3.3, the distance proposed by Chang et al. (2007) was originally devised for analysing trajectory data rather than social media posts. Further extensive research into such spatio-temporal modelling issues is clearly needed. Despite the fact that there is a wide literature focusing on spatio-temporal analysis of social media data, I argue my paper is the first which embeds spatio-temporal distance in deep learning approach to achieve semantic understanding on the content analysis. How to best model spatio-temporal distance in this context is an interesting research area that I hope to explore further in my future work.

4.4.3 Framework Robustness

As mentioned in Section 4.2, the dataset used for the experiments presented in this paper is noisier and more imbalanced than classic benchmarks used in traditional classification tasks. It is therefore important to explore the effects of such imbalances on the classification task, and evaluate the robustness of my framework against variations in training data.

Therefore, I designed an additional experiment using five different samples from my datasets. Each sample has at least four tweets for each category, but the proportion of tweets in the different categories is slightly adjusted. The experiment is conducted using the best performing approach in Table 4.1, that is a weighted graph constructed using the temporal weighted Euclidean distance with weighted minimum spanning tree (4 km).

The results are shown in Figure 4.8. Although the model performance is slightly affected by the variation in the sample, the classification results are reasonably consistent and stable. The results illustrate the robustness of my proposed framework on heavily imbalanced datasets such as "live" social media streams, and thus its relevance for applications in digital geographies.

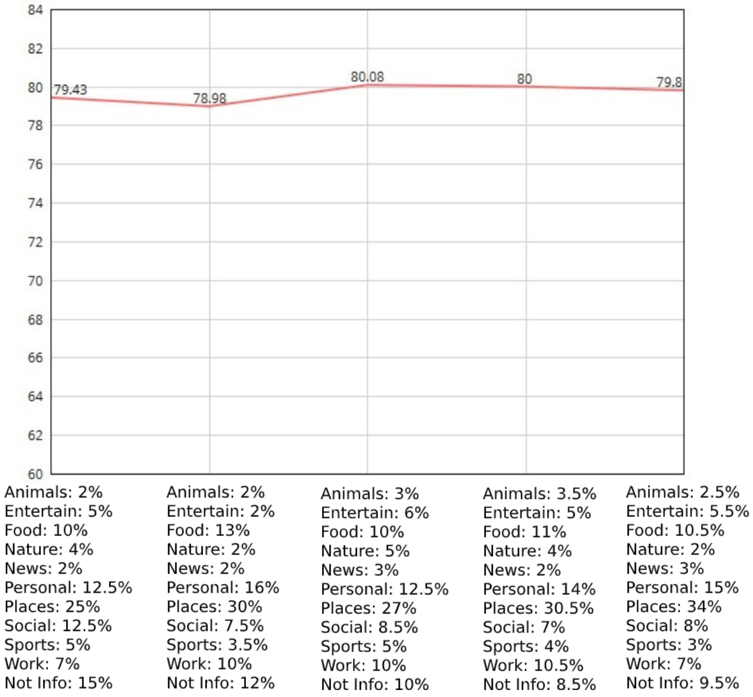


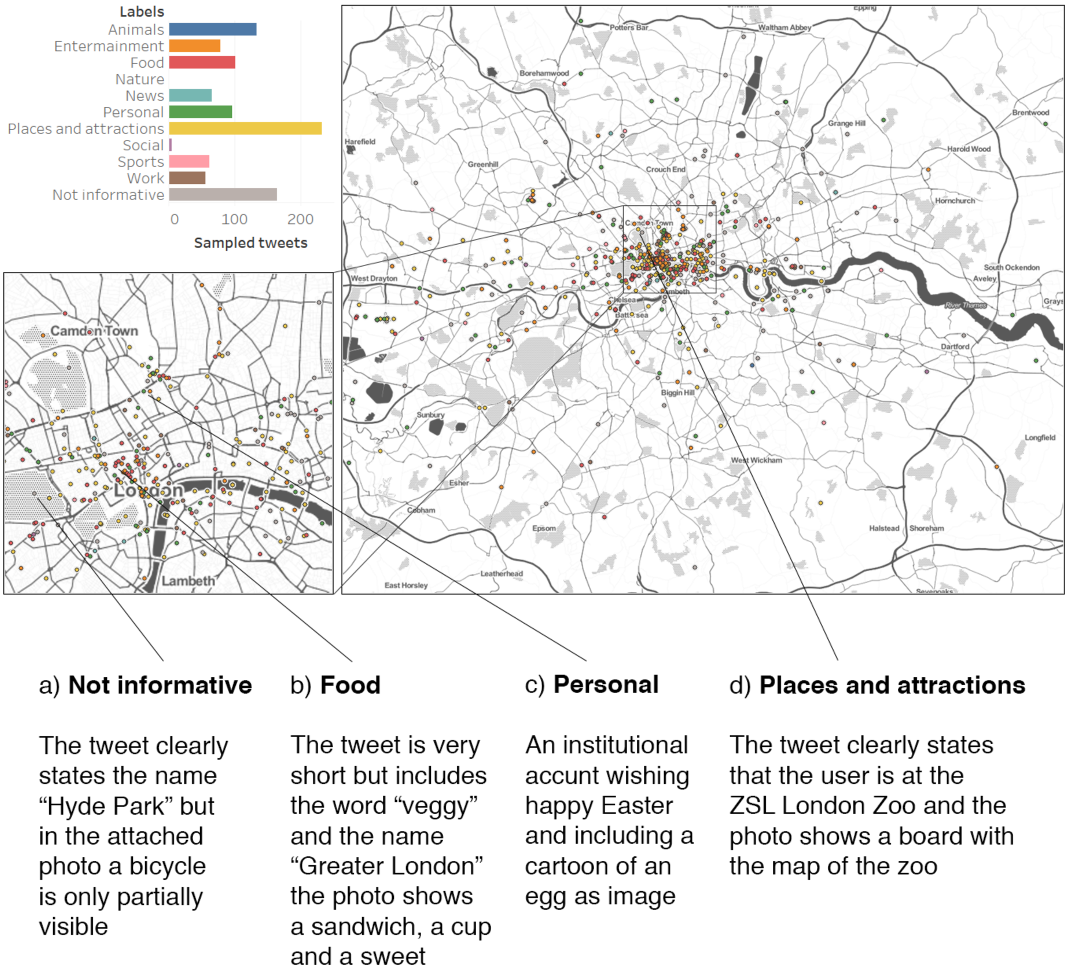
FIGURE 4.8: Performance comparisons on different training data.

4.4.4 Results showcase

Figure 4.9 illustrates how a model trained for the case study above can be used to classify a further sample of unlabeled data with the spatially constructed graph using the minimum spanning tree (3 km). As discussed above, the classification is noisy and, for instance, the tweet in Figure 4.9a is classified as Not informative, as the classifiers struggle to reconcile the location in a park, with a text that might indicate a focus on an attraction and the image of a bicycle. At the same time, the remaining three tweets showcased in Figure 4.9 seem to have been assigned a fairly accurate label among those I defined for the case study and considering that the aim of the tool is to allow users to defined their categories.

4.5 Discussion

In the sections above, I introduced a semi-supervised learning framework based on geographic adjacency networks to categorise social media posts based on their textual and visual content, as well as spatial and temporal aspects. The results demonstrate that taking into account the geography of each post is crucial to achieve a semantic understanding of the content and enable classification. In particular, while the labels used in the experiment were not assigned based on the location of social media posts, spatially-enabled classifiers performed better than a-spatial ones. The temporal component was also established as a key aspect in encapsulating the concept of place, and taking into account spatio-temporal relationships between social media post led to better classification. The results show that my framework can produce good classification results with partially labelled data, even on noisy and imbalanced data such as the one used for the case study presented above. Although I used Twitter as my case study, my framework has the flexibility to be extended to any other social media platform providing location-based services. As such, my



a) **Not informative**

The tweet clearly states the name “Hyde Park” but in the attached photo a bicycle is only partially visible

b) **Food**

The tweet is very short but includes the word “veggy” and the name “Greater London” the photo shows a sandwich, a cup and a sweet

c) **Personal**

An institutional account wishing happy Easter and including a cartoon of an egg as image

d) **Places and attractions**

The tweet clearly states that the user is at the ZSL London Zoo and the photo shows a board with the map of the zoo

FIGURE 4.9: Results of the prediction test on a further unlabeled sample. Map tiles by Stamen Design, under CC BY 3.0. Data by OpenStreetMap, under ODbL.

approach has the potential to be developed into a flexible tool for the study of digital geographies.

The majority of quantitative research on social media analysis in geography focuses on the text, whereas qualitative research maintains the importance of visual content (Ash et al., 2018b). As such, I based my work on the assumption that including the visual component of a post provides key information in understanding its content. To test that assumption, I designed a set of experiments to compare the classification resulting from including both text and images, only text and only images, using my GCN model, as well as the semi-supervised approach Label Propagation (Zhu and Ghahramani, 2002) as my baseline. The outcomes show that the best results are provided by GCN on weighted Minimum Spanning Tree (3 km), which takes into account the geographies of social media content (more on this below), as well as text and image. That indicates that including both the textual and media component improves the classification results compared to traditional text-based social media analysis, confirming my assumption above. These results are particularly important in a time where visual content such as images have become an integral and growing part of social media communication, as users shift from text-based posts to multimedia content (Weller et al., 2014). By taking advantage of recent developments in deep learning technologies, my research presented in this chapter is a first step towards bridging the gap between text-based quantitative analysis and visual methodologies in digital geographies.

To explore how to best encode the spatial information of social media posts in my model, I tested the effect of different graph structures on the performance of the GCN. I implemented my proposed GCN model on different structures, from semi-spatial graphs (e.g., weighted path graph and weighted complete graph) to complex structures taking different approaches to encode the geographies of posts as network links and distances. The results show that constructing a geographic graph taking into account the distances between posts and with an appropriate density of connections (e.g., Minimum Spanning Tree) can significantly improve the performance of my graph-based semi-supervised framework compared to random or complete graph structures. The performance of my model is clearly superior to the traditional machine learning approach SVM (Cortes and Vapnik, 1995), which does not take into account spatial graph structure, and classifies tweets based solely on the extracted feature representations. The comparison with the results obtained by GCN on three a-spatial graphs (i.e., random path graph, cycle graph, and complete graph) demonstrate that a graph-based deep neural network which takes into account the geographies of social media posts provides not only better classification results compared to traditional machine learning methods, but also better results compared to itself on the graphs with no geographies encoded in. Furthermore, the outcomes obtained by using different spatial graphs demonstrate that selecting an appropriate spatial (topological) structure can significantly improve the classification results.

The results ultimately highlight the importance of understanding social media content geographically. The geotag specifying the location in the space of a post is not merely a point, but it is an integral part of the augmentations that bring the place into being (Graham et al., 2015a). As such, taking into account the spatial relations between posts via the convolution of content through the spatial graphs allows me to go beyond the geotag (Crampton et al., 2013), and provides the GCN with key contextual information, that is crucial in the semantic understanding of social media content and thus the digital representations of the city (Ballatore and De Sabbata, 2019).

However, places do not merely exist in space, but they are "specific time-space configurations made up of the intersection of many encounters between 'actants' (people and things)" (Agnew and Livingstone, 2011, p. 325). My experiments indicate that the semantic categorisation of social media posts benefits significantly from including not only the spatial but also the temporal aspects of social media content. I experimented with graphs based on spatio-temporal distances which take the temporal element of tweets into account during the construction of graph. I proposed two distance calculation approaches, one based on a spatio-temporal Euclidean distance and one based on a temporal weighted Euclidean distance. The former simply considers the temporal element as a third, separate dimension, whereas the latter uses a mathematical weight to equate space and time, to control the impact of time on distance. These versions of the GCN thus take into account not merely the spatial neighbours of a tweet to understand the local context, but its spatio-temporal neighbours. The results show that taking into account the temporal component improves the quality of the categorisation and the stability of the model. The GCN model on the graph constructed using temporal weighted Euclidean distance also achieves the overall best results, which does not only illustrate the effectiveness of my distance calculation approach but also indicates that a social media analysis requires sophisticated modelling of the temporal element. The GCN seems to successfully capture the in-depth connections between similar events that might be spatially distant from each other but temporally close, and vice versa.

As such, a GCN on a well-defined spatio-temporal graph achieves better results through a deeper understanding of places as "time-space configurations" (Agnew and Livingstone, 2011, p. 325) and social media posts as "intersection of many encounters between 'actants'" (Agnew and Livingstone, 2011, p. 102), thus contextualising each post within its spatio-temporal neighbours. To the best of my knowledge, this is the first paper to embed a spatio-temporal distance into a deep learning approach to achieve a semantic understanding of social media content. While my approach in this paper has achieved reasonable performance, I suggest that further research is necessary regarding this aspect.

Finally, I tested the robustness of my framework and evaluated whether data variability (e.g., variations in the proportion of data for each category in training data) might affect the classification results. The experiments demonstrate that my framework is robust and can produce stable, consistent classifications. As such, I argue that my proposed framework has the potential to be developed into a powerful tool for the analysis of noisy and imbalanced social media datasets in digital geographies.

4.6 Summary

In this chapter, I proposed a new GeoAI tool that aims at bridging qualitative and quantitative approaches to understand place representations, which can learn a set of arbitrary labels from a small, manually created sample of geo-located social media posts and apply the same labels on a larger set, based on textual and image content, as well as the geographical and temporal aspects of the posts. The findings in this chapter provide evidence that analysing spatial information of social media posts is crucial for the semantic understanding and classification of their content. Temporal analysis is another important aspect in digital geography studies and in this research, the "space-time" relation (Thrift, 1983) can contribute to a better graph representation of social media distribution, and benefit the understanding of the content that users produced and posted online, and eventually help researchers to understand the digital representation of the places. (Pereira et al., 2013).

However, as mentioned in Chapter 1, the understanding of the role played by UGC in place representations has been so far limited by the fact that only a small percentage of social media posts are precisely geolocated Sloan and Morgan, 2015. Social media platforms such as Twitter are increasingly moving towards using platform-specific POIs for location information rather than allowing users to use precise geo-coordinates. Regardless of the reasons behind why Twitter is adopting the policy to restrict the use of precise geo-tagging services for users, it shows a trend of increasing difficulty in collecting UGC with geo-coordinates, even for the academic research purpose (Hu and Wang, 2020). Such a policy has created a significant challenge of modelling place with UGC, where the connection between physical space and the interpretation of users' lived experiences through the content analysis can no longer be easily observed. Aiming to explore this issue, in the next chapter, I will introduce my proposed framework that aims to estimate the locations of social media content, which benefits the understanding of digitally coded spaces and online place representations.

Chapter 5

Location Estimation of Social Media Content through a Graph-based Link Prediction

Part of this work presented in this chapter has been published as:

- **Pengyuan Liu and Stefano De Sabbata**, *2019*. Location Estimation of Social Media Content through a Graph-based Link Prediction. In *13th Workshop on Geographic Information Retrieval*.

The extended journal version of this paper is going to submitted to the Journal of Spatial Information Science [1].

5.1 Introduction

The majority of digital platforms such as Twitter, Instagram, Flickr provide services that allow users to explicitly attach location information to their posts. Based on the concept that users of social media can be considered as sensors of places (Goodchild, 2007), research in GIScience has thereby focused on the emergence of places from the lived experience of people in space through the analysis of how people live in places based on their content production. Goodchild (2011) discusses the idea of formalising place in the digital world. He addresses the relationship between the informal world of human discourse and the formal world of digitally represented geography where *place* stands at the centre of such platial studies within GIScience. Such an academic idea later encourages a wide range of studies towards embedding the digitalised human dynamics and their interaction (i.e., emotions, sentiments, place descriptions) with space into GIScience research, such as place-based GIScience (Gao et al., 2013) and pace–place (splatial) GIScience framework (Shaw and Sui, 2020). Scholars accept the spatial distribution of UGC within digital geographies and GIScience as a valuable resource to advance research on specific urban aspects (Anselin and Williams, 2016; Shelton et al., 2015; Arribas-Bel et al., 2015). The representation and interpretation of data retrieved from social media provide means to assess different urban dynamics and create socio-demographics of the cities. In turn, such information enables us to analyse everyday spatial processes and to gain knowledge about places, especially with respect to collective human dynamics (Steiger et al., 2016).

In the previous chapter, I introduced a framework to classify UGC based on their images, text, as well as geographical and temporal information on digital platforms.

[1] Code to reproduce my experiments is available at: `https://github.com/PengyuanLiu1993/PhD_Thesis_Codes_PengyuanLiu/tree/master/GAE_Location_Estimation`

However, despite the availability of geotagging technology that allows users to share location information online, research in digital geographies has been limited by the fact that only a small percentage of social media posts are geolocated explicitly. Understanding spatial activities of precisely geotagged social media posts can only portrait limited number of users' activities regarding the use of space. Although precise geo-coordinates are no longer easy to access (Hu and Wang, 2020), Twitter remains the ability for users to geotag their content using *Twitter Place* (in form of Twitter's pre-defined bounding boxes, see the introduction in Section 3.1), studies dealing with the data aggregated in a given size of "district" (in the form of bounding boxes) other than geolocations are facing the issues related to the Modifiable Areal Unit Problem (MAUP) during the geographical analysis (Wong, 2004). Although this issue is beyond the scope of this thesis, it sets forth future research directions when analysing uncertainties and bias related to UGC studies. Further discussions and research vision will be provided in Chapter 7.

This study is akin and complementary to the work aimed at estimating the location of geotagged social media content using classic geographic information retrieval approaches. The study is based on an understanding of the geographical concept of *place*. *Place* (see Section 2.1) is a term in geography which refers to the "locales" where human's everyday activities take place (Agnew and Livingstone, 2011), the following assumptions and two hypothesises are proposed in this chapter:

> ***Assumption***: users on social media platforms tend to geolocate content with tags referring to places where the content belongs to. In other words, content shared about a place will reflect the use of space and the activities carried out there, and the same types of activities are more likely distributed in similar places.
>
> ***Hypothesis 1***: the semantic content of social media posts (i.e., categorisation of users' activity types) can aid the process of estimating the location associated with a social media post.
>
> ***Hypothesis 2***: the spatio-temporal patterns of social media posts can provide an insight into the location estimation process.

This chapter investigates two different topological structures to estimate the location of UGC: topological modelling which is used for *geolocated tweets* (with precise coordinates), and hierarchical modelling for *placed tweets* (where a place is represented using a bounding box).

One of the novelties of this research is that it models the spatial structure of UGC using bounding boxes hierarchically. As discussed above, after Twitter's shift away from geolocated tweets in favour of places and POIs, bounding boxes are becoming a more prominent approach to encoding location in UGC. The bounding boxes not only function as geographical containers which contain social media posts located in (physical) spaces, but they are also associated with social practices illustrated through UGC. As such, the bounding boxes in this study can be seen as representing actual *places* that capture the "localities" of users' everyday activities and augment spatial experiences (Elwood and Leszczynski, 2013), and shape the representations of urban structures at different scales.

As illustrated in Section 2.6, existing studies on location estimation focusing on developing or applying GIR methods (e.g., geoparsing on placenames) on the text content of social media posts (Huang and Carley, 2017; Wallgrün et al., 2018; Wang et al., 2020b). However, such studies are limited when placenames are unclear, missing or vernacular in the text content. Although some studies have explored the use of visual content (e.g., Google Street View) to identify locations (Suresh et al., 2018; Sun et al., 2018), those studies are task-specific (e.g., location estimation based on architecture imagery (Doersch et al., 2015)) and not be generalisable to social media studies where images are more diverse and can be posted without location-indicative objects in it. Thus, it remains as a challenging task when estimating the location for a geotagged social media post which has no location-indicative placenames or objects in its text or image content. Existing research has identified that the location of a social media post strongly relates to its user's personal interests and activities (Chen et al., 2013); thus, it is worth to explore if the use of labels of the activities can be a useful tool for location estimation task. In Chapter 4, I proposed a framework to quantitatively label the activity types of users' multimedia content. Following up the study presented in Chapter 4, I propose an approach to estimate geolocations of tweets based on a semantic understanding of tweets' content (i.e., activity types labelled by the framework in the previous chapter) and their spatio-topological structures, which can be useful in the scenario where no placenames can be found in the text.

It is important to notice that my framework is not intended to predict the exact locations of the content. Instead, it provides an approximate area of where the content might be referred to. Despite being an approximation, this approach is still relevant in the context of GIScience and digital geographies studies, for two reasons. First, from GIScience and digital geographies perspectives, the quantitative analysis and summarisation to study the emergence of place from space through content production (Graham et al., 2015a) is considered as a necessary step in both disciplines. Such a step often takes into account the amount of similar UGC produced by users about a geographical area in general rather than focusing on each individual content. Thus, the precise location of the content can be superfluous as long as the framework can provide the estimation of the content's location in a relatively small geographic area. Second, as discussed above, although there is research in the disciplines of geographic information retrieval and location-based service targeting at predicting the precise locations of the content (Moncla et al., 2014a; Memon et al., 2015), such methods require more explicit locational information in the content which can be extracted and modelled (e.g., personal travel history, placenames in the text content or hashtags). However, only 10% tweets include the reference to the location in their text (MacEachren et al., 2011), the approaches mentioned above potentially are not applicable to a large number of UGC regarding social practices of users with no preliminary information of explicit placenames. The framework presented in this chapter can be seen as complementary to the GIR-based approaches, it aims at providing approximate areas of the content where they are referring to with the semantic understanding of the content (i.e., the labels) and their spatio-temporal structures. The results indicate that my proposed framework can produce reasonable estimations using information from manually assigned labels and the geographic information of social media at the urban scale, and can further help with our understanding on the place representations based on the spatial distribution of social media posts.

The novelties of this approach are:

- this approach is akin and complementary to the existing GIR-based text analysis, and it is specifically to address the issues of no placenames are explicitly existed in the text content;

- this chapter discusses a graph-based framework for location estimation based on the semantic understanding of the content (i.e., the labels) and their spatio-temporal topological structures;

- this chapter presents a novel approach to modelling place-related information attached to UCG in the form of bounding boxes.

5.2 Data Labelling

This section introduces the data I adopt as case studies in this chapter. The case studies consist of geolocated tweets and placed tweets.

5.2.1 Geolocated Tweets

Similar to Chapter 4, the case study proposed in this section analyses a specific set of topics (e.g., posts about food, entertainment, or sports) geographically in London, and I am interested in how geolocated social media posts reflect the digital representations (Graham et al., 2013b) of the city. For the scope of this Chapter, I sample 1200 tweets from the dataset introduced in Section 3.1 in two randomly selected subsequent months between July 1st and August 31st, 2018, and sort the tweets by the time when they were posted.

Similar to Chapter 4, because the labelling process of the Twitter data is challenging and time-consuming, I set myself a limit of seven working days to manually label a set of tweets randomly sampled from the dataset introduced in Section 3.1, thus reaching a total of 1200 tweets manually labeled into 7 different categories: *Food*, *Transportation*, *Places and attractions*, *Sports*, *Social*, *Personal* and *Not informative* shown in Figure 5.1. The labelling process is the same as the process used in Chapter 4, where tweets have been labelled based on their text and attached images, rather than their locations or geographic content explicitly. *Not informative* includes content related to advertisement and other content which is difficult to interpret. The category *Personal* includes content related to daily activities (e.g., shopping or selfies), whereas tweets in the category *Social* are related to parties and other social activities. *Transportation* includes tweets regarding public transportation (e.g., underground or airport).

As discussed above, the labels of the tweets are not based on their locations or geographic content. As in Chapter 4, the labelling process used in the case study does not take into account geographical location of tweets. Thus, although the predefined labels are relatively generic, they are still very subjective and reflect the interests and understanding of the authors. Other authors might have the preference to distinguish a diverse set of *Transportation* (ways of users using public transportation) (e.g., train stations or bus stops), or create a separate category for airports. Nonetheless, this is not an issue in the scope of this thesis.

As discussed in the previous section, this study is akin and complementary to the existing GIR-based location estimation or identification, where there are no explicit placenames in the text that can be identified. I aim to use the semantic understanding of their content to explore the digital coded space and place representations, and to study how such representations associate with the spatial clustering of content through space and time. The objective of this chapter is to provide a tool to estimate the geolocations of new posts based on the previous content production in a place and the spatial structures of the posts.

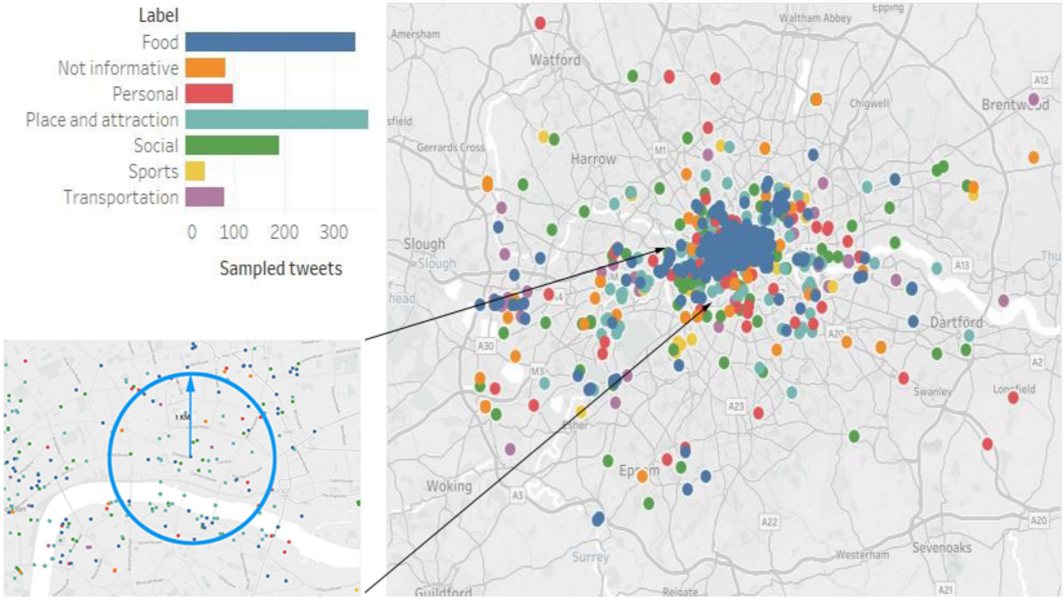


FIGURE 5.1: Geolocated tweets. Map tiles by Stamen Design, under CC BY 3.0. Data by OpenStreetMap, under ODbL.

5.2.2 Placed Tweets

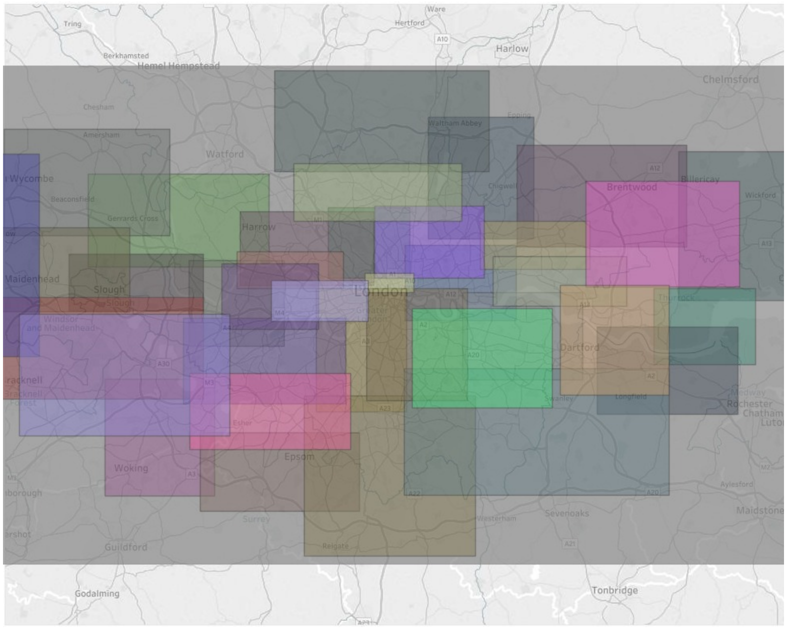


FIGURE 5.2: Placed tweets. Map tiles by Stamen Design, under CC BY 3.0. Data by OpenStreetMap, under ODbL.

The removal of the possibility of precise geolocation on Twitter has caused an increased awareness of the growing difficulty understanding place representations through geolocated users' posts (Hu and Wang, 2020). As discussed in the previous section, *Twitter Places* in forms of bounding boxes representing places have become an option which provides approximate areas indicating the locations of content. Thus, the dataset discussed above is adopted in this case study to test novel ways of modelling spatio-temporal structures using bounding boxes and estimating the place where the content of a tweet relates to.

I randomly sampled an additional 824 tweets in the second half of July between 17th July 2018 and 31st July 2018 to create a similar case study as described in Section 5.2.1. The tweets are *placed tweets* which have bounding boxes to indicate the approximate areas where the tweets were posted. The sizes of bounding boxes are defined by Twitter (see Section 3.1.1).The distribution of the tweets is shown in Figure 5.2. There are 78 bounding boxes in various sizes, and the largest bounding box is the *admin* level of London and all other bounding boxes inside it represent inner boroughs of London, smaller areas in London or points of interests (e.g., shops such as *Tesco Stores*, tourist attractions such as *Queen Elizabeth Olympic Park*).

5.3 Methodology

My proposed methodology estimates the locations of UGC through a link prediction algorithm testing a diverse set of spatio-topological structures. I investigate the variational graph autoencoder introduced in Section 3.4.2 for link prediction tasks on two types of graph constructions: the first one is a series of distance-based graph structures, which will be detailed in Section 5.3.1; the second one is a series of hierarchical graph structure, which will be explained in Section 5.3.2. In this section, I also propose a set of ranking schemes that will be used to assess the performance of the two approaches. Each of the two datasets used to construct topological structures is split into a graph construction set and a prediction set. The graph construction set is used for constructing the spatio-topological structures, whereas the variational graph autoencoder uses the knowledge learned from the graph construction set to make predictions for the prediction set.

5.3.1 Topological Structure Construction

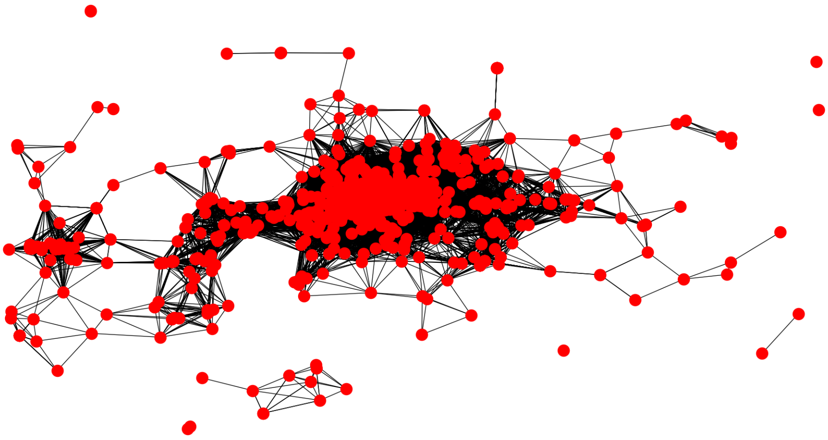


FIGURE 5.3: Topological structure example.

Similar to the spatial graph construction process introduced in Section 4.3.1, I construct the distance-based topological structure based on a spatial adjacency (O'Sullivan and Unwin, 2010). The graph is constructed loading a tweet after a tweet in historical order based on the time when each tweet was posted. As such, the algorithm learns the historical distribution of the geographies of the dataset, and this is used for the estimation of locations for the new tweets. An edge is created between nodes (represent the tweets) that are within a predefined distance between one another. A set of distances is tested, from 50 meters to 1000 meters (defined as *Dist*). Figure 5.3 shows an example of the constructed topological structure. In the spatial adjacency matrix, $A_{ij} = 1$ if two nodes are connected to each other, and $A_{ij} = 0$ otherwise.

When estimating the location of a new post, with no preliminary assumptions about its location, I first randomly connect the node which represents the new post to the graph and then using the variational graph autoencoder (VGAE) introduced in Chapter 3 as a link prediction approach to predict its most likely connected node in the graph to estimate the proximate area where the new post belongs to.

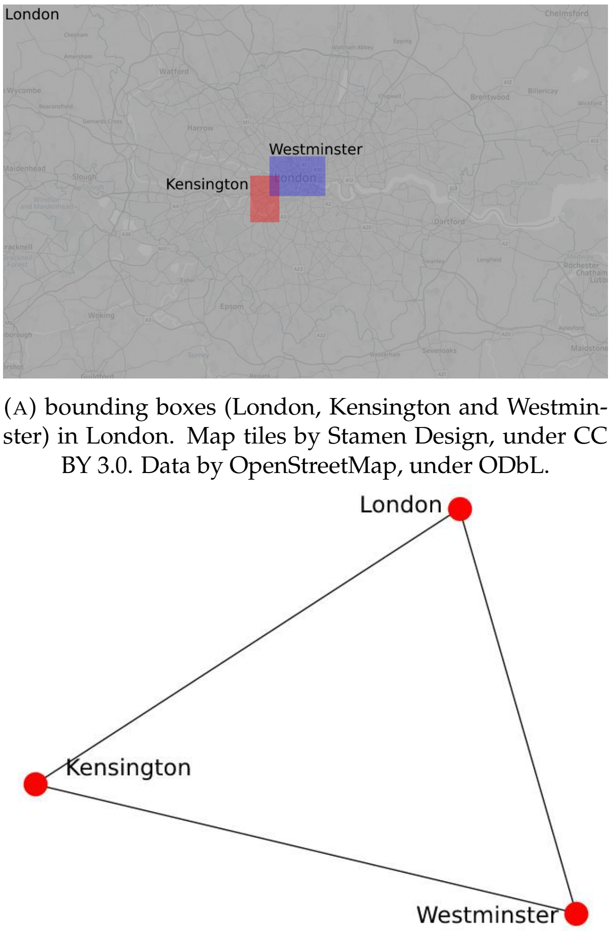
5.3.2 Hierarchical Modelling

Social media platforms are incredibly valuable as it enables scientific research to establish the connections between geographies and contextual data. As mentioned in the previous sections, a larger number of geotagged social media posts are geotagged with *Twitter Places* (in the form of bounding boxes) in their meta-data which indicate the approximate areas where the posts are generated. As discussed above, Twitter has decided to remove the ability to allow users to geolocate the content of their posts using precise geolocation, and future research will have to focus on the exploration of places represented by bounding boxes. The bounding boxes not only function as geographical containers which contain social media posts located in spaces but also can be seen as "locales" which associates with the social practices of users and help scholars to explore place representations in space.

To address the problem, I design three different hierarchical modellings of places using bounding boxes: a general hierarchical modelling (*VGAE hierarchical structure*), and two tree structures (*VGAE tree structure* and *VGAE dense tree structure*).

Hierarchical Structure

As introduced in Section 3.1.1, the types of bounding box in the Twitter data structure are organised in a hierarchical level where the granularity of geotags must be one of the five types: *poi*, *neighborhood*, *city*, *admin* or *country*. The hierarchical graph (*VGAE hierarchical structure*) is constructed as follows: if the bounding boxes of two different tweets intersect, an edge is added between them. A simple example of the graph construction is provided in Figure 5.4 where the node which represents the *admin* level of London is linked to the other two nodes which represent Kensington and Westminster respectively; there is also an edge between Kensington and Westminster as their bounding boxes intersect. The graph constructed for the dataset used in this case study is shown in Figure 5.5 (A), where the nodes which are distributed in the middle of the graph are the tweets with *Twitter Place* representing *admin*, they represent bounding box of the Greater London. As the latter contains all other smaller areas so that they are connected to all the other nodes in the graph. Other clusters of nodes are tweets with *Twitter Place* types as *city* or lower levels, and the graph clearly illustrates how tweets are distributed within different inner boroughs of Greater London. Note that in this hierarchical modelling, each node in the graph represents a bounding box of a tweet, the same as in the graphs discussed in the previous chapter and section.



(A) bounding boxes (London, Kensington and Westminster) in London. Map tiles by Stamen Design, under CC BY 3.0. Data by OpenStreetMap, under ODbL.

(B) A simple graph constructed based on three bounding boxes of their tweets.

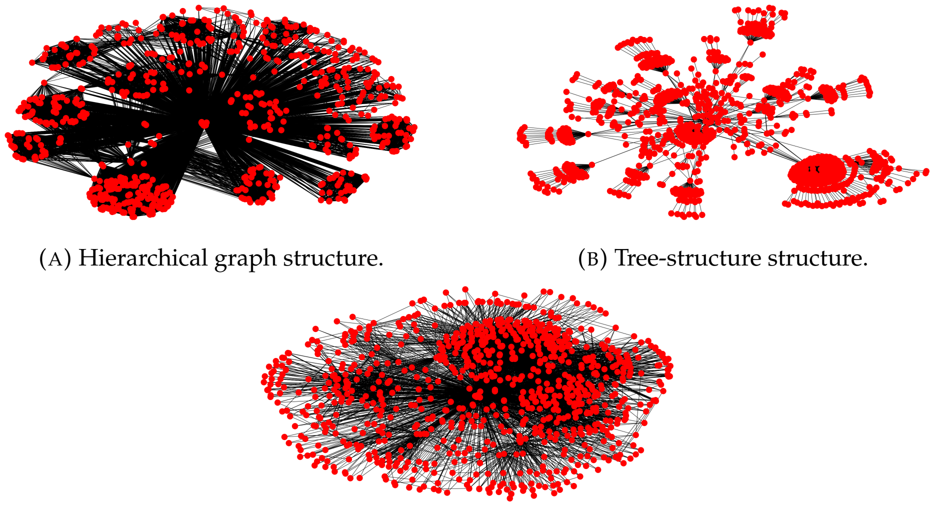
FIGURE 5.4: An example of the graph construction for the *VGAE hierarchical structure*.

Tree-structure Modeling

Inspired by the fact that tree structures have been heavily adopted to manage and structure spatial data within GIScience (Ooi, 1987), I construct a tree-structure graph to model the bounding boxes of tweets using the following steps:

1. create a graph with one node per tweet but no edges.
2. retrieve all the bounding boxes from the granularity level (Hollenstein and Purves, 2010) of Greater London to the level of points of interests.
3. introduce additional nodes representing the places as retrieved in step (1), rather than actual tweets.
4. add an edge between each actual tweet and the additional node representing the bounding box associated with the place of the tweet when their bounding boxes are the same.

An example of such a tree structure modelling is provided in Figure 5.6, I first model the nodes which represent bounding boxes as a hierarchical tree structure. Note that at this step, those nodes are simply representing the hierarchical level of bounding



(A) Hierarchical graph structure.

(B) Tree-structure structure.

(C) dense Tree-structure graph.

FIGURE 5.5: Hierarchical modelling.

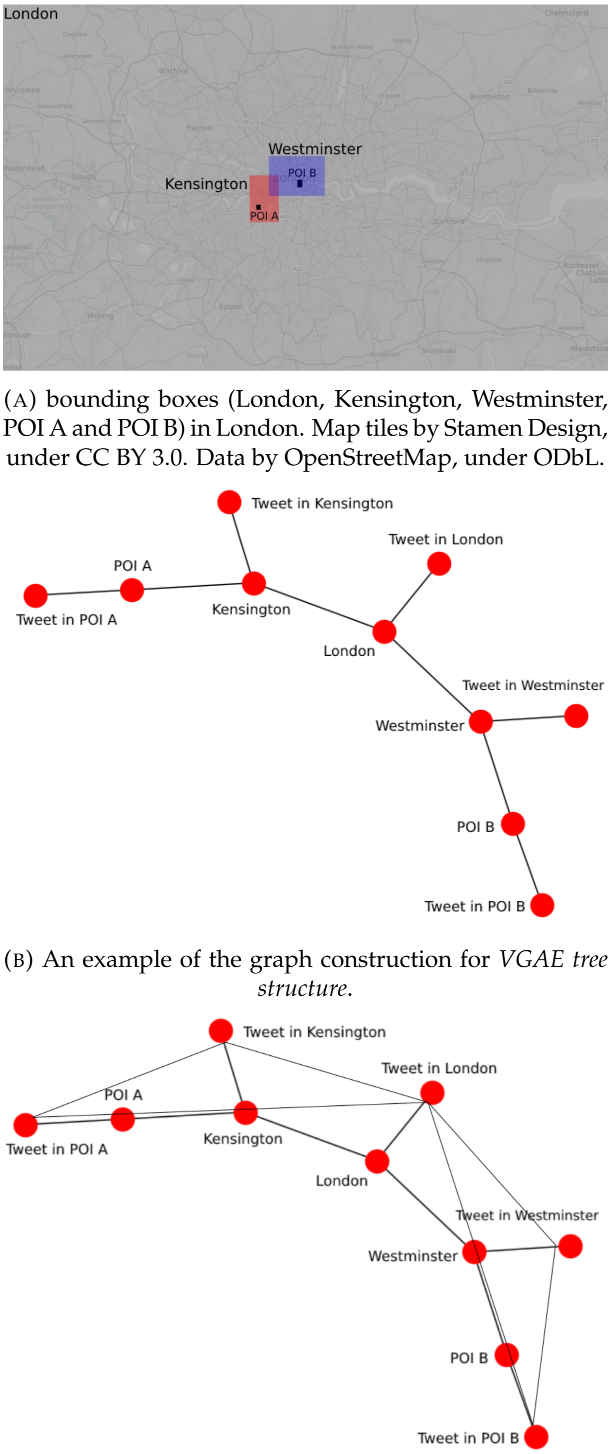
boxes (not the tweets), and they are the *additional nodes* introduced at the step (2). Secondly, I add links between those nodes and the nodes which represent tweets which have the same bounding boxes at each of the hierarchy. The constructed tree structure graph is shown in Figure 5.5 (B), and I consider such modelling of bounding boxes as the baseline for the tree-structure graph (*VGAE tree structure*).

Dense Tree-structure Modeling

As discussed above, *Twitter Places* are organised in a hierarchical level, where one bounding box may contain several smaller bounding boxes representing different granularity levels. Starting from that hierarchical structure, I expand the tree modelling into a dense structure, where a link is added between one bounding box and its higher granularity level of the bounding box. I construct the graph following similar steps :

1. create a graph with one node per tweet but no edges.
2. retrieve bounding boxes from the granularity level of Greater London to the level of points of interests.
3. introduce additional nodes representing those bounding boxes as retrieved in step (1), rather than actual tweets.
4. add an edge between each actual tweet and the additional node representing the bounding box associated with the place of the tweet when their bounding boxes are the same or overlapped.

An example of the graph construction can be seen in Figure 5.6 (C) which has a similar structure as in Figure 5.6(B) but more links are presented between the nodes which represent tweets. The major difference between *VGAE tree structure baseline* and *VGAE dense tree structure* is at the step (3). Instead of simply adding links between the additional nodes and the nodes which represent tweets which have the same bounding boxes at each of the hierarchy, *VGAE dense tree structure* also has links that across the hierarchy. That is, if two bounding boxes are overlapped, an edge is added. The constructed tree structure graph is shown in Figure 5.5 (C).



(A) bounding boxes (London, Kensington, Westminster, POI A and POI B) in London. Map tiles by Stamen Design, under CC BY 3.0. Data by OpenStreetMap, under ODbL.

(B) An example of the graph construction for *VGAE tree structure*.

(C) An example of the graph construction for the *VGAE dense tree structure*.

FIGURE 5.6: Tree-structure modellings.

5.3.3 Evaluation Schemes

Location Ranking Schemes for Topological Modellings:

As discussed in Section 5.1, my framework is not intended to predict the exact coordinates of the content. Instead, it provides approximate areas where the content might be located. With no preliminary information of that explicit locational information, this framework targets to only provide approximate areas of the content where they are generated from with the semantic understanding of the content (i.e., the labels) and their topological structures.

As a link prediction algorithm, when VGAE conducts location estimation for a new node (representing a tweet), it generates the likelihoods (possibility) of how likely it is for a link (or edge) between the new node and all other nodes in the graph to existing. I define the predicted location as the location of the node in the graph that has the highest possibility of linking with the new node (or *Top-1* prediction). The framework considers an estimation is correct when the distance between an actual location of a post in the test dataset and the location of its *Top-1* prediction suggested by the VGAE are in a predefined distance (*Range*). Thus, the first location estimation ranking scheme is defined as:

- Top-*k* Accuracy: for each new tweet that has a location to be estimated, the algorithm will produce the possibilities for the links between this new tweet and every existing tweets which are already in the graph. Top-K evaluation is a widely used approach in Link Prediction tasks within Computer Science (Yang et al., 2015). I select a couple of links ($k = 1, 3, 5, 10$) that have the highest probabilities produced by the link prediction algorithm. Then, I evaluate the accuracy by investigating whether the distances between (most likely connected) existing tweets (in the graph) that are associated with the chosen links and the new tweet are within a certain distance (*Range*). If they are within a defined distance, then I mark the location estimation is correct, otherwise, it's incorrect.

As introduced in Chapter 3, VGAE can perform link prediction with or without taking into account the labels of the nodes. To analyse the level at which the semantics of the tweet content (i.e., the labels) are driving the estimation, I define an activity match ranking scheme. These values will explore the relationship between the activity type of the predicted node and the activity type of its most likely connected node as:

- Topicality rate: percentage of how many predicted nodes and their *k* most likely connected nodes have the same activity types. That is, the *Topicality rate* is to investigate whether the existing K (1, 3, 5, 10) tweets which are likely to be connected to the new tweet and the new tweet have the same activity types.

Location Ranking Schemes for Hierarchical Modelling:

Since the *placed tweets* have no precise location, the location ranking scheme introduced in the previous section is no longer relevant. The goal in this section is to estimate the bounding box where a tweet is generated from. When estimating the location of a new post, with no preliminary assumptions about its location, I first randomly connect the node which represents the new post to the spatial graph constructed using the hierarchical modellings of previous tweets' *Twitter place* types. Then the framework predicts its most likely connected node in the graph to estimate whether the new post's bounding box and its mostly connected node's bounding box are the same or overlapped. Thus, I suggested a separate location ranking scheme as followed:

- Top-*k* *place* accuracy: accuracy for the case in which the actual bounding box of the tweet's *Twitter place* and the bounding box of one of the top-*k* ($k = 1$, 3, 5, 10) most likely connected node according to the prediction are same or overlapped.

5.3.4 Baselines

To assess the capabilities of the proposed graph autoencoder framework, I set up a series of baselines based on a diverse set of link prediction algorithms. The algorithms designed for the baseline comparisons including a series of conventional machine learning approaches (*Adamic-Adar index*, *Jaccard Coefficient* and *Spectral Clustering*) and a neural network-based approach (*Node2Vec*). All the baseline methods take the same graph structures defined in Section 5.3.1 as input, and I use the same Top-*K* accuracy introduced in Chapter 5.3.3 for assessing location estimation qualities produced by each approach. Detailed introduction is as followed:

- Activities Clustering: such a baseline firstly clusters the locations of the tweets in each activity types into several clusters using K-means clustering algorithm (MacQueen et al., 1967) and calculates the locations of the centres of the largest clusters for each activity types. For a tweet with a specific activity type, this baseline is a very simple operationalisation of *hypothesis 1* and checks whether the actual location of such a tweet and the centre of the largest cluster of its activity type is within a specific distance range.

- Adamic-Adar index: this approach was originally introduced to predict links in a social network by Adamic and Adar (2003). It refines the simple counting of common neighbours by assigning the less-connected neighbours more weight. This method performs predictions on the topological structure of the graph with no knowledge of the labels of tweets.

- Jaccard Coefficient: a measure proposed by Jaccard (1901) to assess the similarities between sets of data. It is measured by considering the number of common neighbours divided by the union of neighbours of both vertices. Similar to Adamic-Adar index, it also performs predictions on the topological structure of the graph with no knowledge of the labels of tweets.

- Spectral Clustering: an approach developed based on graph theory, to identify communities of nodes in a graph using the information of the edges that are connecting them. It also performs link predictions on the topological structure of the graph with no knowledge of the labels of tweets.

- Node2Vec: an algorithmic framework for representational learning on graphs introduced by Grover and Leskovec (2016). Such an approach can perform link prediction with or without the knowledge of the labels of tweets. For the purpose of comparison, I enable the *Node2Vec* to take label information of the tweets.

Activities Clustering is a simple operationalisation of *hypothesis 1* and is the simplest baseline adopted in this study to assess the performance of my proposed framework. Such a baseline is designed based on the assumption that the content shared about the place will reflect the use of space and the activities carried out there, and the same types of activities are more likely distributed in similar places. Thus, given a tweet and its activity type, without further modelling, this baseline assumes that the tweet's location is close to the centre of the cluster of the same activity types.

Adamic-Adar index and *Jaccard Coefficient* are two similar approaches focusing on common neighbours in a topological structure between node pairs. That is, if two nodes share common neighbours, a link likely existed between such a node pair. There are two reasons to include these two approaches in the baselines: first of all, both methods have already been widely adopted in social media topological studies (Rawashdeh and Ralescu, 2015), despite most of those topological structures constructed focusing on social interactions between users (e.g., relations between personal home pages), it would still be interesting to explore if such conventional machine learning approaches are suitable for the link prediction tasks formalised in this chapter on spatially constructed graphs. Secondly, these two approaches focus on node-level information aggregation (i.e., common neighbours shared between nodes) but ignoring the impact of the observed graph structures. One of the essential elements in my proposed VGAE framework is to estimate the location of social media posts based on the spatial distribution of the content production from previous posts. Thus, the comparison of the results between VGAE and these two approaches is crucial to justify the use of such spatial distribution.

From this perspective, *Spectral Clustering* is a more sophisticated approach that can perform link predictions with node-level aggregation as well as taking advantage of observed graph structures. *Spectral Clustering* is chosen as it is a more advanced machine learning approach compared to *Adamic-Adar index* and *Jaccard Coefficient*, and also it is an approach which has been widely adopted in social media network analysis (Gupta et al., 2012). I use these three approaches (*Adamic-Adar index*, *Jaccard Coefficient* and *Spectral Clustering*) conducting link predictions without the labels of tweets in the comparisons to test my *Hypothesis 2*. That is, the spatio-temporal structures of the tweets can provide an insight of the location estimation task about where the next tweet will be without any other information.

Node2Vec is a state-of-the-art neural network-based approach to perform link prediction task on graphs, and it is adopted in this chapter as one of the baselines comparing to VGAE which performs link prediction with the graph structure as well as tweets' labels. It is set up as a direct comparison to the proposed VGAE framework in this chapter.

Through the comparisons between VGAE and all these four approaches (*Adamic-Adar index*, *Jaccard Coefficient*, *Spectral Clustering* and *Node2Vec*), I aim to understand at which level the observed graph structures and the information of labels can impact the location estimation Top-*k* accuracy. Thus, the baseline comparisons can test two of my research hypothesis that both the spatio-temporal structure of tweets and their labels can aid the location estimation task of social media posts.

5.4 Model Training

The VGAE takes both the adjacency matrix of the spatial graph and labels of the dataset as input. There exist 7 different categories as the labels, the labels are encoded using one-hot encoding approach, each dimension of the feature vector of every label is $(1, 7)$. For the baseline *VGAE_no_label*, I adopt the same approach as described in Kipf and Welling (2016b), which dropped the dependence on the feature vector matrix $\mathbf{X}$ by replacing it with the identity matrix in the GCN (Kipf and Welling, 2016a). The validation and test sets contain 5% and 10% of links from the graph constructed based on the graph construction set. I initialise the weights as described in Glorot and Bengio (2010), and train the VGAE network for 300 iterations using Adam (Kingma and Ba, 2014) with a learning rate of 0.01. I keep a 32 dimension hidden layer and 16 dimension latent variables same as in Kipf and Welling

(2016b). The model is designed in Tensorflow (Abadi et al., 2015), and the training procedure was performed using Nivida GPU Geforce GTX 1070 (NVIDIA et al., 2020).

5.5 Results

5.5.1 Geolocated Tweets Estimation

Table 5.1 summarises the results on the location estimation task with geolocated tweets. As defined in the previous section, *Dist* is the distance range used for constructing the topological structures. That is, the distance of any two tweets are within a given *Dist*; they are connected in the spatially constructed graph. *Range* is used for testing whether the distance between the actual location of a predicting tweet and its most likely connected node in the constructed graph is within a certain distance radius.

VGAE achieves 30% Top-1 accuracy in a graph with *Dist* as 50 meters when tested with *Range* as 3000 meters. In other words, the results are interpreted as the framework achieves 30% accuracy when the actual location of the tweet and its mostly connected node in the graph are within 3 kilometres. That is, identifying an area of about 28.27 square kilometres, considering that Greater London covers an area of about 1,569 square kilometres. The results are significantly higher than the results achieved by the simplest baseline *Activities Clustering*. Also, VGAE achieves the best Top-*k* (k=3, 5, 10) results on the graph constructed with *Dist* as 100 meters (Top-3 accuracy: 46.5%, Top-5 accuracy: 56.5% and Top-10: 62.5%). By increasing the graph complexity (achieved by increasing the *Dist* when constructing the graph so that more tweets are connected together), the estimation results do not improve. In other words, The VGAE seems to perform better on a simple graph structure, and the best results are obtained with a graph structure constructed using a 100-meter distance as distance range, whereas the Top-*k* prediction accuracy drops as the *Dist* increases.

In comparison, *VGAE_no_label* also achieves the best Top-*k* (k=3, 5, 10) results on the graph constructed with *Dist* as 100 meters (Top-3 accuracy: 38.5%, Top-5 accuracy: 46% and Top-10: 57%), whereas the best Top-1 accuracy (26.5% tested with *Range* as 3 kilometers) is achieved on the graph constructed with *Dist* as 500 meters. The results indicate that the *VGAE_no_label* can already produce reasonable estimations based on the graph structures but not taking into account the labels of the tweets. Such reasonable results achieved by *VGAE_no_label* confirms the research hypothesis *Hypothesis 2* raised in Section 5.1 that the spatio-temporal patterns of the social media posts can provide an insight of the location estimation process even without any further information. However, comparing the results obtained by using *VGAE_no_label* and VGAE, it is clear that the VGAE, which includes the information encoded in the labels outperforms the VGAE that has no semantic information. This shows the semantic information which is encoded in the labels provide an advantage in the location estimation task. These findings are particularly interesting from a geographic and location-based services perspectives. As the labels used are not determined by tweets' geographical locations and have not been assigned based on the tweets' locations, it indicates that knowing the previous content produced in a geographic area can help the proposed framework better estimating the geolocations of the emerging activities. Thus, the comparisons confirm the *Hypothesis 1* presented in Section 5.1.

Models	Input	Dist(m)	Range(m)	Top-1 (±2%)	Top-3 (±2%)	Top-5 (±2%)	Top-10 (±5%)
Activities Clustering	Geolocated Tweets	-	1000	1.0	-	-	-
		-	2000	3.0	-	-	-
		-	3000	8.0	-	-	-
Spectral Clustering	Topological Structure	50	3000	24.0	38.5	48.0	51.5
		100	3000	19.0	42.0	51.0	61.0
		500	3000	24.5	28.0	30.5	36.5
		1000	3000	20.5	24.5	27.0	31.0
Adamic-Adar	Topological Structure	50	3000	3.0	3.0	3.0	3.0
		100	3000	3.0	3.0	3.0	3.0
		500	3000	4.0	4.5	4.5	4.5
		1000	3000	10.0	10.5	10.5	10.5
Jaccard Coefficient	Topological Structure	50	3000	3.0	3.0	3.0	3.0
		100	3000	3.5	3.5	3.5	3.5
		500	3000	3.5	3.5	3.5	3.5
		1000	3000	10.0	11.0	11.0	11.0
VGAE_no_label	Topological Structure	50	1000	5.5	7.0	11.5	15.0
		50	2000	10.0	26.0	35.0	47.5
		50	3000	23.5	39.5	50.5	60.0
		100	1000	5.0	10.0	14.0	18.0
		100	2000	15.0	27.0	34.5	38.5
		100	3000	25.5	38.5	46.0	57.0
		500	1000	8.5	10.0	11.5	13.5
		500	2000	11.5	17.0	19.5	24.5
		500	3000	26.5	35.0	37.5	40.0
		1000	1000	4.0	7.5	8.0	8.5
		1000	2000	6.5	8.0	9.0	10.5
		1000	3000	6.0	13.0	16.0	26.5
VGAE	Topological Structure and labels of activity types	50	1000	6.0	8.5	12.0	16.0
		50	2000	23.5	45.5	45.5	48.0
		50	3000	**30.0**	45.5	55.5	62.0
		100	1000	5.5	9.5	15.0	18.0
		100	2000	28.0	31.0	34.5	40.5
		100	3000	29.5	**46.5**	**56.5**	**62.5**
		500	1000	5.0	10.0	12.5	14.0
		500	2000	10.0	17.5	22.0	28.5
		500	3000	28.0	32.5	35.5	43.0
		1000	1000	4.0	7.5	8.0	9.5
		1000	2000	4.0	10.0	10.5	12.5
		1000	3000	8.5	14.0	18.0	28.5
Node2Vec_label	Topological Structure and labels of activity types	50	3000	24.5	39.5	48.0	61.0
		100	3000	26.0	41.5	47.5	59.5
		500	3000	25.5	37.0	41.0	48.5
		1000	3000	5.0	9.5	15.0	24.5

TABLE 5.1: Baseline comparisons.

Table 5.1 also summarises the results obtained using the baseline methods. Conventional machine learning link prediction methods *Adamic-Adar* and *Jaccard Coefficient* achieve comparably low accuracy in all experiments (Top-*k* accuracy are all below 11%). As introduced in the previous section, both *Adamic-Adar* and *Jaccard Coefficient* perform link prediction based on common neighbours on nodes level aggregation rather than properly encoding the graph structures in the algorithms. The results show that these node-level similarity-based link prediction approaches without taking into account labels of tweets are less effective in my proposed location estimation task. *Spectral clustering*, which works by partitioning a graph into subgroups, where the nodes in one group are similar, and nodes in different groups are dissimilar, achieves competitive results. For example, it achieves 24.5% Top-1 accuracy on the graph constructed with *Dist* as 50 meters, which is 1% higher than the results achieved by *VGAE_no_label* on the same graph. All Top-*k* accuracy achieved by *Spectral clustering* are close to the the results produced by *VGAE_no_label*. Despite there are differences in their mathematical details and implementations, both *Spectral clustering* and *VGAE_no_label* make use of the spectrum (eigenvalues) of the similarity matrix of the graph structure data. Thus, their results are mostly similar to each other. The baseline comparisons conducted on *Adamic-Adar*, *Jaccard Coefficient*, *Spectral clustering* and *VGAE_no_label* illustrate the fact that both the graph structures and semantic understanding of the tweets (i.e., labels) are essential for the frameworks.

Node2Vec, as a comparison in the baselines which perform link prediction taking into account both nodes' labels and graph structures. It generates nodes' embeddings for the spatially constructed graphs through a mapping of nodes to a low-dimensional space of features that maximises the likelihood of preserving network neighbourhoods of nodes, and use the embeddings for the downstream link prediction task. Comparing the results produced by *Node2Vec_label* and *VGAE_no_label*, it clearly shows that *Node2Vec_label* outperforms the *VGAE_no_label*. For example, the best results achieved by *Node2Vec_label* is on the graph with *Dist* as 100 meters (26%), which is 0.5% higher than the *VGAE_no_label*. Such a comparison further justifies the *Hypothesis 1* that the semantic information which is encoded in the labels provide an advantage in the location estimation task. Comparing the estimation accuracy produced by VGAE and *Node2Vec_label*, it demonstrates that the VGAE framework is superior to the *Node2Vec_label*, in particular within the comparisons on Top-1 accuracy.

Models	Dist (m)	Range (m)	Top-*k* Accuracy	Topicality_rate(%)
VGAE	50	1000	Top-1	66.67
			Top-3	61.22
			Top-5	62.34
VGAE_no_label	50	1000	Top-1	14.29
			Top-3	29.41
			Top-5	35.67

TABLE 5.2: Results of Topicality rate.

Looking at the differences between VGAE and *VGAE_no_label* in Table 5.1, it is difficult to understand the impact of the semantic information encoded in the labels has on the estimation. As introduced in Section 5.3.1, I propose a *Topicality rate* to measure at which level the semantic understanding on the labels of tweets impact the location estimation on the content. Table 5.2 shows significant differences

Model	Top-*K place*	Accuracy (±5%)
VGAE_hierarchical_structure	Top-1	10.96
	Top-3	24.20
	Top-5	26.48
	Top-10	38.36
VGAE_tree_structure	Top-1	15.53
	Top-3	28.31
	Top-5	35.62
	Top-10	46.58
VGAE_dense_tree_structure	Top-1	**30.14**
	Top-3	**30.59**
	Top-5	**42.47**
	Top-10	**54.79**

TABLE 5.3: Comparisons for Top-*k place* accuracy of hierarchical modeling of bounding boxes.

between the *Topicality rates* produced by VGAE and *VGAE_no_label*, and it demonstrates a strong association exists between labels of the predicted nodes and the labels of its most likely connected node in the graph when conducting the location estimation using VGAE. Such an association enables the framework to provide a better estimation. Comparing the results in Table 5.1 and Table 5.2, it further confirms my hypothesis that the semantic understanding of social media posts can contribute to the location estimation of the non-geographic tweets.

5.5.2 Tweets Hierarchical Modeling

As discussed in Section 5.3.3, I proposed various ranking schemes to assess the performance of hierarchical modelling approaches. Table 5.3 summarizes the results for Top-*K place* accuracy. It assesses whether the predicted tweet and its most likely connected node have the same or overlapping bounding boxes. When estimating the location of a non-geographic tweet, the possible area where the tweet might be geotagged is contained within the bounding box of its most likely connected node. *VGAE_dense_tree_structure* produces the best results among three hierarchical modelling approaches (Top-1: 30.14%, Top-3: 30.59%, Top-5: 42.47% and Top-10: 54.79%). In particular with the Top-1 *place* accuracy, *VGAE_dense_tree_structure* achieves a significant performance improvement (over 15%) compared to *VGAE_hierarchical_baseline* and *VGAE_tree_structure_baseline*.

It is important to notice that when estimating a location for a tweet, if the tweet's most likely connected node in the graph has *Twitter Place* as London, the Top-*K place* accuracy will consider the estimation is correct regardless the actual geotag of the predicted tweet because the bounding box of London always overlaps with smaller bounding boxes. Although such an issue does not occur in my experiments presented above, this is potentially problematic if the framework is deployed. In the future development, those tweets which have *Twitter Place* as London must be handled separately, e.g., excluding them as a possible link for the predicted tweets.

5.6 Discussion

Our understanding of the role played by UGC in place representations has been so far limited by the fact that only a small percentage of social media posts are precisely geolocated. In this chapter, I focused on harnessing the dynamics of overall content production from multiple users in a single place to estimate the location of content not explicitly labelled by the user. My study presented in this chapter can benefit the understanding of places by exploring the number of users' activities that could be related to a place of interests (i.e., by estimating a location for content that has not been explicitly geolocated by the users).

The representation and interpretation of data retrieved from social media provide means by which to assess different urban dynamics and has the potential to contribute to the creation of socio-demographic analysis of the cities (Shelton et al., 2015). In turn, such information enables us to analyse daily spatial processes and to gain knowledge about places, especially with respect to collective human dynamics (Steiger et al., 2016). Content shared about the place reflects the use of space and the activities carried out, and similar activities are more likely distributed in similar places (Chaniotakis and Antoniou, 2015; Lansley and Longley, 2016). Thus, knowing the activity of a post and the spatio-temporal structures of the previous content production, it is theoretically possible to estimate where the post is generated from. Motivated by such a research hypothesis, I propose a location estimation framework based on two essential elements: first is the spatial distribution of the content production of previous social media posts; second is the semantic understanding (i.e., labels of activity types) of the social media posts.

Existing studies on location estimation are focusing on developing or applying geographic information retrieval (GIR) methods (e.g., geoparsing on placenames) on the text content of social media posts. For example, NeuroTPR (Wang et al., 2020b) achieves 82.1% accuracy on fully geo-annotated texts. However, only a small proportion of tweets include references to geolocations in the text (MacEachren et al., 2011), and the existing text-based studies are limited when placenames are unclear, missing or vernacular in the text content. This study is akin and complementary to the work aimed at geotagged social media content using classic geographic information retrieval approaches. The novel aspect of my proposed framework is the VGAE component which is capable of predicting the link between an unknown node (a new social media post) and its most likely connected node in the graph (previous posts). I investigate two main approaches to modelling the geolocations of social media posts, spatial modelling and hierarchical modelling. For spatial modelling, I measure the quality of the estimation based on whether the geolocation of the new social media post is within a predefined distance from the location of the post represented by its most likely connected node. The results indicate that the spatio-temporal structure of previous content can provide insightful information on the location of new content, and using the semantic information about the content brings a significant advantage. It confirms my proposed hypothesis that a semantic understanding of the content of social media posts (beyond their explicitly geographic content, such as placenames) can aid the estimation of the location of a social media post. Comparing VGAE with a series of baseline methods using spatial modelling graphs, the results prove VGAE can better estimate the locations of social media posts in my defined location estimation task.

Bounding boxes that are attached to social media posts not only function as geographical containers which contain the social media posts located in (physical)spaces. Instead, they are also a host of associated social, economic, and political practices carried by users' everyday spatial activities. In this chapter, I propose hierarchical approaches to model the bounding boxes to measure the performance of the location estimation. The results show that the hierarchical modelling of the places of social media posts can provide reasonable estimations, especially when using the two approaches based on tree structures. Introducing additional nodes that represent the hierarchy of bounding boxes in the tree structures enables the information regarding tweets in the same place to be aggregated and re-distributed during the learning process, and thus benefiting the VGAE on the location estimation task. Dense tree-structure modelling creates more links from top-level nodes to the bottom-level nodes when one bounding box is overlapped within another bounding box, and an edge will be added. During the learning process, the encoded semantic information can be exchanged from top to bottom level of the tree structure, and hence aid VGAE to perform a better estimation on the approximate locations of bounding boxes.

Although the performance achieved so far is reasonable considering the input data are only the topological structures of tweets and labelled users' activity types, it is important to highlight that the current results of the estimations are proven to be unstable. As can be seen from Table 5.1, for VGAE on topological structures, there exits 2% variations for Top-1, 3, 5 accuracy and 5% variation for Top-10 accuracy. For VGAE adopting hierarchical modelling of places, there are 5% variations for each result of the proposed models shown in Table 5.3. The results presented above are the average accuracy after 20 runs of VGAE on each spatio-topological structure. That is, I run each model 20 times and shows the averaged results concluded from those 20 experiments. Despite the variability of the accuracy, the results generally follow the pattern presented in each table. For example, for the comparisons on Top-*k place* accuracy, although there exists 5% variations in their results, *VGAE dense tree structure* proves to be always outperforming *VGAE tree structure baseline*; for geolocated tweets, VGAE proves always outperforming other baselines and *VGAE_no_label*.

As illustrated in Chapter 2, starting from a conceptualisation of users as sensors of places (Goodchild, 2007), GIScience research has thereby focused on questions regarding how corresponding spatio-temporal patterns from social media networks and heterogeneous data streams can be aggregated to study the digitally coded space (Dodge and Kitchin, 2005). By exchanging the social media activity types' information at the node-level aggregation, and propagating through a spatially constructed graph with the GCN component of VGAE, VGAE takes advantage of utilising the semantic information encoded in the labels and exchanges the information with other nodes in the graph during the learning process. As such, VGAE can produce a more precise estimation during link prediction stage, and thus benefit the location estimation task. Although previous research has proven the association between the geographic location of social media posts and the social events (Gurevich and Ghosh, 2014), and constructed a bridge between content and locations (Chen et al., 2013), those studies are developed based on the enriched information collected from social media platforms, such as using geoparsing on text content and metadata to estimate the locations of users. My proposed approach instead only focuses on the semantic categories of the users' contents and the spatial distribution of the content production from previous social media posts. I show that my approach can estimate users' locations without further analysis on more detailed text content and metadata. It is likely that combining advanced geoparsing approaches and the method here presented, possibly coupled with a category classification process such as the one presented in Chapter 4, could lead to high-quality location estimation of social media posts. Such an assumption will be a part of my future research.

In conclusion, as already mentioned in Chapter 2, current studies on social media location prediction problems rely heavily on content analysis (Zheng et al., 2018), which requires extensive research on placenames usage and identifying location-indicative words (e.g., geoparsing methods introduced in Chapter 2). However, despite the fact that such approaches have achieved high accuracy in the disciplines such as disaster management (Wang et al., 2020b), they potentially ignore a large amount of data which do not include a spatial element explicitly in the text when studying social practices to understanding place representations. Thus, this work is complementary to the existing GIR methods to estimate locations of tweets without explicit placenames in the text. From GIScience and digital geographies perspectives, the quantitative analysis and summarisation to study the emergence of place from space through content production (Graham et al., 2015a) often takes into account the amount of similar UGC produced by users from a geographical area in general rather than focusing on each individual content. Thus, the precise location of the content can be trivial as long as the framework can provide the estimation of the content's location in a relatively small geographical area. My proposed methodology can estimate locations at urban scale with little information (labels of social media activity types and spatial topological structure) required from the social media platforms to provide approximate areas where social media posts are generated from.

It is important to highlight that both the spatio-temporal topological structure of the graphs and labels of social media activity types play a vital role in my proposed framework. The interaction between the two elements (i.e., the labels aggregated and propagate through the topological structure) contributes to a better estimation accuracy. Meanwhile, as the VGAE framework only takes labels of activities and spatio-temporal structures of the previous content production as the input, the methodology requires no changes when it is adapted to other platforms such as Facebook, Foursquare, Flickr, etc. My research shows the potential of applying deep learning methods directly to digital geographies studies, and suggests that adopting an appropriate combination of social media properties and deep learning techniques to understand online places deserves further research.

5.7 Summary

This study has proposed a novel location estimation method based on the spatio-temporal distribution of the content production from a social media platform, Twitter, to estimate the future social media post distribution in a city. The proposed GeoAI tool is akin and complementary to the work that estimates the location of social media content using classical geoparsing-based geographic information retrieval approaches where location-indicative words are missing in the text. The findings demonstrate that by knowing the content of a post and the spatio-temporal structures of the previous content production distributed in the places, it is possible to estimate where the post is generated from. In other words, knowing the representations of the places in a given space and how the content about the places are distributed, one can use my proposed VGAE-based framework to aggregate more data at a high spatial resolution in the cities.

Traditionally, the study of place representation is often associated with various official spatial statistics. Given the increasing popularity of digital content and emerging tools to aggregate spatial data from digital platforms, the demographics of the spatial distribution of UGC and how they are describing the social space (Wilken,

2012) in urban areas have also played a significant role in understanding cities. Spatial distribution of UGC, including social media, is accepted by scholars as a valuable resource to advance research on specific urban aspects (Anselin and Williams, 2016; Arribas-Bel et al., 2015). In next chapter, I will discuss how UGC can be used as a proxy indicator of urban development and propose a novel spatially explicit GeoAI tool that can use place representations described by UGC and official statistics to predict urban deprivation changes.

Chapter 6

Urban Change Modelling with Spatial Knowledge Graphs

Part of this work presented in this chapter is published as:

- **Pengyuan Liu and Stefano De Sabbata**, *2020*. Modeling Urban Socio-demographic Change using Knowledge Graph. In *28th Geographical Information Science Research UK conference (GISRUK)*.

The extended journal version of this paper is going to be submitted to the International Journal of Geographical Information Science[1].

6.1 Introduction

The study of the social, economical and spatial structure of cities and its evolution over time has always been a key component of geographical analysis. Mapping urban change is fundamental to inform our understanding of cities and places. As mentioned in Chapter 2, the socio-spatial structure of cities and metropolitan areas changes over time facing the rapid development of urbanisation and the increasing demands of understanding socio-economic structures of the society. One of the common approaches in analysing these dynamics is to observe change at the level of individual neighbourhoods (Modai-Snir and Ham, 2018). However, the majority of socio-demographic data are commonly collected periodically. For example, census data are collected every 10 years, but neighbourhoods are dynamic and may undergo changes which might not be captured by decadal censuses (Gray et al., 2018). The constant changes of neighbourhood's socio-demographic profiles and their spatial extent (Rey et al., 2011) interject levels of spatial uncertainty within classifications. Therefore, the socio-demographic classifications are unstable during the development of cities (Singleton et al., 2016), and that can lead to potential uncertainties when adopted to understand urban spaces (Gale and Longley, 2013; Fisher and Tate, 2015).

Given the increasing popularity of digital content, the socio-demographics of the spatial distribution of UGC (such as Twitter, Wikipedia etc.), and how they are describing the social space (Wilken, 2012) in urban areas have also played a significant role in understanding cities. Spatial distribution of social media is accepted by scholars as a valuable resource to advance research on specific urban aspects (Anselin and Williams, 2016; Arribas-Bel et al., 2015). People live in or visit an area for a variety of factors (e.g., economic, social, cultural) might produce data about that area. That is,

[1] Code to reproduce my experiments is available at: https://github.com/PengyuanLiu1993/PhD_Thesis_Codes_PengyuanLiu/tree/master/SKG_Urban_Dynamics

UGC produced in an area have associations to how city infrastructure is being developed and used (Mora et al., 2018). Therefore, the aggregated activities generated from UGC platforms could indicate how social activities in a city are distributed, revealing fine-grained spatial patterns evident in the social life of cities and informing our understanding concerning how cities are developed (Bawa-Cavia, 2011; Abbar et al., 2018). The representation and interpretation of data retrieved from social media provide a means by which to assess different urban dynamics, and understand the underlying socio-demographies of the cities.

Two datasets that are adopted in this chapter are Twitter and Wikipedia, as introduced in Chapter 3. Twitter has become an important source of content about how people want to represent place and their interaction with the physical environment (Frias-Martinez and Frias-Martinez, 2014). Previous research on characterising urban change using Twitter data focused on analysing changes in the number of tweets sent from a geographic location over time (Soliman et al., 2017). For example, it has been observed that residential zones on the periphery of cities generate more tweets in the evening, when people have returned to their homes, whereas areas of activity in the city centre are especially active during the day, when people visit them to undertake activities such as work or shopping (Ciuccarelli et al., 2014). The spatio-temporal pattern of tweets posted from different parts in the city indicates the existence of the inherent link between the human urban activity patterns and the underlying land use (Zhan et al., 2014), and can help to understand urban structure and related socioeconomic performance (Shen and Karimi, 2016; Martí et al., 2017; García-Palomares et al., 2018). Content posted on UGC platforms such as Twitter can be seen as a signal of the emergence of "urban buzz" in cities or urban districts which by definition are the "powerhouses of innovation, creativity, and unconventional lifestyles" (Arribas-Bel et al., 2016, p. 190). Similarly, Wikipedia editing activities have also been used to investigate the geographies of content production. For example, studies on participation and geographical distribution of Wikipedia articles found spatial clusters in knowledge production which lead to a digital underrepresentation of certain parts of the world (Graham et al., 2015a). Ballatore and De Sabbata (2019) illustrate how the spatial distribution of Wikipedia and Twitter data is related to population density, education level, and income, although every city and platform has its own idiosyncrasies. That uneven distribution is both an issue and an opportunity, which might indicate how the digital place representation emerging from those platforms could be used as a proxy to estimate socio-demographic dynamics, thus benefiting the understanding of places and urban dynamics, such as gentrification (Boy and Uitermark, 2016; Gibbons et al., 2018; Reades et al., 2019).

Many quantitative machine learning models, which have been widely adopted in existing studies towards understanding urban dynamics, are a-spatial. Reades et al. (2019) used a random forest approach to learn and then predict neighbourhood changes, based on a score derived from census variables, and use visual analysis as one of the tools to understand the model and the predictions. Similarly, Alejandro and Palafox (2019) used a random forest approach to predict the likelihood of gentrification, and use visual analysis to explore the results. However, urban development can be a spatial process, whereby once an area gentrifies, neighbouring areas can be affected by that gentrification process independently or in conjunction with other factors. Reades et al. (2019) explicitly discuss how the addition of a spatial component to their model would likely improve the output by accounting for edge effects. More generally, many of the variables that are usually used to model urban changes (from population age to house prices) are frequently spatially autocorrelated – that is, similar values are found in neighbouring areas. Therefore, there is an opportunity for devising spatial models to better model urban development.

In this chapter, I explore the use of knowledge graphs to model urban socio-demographic change using various sources of data and volunteered geographic information. Taking into account (London) Output Area Classification, UK Indices of Deprivation, and the distribution of geotagged social media data and Wikipedia articles at different geographic (inner-city borough-, urban- and nation-level) scales, the results show that my proposed machine learning approach using a knowledge graph not only successfully captures the changes of deprivation between the year of 2015 and 2019 in the city but also has the potential to be developed into a powerful tool in modelling urban changes. It is worth noticing that the study provided in this chapter does not account for the impact of population change on the model. For the scope of this chapter, this study adopts OAs and LSOAs as the area units to aggregate Twitter and Wikipedia, and OAs and LSOAs are designed to be consistent in terms of their population size (an average population of about 310 residents for each OA and an average population of 1500 people or 650 households for LSOA). This chapter is investigating whether the content production of UGC platforms can be a useful proxy of urban change. Further discussion on data at fine spatial and temporal granularity (e.g., population, property value, crime, etc..) that can be adopted for this study will be provided in Chapter 7.

The novelties of this study are:

- this study combines official spatial statistics with place representations described through UGC to understand and predict the dynamic changes of the socio-economic characteristics of places;
- this chapter presents two different ways constructing spatial knowledge graphs to predict socio-demographic changes (i.e., deprivation level) at different geographic scales;
- spatial neighbouring information of areas is a key aspect of the proposed spatial knowledge graphs. The case studies will show that the encodings of spatial information in the graphs are crucial in the task of socio-demographic prediction.

The rest of this chapter is structured as followed: I will provide a detailed introduction to the different datasets used in the different proposed case studies in Section 6.2. Section 6.3 will introduce two different approaches to construct the knowledge graph for the task of socio-demographic prediction. Then, Section 6.4 will present the experiment results concluded from different case studies. Finally, I will provide a conclusion for this study in Section 6.5.

6.2 Study Area

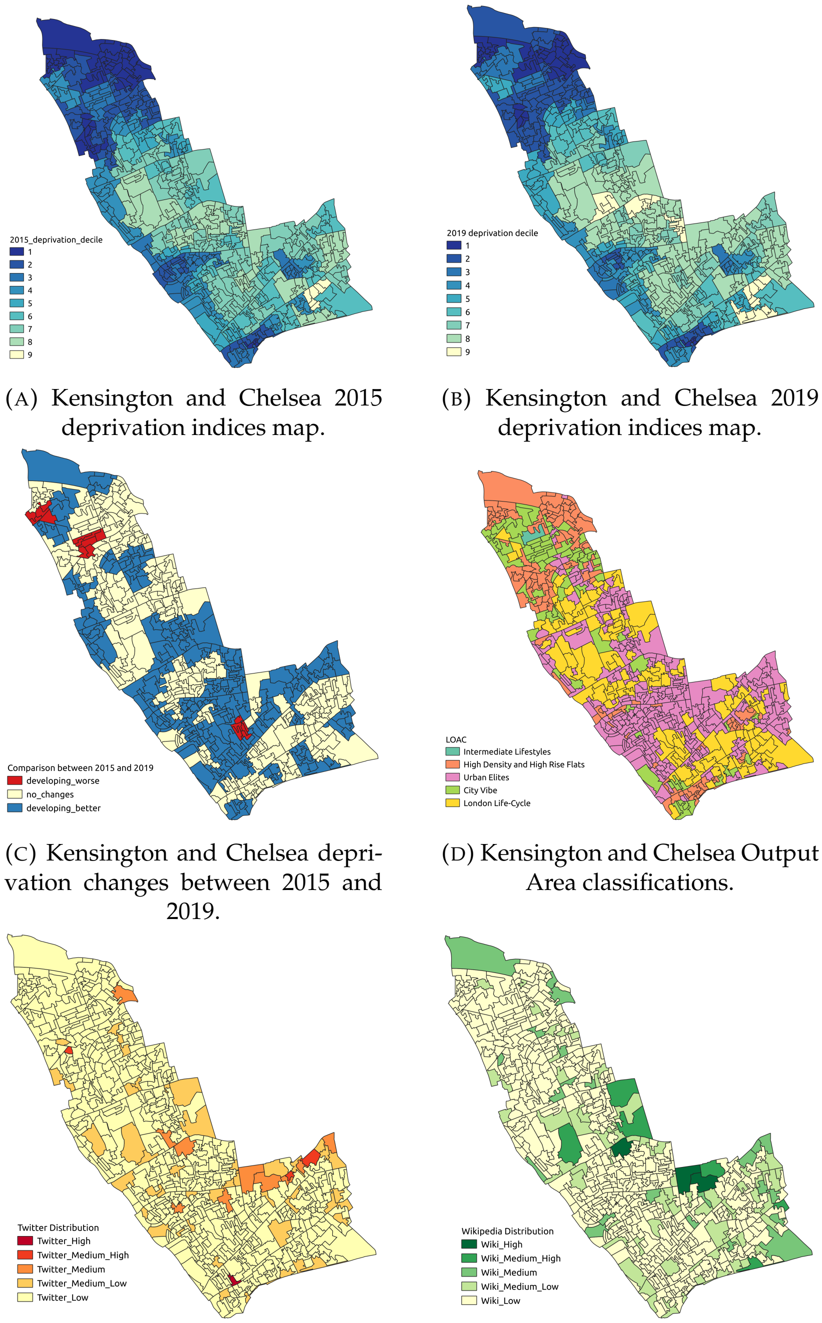
As mentioned in Chapter 2, despite the risk that some neighbourhood processes at a granular level cannot be observed through quantitative data (Barton, 2016), there remains the challenge of defining a neighbourhood in the first place (Reades et al., 2019). According to Knaap et al. (2019), there is no precise definition of "neighbourhood" in either spatial extent or social composition. For the scope of this thesis, I take the definition by Galster (2001, p. 2112) as the starting point to define the term *neighbourhood* in the context of this chapter as: "the bundle of spatially-based attributes associated with clusters of residences, sometimes in conjunction with other land uses". As discussed by Reades et al. (2019, p. 923), such a definition "does not establish neighbourhoods as discrete, bounded entities as it does not directly provide the size of the neighbourhood, but it provides a basis for defining neighbourhoods on different spatial scales through the 'bundling' of attributes". Following such a definition, neighbourhoods, in the context of this chapter are the spatial units defined by ONS (in particular, output areas and lower-layer super output areas) which underpin the the operationalisation of the 2001 and 2011 Output Area Classifications in the UK (Gale, 2014; Gale et al., 2016).

In this chapter, I present three case studies using knowledge graphs to model urban socio-demographic change using various sources of data and volunteered geographic information, at three different scales from an inner-city borough level to a national level: Kensington and Chelsea borough, Greater London and England. The datasets applied to the case studies include the spatial distributions of two UGC data (Twitter and Wikipedia), English Indices of Deprivation (IMD) deciles in 2015 and 2019, London Output Areas Classification (LOAC) for case studies at the urban scale, and Output Areas Classification (OAC2011) in the UK at the national level of the case study (IMD, LOAC and OAC2011 have been described in Chapter 3). In the rest of this section, I will provide a detailed introduction to the datasets used for each case study area and illustrate the socio-demographic patterns of the areas carried out through each dataset.

6.2.1 Kensington and Chelsea

Kensington and Chelsea is an inner London borough located on the north bank of the River Thames in the centre of London. It is a borough with some very wealthy areas, as well as the highest average income in London. However, Kensington and Chelsea is an area embodying "gross level[s] of economic inequality" (Shildrick, 2018, p. 784) with the poorest and richest living in close spatial proximity. MacLeod (2018) identifies an astonishing landscape of inequality (in health, income, etc) across the borough, which renders Kensington and Chelsea an interesting case study to study urban change at a small scale. The study area includes 631 OAs within the Kensington and Chelsea borough of London. In the 2011 LOAC, the borough's OA are classified into five supergroups (Longley and Singleton, 2014): *Intermediate Lifestyles*, *High Density and High Rise Flats*, *Urban Elites*, *City Vibe*, *London Life-Cycle*. As shown in Figure 6.1(D), the LOAC classifies the majority of OAs within Kensington and Chelsea borough as *Urban Elite*, *London Life-Cycle* and *High Density and High Rise Flats*, which indicates that this borough is well developed and highly densely populated.

As introduced in Chapter 3, the 2015 IMD (*2015_dep*) and 2019 IMD (*2019_dep*) datasets ranks the 32,844 Lower-layer Super Output Areas (LSOAs) in England from the most to the least deprived and include a classification based on the decile of ranking, ranging from 1 (most deprived) to 10 (least deprived). In the case study area of Kensington and Chelsea, the deprivation deciles were extracted from the national deprivation ranks in both years of 2015 and 2019. The LSOA in this borough all fall into the first 9 of the 10 deciles, and none in the tenth decile as shown in Figure 6.1 (A) and (B). Between 2015 and 2019, 35.05% of LSOAs in England moved from one decile category to another (17.56% LSOAs had positive changes while 17.49% of LSOAs had negative changes), and 48.81% of LSOAs in Kensington and Chelsea changed as shown in Figure 6.1 (C) and Figure 6.2 (B) (46.12% of LSOA had positive changes while 2.69% of LSOAs had negative changes). Note that LOAC and IMD deciles are organised at different area units where LOAC was created at OA-level, and IMD was created at LSOAs-level. The spatial extent of LSOAs is larger than the OAs. In my model, the deprivation decile of an OA is simply defined as the decile of



(A) Kensington and Chelsea 2015 deprivation indices map.

(B) Kensington and Chelsea 2019 deprivation indices map.

(C) Kensington and Chelsea deprivation changes between 2015 and 2019.

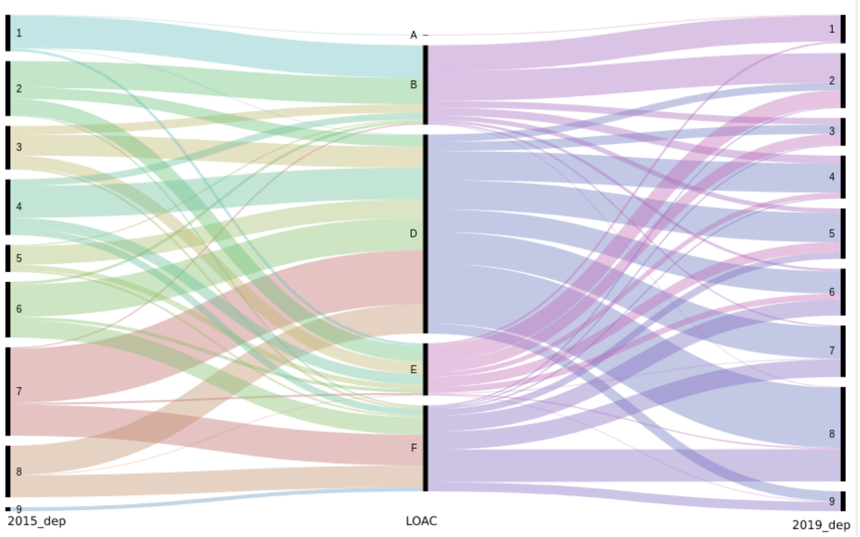
(D) Kensington and Chelsea Output Area classifications.

(E) Kensington and Chelsea Twitter distribution.

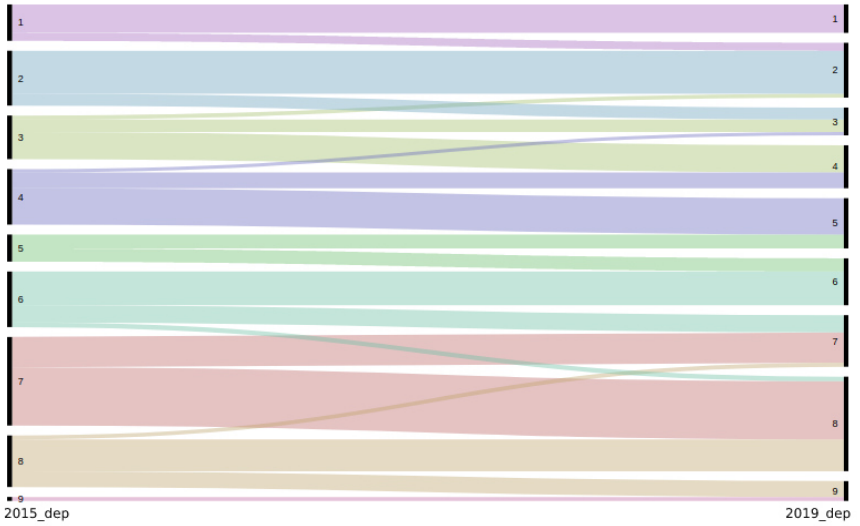
(F) Kensington and Chelsea Wikipedia distribution.

FIGURE 6.1: Socio-demographics of Kensington and Chelsea. Map boundaries source: Office for National Statistics licensed under the Open Government Licence v.3.0. Contains OS data © Crown copyright and database right 2021.

the LSOA that contains it. However, I am aware that this is in part problematic due to uncertainties, and the local variation-aggregated value calculated at one area unit (LSOA) might not necessarily apply equally to all parts of the area (OA). Detailing the implications of this issue is beyond the scope of this dissertation, and it will be considered part of the known unknowns of the models here presented. The issue will be discussed further in the concluding chapter and future studies.



(A) Comparisons between LOAC and 2015 & 2019 deprivation deciles (A-*Intermediate Lifestyles*; B-*High Density and High Rise Flats*; D-*Urban Elites*; E-*City Vibe*; F-*London Life-Cycle*).

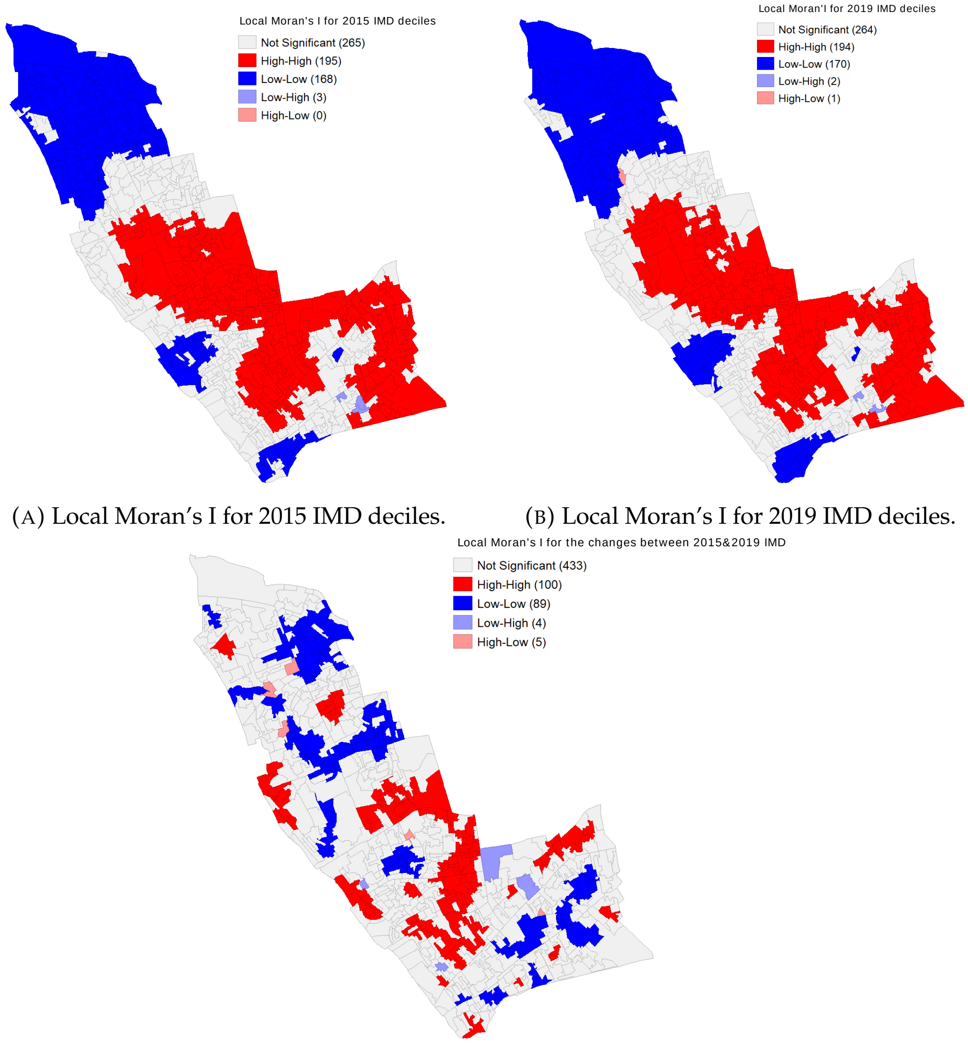


(B) Comparisons between UGC distributions and 2015 & 2019 deprivation deciles.

FIGURE 6.2: Comparisons between deprivation deciles of Kensington and Chelsea.

Geotagged UGC data from Twitter and Wikipedia were previously introduced in Chapter 3. I calculated the count of the UGC data distribution within each OA in Greater London. The reason behind using count rather than density of the UGC data is due to the fact that OAs and LSOAs are statistical area units that have been devised to be homogeneous and of smaller size in terms of population. Since the knowledge graph approach adopted in this chapter and its baseline machine learning comparisons can be understood as labelling methods (will be introduced in Section 6.4.4), the models require categorical data of the spatial distributions of UGC based on the counts of data in each OA as input. As the count of the spatial distribution of two UGC datasets are geographically uneven, I use Jenks natural breaks classification method (Jenks, 1967) to generate a five-level classification for both UGC platforms: *high*, *medium high*, *medium*, *medium low* and *low*.

The deprivation indices maps for the 2015 and 2019 IMD illustrate that despite the presence of areas with a high level of deprivation concentrating in the north part of the borough, the overall deprivation within the borough is between deciles 5 and 9. The deprivation deciles in Kensington and Chelsea in both 2015 and 2019 are positively global spatial autocorrelated (2015 deprivation deciles: Moran's I=0.838, p=0.001; 2019 deprivation deciles: Moran's I=0.846, p=0.001). Both local spatial clusters of deprivation deciles shown in Figure 6.3 (A) and (B) demonstrate that most areas with highest deprivation levels are distributed in the north and east part of the borough. The changes of deprivation levels can be seen in the Figure 6.1(C) and Figure 6.3 (C), where a majority of positive changes occurred in the central part of the borough, and few neighbourhoods are developing worse through the years. The changes in deprivation levels are also positively global spatial autocorrelated (Moran's I=0.541, p=0.001).



(A) Local Moran's I for 2015 IMD deciles.

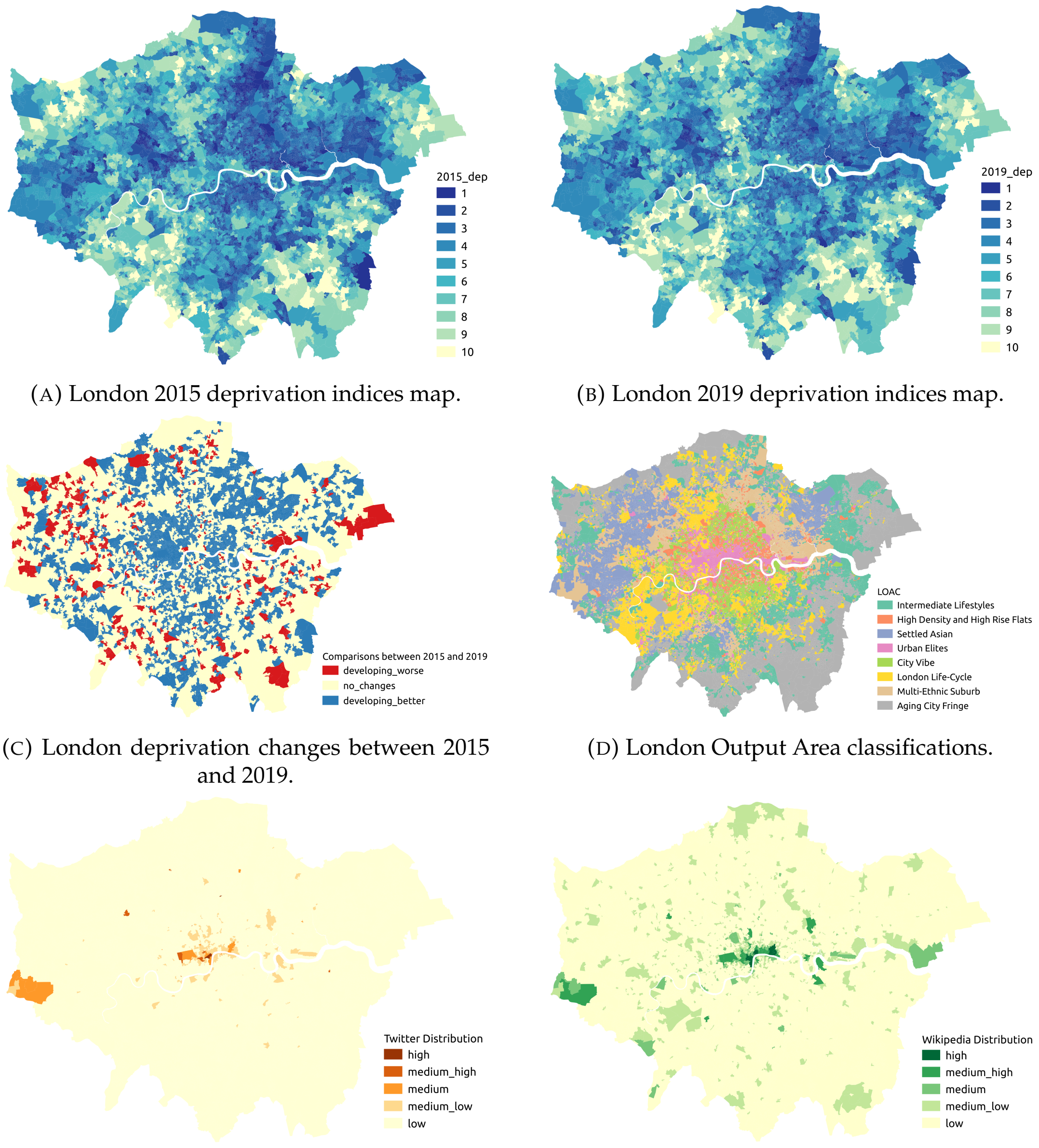
(B) Local Moran's I for 2019 IMD deciles.

(C) Local Moran's I for the changes between 2015 and 2019 IMD deciles.

FIGURE 6.3: Local Moran's I for 2015 and 2019 IMD deciles in Kensington and Chelsea. Map boundaries source: Office for National Statistics licensed under the Open Government Licence v.3.0. Contains OS data © Crown copyright and database right 2021.

Comparing the deprivation maps with the map of LOAC classification as shown in Figure 6.2(A), it is clear that there is some level of correlation between areas with high deprivation level and areas which are in categories *City Vibe* and *High Density and High Rise Flats* – as one would expect. Thus, such a correlation between high deprivation level and particular geodemographic categories seems to indicate that LOAC can provide an insight into the deprivation changes. In Figure 6.1(C) and 6.1(D) illustrate how the overall distribution of Twitter and Wikipedia in the borough is relatively low compared to the whole of London.

6.2.2 Greater London



(A) London 2015 deprivation indices map.

(B) London 2019 deprivation indices map.

(C) London deprivation changes between 2015 and 2019.

(D) London Output Area classifications.

(E) London Twitter distribution.

(F) London Wikipedia distribution.

FIGURE 6.4: Socio-demographics of London. Map boundaries source: Office for National Statistics licensed under the Open Government Licence v.3.0. Contains OS data © Crown copyright and database right 2021.

The second study area is conducted on the urban level, including 25,053 OAs in Greater London. The purpose of this case study is to assess the scalability and

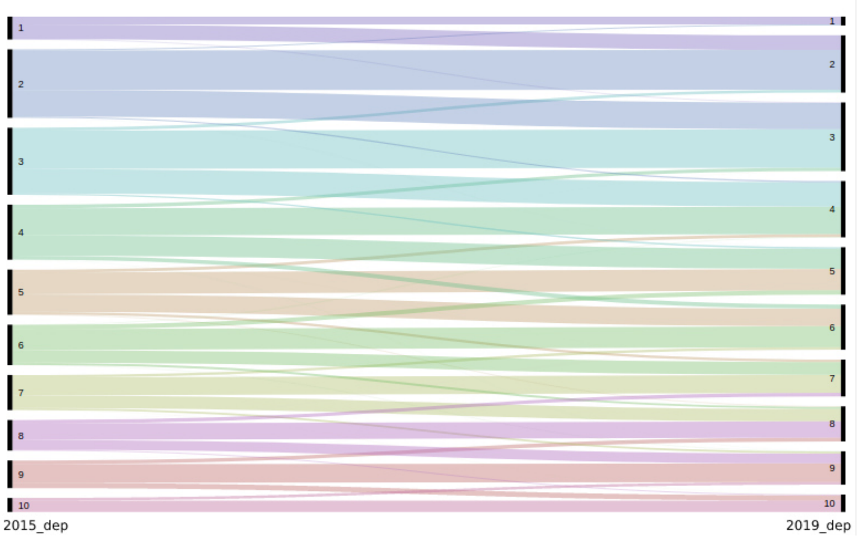
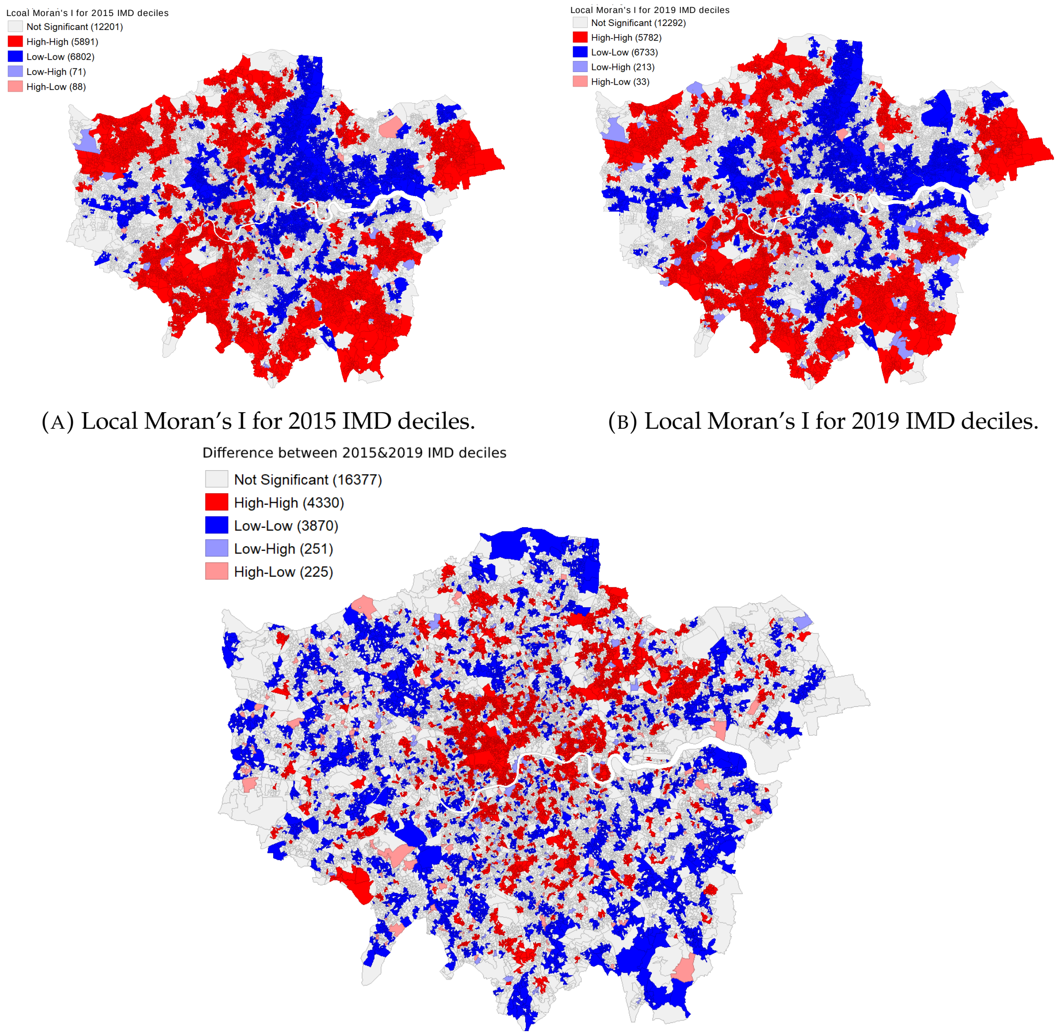


FIGURE 6.5: Comparisons between deprivation deciles of London.

robustness of the proposed framework. In the 2011 LOAC for Greater London in Figure 6.4(D), in addition to the supergroups mentioned in Kensington and Chelsea, there are three more supergroups introduced in Longley and Singleton (2014): *Aging City Fringe*, *Settled Asians* and *Multi-Ethnic Suburbs*. In Kensington and Chelsea, IMD deciles in 2015 and 2019 are both between 1 and 9; however, in Greater London, IMD deciles are both between 1 and 10, as shown in Figure 6.4(A) and (B). Similar to Kensington and Chelsea, the deprivation level in London in both 2015 and 2019 are positively global spatial autocorrelated (2015 deprivation deciles: Moran's I=0.829, p=0.001; 2019 deprivation deciles: Moran's I=0.819, p=0.001). The spatial clusters in Figure 6.6 (A) and (B) show that areas in London which reported to have high deprivation levels are clustered in the east of London and south bank of Thames River. The changes of deprivation levels can be seen in the Figure 6.4(C), Figure 6.6 (C) and Figure 6.5, where a majority of positive changes occurred in the central part of London, and neighbourhoods that are falling in the rankings through the years are mostly distributed in the suburbs of London. The changes in deprivation levels are also positively global spatial autocorrelated (Moran's I=0.548, p=0.001). Figure 6.4(E) and Figure 6.4(F) illustrate the overall spatial patterns of Twitter and Wikipedia presented in London. Despite the fact that each data dataset has its unique pattern distributed in the city, they share the same pattern that central London has a large number of UGC (Pearson's correlation test result: r=0.61, p<0.01). A detailed discussion about such a statistically correlation will be presented in Section 6.3.

6.2.3 England

As introduced above, the first two case studies on Kensington and Chelsea borough and London are at the borough-level and urban-level. To further test the scalability of the framework, I conduct a further case study on the national level of England. Instead of using OAs, this study area adopts Lower Layer Super Output Areas (LSOAs) as the studying neighbourhoods to reduce the computation costs (England has 32,844 LSOAs). For each LSOA, following the two case studies above, I calculated the density of the two UGC data (Twitter and Wikipedia) distribution within



(A) Local Moran's I for 2015 IMD deciles.

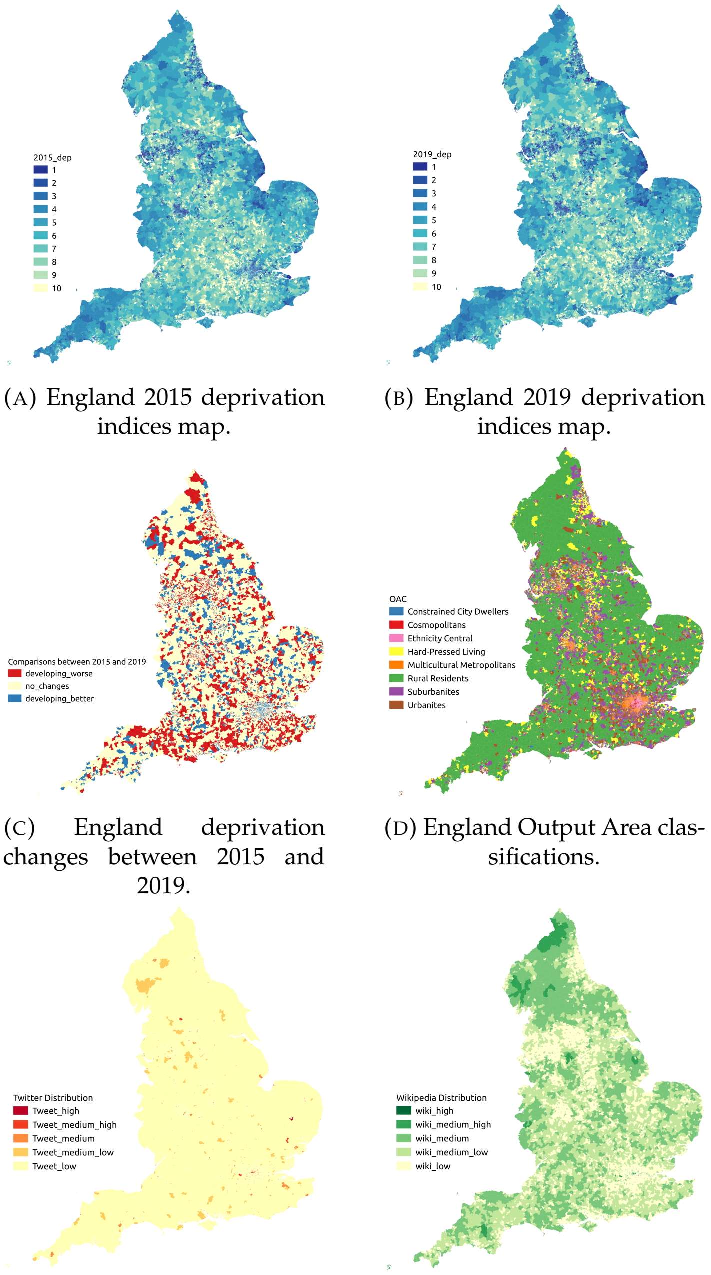
(B) Local Moran's I for 2019 IMD deciles.

(C) Local Moran's I for the changes between 2015 and 2019 IMD deciles.

FIGURE 6.6: Local Moran's I for 2015 and 2019 IMD deciles. Map boundaries source: Office for National Statistics licensed under the Open Government Licence v.3.0. Contains OS data © Crown copyright and database right 2021.

each LSOA, and used Jenks natural breaks classification method (Jenks, 1967) to generate a five-level classification for both UGC platforms: *high*, *medium high*, *medium*, *medium low* and *low*. As shown in Figure 6.7 (E) and (F), despite Wikipedia shows the higher distribution patterns in rural areas across the country compared to the spatial distribution of Twitter, two datasets are still sharing moderate similarities in their distribution patterns (Pearson's correlation test result: r=0.447, p<0.01).

Since this case study is conducted at the national level, LOAC can no longer be used, as it only covers the Greater London area, as such, I choose the OAC2011 (Gale et al., 2016) as the socio-economic descriptors to substitute the LOAC in the framework. The OAC2011 groups include *Rural Residents*, *Cosmopolitans*, *Ethnicity Central*, *Multicultural Metropolitans*, *Urbanites*, *Suburbanites*, *Constrained City Dwellers* and *Hard-Pressed Living*. As illustrated at the beginning of this subsection, this case study is conducted on the LSOA-level neighbourhoods to reduce the computation costs of testings for the proposed framework. However, OAC2011 is conducted on



(A) England 2015 deprivation indices map.

(B) England 2019 deprivation indices map.

(C) England deprivation changes between 2015 and 2019.

(D) England Output Area classifications.

(E) England Twitter distribution.

(F) England Wikipedia distribution.

FIGURE 6.7: Socio-demographics of England. Map boundaries source: Office for National Statistics licensed under the Open Government Licence v.3.0. Contains OS data © Crown copyright and database right 2021.

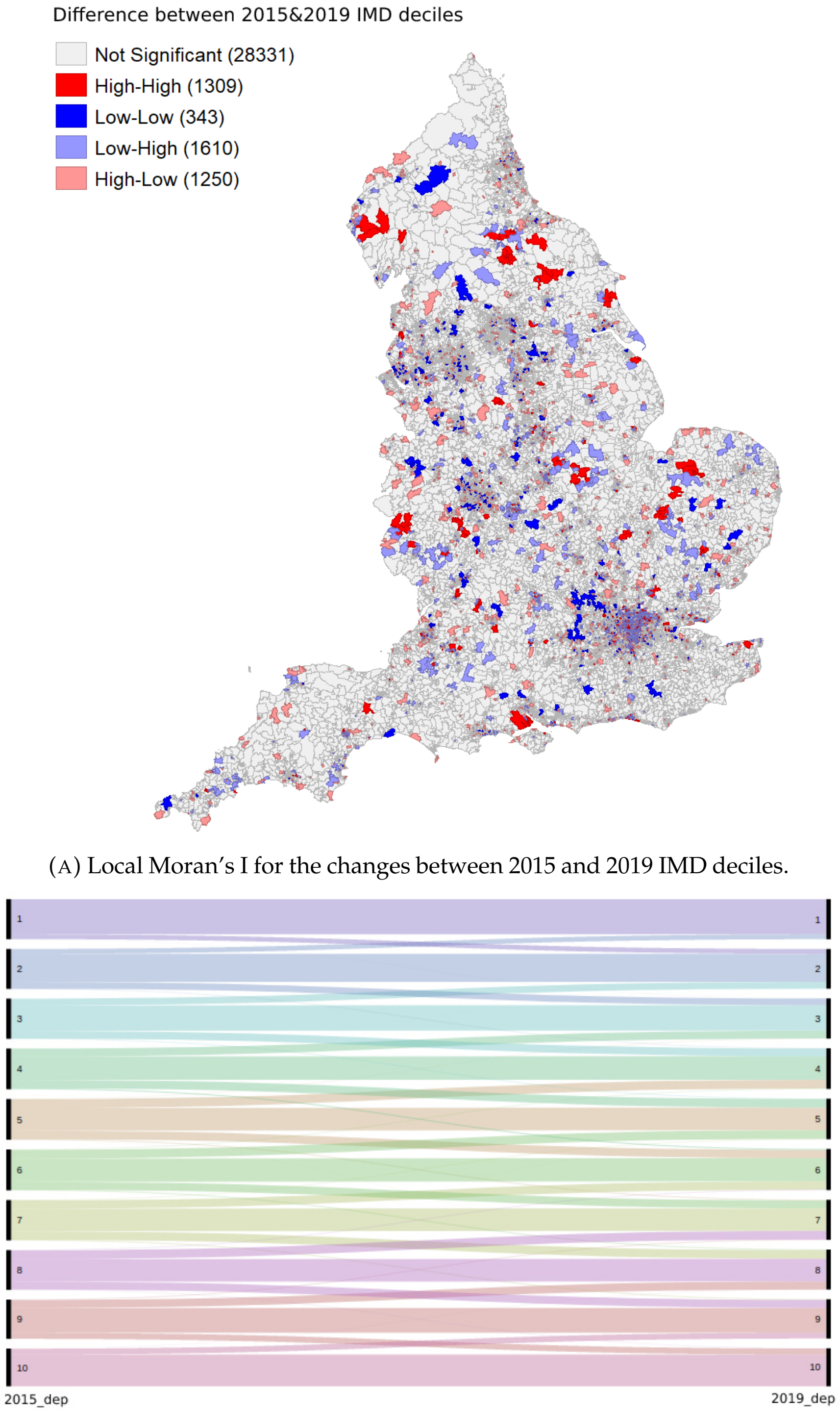
the OA level which is a smaller geographic level compared to LSOA; thus, it requires a further summarisation and characterisation of the socio-economic characteristics for LSOA based on the original classifications on the OAs. As such, I first aggregate the counts of OAC2011 categories of in each LSOA, and then choose the most frequent category as the socio-economic descriptor for the LSOA. For example, if the majority of OAs within a LSOA are *Ethnicity Central*, the socio-economic descriptor for such a LSOA is *Ethnicity Central* in the prepared dataset. If an LSOA has two OAC2011 categories with the same counts, I randomly pick up one as the socio-economic descriptor for this LSOA. It is important to notice that, again, I am aware that such a socio-economic descriptor chosen process might lead to uncertainties at some level as LSOAs and OAs are two area units with different sizes. These uncertainties will be considered in the final chapter. Figure 6.8 (B) shows the comparisons between deprivation deciles in 2015 and 2019. As discussed in Section 6.2.1, 35.05% of LSOAs in England have their deciles changed (17.56% LSOAs have positive changes, while 17.49% of LSOAs have negative changes).

As shown in Figure 6.7(C) and Figure 6.8 (A), the deprivation changes between 2015 and 2019 in England indicate that most urban areas are moving upwards in the rankings, whereas the spatial distribution of the deprivation deciles changing patterns in the rural areas in the country are showing the opposite. Statistically, the overall deprivation deciles changes in England are positively spatial autocorrelated (Moran's I=0.111, p=0.001), which shows that if one neighbourhood's deprivation level is improved, so does its surrounding neighbourhoods.

This case study aims to explore the scalability of the framework. Such a goal is achieved by assessing the prediction quality for the rest of the country when knowing the IMD deprivation patterns in a specific city. For this purpose, I select three different cities or regions in particular as the known data for the framework (also known as training data, see next section): Greater London, Leicester and the county of Cumbria. Each one of the three cities or regions in the UK has its unique socio-economic characteristics, and it is interesting to assess how knowing the unique socio-economic patterns of a specific city can impact the socio-demographic prediction for the rest of the country with my proposed framework. A brief introduction is provided as followed:

- Greater London: London is a thriving and highly prosperous city with great diversity. It is one of the richest cities in the world, with a growing economy, but is also home to some of the poorest communities in the UK. Many socio-economic patterns in London are unique and different from the rest of the UK. For example, according to Economics (2016), poverty levels among London's population after taking account of housing costs are much higher than the UK as a whole. Up to a third of all inner London residents and nearly a quarter of outer London residents are in poverty, which is higher than for any other UK region. Such economic diversities in London render the city unique in the UK; thus it is chosen here specifically to assess the framework's performance by encoding the deprivation level in 2019 in London as the training data.

- Leicestershire and Rutland: Leicestershire is a landlocked county in the English Midlands, being within the East Midlands with 396 LSOAs. It takes its name from the city of Leicester which is located at its centre and administered separately from the rest of the county. Leicester is unique in the county and even in the UK in terms of its ethnic diversity. Leicester residents have origins from over 50 countries from across the globe, making the city one of the most ethnically and culturally diverse places in the UK (55% population are from black and minority ethnic backgrounds) (Jivraj and Finney, 2013). Leicester is the 14th most deprived local authority of the 152 upper-tier authorities and is therefore in the bottom decile nationally, whereas Leicestershire is ranked 136th most deprived of the 152 upper-tier local authorities (Information Policy Team, 2017). Rutland is also a landlocked county in the East Midlands of England, bounded to the west and north by Leicestershire. It is the smallest historic county in England and the fourth smallest in the UK as a whole (23 LSOAs). Rutland has a very low population density, at 98 people per square kilometre,



(A) Local Moran's I for the changes between 2015 and 2019 IMD deciles.

(B) Comparisons between deprivation deciles of England.

FIGURE 6.8: Comparisons between 2015 and 2019 deprivation deciles. Map boundaries source: Office for National Statistics licensed under the Open Government Licence v.3.0. Contains OS data © Crown copyright and database right 2021.

compared to a national average of 413. Rutland is one of the most affluent counties in England; of 152 Upper Tier Local Authorities, Rutland ranked 148 (The Scrutiny Commission, 2018). Rutland was a district within Leicestershire for 23 years under the Local Government Act 1972, which took effect on 1 April

1974 (Local Government Act, 1972); thus, even until recent time, two counties are commonly jointly studied in various government reports (e.g., Leicester and Leicestershire Economic Assessment (Leicester-Shire Rutland Statistics Research, 2010)).

- Cumbria: Cumbria is a ceremonial and non-metropolitan county in North West England. It is predominantly rural and contains the Lake District National Park which is considered as one of England's finest areas of natural beauty, serving as an inspiration for artists, writers, and musicians; thus, it attracts enormous tourists visiting every year. In 2018, Cumbria and the Lake District received over 47 million visitors, made up of 40.4 million day-trippers and 6.6 million overnight visitors (Tate, 2018). The county itself has at almost 7,000 square kilometres, it is the second-largest county, but with just under 500,000 people living here, it is also one of the sparsest populated counties in the UK (Cumbria Vision, 2009). Cumbria's average IMD score ranked as 83rd nationally with only 8.1% of LSOAs sat within IMD decile 1 (the most deprived 10% of LSOAs nationally) (Cumbria Intelligence Observatory, 2019). As a county predominantly rural and well-developed, Cumbria is also chosen in this case study to be encoded as training data in the graph for assessing how it impacts the prediction quality for the rest of the country.

6.3 Statistical Inference on IMD

Having presented the datasets in the previous section, the primary hypothesis is that deprivation does not change considerably from one estimation (e.g., 2015) to the next (e.g., 2019). That is, the urban change in the city is not dramatic. Therefore, IMD in 2015 can be a proxy indicator of how likely the IMD would be in 2019. Although in this thesis, I chose IMD deciles to be the indicators of how deprivative an area is, one may choose IMD scores (see Chapter 3) to perform the analysis. Regardless of the choices, a high correlation between the IMD 2015 scores and the IMD 2019 scores is expected, as well as the IMD 2015 deciles and the IMD 2019 deciles.

In studying gentrification processes, Reades et al. (2019) discussed how the model output would be improved when including the spatial component, and they highlighted how social media such as Twitter could be a useful proxy to indicate the neighbourhood change at the urban scale. Ballatore and De Sabbata (2018, 2019) also argued that the content production of UGC is strongly related to the underlying socio-demographics (e.g., education and wealth), which are commonly associated with gentrification within cities. Therefore, the second hypothesis is that the change of content produced on UGC platforms is another proxy for urban change.

The analysis in this subsection explores the two hypotheses mentioned above. The study first examines the strength of the relationship between IMD scores in 2015 and 2019, and IMD deciles in 2015 and 2019. The scatter plots in Figure 6.9 illustrate the linear relationship between the IMD Scores in 2015 and 2019, and between the IMD Deciles 2015 and 2019. Then, to further test the hypothesises, I adopted statistical modelling (both a-spatial model and spatial model) to explore the strength of the relationship between IMD scores in 2015 and 2019, and IMD deciles in 2015 and 2019. Then, I used the data (normalised using inverse hyperbolic sin) from two UGC platforms, Twitter and Wikipedia, as a proxy indicator of urban change to test the second hypothesis.

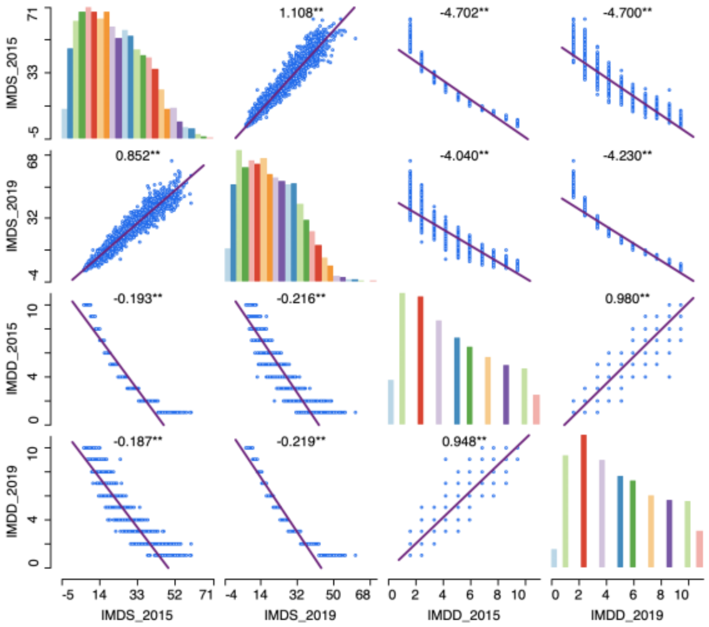


FIGURE 6.9: Case study: London

6.3.1 A-spatial model, Deprivation Scores and Deciles

The first study is to test the first hypothesis that IMD in 2015 can be a proxy indicator of how likely the IMD would be in 2019. First, the hypothesis is tested with an a-spatial regression model using deprivation scores (*Score model*), and the results are shown in Figure B.1. Then I adopted the deciles (*Deciles model*) as in the same a-spatial regression model, and the results are presented in Figure B.2.

As shown in the figures, both the intercept and the coefficient related to the 2015 IMD are highly significant. In addition, the R-square values are high, demonstrating a strong correlation between 2015 IMD and 2019 IMD in both deprivation scores and deciles. Moreover, the Jarque-Bera test for both models is significant, indicating a high degree of non-normality of the residuals. For *Score model* which takes deprivation scores as input, the Breusch-Pagan test and the Koenker-Bassett test are significant, indicating that the residuals are heteroskedastic. However, for the *Deciles model* with input data as deprivation deciles, both the Breusch-Pagan test and the Koenker-Bassett test are not significant, indicating that the residuals are homoscedastic. Therefore, *Deciles model* is proven to be more robust, which justifies my choice of using deprivation deciles in the knowledge graph construction in this chapter (see Section 6.4).

For both models, both the Lagrange Multiplier (lag) statistics and the Lagrange Multiplier (error) statistic are significant, indicating the presence of spatial autocorrelation in the data. Thus, a model which incorporates the spatial component may provide a more robust prediction for the 2019 IMD using the 2015 IMD as a proxy. Furthermore, because the Robust LM (lag) statistic is not significant, but the Robust LM (error) is significant for both models, a spatial error model can be a better specification for the analysis. In the study below, I carried out a series of spatial error models (maximum likelihood estimation) for both IMD scores and IMD deciles. The models further explore incorporating two UGC variables (Twitter posts and Wikipedia pages) as independent variables to predict the change of 2019 IMD.

6.3.2 Spatial Models, Deprivation Scores and Deciles

The spatial lag regression model is adopted with both deprivation scores and deciles, and the results are summarised in Figure B.3 and B.4. For both models, the spatially autoregressive parameter (LAMBDA) and the Likelihood Ratio Test are significant, confirming that neighbouring values positively affect the model. Similar to the previous section, the Breusch-Pagan test of the *Scores model* is significant, indicating that the residuals are heteroskedastic. However, for the *Deciles model*, the Breusch-Pagan test is not significant, meaning that the residuals are homoscedastic. The results show that the *Deciles model* is more stable, which again justifies my choice of using deprivation deciles in the knowledge graph construction in this chapter.

6.3.3 Combining User Generated Content

To test the second hypothesis that data from Twitter posts and Wikipedia pages can be used as a proxy indicator of urban change, I present six models incorporating the distribution patterns of Twitter and Wikipedia separately or together into the spatial lag regression model. For the first model (*Twitter-Score model*), I combined 2015 IMD deprivation scores with the distribution pattern of Twitter in the spatial lag regression model to predict 2019 IMD scores. The results are presented in Figure B.5. Similar to *Twitter-Score model*, the second model (*Wikipedia-Score Model*) combined 2015 IMD deprivation scores with the distribution pattern of Wikipedia pages in the spatial lag regression model to predict 2019 IMD scores, and the results are in Figure B.7. The third (*Twitter-Decile model*) and fourth model (*Wikipedia-Decile model*) combined 2015 IMD deprivation deciles with Twitter and Wikipedia distribution patterns separately to predict IMD 2019 deprivation deciles, and results are presented in Figure B.6 and B.8. The fifth (*UGC-Score model*) and sixth model (*UGC-Decile model*) combined both the distribution patterns of Twitter and Wikipedia with IMD 2015 deprivation scores and deciles to predict IMD 2019 deprivation scores and deciles separately, and results are shown in Figure B.9 and B.10.

The spatially autoregressive parameter (LAMBDA) and the Likelihood Ratio Test are significant for all six models. Moreover, the UGC-related independent variables are all significant, except the Wikipedia-related variable when both Twitter and Wikipedia are included in the model. Such an exception may be due to the co-linearity existing in both datasets (Ballatore and De Sabbata, 2018) (see discussion in the next paragraph). The two tables B.1 and B.2 summarise the pseudo R-squared, Log-likelihood, and AIC values. In both cases, it is clear that the addition of UGC-related independent variables marginally increases the pseudo R-squared value and the Log-likelihood value while decreasing the AIC value. Thus, the results indicate a marginally better fit of the model when including UGC-related variables and neighbouring information of the study areas in London.

In conclusion, the results in this section indicate that in London, taking into account the neighbouring values of the IMD 2015 and UGC-related variables provide a marginally better model for the inference of the IMD 2019 than taking into account the IMD 2015 alone. The models using deciles seem to be marginally more robust, which justifies adopting IMD deciles in this chapter rather than using IMD scores. Such results also justify and support the use of ComplEx based on the labelling of IMD deciles (see Section 6.4). Furthermore, the analysis demonstrates that the model using the Twitter-related variable seems to provide a better fit than those using the Wikipedia-related variables. The models using both the Twitter and Wikipedia-related variables seem to provide the best fit but suffer from the co-linearity of the two (Ballatore and De Sabbata, 2018). However, the co-linearity is not an issue for the adopted knowledge graph approach, the co-occurrence between entities and relations of Twitter and Wikipedia captured by ComplEx is expected to lead to more accurate predictions on the 2019 IMD deciles.

Such a statistical inference study contextualises the research hypothesis in the study of urban change that UGC can be adopted as a proxy for modelling urban dynamics. Thus, it further motivates the study presented in this chapter to develop a more sophisticated model to predict urban change.

6.4 Methodologies

6.4.1 Spatial Knowledge Graph

In this chapter, I frame my proposed socio-demographic change modelling as a link prediction task. Considering that the source data discussed in the section above was not already encoded as a knowledge graph, to fully explore the potential of the knowledge graph approach on the socio-demographic change modelling task, I developed and tested two distinct to constructing spatial knowledge graphs for modelling deprivation change.

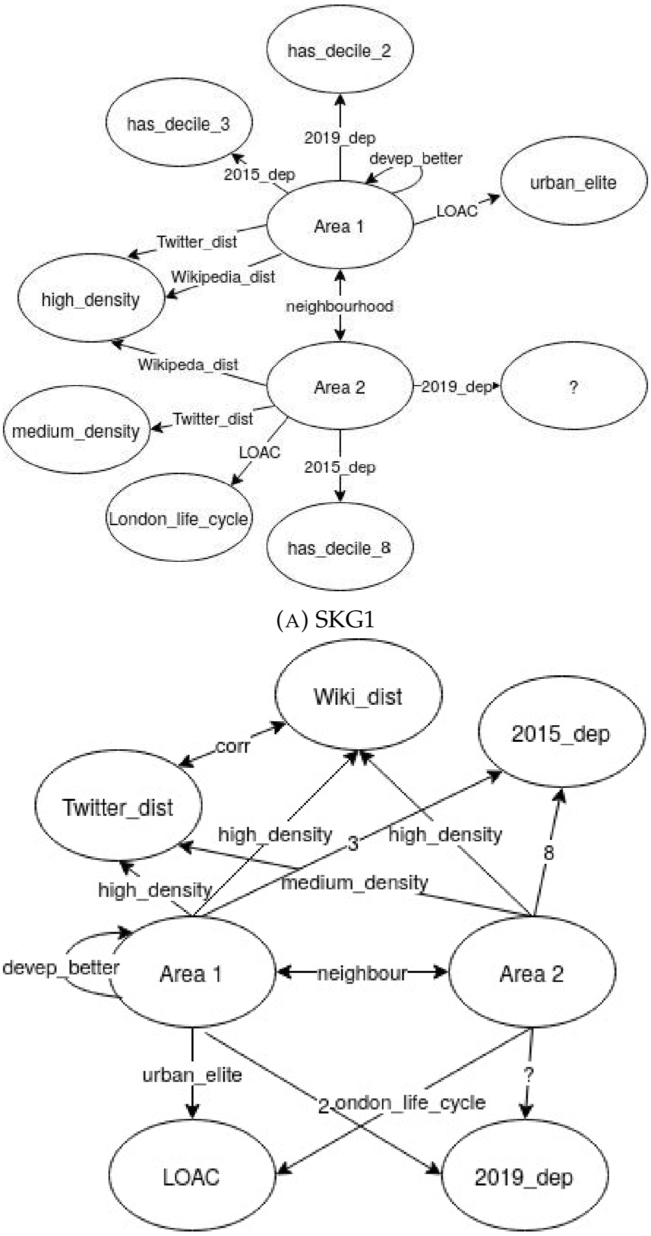
SKG1

Following the common knowledge graph triples <subject, relation, object> defined in databases such as DBpedia (Auer et al., 2007) and Wikidata (Vrandečić and Krötzsch, 2014), my first spatial knowledge graph (*SKG1*) is proposed as follows. The entities of the spatial knowledge graph represent the OAs and socio-demographic representations described in each dataset. For example, as shown in Figure 6.10(A), if "*Area1*" is classified as "*Urban Elites*" in the LOAC, a relation "*has_LOAC_value*" is created between "*Area1*" and "*urban_elites*". As such, this triple in *SKG1* is understood as <*Area1*, *has_LOAC_value*, *urban_elites*>. In such a setting, when the link prediction algorithm infers the 2019 deprivation decile for "*Area2*", it generate possibility scores between entity "*Area2*" and all entities "*has_decile_n*"(*n* ranges from 1 to 10), and the prediction framework chooses the entity with highest score as the predicted 2019 deprivation decile for "*Area2*".

SKG2

The design of *SKG1* is consistent with most existing knowledge graph bases in which places' status in each data source is encoded as an entity in the graph. Such knowledge graphs are mainly design for (geographic) information retrieval (e.g., Wikidata) or (geographic) questions answering (e.g., CrowdGeoKG (Chen et al., 2017)), and most often the entity description contains a classification of the entity with respect to a class hierarchy. However, my proposed socio-demographic prediction is fundamentally different from the above-mentioned tasks, in which the framework is only targeted on predicting one type of socio-economic changes (deprivation changes in this study), and datasets have no distinct hierarchies. Thus, it can be assumed that it requires no sophisticated capabilities for further modelling. As such, the design choice on how to design the triplets in the knowledge graph can be more flexible. As the source data discussed in the section above was not already encoded as a knowledge graph, to fully explore the potential of knowledge graphs, another graph design is introduced in Figure 6.10(B).

The second spatial knowledge graph (*SKG2*) differs substantially from *SKG1*, and the entities of the *SKG2* represent the OAs and each class of the different datasets. If "*Area1*" is classified as *"Urban Elites*" in the LOAC, a relation *"urban_elites*" is created between "*Area1*" and "*LOAC*". Therefore, this triple in *SKG2* is understood as <*Area1*, *is_urban_elites_in*, *LOAC*>. Such a graph construction choice enable the link prediction algorithm to directly predict the link's label between *"Area2*" and *"2019_dep.*". Therefore, the fundamental difference between two means of spatial knowledge graph construction is that *SKG2* formalises the task of deprivation prediction as a link prediction task, while *SKG1* considers the task as a node classification task as shown in Figure 6.10.



(A) SKG1

(B) SKG2

FIGURE 6.10: Different spatial knowledge graph.

As mentioned in Chapter 2, despite existing research demonstrating that the geographical encoding in the knowledge graph has a significant performance increase when dealing with geographic questions, how to properly encode such geographic knowledge (i.e., locations) remains as a domain-specific challenge and is still in the exploratory stage (Wang et al., 2019b; Yan, 2019).

The spatial differences of the 2015 and 2019 deprivation deciles are spatially autocorrelated in all case study areas as introduced in the previous section; thus, such correlations indicate the fact that if one neighbourhood's deprivation level improves, its surrounding neighbourhoods might also improve. That seems to indicate that spatial neighbouring information among areas can significantly benefit the prediction of the deprivation changes. As such, there are relations that represent spatial neighbouring information among areas, which are included in both graphs SKG1 and SKG2. For example, as shown in Figure 6.10(B), if *"Area1*" and *"Area2*" are neighbours, a *"neighbour*" relation is created between two nodes representing the two OAs. Such neighbouring information in GIScience can be described with various spatial weight matrices, such as Rook and Queen contiguity-based spatial weights, or distance-based spatial weights. For the scope of this study, the spatial structures of the neighbourhoods in all case studies are decided by the Queen contiguity-based spatial weight matrices. The Queen contiguity-based spatial weight matrix is one of the most popular and simplest approaches in GIScience to determine the spatial structure of areas, and two areas are the spatial neighbours if they share the same boundary or a vertex. Explorations on other options which determine the spatial structure of areas in the knowledge graph form the basis of future research objectives in analysing boundary effects on the prediction qualities.

As discussed in the previous section, the results in Section 6.3.3 suggest that using both the Twitter and the Wikipedia-related variables seem can provide the best insight of 2019 IMD prediction, but suffer from the co-linearity of the two. Therefore, I add a relation between entities "*Twitter_dist*" and "*Wikipedia_dist*" as "*correlation*" in *SKG2*, to capture the existence of a correlation between two data distribution and enrich the information contained in the knowledge graph.

In addition, in my experiment, I randomly select n areas and their corresponding 2019 deprived deciles into training data as the known information for the model. For the randomly selected *"Areas*", I compare their deciles between *"2015_dep*" and *"2019_dep*", and further define a self-pointed relations to themselves from one of the following as: *"develop_better*", *"more_deprived*" and *"remain_same*". For example in Figure 6.10, *"Area1*" was in the first decile of the IMD 2015, but in the second decile in 2019. Therefore, a self-pointed relation "*less_deprived*" is added in the graph.

For *SKG1*, ComplEx is to infer the highest probability score of the relation *"2019_dep*" between each *"Area*" and 2019 IMD deciles, I choose the decile which has the highest probability score of *"2019_dep*" as the model prediction. For *SKG2*, ComplEx is to infer the missing relations between each *"Area*" and *"2019_dep*", and I choose the relation of deciles which has the highest probability score as the model prediction.

6.4.2 Evaluation

In this chapter, I adopt the state-of-art method ComplEx introduced in Section 3.5.2 to create graph embeddings for my proposed spatial graphs. It is important to highlight that unlike a classification task, where algorithms provide a "predicted" category which can be correct or not, the output of a link prediction algorithm is the likelihood of relationships between nodes in the data. To assess the performance of ComplEx on my proposed spatial knowledge graph, I design an independent assessment on the graph on both *SKG1* and *SKG2* using the knowledge graphs for the case study area Kensington and Chelsea, and then I calculate *Hit at N* (*H@n*) (Bordes et al., 2013) as the proportion of correct predictions in top *n* relations suggested (calculated by the probabilities of whether the link is existed between the nodes in the graph) by the algorithm to evaluate the performance of ComplEx. During this step, I evaluate the robustness of the spatial knowledge graph approach on the data, which are randomly selected for omission and can contain any relations between different entities (*LOAC*, *Twitter_dist*, etc.). ComplEx achieves 76.0% (*SKG1*) and 72.3% (*SKG2*) accuracy on *H*@1, which indicates that more than 70% of omitted relations are correctly predicted as top preferences suggested by the algorithm. Therefore, in my further urban changes modelling experiments, I consider one relation is correctly predicted if the probability score of this relation generated by the algorithm is over 70%. I define *confidence ratio* as the percentage of correctly predicted relations over all the edges in the test data. And *confidence ratio* will be adopted to evaluate the performances for both *SKG1* and *SKG2*.

6.4.3 Comparison Experiments

To evaluate which factors drive most in the process of inferring the dynamics of urban deprivation level. I design a series of ablation studies (an ablation study studies the performance of an AI system by removing certain components, to understand the contribution of the components to the overall system) using ComplEx with fewer data encoded in the knowledge graphs as followed:

- no Spatial Neighbours, I keep all entities and relations except the "*neighbour*" information between areas in both graphs;
- no UGC, I omit the information regarding Twitter and Wikipedia spatial distributions in both graphs;
- UGC and Spatial Neighbours, I omit the information regarding LOAC and 2015 IMD, but keeping the entities and relations regarding Twitter and Wikipedia distributions, and spatial neighbours between areas.
- LOAC and Spatial Neighbours, I omit information about UGC data and 2015 IMD but keeping the entities and relations about LOAC and spatial neighbours between areas.
- 2015 IMD and Neighbours, I omit information about LOAC and UGC distribution but keep the entities and relations about 2015 IMD, and spatial neighbours between areas.

6.4.4 Baseline Comparisons

To test the capability of my proposed framework using knowledge graphs, I compare the results with two conventional machine learning approaches and a baseline that assume that the IMD deciles do not change between 2015 and 2019. The socio-demographic categories from each dataset are encoded with *One-Hot encoding* which is a commonly adopted method to quantify categorical data in the models:

- *No Changes*: This is the simplest baseline which assumes that the IMD deciles do not change between 2015 and 2019. The primary analysis of the datasets (IMD deciles in 2015 and 2019) shows 51% of OAs in Kensington and Chelsea, and 65% of LSOAs in England do not change through the years. Thus, such an assumption is useful to be set up in comparison with my proposed framework assessing if my framework can perform reasonable predictions.

- *Random Forest*: I adopt a traditional machine learning approach Random Forest (Kam, 1995) as a classifier on randomly selected 300 of 631 OAs in Kensington and Chelsea borough (the largest number chosen in the experiment for knowledge graph approaches) as the training dataset to predict the labels for 2019 IMD deciles in the test dataset. Random Forest is a state-of-the-art machine learning approach, and it has been successfully applied to various socio-demographic studies within geography as well sociology, such as urban socio-economic deprivation analysis (Zhou et al., 2017; Niu et al., 2020), urban crime studies (Bowen et al., 2018) and geo-political studies (Hao et al., 2019). Therefore, it is appropriate to adopt such a state-of-the-art approach as one of the baselines.

- *Decision Tree*: I adopt Decision Tree as another baseline comparison with the same experimental data used in *Random Forest*. Decision Tree is another state-of-the-art machine learning approach, and it has been successfully applied to various socio-demographic studies within geography as well sociology, such as urban planning (Venerandi et al., 2015), urban deprivation analysis (Akinyemi and Elias, 2009) and geo-economic studies (Wu et al., 2020a). Therefore, it is appropriate to adopt such a state-of-the-art approach as one of the baselines.

The baselines are conducted on the case study area of Kensington and Chelsea borough. For *Decision Tree* and *Random Forest*, the deprivation prediction task is formalised as a classification problem, in which two algorithms target on predicting 2019 deprivation deciles for each OAs based on the spatial density of Twitter and Wikipedia data (the same five-level classifications used for spatial knowledge graphs), 2015 deprivation deciles and LOAC categories within the borough.

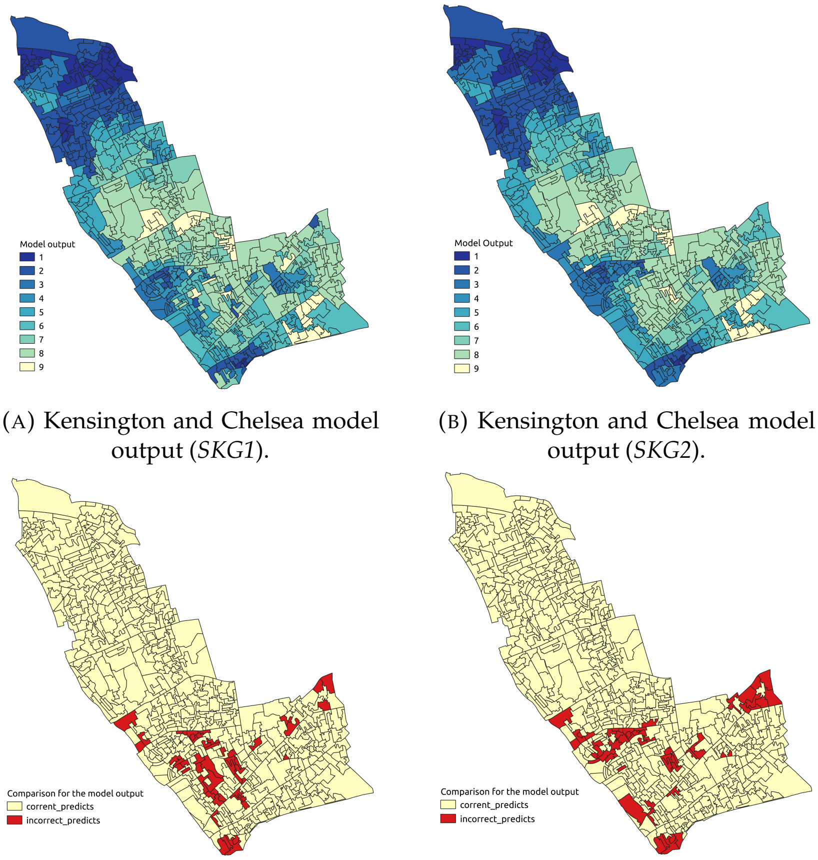
6.5 Results

6.5.1 Kensington and Chelsea Borough

No Changes	Number of n	SKG1	SKG2	Random Forest	Decision Tree
51.00	10	63.23	54.68	-	-
	100	77.34	73.74	-	-
	200	85.10	84.79	58.25	61.08
	300	90.70	91.13	57.48	60.76

TABLE 6.1: Summarized results (Confidence Ratio (%)).

Table 6.1 summarises the results of the experiments. ComplEx achieves a reasonable confidence ratio at 63.23% (*SKG1*) and 54.68% (*SKG2*) when the framework is trained on 10 OAs – for which the 2019 IMD decile is provided to the knowledge graphs. As introduced in Section 6.3.3, the simplest comparison *No Changes* achieves 51% accuracy, which assumes the deprivation level of OAs in Kensington and Chelsea has no changes through the years. Both of *SKG1* and *SKG2* outperform such a baseline even the knowledge graphs have only 10 OAs with 2019 deprivation deciles, which illustrate the usability of my proposed framework in such a deprivation prediction task. The more information the algorithm has, the higher the



(A) Kensington and Chelsea model output (*SKG1*).

(B) Kensington and Chelsea model output (*SKG2*).

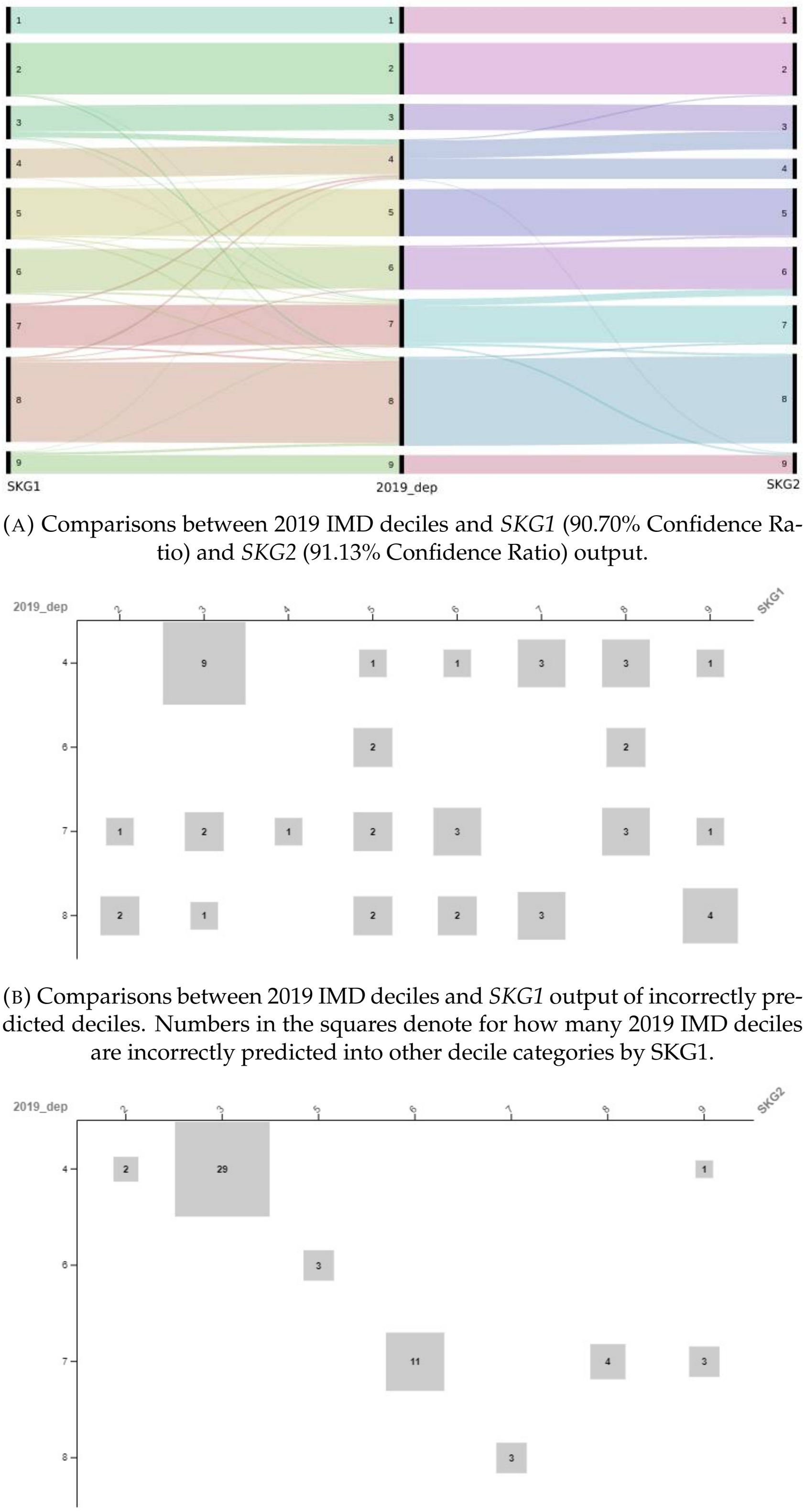
(C) Comparison between the 2019 deprivation deciles and model output (*SKG1*).

(D) Comparison between the 2019 deprivation deciles and model output (*SKG2*).

FIGURE 6.11: Maps of 2019 IMD deciles and model outputs. Map boundaries source: Office for National Statistics licensed under the Open Government Licence v.3.0

confidence ratio. When the model is trained on nearly half of areas (i.e., 300), the outputs for both graphs are over 90% accurate, which are significantly higher than the outputs produced by the machine learning baseline methods (*Random Forest* and *Decision Tree*). A preliminary analysis of the results seems to indicate that the link prediction algorithm has the ability to capture socio-demographic change based on a spatially constructed knowledge graph with small samples of data. It is also interesting to see that despite when the model trained on 10 OAs, ComplEx on *SKG1* achieves much higher confidence ratio than *SKG2*, the performance differences are smaller when more areas are included in the training data.

As described in the section above, I consider a decile is predicted incorrectly if the probability generated by the model is below 70%. For deciles which are incorrectly predicted, I further explored the most likely deciles suggested by the model with n=300. Comparing deciles in IMD 2019 and deciles suggested by the model in Figure 6.11(C) and (D), it suggests that small differences are concentrated in the southwest part of the borough. Figure 6.12 shows that for the deciles that are not predicted correctly with n=300, the differences between the deciles suggested by *SKG2* and 2019 IMD deciles are mostly within the two deciles differences, but the errors produced by *SKG1* seem to be more random, but most errors are still shown within three deciles. Therefore, it indicates that despite the fact that some of the



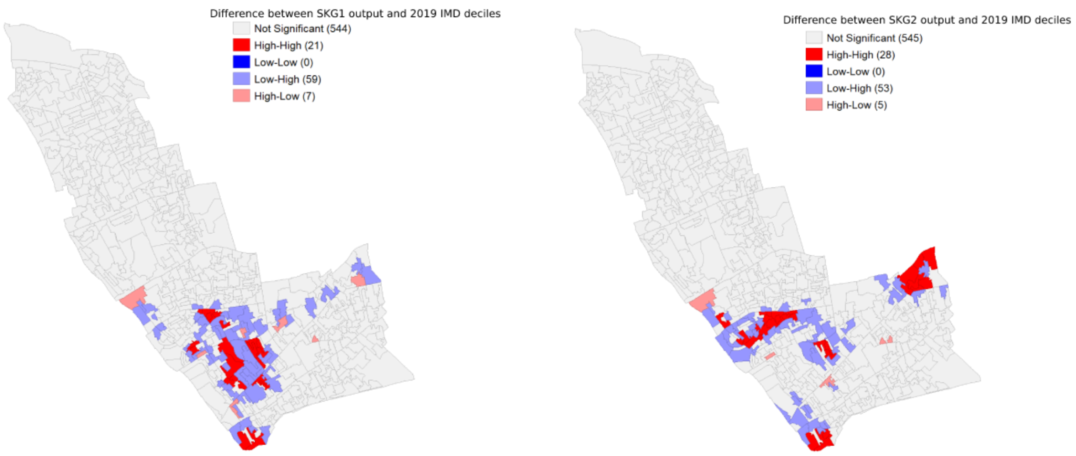
(A) Comparisons between 2019 IMD deciles and *SKG1* (90.70% Confidence Ratio) and *SKG2* (91.13% Confidence Ratio) output.

(B) Comparisons between 2019 IMD deciles and *SKG1* output of incorrectly predicted deciles. Numbers in the squares denote for how many 2019 IMD deciles are incorrectly predicted into other decile categories by SKG1.

(C) Comparisons between 2019 IMD deciles and *SKG2* output of incorrectly predicted deciles. Numbers in the squares denote for how many 2019 IMD deciles are incorrectly predicted into other decile categories by SKG2.

FIGURE 6.12: Comparisons between 2019 IMD deciles and model output.

deciles are predicted incorrectly, the difference between the deciles suggested by the model and the real deciles is relatively small. However, the errors of the predictions produced by both models present identifiable spatial clusters in Figure 6.13, which indicates that ComplEx and SKGs do not fully capture the spatialities of the data, and further research may be required in my future studies (see discussion in Section 7.4.3).



(A) Local Moran's I on the comparison between the 2019 deprivation deciles and SKG1 output.

(B) Local Moran's I on the comparison between the 2019 deprivation deciles and SKG2 output.

FIGURE 6.13: Local Moran's I on the comparison between the 2019 deprivation deciles and model output. Maps of 2019 IMD deciles and model outputs. Map boundaries source: Office for National Statistics licensed under the Open Government Licence v.3.0.

TABLE 6.2: Summarized comparisons on confidence ratio (%).

Type of Comparisons	*SKG1*(n=10)	*SKG1*(n=100)	*SKG2*(n=10)	*SKG2*(n=100)
All information	**63.32**	**77.34**	**54.68**	**73.74**
No_UGC	59.75	72.43	50.00	68.63
No_Neighbours	45.32	61.98	41.84	59.34
UGC_Neighbours	28.71	40.82	25.20	36.05
LOAC_Neighbours	33.20	50.95	30.27	47.57
2015_dep_Neighbours	40.83	53.64	38.35	51.80

As discussed above, to evaluate which factors play a greater role in the process of estimating changes in urban deprivation levels, I conducted further comparisons with simpler knowledge graphs and included information about the 2019 IMD decile for 10, and 100 randomly selected OAs. The results are summarised in Table 6.2, which indicates that the less information the knowledge graph has, the lower is the confidence ratio it can achieve. The table shows that spatial knowledge graphs with information on UGC distribution and spatial neighbours have the lowest confidence ratio. Additionally, the confidence ratios are higher when the graphs are only lacking information about Twitter and Wikipedia. Therefore, it seems to indicate that the distribution of UGC data has the least impact on inferring the dynamics of urban deprivation changes. Still, when compared with the knowledge graphs that encode all the data sources, the confidence ratios of the knowledge graphs which have no spatial neighbour information are over 10% lower than the results shown in Figure 6.2. It shows that having spatial information is crucial in the task of inferring the urban deprivation level. Comparing the results between *LOAC_Neighbours* and *2015_dep_Neighbours*, it seems to indicate that knowing the historical deprivation level has a more positive impact on inferring future deprivation deciles.

6.5.2 Results Showcase: Greater London

Number of n	SKG1	SKG2
500	73.19	69.96

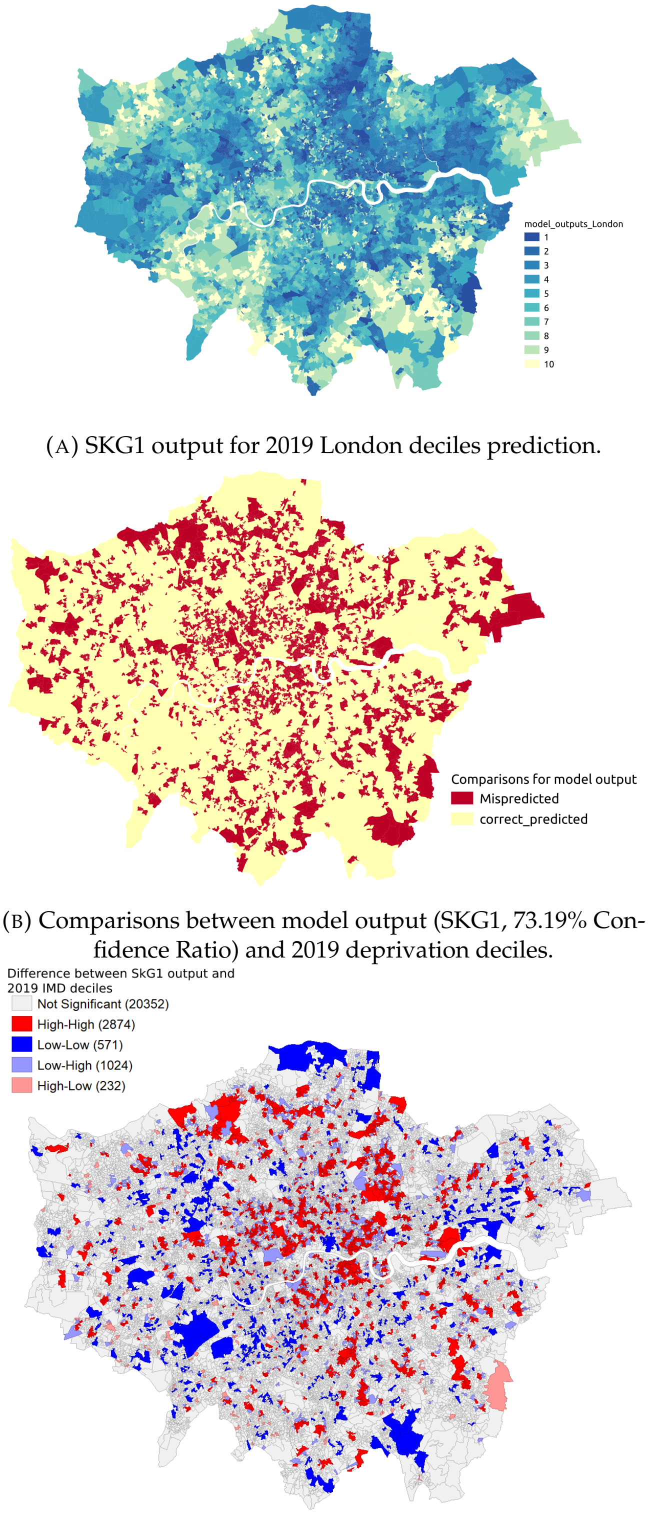
TABLE 6.3: Confidence Ratio for the experiment on the Greater London (Confidence Ratio (%)).

A second case study conducted using *SKG1* and *SKG2* is summarised in Table 6.3. Figure 6.14 and Figure 6.15 illustrate the results obtained using SKG1 on all the 25053 output areas in Greater London. The graph is constructed using the same approach as the graph constructed for Kensington and Chelsea borough, although three further LOAC classes are present (*Aging City Fringe*, *Settled Asians* and *Multi-Ethnic Suburbs*). The spatial knowledge graph is provided with the information regarding 500 randomly selected OAs and their relations between "*2019_dep*". As shown in Table 6.3, ComplEx achieves nearly 70% confidence ratio in the experiment and the actual 2019 IMD deciles map (Figure 6.7(B)) and the map showing the deciles suggested by the model (Figure 6.14(A)) present very similar patterns. Detailed visualisation of the differences is presented in Figure 6.15 (A) and (B). Although some inaccuracies are clearly visible – for example, some deciles in 3 incorrectly predicted as 4 in the model output, also some deciles in 5 incorrectly predicted as 4 – the model clearly provides a rather robust prediction. As shown in Figure 6.14 (B), similar to the results on Kensington and Chelsea, the errors of the predictions based on the model *SKG1* presents identifiable spatial clusters, which indicates that the model do not fully capture the spatialities of the data.

Overall, the two case studies presented above indicate that the proposed framework is able to provide an accurate prediction socio-demographic changes using small data samples in a scalable and robust manner.

6.5.3 Results Showcase: England

The case studies on both Kensington and Chelsea borough and Greater London have already demonstrated the scalability and robustness of my proposed knowledge graphs. The experiments have illustrated that both *SKG1* and *SKG2* have a strong capability in predicting socio-economic changes at a relatively small geographic scale (e.g., urban scale or inner-city areas). Note that there exists a significant change on *SKG2*. As introduced in Chapter 2, the theoretical background and support for the correlation between two UGC data are at the urban level in London (Ballatore and De Sabbata, 2018). Although the Chi-Square test shows that the correlation between the spatial distribution of Twitter and Wikipedia are still significant, the Pearson's correlation test r value is 0.447 which indicates a weak correlation existed between two UGC dataset, and the correlation presented weaker comparing to the spatial distribution of two datasets in London (Pearson's correlation test result: r=0.61, p<0.01). Therefore, I drop the relation "*correlation*" in *SKG2*. For the case study in this subsection, I will demonstrate our framework is scalable by predicting the 2019 IMD deciles at the national scale (England). The results are summarised in Table 6.4. By encoding 2019 IMD deciles in Greater London, Leicester and Cumbria in the knowledge

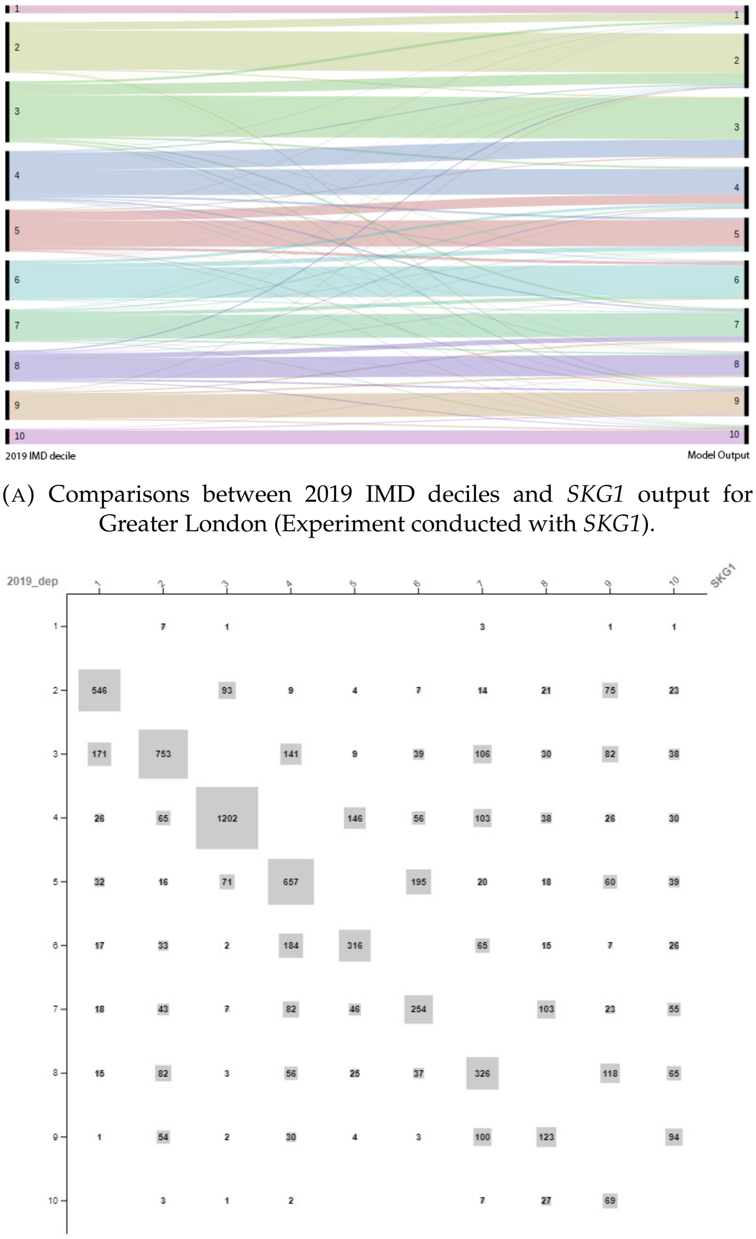


(A) SKG1 output for 2019 London deciles prediction.

(B) Comparisons between model output (SKG1, 73.19% Confidence Ratio) and 2019 deprivation deciles.

(C) Local Moran's I for the difference between 2019 IMD deciles and SKG1 output.

FIGURE 6.14: Maps of 2019 IMD deciles and model output in Greater London. Maps of 2019 IMD deciles and model outputs. Map boundaries source: Office for National Statistics licensed under the Open Government Licence v.3.0



(A) Comparisons between 2019 IMD deciles and *SKG1* output for Greater London (Experiment conducted with *SKG1*).

(B) Comparisons between 2019 IMD deciles and *SKG1* output of incorrectly predicted deciles. Numbers in the squares denote for how many 2019 IMD deciles are incorrectly predicted into other decile categories by SKG1.

FIGURE 6.15: Comparisons between 2019 IMD deciles and model output for London.

graph as part of the training data respectively, the results produced by both *SKG1* and *SKG2* are both stable and reasonably accurate. The results indicate good scalability of my framework. That seems to indicate that, as long as the framework has information regarding IMD for one city, it can predict the socio-economic changes for the rest of the country.

Interestingly, London does not provide a significant advantage despite having approximately ten times the number of LSOAs for training compared to Cumbria and the combination of Leicestershire and Rutland. That might be due to the fact

City/region	SKG1	SKG2
Greater London (4,836 LSOAs)	76.13	66.96
Leicestershire and Rutland (419 LSOAs)	77.56	68.01
Cumbria (321 LSOAs)	77.98	66.13

TABLE 6.4: Confidence Ratio for the experiment on England (Confidence Ratio (%)).

that many socio-economic patterns in London are unique and different from the rest of the UK. As such, London may do not represent the rest of England properly in a more general socio-demographic and geographic sense. Adopting LSOAs in Cumbria as training data works best for *SKG1* but worse for *SKG2*, this might because most of Cumbria are rural and SKG1 is more capable of capturing the internal associations between rural socio-demographics and deprivation conditions, thus benefiting to the country-level predictions of deprivations. Using LSOAs in Leicestershire and Rutland as training data seems to indicate the importance of having both urban and rural areas in the training data as both *SKG1* and *SKG2* achieve reasonable performance.

6.6 Discussion

While a large part of research on socio-demographic classification focuses on static representations of cities (Gale and Longley, 2013; Singleton et al., 2016), temporal modelling and predictive geodemographic classification have attracted a growing interest within the field of GIScience and quantitative geography. Most current research in that field models focuses on the analysis of temporal differences using cluster reassignments (Singleton et al., 2016; Gray et al., 2018). In this chapter, I introduce a novel framework for modelling socio-demographic changes in urban deprivation levels using a spatial knowledge graph and a link prediction process. The results show that my proposed framework can provide robust predictions with a small sample of data. As such, my approach is capable of modelling urban dynamics for the study of predictive geodemographics. The key findings are summarised as follows:

- Spatial neighbouring information of areas is a key aspect of the proposed spatial knowledge graphs for the task of IMD prediction.
- The knowledge graph (*SKG1*) following the conventional knowledge graph triples <subject, relation, object> defined in DBpedia and Wikidata achieves better results in socio-demographic prediction.
- Despite the spatial distributions of UGC (Twitter and Wikipedia) have correlations to the deprivation deciles and contribute to the prediction at some level, and my results show that UGC is not the most important factor that drives the prediction. In comparison, the results suggest that the underlying socio-demographics (here represented by LOAC and OAC) in my case studies contribute the most to a more accurate prediction.
- The knowledge approach proves to be scalable and robust when analysing larger datasets in both urban- and nation-level.

As introduced in Chapter 2, the term *Knowledge Graph* (KG) represents a collection of labelled and interlinked descriptions of entities (called triples) – real-world objects, events, situations or abstract concepts, where such descriptions have a formal structure that allows both people and computers to process them in an efficient and unambiguous manner. Taking geographic information into account, KG has played a vital plays in answering geographic research question from various perspectives, such as geographic knowledge graph completion (Qiu et al., 2019), geographic ontology alignment (Zhu et al., 2016), geographic question answering (Mai et al., 2019; Mai et al., 2020), etc.. Nevertheless, how to encode geographic knowledge (i.e., locations) into a knowledge graph remains as a domain-specific challenge and is still in the exploratory stage (Wang et al., 2019b; Yan, 2019).

As discussed in Section 6.1, many machine learning models describing urban change do not explicitly embed space in their models. However, the spatial differences of the 2015 and 2019 deprivation deciles are spatially autocorrelated in all case study areas. Such correlations indicate the fact that if one neighbourhood's deprivation level is improved, it is likely that its neighbours have also improved. That seems to indicate that embedding spatial neighbouring information within the model can significantly benefit the prediction of the deprivation changes. In this study, the geographic information is encoded as the spatial adjacency of areas defined by Queen contiguity-based spatial weights. The results presented in the ablation studies strongly suggest the importance of such geographic components in the graph and the developed knowledge graph-based framework, and they also support what Reades et al. (2019) suggested that the addition of a spatial component to a quantitative urban analysis model would likely improve the output. Explorations on other options which determine the spatial structures of areas in the knowledge graphs form the basis of future research objectives in analysing boundary effects on the prediction qualities (Reades et al., 2019).

This study focused on three case studies in the Kensington and Chelsea borough of London and Greater London and England. For the first two case study areas, I construct spatial knowledge graphs using various sources of official statistics data, including London Output Area Classification and the UK Indices of Deprivation 2015, as well as and volunteered geographic information, incorporating the distribution of geotagged Twitter data and Wikipedia articles in the process. I demonstrate that my proposed framework is able to predict socio-demographic changes with two different spatial knowledge graph structures (*SKG1* and *SKG2*), in particular the UK Indices of Deprivation 2019, with a small sample of data. I show that the link prediction algorithm on my proposed knowledge graph structures of data encoding can make an accurate prediction for about 70% of the OAs when trained using 100 over 631 OAs in Kensington and Chelsea, and 500 over 25,053 OAs in Greater London. I also illustrate how most errors are still within a reasonably small bracket and that the more is consistent and robust in its prediction when changing the number of OAs on which it is trained. The results suggest that a knowledge graph which follows a more conventional knowledge graph defining strategies (*SKG1*) is a better designing choice. Such an approach follows the common designs in conceptual modelling, and the classifications of places in different datasets are represented as entities to give additional manipulation capabilities to the modelling approach.

One of the novelties of this study mentioned at the beginning of this chapter has been to combine official spatial statistics with place representations described through UGC to understand and predict the dynamic changes of the socio-economic characteristics of places. In the case studies, I adopt the spatial distributions of Twitter and Wikipedia as two UGC datasets used in the graph. The ablation studies suggest that the spatial distribution of UGC data has the least impact on the process of predicting deprivation levels, although their use still provides a very significant advantage in making the prediction. Instead, the use of the LOAC (representing the underlying socio-demographics of the area) has the most significant impact on the deprivation level predictions. These results corroborate the findings from various sociological studies which also demonstrate the potential of Local Authority geodemographic classifications as valuable alternative tools for targeted neighbourhood's socio-economic interventions in England (Petersen et al., 2011; Wami et al., 2019), and the added information from UGC provides a valuable overall improvement of the model.

My proposed framework in this chapter shows a great advantage when analysing studying areas at various geographical scales, where the experiments demonstrate its scalability and robustness regardless of the changes on the size and scale of the studied areas. The results achieved by the framework on the scale of Greater London and England strongly prove that the framework requires nearly no changes at all to the settings of the algorithms as well the spatially constructed graphs in the task of deprivation prediction. Thus, it has the potential to be developed into a powerful tool within geography to studies socio-demographic changes in the urban or larger geographical settings.

6.7 Summary

This chapter has proposed a novel GeoAI tool to predict urban change and highlighted two key findings. First of all, UGC is a useful proxy indicator of urban change and can be associated with official statistics to predict the cities' development. Secondly, a model which incorporates the spatial component can better predict urban change. By proposing two different ways of constructing spatial knowledge graphs, the results of the case studies indicate that my proposed framework is preferable to those a-spatial machine learning models which are commonly adopted in urban studies regarding both prediction accuracy and stability.

In next chapter, I will summarise the research findings in each analysis chapter, discuss the current imitations and challenge and point out my future research directions.

Chapter 7

Discussion

7.1 Introduction

This chapter discusses and summarises the results obtained in the analysis chapters (i.e., Chapters 4, 5 and 6), which study geotagged UGC through graph-based deep learning and machine learning methods (GCN, VGAE and ComplEx) on place understanding, location estimation and dynamic urban change prediction. Each analysis chapter targets at the different research questions proposed in Chapter 1.

The key findings and contributions of this thesis can be summarised as followed:

- the spatio-temporal component of UGC provides insights into the semantic interpretation of their content and benefit the understanding of places, the use of space, and people's experience of landscape;
- a combination of the geographic distribution of online participation and the resulted coded space with various official spatial statistics can be used as a signal in the process of socio-economic change in cities;
- *Spatial is Special*: spatial proximity is a key component in developing a viable quantitative GeoAI model to address the questions in geography.

The information created and distributed through digital platforms is now a significant source for scholars to understand the human cognition of intra-urban spatial heterogeneity (Liu et al., 2020) and the reproduction of urban spaces (Shaw and Graham, 2017). The intersections between the "code" (Dodge and Kitchin, 2004) of digital platforms and space capture the "localities" of users' everyday activities, augment spatial experiences (Elwood and Leszczynski, 2013) and shape the representations of places. This thesis provides tools and mathematical models that operationalise a conceptualisation of users' online multimedia posts (image and text) as "augmentations" (Graham et al., 2015a) of places, understood as "time-space configurations "(Agnew and Livingstone, 2011), to understand place representations carried out through users' spatial activities. Supporting the research assumption proposed by Liu et al. (2020), this thesis also proves that a combination of UGC and classic socio-demographic data provides complementary information of places; thus benefiting the understanding of their local socio-economic characteristics.

Many machine learning and deep learning models currently adopted in GIScience and quantitative geographies are a-spatial (e.g., Reades et al. (2019) and Alejandro and Palafox (2019) in gentrification studies; Huang et al. (2018) in disaster management; clustering methods in Gale (2014) and Longley and Singleton (2014) for geodemographic classifications). That is, models do not explicitly incorporate a spatial component. In this thesis, I explore the use of a spatial component encoded as graph structures and introduce various models that can be directly conducted on the graph generalisations of the spatial data. This thesis highlights the importance of incorporating a spatial component in the machine learning or deep learning models; thus, setting forth the future research directions and objectives to devise spatial models. It is important to emphasise that in the research carried on in this thesis, all the datasets that are used were example datasets, which were used to demonstrate the feasibility of the frameworks. The aim was to create flexible frameworks that one can easily modify and apply to answer specific research questions in the broader field of geography.

Chapter 4 and 5 mainly focus on geotagged (geolocated and placed) multimedia UGC (text and image) analysis to answer research questions **RQ1-3**, meanwhile, Chapter 6 utilises the spatial distribution of geotagged social media along with geotagged Wikipedia data and official spatial statistics (i.e., English Indices of Multiple Deprivation and London output area classification) to answer the research question **RQ4**. This chapter highlights the key findings that emerged during this thesis, stressing the contribution to the existing body of knowledge within GIScience, quantitative geography and digital geographies. The chapter also highlights the summary findings related to the various research questions, the limitations encountered in this research, and finally, some areas of possible future work.

7.2 Answering Research Questions

In this subsection, I will provide a detailed summary of my research findings which help answering research questions *RQ1-4* proposed in Chapter 1.

7.2.1 Understanding Places through Users' Spatial Activities

- **RQ1:** *How can we combine information extracted from both text and images from multimedia UGC to better understand places through users' spatial activities in a given geographical area?*

- **RQ2:** *How can spatial or spatio-temporal distributions of UGC benefit our understanding of places and their representations?*

RQ1 and **RQ2** are jointly answered in Chapter 4. Based on the experiments and results presented in Chapter 4, I argue that by extracting combined representations from images and text from social media posts using my proposed multi-modal autoencoder, together with properly encoded spatial or spatio-temporal information, the proposed deep learning framework has the ability to categorise users' spatial activities despite the noisy and imbalanced dataset. The proposed framework contributes to the studies of digital geographies with a useful tool that provides a quantitative mean to study users' activities; thus, it can lead to a better understanding of how users' spatial experiences shape and augment the place representations when studying with a huge volume of data which traditional qualitative methods are unable to handle.

By implementing a multi-modal autoencoder to extract combined representations from images and text on social media platforms, and a graph convolutional network on spatially and spatio-temporally constructed graphs to categorise users' spatial activities, I introduce a semi-supervised learning framework based on geographic adjacency networks to categorise users' online activities based on their multimedia content on the case study platform Twitter. Based on the literature, where the majority of research regarding social media analysis in the field of GIScience and digital geographies focus on the text, I assume that adding in visual content can provide a more comprehensive understanding on user's online content. To test my assumption, I design and present various experiments in Chapter 4. The results show that my framework can produce good classification results with partially labelled data, even on noisy and imbalanced data such as the one used for the case study presented in Chapter 4.

To explore how to understand social media posts as "time-space configurations" (Agnew and Livingstone, 2011) and best encode spatial and spatio-temporal information in the graph, I designed various topological structures on both spatially and spatio-temporally constructed graphs. I test the spatial graphs constructed from a diverse set of a-spatial, semi-spatial, spatial and spatio-temporal graphs. The results presented in Chapter 4 show that a GCN model employing a spatially or spatio-temporally constructed graph achieves reasonably good classification accuracy and outperforms traditional machine learning approaches (SVM and Label Propagation) and two state-of-art deep neural networks (DNN and VTCNN, see, Huang et al., 2018).

The dataset presented in the proposed case study is directly downloaded from Twitter; thus, the dataset is noisy and heavily imbalanced. Despite that complication, which is intrinsic in the usage scenario under consideration, acknowledging that the tool aims to allow users to define their own categories, but uncommon compared to computer science benchmarks, the classification results are still robust. As illustrated through the fact that although model performance can be affected by the variation of the tweets in each category in the training data, classification results are generally consistent and stable. The results demonstrate the robustness of my proposed framework on a heavily imbalanced dataset, which can be most useful for studies with "live" datasets. The scalability of the framework is shown by presenting how the model trained for the case study above can be used to classify a further sample of unlabeled data using a GCN model employing a spatial graph. As discussed above, the dataset is noisy (see Section 4.2) and contains data which are challenging for human annotators to assign labels; however, the framework seems to have been assigned rather accurate labels among those defined for the case study.

7.2.2 Location Estimation

- **RQ3:** *Can the users' activity type of social media posts reflect the location of users and further benefit the understanding of place?*

RQ3 is answered in Chapter 5 which has proposed an approach to estimate geolocations of tweets based on a semantic understanding of tweets' content (i.e., activity types labelled by the framework in the previous chapter) and their spatio-topological structures. I demonstrate that the spatio-temporal patterns of the UGC can provide an insight into the location estimation process, and knowing the activity types of the social media posts in the graph can significantly increase the estimation accuracy.

As discussed in Chapter 5, the understanding of the role played by social media in exploring place representations has been so far limited by the fact that only a small percentage of social media posts are geolocated. As such, the experiment presented in Chapter 5 aimed to harness the dynamics of overall content production from multiple users in a single place to estimate the location of new non-geotagged content. Most existing studies are aimed at estimating the location of content focusing on a coarse level of prediction, and they mainly provide the predictions at the country level or city level (Lau et al., 2017) using text-based GIR methods (e.g., geoparsing). My study is akin and complementary to GIR methodology where placenames are not explicitly in the text and can benefit the understanding of places through users' spatial activities and their spatio-temporal patterns that could be related to a place of interest inside the city at the urban scale.

The proposed location estimation framework can be useful in a scenario where no placenames can be found in the text. The experimental results indicate that different places attract different types of content and prove the assumption that social media content is part of the social construction of the place, and certain places tend to be associated with a certain type of content based on the roles that those places have for the users of the platform. Therefore, this allows explore the hypothesis proposed in Chapter 5 that the semantic understanding of social media posts can contribute to the location estimation of non-geotagged content.

Geotagged social media data not only contains geolocated content (geotagged to a specific coordinate point) but also has placed posts (geotagged with a placename and an associated bounding box). Nowadays, accessing the geolocated social media data is increasingly difficult due to the increasing awareness of privacy and ethical concerns, which leaves the modelling of data only with bounding boxes a vital research objective to explore. Spatial modelling with bounding boxes are far less intuitive comparing to the modellings using coordinates as there is no information about the absolute locations of the content to construct the spatial graphs explicitly. As such, I explored the use of bounding boxes with the proposed hierarchical modelling, and each node in the graph represents a bounding box of a social media post. The GAE-based framework based on the best hierarchical modelling choice which is proposed in Chapter 5 can provide over 30% Top-*1* location estimation accuracy and over 50% Top-*10* accuracy. The hierarchical modellings focus on the nature of the hierarchy of the bounding boxes defined by the digital platform (Twitter in the case study). This study sets forth the future research direction for new possibilities of spatial modellings using UGC, considering the increasing difficulty of accessing geolocated information of social media posts.

7.2.3 Dynamic Socio-demographic Prediction

- **RQ4:** *How can the distribution of UGC benefit the modelling of urban socio-demographic change and inform our understanding of places?*

To answer **RQ4**, Chapter 6 explores the use of knowledge graphs to model urban socio-demographic change using various datasets including geodemographic classifications (OAC and LOAC), UK Indices of Multiple Deprivation (IMD), and distribution of geolocated Twitter data and Wikipedia articles in England with the adopted machine learning algorithm (i.e., ComplEx). The results highlight that a combination of the geographic distribution of online participation and the resulted coded space with various official spatial statistics can be used as a signal in the process of socio-economic change in cities.

The socio-demographic classification has a longstanding history in being used to describe places using various spatial statistics at defined (physical) spaces. While the majority of research on socio-demographic classification focuses on static representations of cities, temporal modelling and predictive geodemographic classification have attracted a growing interest within the field of GIScience. In the study of gentrification and urban dynamics, most of the models currently discussed seem to be fairly a-spatial (e.g., Reades et al., 2019; Alejandro and Palafox, 2019). However, urban development can be a spatial process, whereby once an area gentrifies, neighbouring areas might be affected by that gentrification process independently or in conjunction with other factors.

A key, novel aspect of my approach is the use of VGI data (Twitter and Wikipedia articles) in modelling urban change. Compared to the conventional official spatial statistics, VGI data are often huge in volume and high in velocity. Such characteristics may enable VGI to capture the dynamics of neighbourhood change that are not easily understood with decadal censuses. Ballatore and De Sabbata (2019) identified a number of associations between the spatial distributions of VGI data and socio-demographic characteristics of urban areas. Reades et al. (2019) also suggested VGI can be a critical complementary benefiting the understanding of urban development. On such basis, in Chapter 6, I start from the assumption that the spatial concentration of VGI data can help predict urban socio-demographic changes. That is, a high concentration of VGI data correlate with less deprived areas.

Taking geographic information into account, the study area of knowledge graph has played a vital play in answering geographic research questions from various perspectives, such as geographic knowledge graph completion (Qiu et al., 2019), geographic ontology alignment (Zhu et al., 2016), geographic question answering (Mai et al., 2019; Mai et al., 2020), etc.. This study explores the use of knowledge graphs encoding various geographic information (spatial neighbouring information, geodemographics (OAC and LOAC), IMD and spatial patterns of VGI) in the task of deprivation deciles prediction using three cases studies in the Kensington and Chelsea borough of London, Greater London and England. I demonstrate that my proposed framework can predict socio-demographic changes with a high accuracy (over 70% accuracy with only 15% sample data) and also illustrate how most errors are still within a reasonably small bracket and that the more is consistent and robust in its prediction when changing the number of OAs. The experiment conducted on Greater London and England demonstrates the robustness and scalability of the proposed framework on a larger dataset. It shows that my proposed framework has the potential to be developed into a useful tool predicting socio-demographic changes within the study of GIScience.

Furthermore, I explore which factors have the most significant impact on the process of inferring the dynamics of urban deprivation levels. As expected, I show that the more information the knowledge graph has, the more accurate the predictions are. The results suggest that spatial neighbouring information is crucial for my proposed framework, which highlights the geographic nature of the phenomenon under study, whereby urban development can be a spatial process, and many of the variables usually are spatially autocorrelated. Once an area gentrifies, neighbouring areas might be affected by that gentrification process independently or in conjunction with other factors. The study demonstrates the fact that *spatial is special*, where the addition of a spatial component to the model would likely improve the output.

7.3 Conclusions

This section starts with answering the main research objective pursued in this dissertation was proposed in Chapter 1:

- *How can the use of content production of UGC inform our understanding of place representations and socio-economic characteristics?*

This thesis has proposed and tested three GeoAI frameworks to answer the main research question. Through the introduction of two graph convolutional neural network-based frameworks which can incorporate the spatial as well as temporal information of UGC, I demonstrate that by explicitly taking into account the spatio-temporal relations between posts into the deep learning models, it allows us to go beyond the geotag (Crampton et al., 2013) and better understand how users' spatial activities shape the place representations. Thus, this thesis suggests closer attention to the spatio-temporal variations of the social and spatial processes carried out by the UGC content production and emphasises the importance of developing spatio-temporally-awareness frameworks or algorithms when quantitatively studying place representations with UGC in the field of digital geographies and GIScience. The studies on the knowledge graph-based framework investigate the possibilities offered by UGC in the field of GIScience to study city development and highlight that a combination of the use of UGC and conventional official spatial statistics can be a useful approach in studying how online participation impact the understanding of the urban dynamics and neighbourhood change.

This thesis makes four contributions in four different fields of geography: first of all, from a GIScience perspective, it provides frameworks with higher classification and prediction accuracy but requiring fewer sample data, thus, contributing to an advanced framework to summarise spatial characteristics of places. Secondly, from a digital geographies perspective, it shows that multimedia content provides rich information regarding places, the use of space, and people's experience of the landscape; thus, benefiting a better understanding of place representations. Thirdly, this thesis illustrates that the spatial patterns of UGC can be adopted as a useful proxy to understand urban development and neighbourhood change. Finally, this thesis reinforces the concepts that *Spatial is Special*. Spatial processes are commonly spatially autocorrelated while the mainstream of machine learning methods does not explicitly incorporate the spatial or spatio-temporal component to address such a speciality of spatial data. This thesis highlights the importance of explicitly incorporating spatial or spatio-temporal components in the models in geographical analysis and devise various frameworks suitable for spatial analysis.

The quantitative spatial-explicit graph-based frameworks in this thesis contribute to significant technological advancements to the study of digitally-mediated place representations using UGC. As introduced in Chapter 4, given the increasing popularity of image-focused online platforms (e.g., Instagram, Flickr), visual content can also provide rich information regarding places, the use of space, and people's experiences of landscape. This thesis bridges the gap between quantitative textual processing and visual content analysis, which provides a tool to classify large volumes of multimedia UGC that is unrealistic to process manually, based on a set of predefined labels tailored to a specific project or task.

Furthermore, instead of merely focusing on content analysis of geotagged UGC (e.g., Huang et al., 2018), the studies in this thesis demonstrate the importance of understanding the content geographically as well as temporally. The experimental results in the case studies reinforce the concept that spatial as well as temporal information, are two essential components when studying place representations (Agnew and Livingstone, 2011). As introduced in Chapter 2, given the fact that most studies in digital geographies emphases in-depth, often ethnographic and more qualitative research methods, this thesis opens up new research objectives towards quantitative digital geographies by developing interdisciplinary methodologies working across the quantitative and qualitative realms (Sui and DeLyser, 2012) to understand the potential complementary value of the GeoAI approach to UGC studies.

Another important implication of this thesis is its ability to underpin a novel analysis of urban dynamics. The socio-spatial structure of cities and metropolitan areas changes over time facing the rapid development of urbanisation and the increasing demands of understanding socio-economic structures of society. However, the majority of socio-demographic data are commonly collected periodically (e.g., census). UGC is an important new and fast-growing source of information, it not only has become one of the proxies to study the digital representation of the cities based on the digital social practices produced online but also drove our understanding of the modern socio-spatial structures at the urban scale. As mentioned in Chapter 2 and Chapter 6, UGC can be considered as the proxy to understand the correlations between the spatial patterns of UGC and the underlying city's socio-demographics at the urban scale, this thesis highlights that UGC data would be well served by combining it with other data sources such as census and deprivation index to understand and predict the urban changes. Such an implication strongly supports and justifies what Crampton et al. (2013) suggested that ancillary datasets would be essential when utilising UGC and big data methods to study socio-spatial changes or phenomenons.

By encoding space and time into deep learning and machine learning models, the performances of the proposed models in this thesis are superior to the traditional models that are a-spatial. The proposed models may have an essential impact on geographical studies such as Reades et al. (2019) which aims to use machine learning to study urban changes or Huang et al. (2018) which targets to apply deep learning methods for the multimedia content analysis of geotagged social media data. The incorporation of the spatio-temporal element may significantly improve the model accuracy and contextualise the algorithms being suitable for the spatial analysis. This thesis sets forth future research directions as well as highlighting the importance of devising geographically-aware machine learning or deep learning tools within quantitative geographical studies.

As a demonstration of the capabilities of UGC in the context of place representation and urban studies, this thesis can be a useful marker of the need for a rapprochement across the "qualitative-quantitative divide" (DeLyser and Sui, 2013). This thesis is not claiming to have fully explained or "solved" the problem of urban dynamics and digital representations of the places, nor am I suggesting that the quantitative approaches and models proposed supersedes the intensive, on-the-ground or survey-based work undertaken by so many before.

In summary, this thesis opens a new "front" in geographical studies aiming to understand the place and its representations, by bridging UGC and the development of quantitative spatial-explicitly GeoAI methods. I hope that, in the development of quantitative models and algorithms which incorporate UGC as an essential source of information to understand place and urban dynamics, we are ultimately able to identify ways that can benefit our understanding of the online socio-spatial process in the urban context and its impact on the physical environment we are living in.

7.4 Limitations

This section has the purpose of discussing the limitations of the work presented in this thesis, pointing out possible solutions to overcome them. My efforts to explore place representations and their socio-economic characteristics consist of three different graph-based frameworks: a GCN-based framework for understanding place representations, a graph autoencoder framework for location estimation and a knowledge graph urban change framework. The remainder of this section discusses the issues I have identified while conducting this body of research, for each aspect of my contributions summarised in the previous sections.

Before addressing the limitations regarding each analysis chapter, I will provide an introduction to the general limitations rooted in the nature of this thesis, and they will be

- *Scale*: scale is a missing point which was not explicitly discussed in Chapter 4 and 6 of this thesis. In geography, the scale has multiple referents as introduced in Mason (2001): *Cartographic scale* refers to the depicted size of a feature on a map relative to its actual size in the world. *Analysis scale* refers to the size of the unit at which some problem is analysed, such as at the county or state level. *Phenomenon scale* refers to the size at which human or physical earth structures or processes exist, regardless of how they are studied or represented. *Analysis scale* is the concerning point within the scope of this thesis, which raises the challenge of the research and leads to uncertainties in data as well as the results. Detailed discussions regarding the scale will be provided in Section 7.4.1 and Section 7.4.3.

- *Modifiable Areal Unit Problem (MAUP)*: the MAUP is statistical bias that can significantly impact the results of the hypothesis tests, and it was not explicitly discussed in Chapter 5 and 6. MAUP affects results when spatial phenomena in the form of point-based measures are aggregated into districts, and the resulting summary values (e.g., totals, rates, proportions, densities) are influenced by both the shape and scale of the aggregation unit (Wong, 2004). In this thesis, MAUP is also an inevitable issue. For example, in Chapter 5, when understanding place representations using placed UGC, it inherently requires the data aggregated in a given size of "district" (in the form of bounding boxes); in Chapter 6, the spatial distributions of Twitter and Wikipedia are aggregated into OAs and LSOAs. Future research will need to investigate such an issue in the study and devise methods to mitigate its impact on the results.

- *Uncertainties*: the uncertainties in this thesis including quality of UGC, Twitter bots, Wikipedia areas as points, uncertainties in the creation of OAC2011 and LOAC classification and IMD, age of the data (e.g., census data for creating LOAC and OAC2011 was published in 2011). Detailed discussions will be presented in the following subsections.

- *Machine learning or deep learning as black box*: despite many of existing research and the studies in this thesis have proven the utility of machine learning or deep learning approaches in spatial analysis, other researchers are criticising machine learning or deep learning for creating models that are black boxes that produce results but do not explain phenomena, and sometimes can not really be examined (Krishnan, 2019). Moreover, the uncertainties within the data can propagate through the learning process of the models and further impact the accuracy of the results (Xing and Sieber, 2018). The machine learning or deep learning models developed and adopted in this thesis still in the face of such a "black box" issue, which impact the confidence of models' output and lead to uncertainties on the results.

- *Labelling*: this limitation is relevant to Chapter 4 and 5. The labelling process of the UGC was done by me and on a small scale. A larger dataset could have been created through crowdsourcing, e.g., with Amazon Mechanical Turk. Although my proposed approaches still fit the case study scenario in digital geographies, and the experiments in the related chapters are the first test to focus on evaluating the models. Broader tests will be done in the future.

7.4.1 Understanding Place Representations

As illustrated in the related chapter and experiments, despite the fact that I prove that a graph which encodes locations, temporal information as well as distances between nodes can perform comparably the best results, such a framework still has no universal graph structure because the choice on constructing a minimum spanning tree remaining as a hyperparameter. Using 3 kilometres as the radius to construct the spatial graph using the case study dataset achieved the best results among all spatial graphs presented in Chapter 4, and the spatio-temporally constructed graph with 4 kilometres as radius achieved the best results among all comparisons. However, the choice of the radius presented in this case study is task-specific. That is, it requires further investigation into such a hyperparameter when the framework applies to a new dataset. Depending on the scale of the study area, the choice of the radius might change. It refers to the scale issue mentioned before, where the scale of the study area raises research challenges to decide the hyperparameter in the model. Thus, it over-complex the framework for the users by training on different structures and choosing the best one.

Currently, the feature extraction from the multi-modal autoencoder and the graph-based semi-supervised training are separated in two subsequent stages, which is not a so-called "end-to-end" framework. Thus, this framework currently can not be deployed as an off-the-shelf tool for digital geographies and social science researchers to use independently. Thus, it requires further engineering to simplify the complexity of the use of the framework.

7.4.2 Location Estimation

Despite the fact that the result achieved in the location estimation tasks is reasonable, the variation of the estimation accuracy is between 2% to 5% every time running the model due to the model's automatic hyper-parameter optimisation process. This is often because each time the stochastic deep learning algorithm is run on the same data, it learns a slightly different model. Therefore, the model may make slightly different predictions, and when evaluated based on error or accuracy, may have a slightly different performance. Thus, the results produced by the framework are not stable enough. Moreover, the location estimation accuracy achieved by the framework is relatively low. To achieve good results, the framework would need to be combined with complementary approaches in the field of geographic information retrieval using geoparsing or text analysis on the content.

7.4.3 Urban Change

One of the major limitations in the studies presented in Chapter 6 is the uncertainties regarding the data quality in each dataset. As discussed in Section 3.1, although I designed a 2-step process to exclude as many bots as possible in the dataset, it is still impossible to identify every bot. In future research, it requires a more sophisticated process to exclude the bots in the social media posts more accurately; for example, by applying machine learning approaches (Efthimion et al., 2018).

As discussed in Ballatore and De Sabbata (2018) and Chapter 3 of this thesis, geotagged Wikipedia areas are in the form of points pinpointed in the map whereas the decision about where to locate entities is a combination of the platform guidelines and the editors' arbitrary choices. As a result, the same entity can be pinpointed in different locations in different language editions. Such inconsistencies of geolocations of Wikipedia articles might lead to uncertainties when analysing urban change and place representations. Thus, further investigation is needed to reduce the impact of the uncertainties within the geolocated Wikipedia data.

Uncertainties are also rooted in the nature of the creation of geodemographic classifications and IMD, regarding OAC2011 and LOAC adopted in this thesis. As discussed in Slingsby et al. (2011), population profiles of geographical areas may share many or few characteristics with multiple categories, yet each is assigned to one category. Such phenomena result in heterogeneity within these categories, which varies by category and geographical region. Differences in the share of the population classified into each of the categories at a local level affects the discriminating potential of the classifiers. Although efforts (e.g., Fisher and Tate, 2015) are made to mitigate these problems, they cannot be eliminated. Also, the census data was published in 2011, and the OAC2011 and LOAC were published in 2014; thus, it is questionable whether the geodemographic classifications adopted in this thesis can capture the latest geographical phenomena. IMD also has practical issues concerning the data quality of deprivation index creation. Clelland and Hill (2019) pointed out that area-based deprivation measures risk missing out a significant number of people who experience deprivation but do not live in "deprived" areas because IMD approach fails to give much weight to deprivation which is not geographically concentrated. That is, rural areas are inherently less likely to feature amongst those ranked as most deprived.

Another important limitation in this study is relevant to the scale issues of "downscaling" and "upscaling" (Aoyama et al., 2010). In Chapter 6, I illustrated how I downscale the IMD deciles from LSOA-level to OA-level area. However, the aggregated value calculated at one area unit (LSOA) might not necessarily apply equally to all parts of the area (OA), which can lead to uncertainties in the model's output. In my future research, I will adopt the data which for the creation of IMD and create similar IMD deciles at OA-level areas to mitigate such an issue. Also, I demonstrated in Chapter 6 how I upscale OAC2011 categories from OA-level to LSOA-level areas. Such an approach I adopted was relatively naive that ignores the impact of information loss that may markedly under-represent or over-represent spatial variations in an area (Lloyd, 2016). Thus, as suggested by Aoyama et al. (2010), I will adopt more sophisticated methods such as statistical analysis to help with the upscaling of the OAC2011 categories in my future research.

In my current implementation of the spatial knowledge graphs, the information regarding geolocated UGC that encoded in the graph is only taking into account their spatial distribution and counts in each area. More sophisticated modelling can be adopted for the construction of graphs, such as UGC content analysis, social network analysis, etc.

As discussed in Section 6.5, the errors produced by the framework present spatial clusters, which indicates that the ComplEx and proposed SKGs do not fully capture the full spatialities of the data. Such an issue may be the key aspect to further improve the performance of the proposed method. Further modelling about how to better incorporate the correlation between datasets and spatial autocorrelation within each dataset will be a fundamental research objective that will enhance the model performance and help the method to produce more robust results.

The knowledge graph represented here lacks what are known to be the socio-demographic correlates of urban neighbourhoods. OAC2011 and LOAC and IMD adopted in this thesis are effectively proxy variables rather than measures directly linked to urban change (e.g., number of people displaced). Typical urban socio-spatial models have many more "layers" (nodes here) of data that are often available at fine spatial and temporal granularity, like cars, bikes, pedestrian traffic, property values, crimes, etc.. By including complimentary descriptors based on these data, the performance of the framework is expected to be increased.

While some of these limitations have a manageable solution (e.g., encoding complementary information in the knowledge graph), others present more profound challenges, approachable via more extensive research efforts, such as performing exclusive explorations on the possibilities for graph representations of social media posts. My contributions to place representation understanding within digital geographies should be considered as promising starting points to provide practical quantitative tools in real-world applications, enabling a more effective usage of the information generated through UGC and VGI. The next section outlines my plans for future work, pointing out directions that I deem promising.

7.5 Outlook

This thesis contributes to the discipline of digital geographies, GIScience and GeoAI by providing researchers novel deep learning methodologies and frameworks to quantitatively and efficiently study the digitally coded space and place representations with geotagged UGC data and to understand and predict socio-economic changes within areas combining UGC with survey-based (e.g., census) official spatial statistics. Advancing this research area may have positive outcomes for a number of advanced "real-world" geographic applications and studies, for which the interpolations of UGC are key.

Although my studies mainly adopt Twitter for case studies, it is important to mention that the adopted and developed approaches can apply to any other crowd-sourced datasets, in which content generation can be considered as a social creation process interacting with the use of space. These contributions not only constitute semantic support for social media studies but also can be extended in several promising directions. Among many other possible directions for future research, I identify the following as particularly important:

- In the scope of my research proposed in this thesis, each social media post are only labelled into one category of user's behaviour. However, social media posts could be classified into multiple categories. For example, a social media post regarding *Food* might also be classified into *Social* when the post is taken during a party. As such, there exist potential uncertainties in the semantic understanding of users' activities by forcing a "one-to-one" match between the social media post and defined categories. In my future research, I am interested in exploring a fuzzy logic classification, where multiple labels can be attached to each tweet. This approach would be well suited with case studies such as the one presented above in the field of digital geographies, where frequently more than one label can be attached to a single piece of text or image during qualitative content analysis. Such a fuzzy logic approach has the potential to mitigate the issue caused by human labelling process as one social media posts can be labelled from different perspectives.

- The location estimation framework is currently based on just the semantic labels of each social media post. Although those labels are categorised on multimedia content in the case study, they are not directly using the features extracted from multimedia content. Future research will aim to expand the graph autoencoder model presented in this thesis, substituting the labels with content features extracted through the multi-modal autoencoder also proposed in this thesis, to achieve a more sophisticated and automated understanding of the semantic content of social media posts.

- The graph representation for placed social media posts in Chapter 5 can be benefited by encoding other soico-spatial information in the graph, such as following-followers network as mentioned in Liu and Huang (2016); or to encode the survey-based data to help with understanding the places by exploring spatial clusters of activities, and further benefiting the location estimation accuracy.

- As mentioned in the previous section, in my current implementation of the spatial knowledge graphs, I used OAC2011 and LOAC and IMD in this thesis. They are effectively proxy variables rather than measures directly linked to urban change (e.g., number of people displaced). In my future research, I will aim to work directly with socio-demographic variables to model urban changes.

Among all above-mentioned research goals, improving the location estimation framework will stand at the centre of my future research. As mentioned in this thesis, accessing the geolocated UGC content is now increasingly difficult. A better location estimation framework will be crucial in understanding the place representations through users' lived experiences using my activity classification framework, as well as benefiting the predictions of my knowledge graph framework on dynamic urban changes.

Appendix A

Showcase of Multi-modal Autoencoder

The proposed multi-modal autoencoder was published as:

- **Pengyuan Liu and Stefano De Sabbata**, *2019*. Learning Digital Geographies through a Correlationbased Autoencoder. In *GeoAI and Deep Learning Symposium: Geo-Text Data and Location-based Social Media, American Association of Geographers Annual Meeting 2019*[1].

The overview of the proposed multi-modal autoencoder is shown in Figure A.1. To illustrate the usefulness of the proposed multi-modal autoencoder, I evaluate the model on a dataset extracted of social media posts collected through the Twitter API between 7th May, 2018 and 20th May, 2018 in the British Isles. I select tweets containing images, texts, and exact coordinates, and limit the amount of tweets per account to 10, obtaining a dataset of 2876 tweets.

Based on the assumption that there is correlation between the contents of the posts (image and text) and their corresponding geolocations, I perform kernel canonical-correlation analysis (CCA) on the created dense, numeric representation from each tweet and its corresponding geo-location. I first employed principal component analysis (PCA) to reduce the dimension of the extracted representations from 399424 to 98 with 98.2% information preserved, and then t-Distributed Stochastic Neighbor Embedding (t-SNE) to reduce the dimension from 98 to 2. The Kernel CCA algorithm is thus used to create new 2-dimensional representations. Finally, agglomerative hierarchical clustering algorithm is used to cluster the newly created representations as 42 different clusters.

A preliminary analysis of the results in this paper seems to indicate that the proposed model with an additional correlation analysis with geo-location information has the ability to capture both content similarities of images and texts and geographical closeness.

[1] Code to reproduce this framework is available at: https://github.com/PengyuanLiu1993/PhD_Thesis_Codes_PengyuanLiu/tree/master/GCN_Activities_Classification/Multi-Modal%20Autoencoder

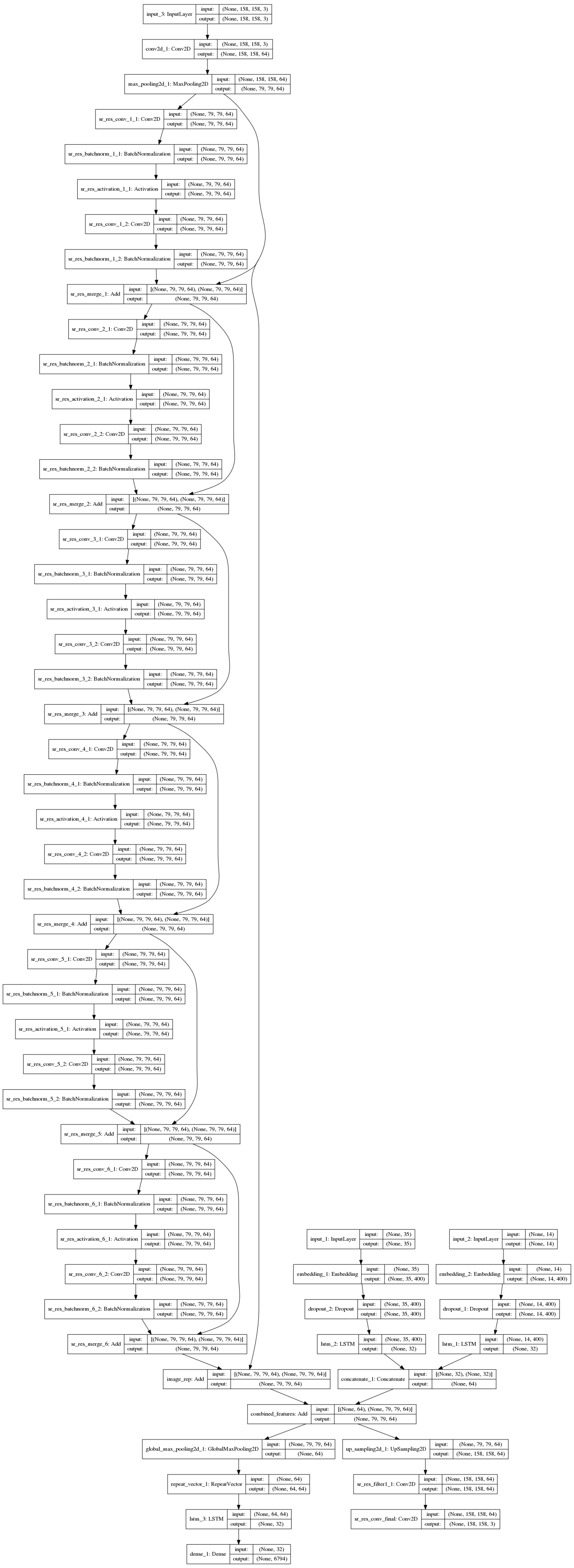
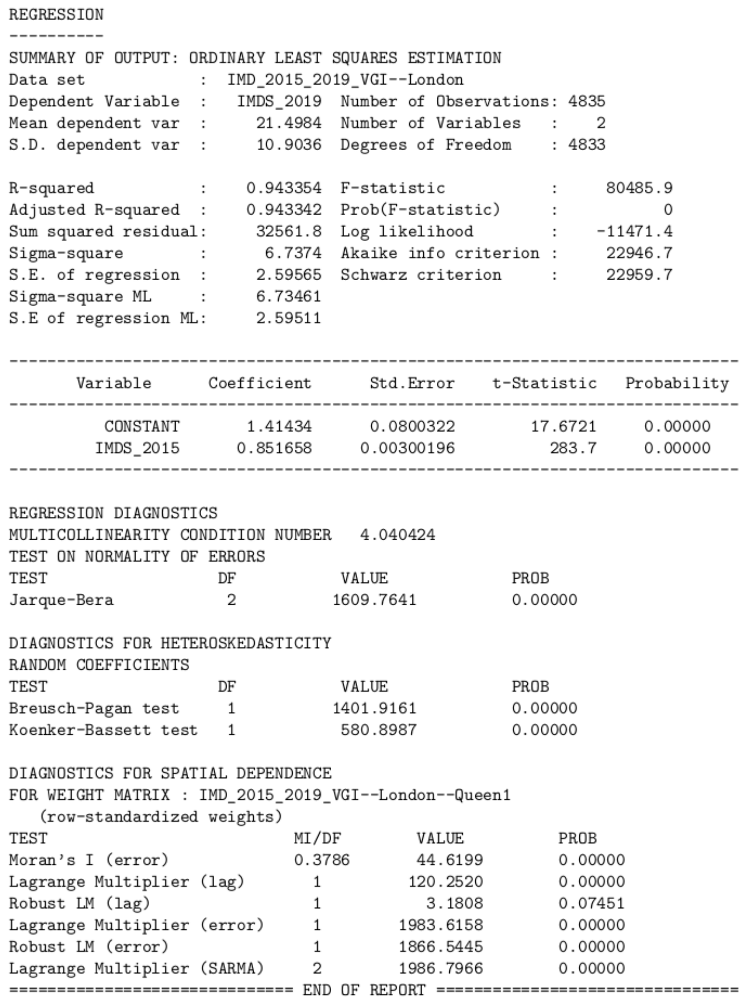

input_3: InputLayer input: (None, 158, 158, 3) output: (None, 158, 158, 3)
conv2d_1: Conv2D input: (None, 158, 158, 3) output: (None, 158, 158, 64)
max_pooling2d_1: MaxPooling2D input: (None, 158, 158, 64) output: (None, 79, 79, 64)
sr_res_conv_1_1: Conv2D input: (None, 79, 79, 64) output: (None, 79, 79, 64)
sr_res_batchnorm_1_1: BatchNormalization input: (None, 79, 79, 64) output: (None, 79, 79, 64)
sr_res_activation_1_1: Activation input: (None, 79, 79, 64) output: (None, 79, 79, 64)
sr_res_conv_1_2: Conv2D input: (None, 79, 79, 64) output: (None, 79, 79, 64)
sr_res_batchnorm_1_2: BatchNormalization input: (None, 79, 79, 64) output: (None, 79, 79, 64)
sr_res_merge_1: Add input: [(None, 79, 79, 64), (None, 79, 79, 64)] output: (None, 79, 79, 64)
sr_res_conv_2_1: Conv2D input: (None, 79, 79, 64) output: (None, 79, 79, 64)
sr_res_batchnorm_2_1: BatchNormalization input: (None, 79, 79, 64) output: (None, 79, 79, 64)
sr_res_activation_2_1: Activation input: (None, 79, 79, 64) output: (None, 79, 79, 64)
sr_res_conv_2_2: Conv2D input: (None, 79, 79, 64) output: (None, 79, 79, 64)
sr_res_batchnorm_2_2: BatchNormalization input: (None, 79, 79, 64) output: (None, 79, 79, 64)
sr_res_merge_2: Add input: [(None, 79, 79, 64), (None, 79, 79, 64)] output: (None, 79, 79, 64)
sr_res_conv_3_1: Conv2D input: (None, 79, 79, 64) output: (None, 79, 79, 64)
sr_res_batchnorm_3_1: BatchNormalization input: (None, 79, 79, 64) output: (None, 79, 79, 64)
sr_res_activation_3_1: Activation input: (None, 79, 79, 64) output: (None, 79, 79, 64)
sr_res_conv_3_2: Conv2D input: (None, 79, 79, 64) output: (None, 79, 79, 64)
sr_res_batchnorm_3_2: BatchNormalization input: (None, 79, 79, 64) output: (None, 79, 79, 64)
sr_res_merge_3: Add input: [(None, 79, 79, 64), (None, 79, 79, 64)] output: (None, 79, 79, 64)
sr_res_conv_4_1: Conv2D input: (None, 79, 79, 64) output: (None, 79, 79, 64)
sr_res_batchnorm_4_1: BatchNormalization input: (None, 79, 79, 64) output: (None, 79, 79, 64)
sr_res_activation_4_1: Activation input: (None, 79, 79, 64) output: (None, 79, 79, 64)
sr_res_conv_4_2: Conv2D input: (None, 79, 79, 64) output: (None, 79, 79, 64)
sr_res_batchnorm_4_2: BatchNormalization input: (None, 79, 79, 64) output: (None, 79, 79, 64)
sr_res_merge_4: Add input: [(None, 79, 79, 64), (None, 79, 79, 64)] output: (None, 79, 79, 64)
sr_res_conv_5_1: Conv2D input: (None, 79, 79, 64) output: (None, 79, 79, 64)
sr_res_batchnorm_5_1: BatchNormalization input: (None, 79, 79, 64) output: (None, 79, 79, 64)
sr_res_activation_5_1: Activation input: (None, 79, 79, 64) output: (None, 79, 79, 64)
sr_res_conv_5_2: Conv2D input: (None, 79, 79, 64) output: (None, 79, 79, 64)
sr_res_batchnorm_5_2: BatchNormalization input: (None, 79, 79, 64) output: (None, 79, 79, 64)
sr_res_merge_5: Add input: [(None, 79, 79, 64), (None, 79, 79, 64)] output: (None, 79, 79, 64)
sr_res_conv_6_1: Conv2D input: (None, 79, 79, 64) output: (None, 79, 79, 64)
sr_res_batchnorm_6_1: BatchNormalization input: (None, 79, 79, 64) output: (None, 79, 79, 64)
sr_res_activation_6_1: Activation input: (None, 79, 79, 64) output: (None, 79, 79, 64)
sr_res_conv_6_2: Conv2D input: (None, 79, 79, 64) output: (None, 79, 79, 64)
sr_res_batchnorm_6_2: BatchNormalization input: (None, 79, 79, 64) output: (None, 79, 79, 64)
sr_res_merge_6: Add input: [(None, 79, 79, 64), (None, 79, 79, 64)] output: (None, 79, 79, 64)
input_1: InputLayer input: (None, 35) output: (None, 35)
input_2: InputLayer input: (None, 14) output: (None, 14)
embedding_1: Embedding input: (None, 35) output: (None, 35, 400)
embedding_2: Embedding input: (None, 14) output: (None, 14, 400)
dropout_2: Dropout input: (None, 35, 400) output: (None, 35, 400)
dropout_1: Dropout input: (None, 14, 400) output: (None, 14, 400)
lstm_2: LSTM input: (None, 35, 400) output: (None, 32)
lstm_1: LSTM input: (None, 14, 400) output: (None, 32)
image_rep: Add input: [(None, 79, 79, 64), (None, 79, 79, 64)] output: (None, 79, 79, 64)
concatenate_1: Concatenate input: [(None, 32), (None, 32)] output: (None, 64)
combined_features: Add input: [(None, 64), (None, 79, 79, 64)] output: (None, 79, 79, 64)
global_max_pooling2d_1: GlobalMaxPooling2D input: (None, 79, 79, 64) output: (None, 64)
up_sampling2d_1: UpSampling2D input: (None, 79, 79, 64) output: (None, 158, 158, 64)
repeat_vector_1: RepeatVector input: (None, 64) output: (None, 64, 64)
sr_res_filter1_1: Conv2D input: (None, 158, 158, 64) output: (None, 158, 158, 64)
lstm_3: LSTM input: (None, 64, 64) output: (None, 32)
sr_res_conv_final: Conv2D input: (None, 158, 158, 64) output: (None, 158, 158, 3)
dense_1: Dense input: (None, 32) output: (None, 6794)

FIGURE A.1: Proposed Multi-modal autoencoder.

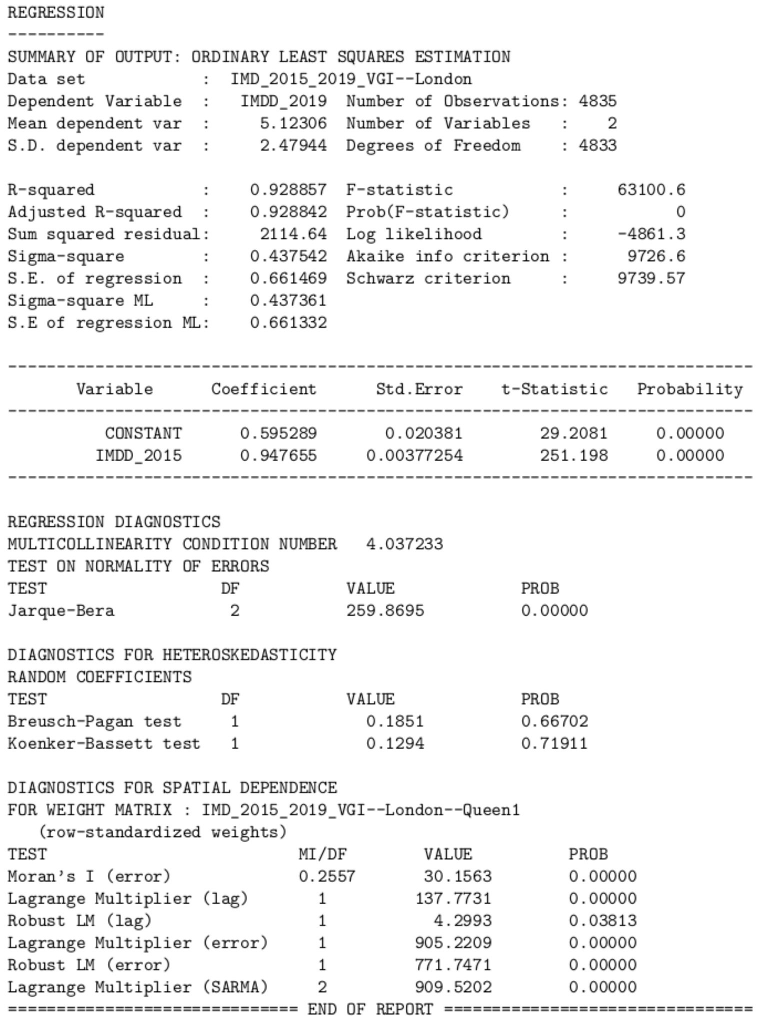
Appendix B

Urban Change and User Generated Content



```
REGRESSION
----------
SUMMARY OF OUTPUT: ORDINARY LEAST SQUARES ESTIMATION
Data set            :  IMD_2015_2019_VGI--London
Dependent Variable  :   IMDS_2019  Number of Observations: 4835
Mean dependent var  :     21.4984  Number of Variables   :    2
S.D. dependent var  :     10.9036  Degrees of Freedom    : 4833

R-squared           :    0.943354  F-statistic           :     80485.9
Adjusted R-squared  :    0.943342  Prob(F-statistic)     :           0
Sum squared residual:     32561.8  Log likelihood        :    -11471.4
Sigma-square        :      6.7374  Akaike info criterion :     22946.7
S.E. of regression  :     2.59565  Schwarz criterion     :     22959.7
Sigma-square ML     :     6.73461
S.E of regression ML:     2.59511

-----------------------------------------------------------------------------
       Variable      Coefficient      Std.Error    t-Statistic   Probability
-----------------------------------------------------------------------------
          CONSTANT       1.41434      0.0800322       17.6721      0.00000
         IMDS_2015      0.851658     0.00300196         283.7      0.00000
-----------------------------------------------------------------------------

REGRESSION DIAGNOSTICS
MULTICOLLINEARITY CONDITION NUMBER   4.040424
TEST ON NORMALITY OF ERRORS
TEST                  DF           VALUE            PROB
Jarque-Bera            2         1609.7641         0.00000

DIAGNOSTICS FOR HETEROSKEDASTICITY
RANDOM COEFFICIENTS
TEST                  DF           VALUE            PROB
Breusch-Pagan test     1         1401.9161         0.00000
Koenker-Bassett test   1          580.8987         0.00000

DIAGNOSTICS FOR SPATIAL DEPENDENCE
FOR WEIGHT MATRIX : IMD_2015_2019_VGI--London--Queen1
   (row-standardized weights)
TEST                          MI/DF       VALUE          PROB
Moran's I (error)             0.3786       44.6199        0.00000
Lagrange Multiplier (lag)       1         120.2520        0.00000
Robust LM (lag)                 1           3.1808        0.07451
Lagrange Multiplier (error)     1        1983.6158        0.00000
Robust LM (error)               1        1866.5445        0.00000
Lagrange Multiplier (SARMA)     2        1986.7966        0.00000
============================= END OF REPORT ================================
```

FIGURE B.1: The a-spatial regression model using IMD 2015 deprivation scores to predict IMD 2019 deprivation scores.



```
REGRESSION
----------
SUMMARY OF OUTPUT: ORDINARY LEAST SQUARES ESTIMATION
Data set            :  IMD_2015_2019_VGI--London
Dependent Variable  :   IMDD_2019  Number of Observations: 4835
Mean dependent var  :     5.12306  Number of Variables   :    2
S.D. dependent var  :     2.47944  Degrees of Freedom    : 4833

R-squared           :    0.928857  F-statistic           :     63100.6
Adjusted R-squared  :    0.928842  Prob(F-statistic)     :           0
Sum squared residual:     2114.64  Log likelihood        :     -4861.3
Sigma-square        :    0.437542  Akaike info criterion :      9726.6
S.E. of regression  :    0.661469  Schwarz criterion     :     9739.57
Sigma-square ML     :    0.437361
S.E of regression ML:    0.661332

-----------------------------------------------------------------------------
       Variable     Coefficient      Std.Error    t-Statistic   Probability
-----------------------------------------------------------------------------
          CONSTANT     0.595289       0.020381       29.2081     0.00000
         IMDD_2015     0.947655     0.00377254       251.198     0.00000
-----------------------------------------------------------------------------

REGRESSION DIAGNOSTICS
MULTICOLLINEARITY CONDITION NUMBER   4.037233
TEST ON NORMALITY OF ERRORS
TEST                     DF            VALUE              PROB
Jarque-Bera               2            259.8695           0.00000

DIAGNOSTICS FOR HETEROSKEDASTICITY
RANDOM COEFFICIENTS
TEST                     DF            VALUE              PROB
Breusch-Pagan test        1              0.1851           0.66702
Koenker-Bassett test      1              0.1294           0.71911

DIAGNOSTICS FOR SPATIAL DEPENDENCE
FOR WEIGHT MATRIX : IMD_2015_2019_VGI--London--Queen1
   (row-standardized weights)
TEST                           MI/DF        VALUE          PROB
Moran's I (error)              0.2557       30.1563        0.00000
Lagrange Multiplier (lag)        1         137.7731        0.00000
Robust LM (lag)                  1           4.2993        0.03813
Lagrange Multiplier (error)      1         905.2209        0.00000
Robust LM (error)                1         771.7471        0.00000
Lagrange Multiplier (SARMA)      2         909.5202        0.00000
============================== END OF REPORT ===============================
```

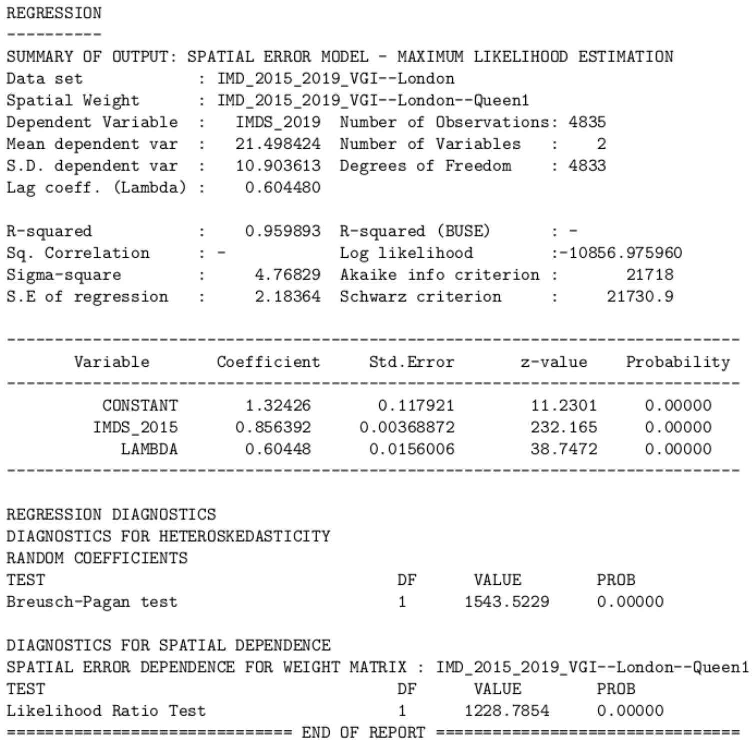
FIGURE B.2: The a-spatial regression model using IMD 2015 deprivation deciles to predict IMD 2019 deprivation deciles.

Model (Spatial Error)	Pseudo R-square	Log likelihood	AIC
Scores	0.959893	-10856.98	21718.00
Scores + Twitter	0.960059	-10838.95	21683.90
Scores + Wikipedia	0.959908	-10853.23	21712.50
Scores + Twitter + Wikipedia	0.960062	-10838.87	21685.70

TABLE B.1: Spatial lag models on IMD scores.

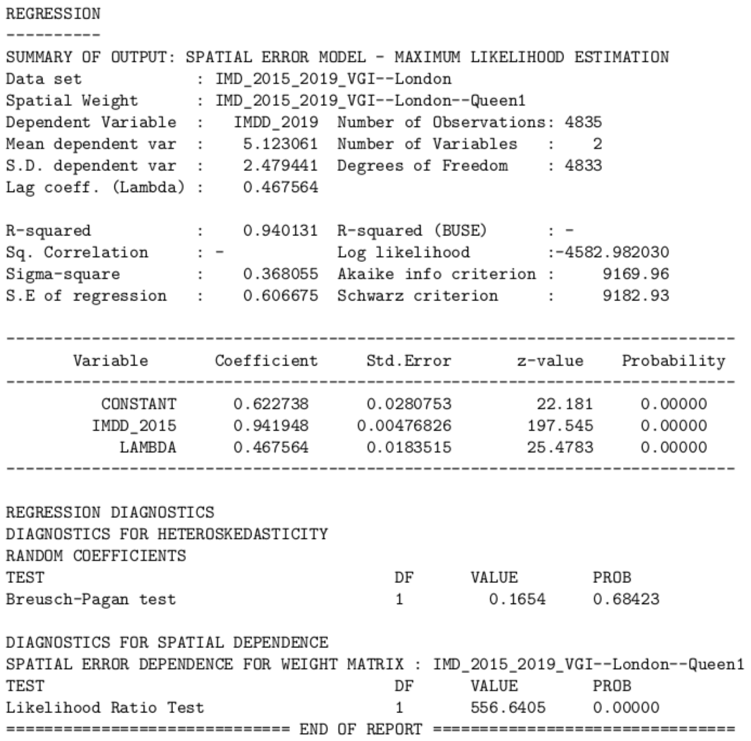
Model (Spatial Error)	Pseudo R-square	Log likelihood	AIC
Deciles	0.940131	-4582.98	9169.96
Deciles + Twitter	0.940423	-4563.29	9132.58
Deciles + Wikipedia	0.940196	-4576.52	9159.05
Deciles + Twitter + Wikipedia	0.940421	-4563.24	9134.48

TABLE B.2: Spatial lag models on IMD deciles.



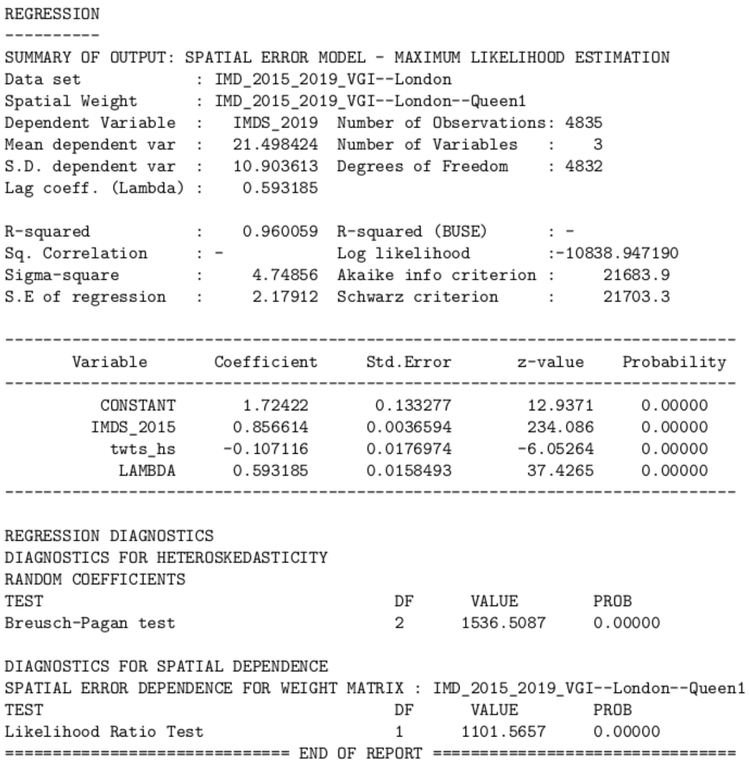
```
REGRESSION
----------
SUMMARY OF OUTPUT: SPATIAL ERROR MODEL - MAXIMUM LIKELIHOOD ESTIMATION
Data set            : IMD_2015_2019_VGI--London
Spatial Weight      : IMD_2015_2019_VGI--London--Queen1
Dependent Variable  :   IMDS_2019  Number of Observations: 4835
Mean dependent var  :   21.498424  Number of Variables   :    2
S.D. dependent var  :   10.903613  Degrees of Freedom    : 4833
Lag coeff. (Lambda) :    0.604480

R-squared           :    0.959893  R-squared (BUSE)      : -
Sq. Correlation     : -            Log likelihood        :-10856.975960
Sigma-square        :     4.76829  Akaike info criterion :       21718
S.E of regression   :     2.18364  Schwarz criterion     :     21730.9

-----------------------------------------------------------------------------
       Variable       Coefficient     Std.Error       z-value    Probability
-----------------------------------------------------------------------------
          CONSTANT       1.32426       0.117921       11.2301     0.00000
         IMDS_2015      0.856392     0.00368872       232.165     0.00000
            LAMBDA       0.60448      0.0156006       38.7472     0.00000
-----------------------------------------------------------------------------

REGRESSION DIAGNOSTICS
DIAGNOSTICS FOR HETEROSKEDASTICITY
RANDOM COEFFICIENTS
TEST                                    DF      VALUE        PROB
Breusch-Pagan test                      1     1543.5229     0.00000

DIAGNOSTICS FOR SPATIAL DEPENDENCE
SPATIAL ERROR DEPENDENCE FOR WEIGHT MATRIX : IMD_2015_2019_VGI--London--Queen1
TEST                                    DF      VALUE        PROB
Likelihood Ratio Test                   1     1228.7854     0.00000
============================== END OF REPORT ================================
```

FIGURE B.3: Spatial lag regression model using IMD 2015 deprivation scores to predict IMD 2019 deprivation scores.



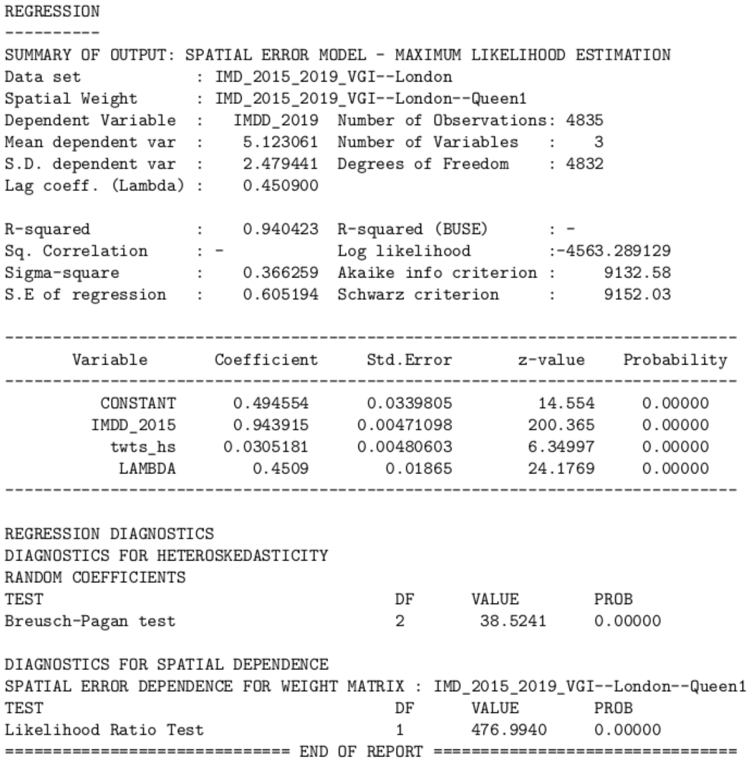
```
REGRESSION
----------
SUMMARY OF OUTPUT: SPATIAL ERROR MODEL - MAXIMUM LIKELIHOOD ESTIMATION
Data set            : IMD_2015_2019_VGI--London
Spatial Weight      : IMD_2015_2019_VGI--London--Queen1
Dependent Variable  :   IMDD_2019  Number of Observations: 4835
Mean dependent var  :    5.123061  Number of Variables   :    2
S.D. dependent var  :    2.479441  Degrees of Freedom    : 4833
Lag coeff. (Lambda) :    0.467564

R-squared           :    0.940131  R-squared (BUSE)      : -
Sq. Correlation     : -            Log likelihood        :-4582.982030
Sigma-square        :    0.368055  Akaike info criterion :     9169.96
S.E of regression   :    0.606675  Schwarz criterion     :     9182.93

-----------------------------------------------------------------------------
       Variable       Coefficient     Std.Error       z-value    Probability
-----------------------------------------------------------------------------
          CONSTANT      0.622738      0.0280753        22.181     0.00000
         IMDD_2015      0.941948     0.00476826       197.545     0.00000
            LAMBDA      0.467564      0.0183515       25.4783     0.00000
-----------------------------------------------------------------------------

REGRESSION DIAGNOSTICS
DIAGNOSTICS FOR HETEROSKEDASTICITY
RANDOM COEFFICIENTS
TEST                                    DF      VALUE        PROB
Breusch-Pagan test                      1        0.1654     0.68423

DIAGNOSTICS FOR SPATIAL DEPENDENCE
SPATIAL ERROR DEPENDENCE FOR WEIGHT MATRIX : IMD_2015_2019_VGI--London--Queen1
TEST                                    DF      VALUE        PROB
Likelihood Ratio Test                   1      556.6405     0.00000
============================== END OF REPORT ================================
```

FIGURE B.4: Spatial lag regression model using IMD 2015 deprivation deciles to predict IMD 2019 deprivation deciles.



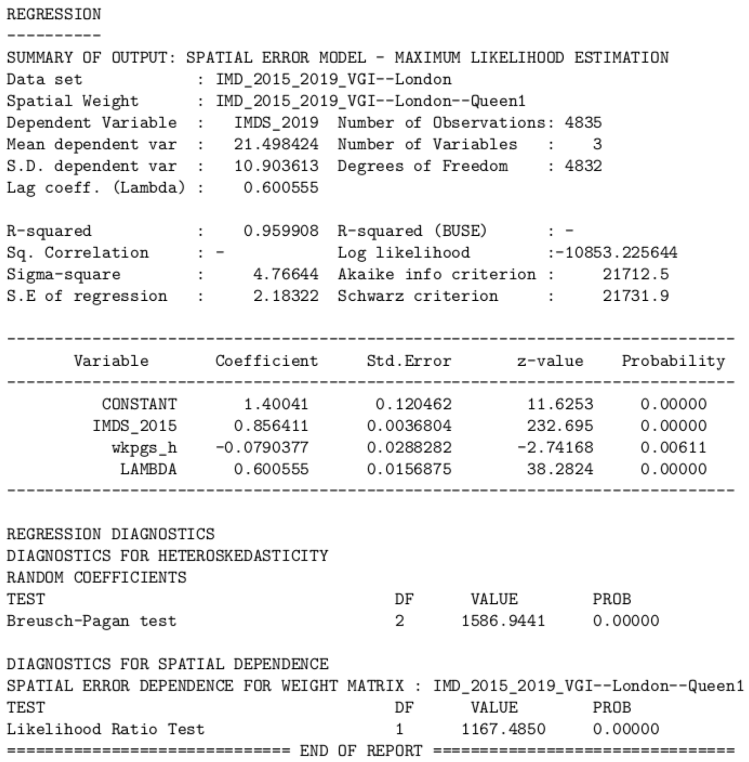
```
REGRESSION
----------
SUMMARY OF OUTPUT: SPATIAL ERROR MODEL - MAXIMUM LIKELIHOOD ESTIMATION
Data set            : IMD_2015_2019_VGI--London
Spatial Weight      : IMD_2015_2019_VGI--London--Queen1
Dependent Variable  :   IMDS_2019  Number of Observations: 4835
Mean dependent var  :   21.498424  Number of Variables   :    3
S.D. dependent var  :   10.903613  Degrees of Freedom    : 4832
Lag coeff. (Lambda) :    0.593185

R-squared           :    0.960059  R-squared (BUSE)      : -
Sq. Correlation     : -            Log likelihood        :-10838.947190
Sigma-square        :     4.74856  Akaike info criterion :     21683.9
S.E of regression   :     2.17912  Schwarz criterion     :     21703.3

------------------------------------------------------------------------------------
       Variable       Coefficient      Std.Error       z-value    Probability
------------------------------------------------------------------------------------
          CONSTANT       1.72422       0.133277        12.9371     0.00000
         IMDS_2015      0.856614      0.0036594        234.086     0.00000
           twts_hs     -0.107116      0.0176974       -6.05264     0.00000
            LAMBDA      0.593185      0.0158493        37.4265     0.00000
------------------------------------------------------------------------------------

REGRESSION DIAGNOSTICS
DIAGNOSTICS FOR HETEROSKEDASTICITY
RANDOM COEFFICIENTS
TEST                                     DF      VALUE       PROB
Breusch-Pagan test                        2     1536.5087     0.00000

DIAGNOSTICS FOR SPATIAL DEPENDENCE
SPATIAL ERROR DEPENDENCE FOR WEIGHT MATRIX : IMD_2015_2019_VGI--London--Queen1
TEST                                     DF      VALUE       PROB
Likelihood Ratio Test                     1     1101.5657     0.00000
============================= END OF REPORT ================================
```

FIGURE B.5: Spatial lag regression model Using IMD 2015 deprivation scores in combination of Twitter distribution patterns to predict IMD 2019 deprivation scores.



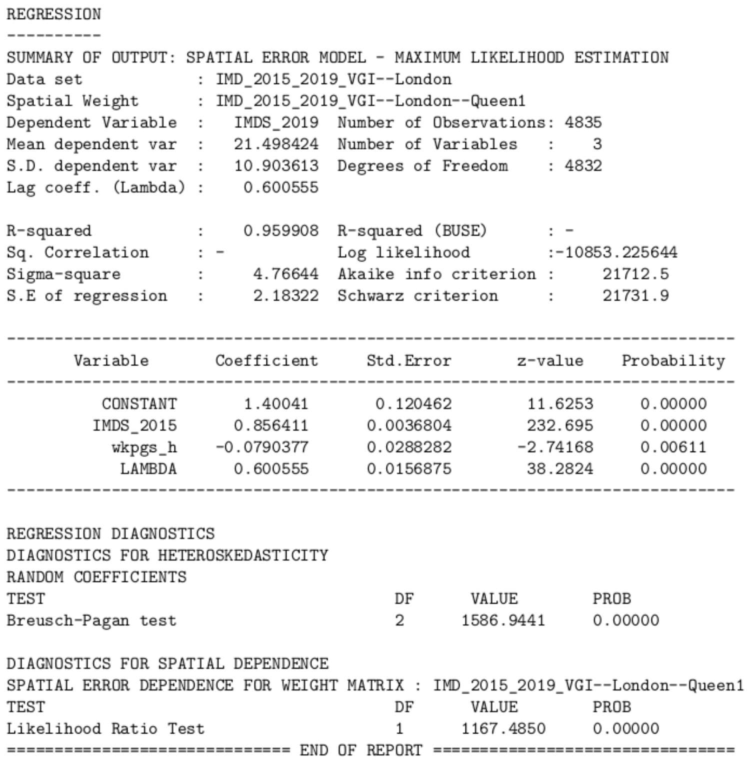
```
REGRESSION
----------
SUMMARY OF OUTPUT: SPATIAL ERROR MODEL - MAXIMUM LIKELIHOOD ESTIMATION
Data set            : IMD_2015_2019_VGI--London
Spatial Weight      : IMD_2015_2019_VGI--London--Queen1
Dependent Variable  :   IMDD_2019  Number of Observations: 4835
Mean dependent var  :    5.123061  Number of Variables   :    3
S.D. dependent var  :    2.479441  Degrees of Freedom    : 4832
Lag coeff. (Lambda) :    0.450900

R-squared           :    0.940423  R-squared (BUSE)      : -
Sq. Correlation     : -            Log likelihood        :-4563.289129
Sigma-square        :    0.366259  Akaike info criterion :     9132.58
S.E of regression   :    0.605194  Schwarz criterion     :     9152.03

------------------------------------------------------------------------------------
       Variable       Coefficient      Std.Error       z-value    Probability
------------------------------------------------------------------------------------
          CONSTANT      0.494554      0.0339805         14.554     0.00000
         IMDD_2015      0.943915     0.00471098        200.365     0.00000
           twts_hs     0.0305181     0.00480603        6.34997     0.00000
            LAMBDA        0.4509        0.01865        24.1769     0.00000
------------------------------------------------------------------------------------

REGRESSION DIAGNOSTICS
DIAGNOSTICS FOR HETEROSKEDASTICITY
RANDOM COEFFICIENTS
TEST                                     DF      VALUE       PROB
Breusch-Pagan test                        2       38.5241     0.00000

DIAGNOSTICS FOR SPATIAL DEPENDENCE
SPATIAL ERROR DEPENDENCE FOR WEIGHT MATRIX : IMD_2015_2019_VGI--London--Queen1
TEST                                     DF      VALUE       PROB
Likelihood Ratio Test                     1      476.9940     0.00000
============================= END OF REPORT ================================
```

FIGURE B.6: Spatial lag regression model Using IMD 2015 deprivation deciles in combination of Twitter distribution patterns to predict IMD 2019 deprivation deciles.



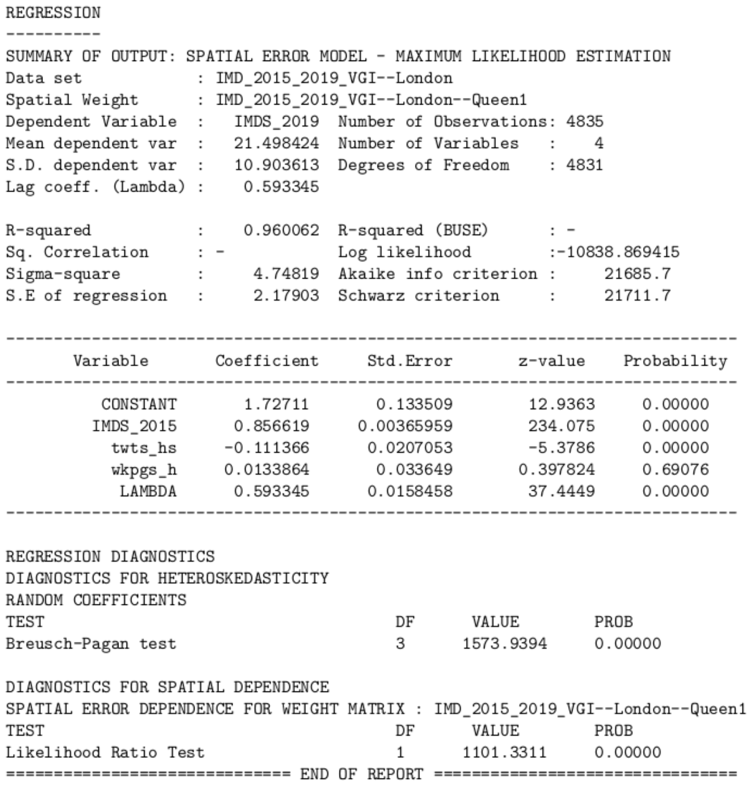
```
REGRESSION
----------
SUMMARY OF OUTPUT: SPATIAL ERROR MODEL - MAXIMUM LIKELIHOOD ESTIMATION
Data set            : IMD_2015_2019_VGI--London
Spatial Weight      : IMD_2015_2019_VGI--London--Queen1
Dependent Variable  :   IMDS_2019  Number of Observations: 4835
Mean dependent var  :   21.498424  Number of Variables   :    3
S.D. dependent var  :   10.903613  Degrees of Freedom    : 4832
Lag coeff. (Lambda) :    0.600555

R-squared           :    0.959908  R-squared (BUSE)      : -
Sq. Correlation     : -            Log likelihood        :-10853.225644
Sigma-square        :     4.76644  Akaike info criterion :     21712.5
S.E of regression   :     2.18322  Schwarz criterion     :     21731.9

----------------------------------------------------------------------------
       Variable       Coefficient     Std.Error       z-value    Probability
----------------------------------------------------------------------------
          CONSTANT       1.40041       0.120462       11.6253     0.00000
         IMDS_2015      0.856411      0.0036804       232.695     0.00000
           wkpgs_h    -0.0790377      0.0288282      -2.74168     0.00611
            LAMBDA      0.600555      0.0156875       38.2824     0.00000
----------------------------------------------------------------------------

REGRESSION DIAGNOSTICS
DIAGNOSTICS FOR HETEROSKEDASTICITY
RANDOM COEFFICIENTS
TEST                                     DF      VALUE        PROB
Breusch-Pagan test                       2      1586.9441     0.00000

DIAGNOSTICS FOR SPATIAL DEPENDENCE
SPATIAL ERROR DEPENDENCE FOR WEIGHT MATRIX : IMD_2015_2019_VGI--London--Queen1
TEST                                     DF      VALUE        PROB
Likelihood Ratio Test                    1      1167.4850     0.00000
============================= END OF REPORT ================================
```

FIGURE B.7: Spatial lag regression model Using IMD 2015 deprivation scores in combination of Wikipedia distribution patterns to predict IMD 2019 deprivation scores.



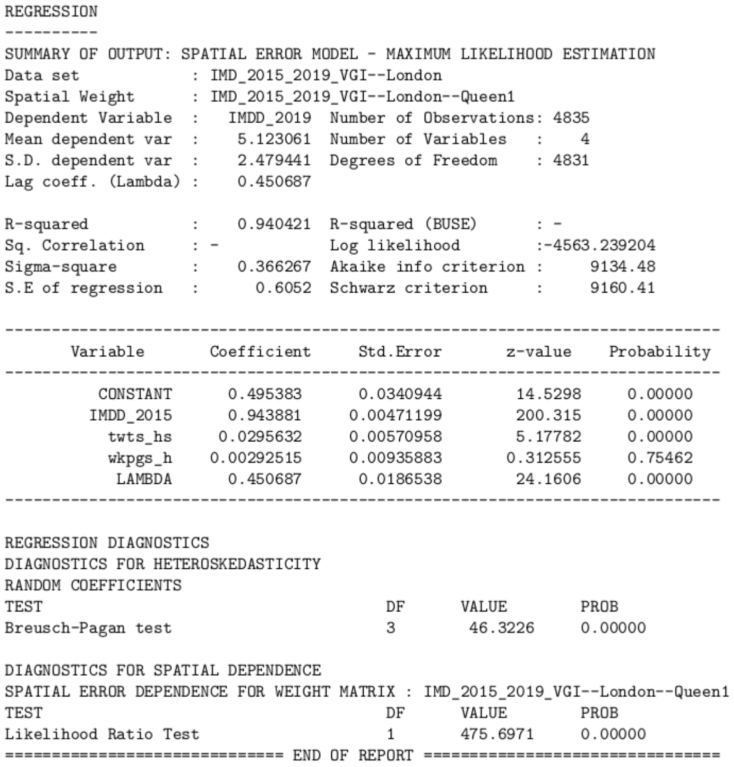
```
REGRESSION
----------
SUMMARY OF OUTPUT: SPATIAL ERROR MODEL - MAXIMUM LIKELIHOOD ESTIMATION
Data set            : IMD_2015_2019_VGI--London
Spatial Weight      : IMD_2015_2019_VGI--London--Queen1
Dependent Variable  :   IMDS_2019  Number of Observations: 4835
Mean dependent var  :   21.498424  Number of Variables   :    3
S.D. dependent var  :   10.903613  Degrees of Freedom    : 4832
Lag coeff. (Lambda) :    0.600555

R-squared           :    0.959908  R-squared (BUSE)      : -
Sq. Correlation     : -            Log likelihood        :-10853.225644
Sigma-square        :     4.76644  Akaike info criterion :     21712.5
S.E of regression   :     2.18322  Schwarz criterion     :     21731.9

----------------------------------------------------------------------------
       Variable       Coefficient     Std.Error       z-value    Probability
----------------------------------------------------------------------------
          CONSTANT       1.40041       0.120462       11.6253     0.00000
         IMDS_2015      0.856411      0.0036804       232.695     0.00000
           wkpgs_h    -0.0790377      0.0288282      -2.74168     0.00611
            LAMBDA      0.600555      0.0156875       38.2824     0.00000
----------------------------------------------------------------------------

REGRESSION DIAGNOSTICS
DIAGNOSTICS FOR HETEROSKEDASTICITY
RANDOM COEFFICIENTS
TEST                                     DF      VALUE        PROB
Breusch-Pagan test                       2      1586.9441     0.00000

DIAGNOSTICS FOR SPATIAL DEPENDENCE
SPATIAL ERROR DEPENDENCE FOR WEIGHT MATRIX : IMD_2015_2019_VGI--London--Queen1
TEST                                     DF      VALUE        PROB
Likelihood Ratio Test                    1      1167.4850     0.00000
============================= END OF REPORT ================================
```

FIGURE B.8: Spatial lag regression model Using IMD 2015 deprivation deciles in combination of Wikipedia distribution patterns to predict IMD 2019 deprivation deciles.



```
REGRESSION
----------
SUMMARY OF OUTPUT: SPATIAL ERROR MODEL - MAXIMUM LIKELIHOOD ESTIMATION
Data set            : IMD_2015_2019_VGI--London
Spatial Weight      : IMD_2015_2019_VGI--London--Queen1
Dependent Variable  :   IMDS_2019  Number of Observations: 4835
Mean dependent var  :   21.498424  Number of Variables   :    4
S.D. dependent var  :   10.903613  Degrees of Freedom    : 4831
Lag coeff. (Lambda) :    0.593345

R-squared           :    0.960062  R-squared (BUSE)      : -
Sq. Correlation     : -            Log likelihood        :-10838.869415
Sigma-square        :     4.74819  Akaike info criterion :     21685.7
S.E of regression   :     2.17903  Schwarz criterion     :     21711.7

-----------------------------------------------------------------------------
       Variable      Coefficient     Std.Error       z-value    Probability
-----------------------------------------------------------------------------
          CONSTANT       1.72711      0.133509        12.9363     0.00000
         IMDS_2015      0.856619    0.00365959        234.075     0.00000
           twts_hs     -0.111366     0.0207053        -5.3786     0.00000
           wkpgs_h     0.0133864      0.033649       0.397824     0.69076
            LAMBDA      0.593345     0.0158458        37.4449     0.00000
-----------------------------------------------------------------------------

REGRESSION DIAGNOSTICS
DIAGNOSTICS FOR HETEROSKEDASTICITY
RANDOM COEFFICIENTS
TEST                                     DF      VALUE        PROB
Breusch-Pagan test                       3      1573.9394     0.00000

DIAGNOSTICS FOR SPATIAL DEPENDENCE
SPATIAL ERROR DEPENDENCE FOR WEIGHT MATRIX : IMD_2015_2019_VGI--London--Queen1
TEST                                     DF      VALUE        PROB
Likelihood Ratio Test                    1      1101.3311     0.00000
============================== END OF REPORT ================================
```

FIGURE B.9: Spatial lag regression model using IMD 2015 deprivation scores in combination of Twitter and Wikipedia distribution patterns to predict IMD 2019 deprivation scores.



```
REGRESSION
----------
SUMMARY OF OUTPUT: SPATIAL ERROR MODEL - MAXIMUM LIKELIHOOD ESTIMATION
Data set            : IMD_2015_2019_VGI--London
Spatial Weight      : IMD_2015_2019_VGI--London--Queen1
Dependent Variable  :   IMDD_2019  Number of Observations: 4835
Mean dependent var  :    5.123061  Number of Variables   :    4
S.D. dependent var  :    2.479441  Degrees of Freedom    : 4831
Lag coeff. (Lambda) :    0.450687

R-squared           :    0.940421  R-squared (BUSE)      : -
Sq. Correlation     : -            Log likelihood        :-4563.239204
Sigma-square        :    0.366267  Akaike info criterion :     9134.48
S.E of regression   :      0.6052  Schwarz criterion     :     9160.41

-----------------------------------------------------------------------------
       Variable      Coefficient     Std.Error       z-value    Probability
-----------------------------------------------------------------------------
          CONSTANT      0.495383     0.0340944        14.5298     0.00000
         IMDD_2015      0.943881    0.00471199        200.315     0.00000
           twts_hs     0.0295632    0.00570958        5.17782     0.00000
           wkpgs_h    0.00292515    0.00935883       0.312555     0.75462
            LAMBDA      0.450687     0.0186538        24.1606     0.00000
-----------------------------------------------------------------------------

REGRESSION DIAGNOSTICS
DIAGNOSTICS FOR HETEROSKEDASTICITY
RANDOM COEFFICIENTS
TEST                                     DF      VALUE        PROB
Breusch-Pagan test                       3        46.3226     0.00000

DIAGNOSTICS FOR SPATIAL DEPENDENCE
SPATIAL ERROR DEPENDENCE FOR WEIGHT MATRIX : IMD_2015_2019_VGI--London--Queen1
TEST                                     DF      VALUE        PROB
Likelihood Ratio Test                    1       475.6971     0.00000
============================== END OF REPORT ================================
```

FIGURE B.10: Spatial lag regression model using IMD 2015 deprivation deciles in combination of Twitter and Wikipedia distribution patterns to predict IMD 2019 deprivation deciles.

Bibliography

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., and Zheng, X. (2015). *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. Software available from tensorflow.org. URL: `http://tensorflow.org/`.

Abbar, S., Zanouda, T., Al-Emadi, N., and Zegour, R. (2018). “City of the people, for the people: Sensing urban dynamics via social media interactions”. In: *International Conference on Social Informatics*. Springer, pp. 3–14.

Abernathy, D. (2016). *Using Geodata and Geolocation in the Social Sciences: Mapping Our Connected World*. Sage.

Abudalfa, S. I. and Ahmed, M. A. (2019). “Semi-Supervised Target-Dependent Sentiment Classification for Micro-Blogs”. In: *Journal of Computer Science and Technology* 19.01, e06–e06.

Acedo, A., Painho, M., Casteleyn, S., and Roche, S. (2018). “Place and city: toward urban intelligence”. In: *ISPRS International Journal of Geo-Information* 7.9, p. 346.

Adamic, L. A. and Adar, E. (2003). “Friends and neighbors on the web”. In: *Social networks* 25.3, pp. 211–230.

Agarwal, A., Singh, R., and Toshniwal, D. (2018). “Geospatial sentiment analysis using twitter data for UK-EU referendum”. In: *Journal of Information and Optimization Sciences* 39.1, pp. 303–317.

Agnew, J. A. and Livingstone, D. N. (2011). *The Sage handbook of geographical knowledge*. Sage Publications.

Ahmad, K. and Conci, N. (2019). “How deep features have improved event recognition in multimedia: a survey”. In: *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 15.2, pp. 1–27.

Akinyemi, F. and Elias, B. (2009). “Modelling urban deprivation in West Africa”. In: *Proceedings: 12th AGILE International Conference on Geographic Information Science*, pp. 1–6.

Alejandro, Y. and Palafox, L. (2019). “Gentrification Prediction Using Machine Learning”. In: *Mexican International Conference on Artificial Intelligence*. Springer, pp. 187–199.

Aloteibi, S. and Sanderson, M. (2014). “Analyzing geographic query reformulation: An exploratory study”. In: *Journal of the association for information science and technology* 65.1, pp. 13–24.

Anderson, C. (2008). “The end of theory: The data deluge makes the scientific method obsolete”. In: *Wired magazine* 16.7, pp. 16–07.

Andrade, S. C. de, Restrepo-Estrada, C., Costa, T. A. da, Ueyama, J., Delbem, A. C., and Albuquerque, J. P. de (2018). “Situational awareness in social media: lessons learned using information entropy in flood risk management”. In:

Andrienko, G., Andrienko, N., Bosch, H., Ertl, T., Fuchs, G., Jankowski, P., and Thom, D. (2013). "Thematic patterns in georeferenced tweets through space-time visual analytics". In: *Computing in Science & Engineering* 15.3, pp. 72–82.

Anselin, L. and Williams, S. (2016). "Digital neighborhoods". In: *Journal of Urbanism: International Research on Placemaking and Urban Sustainability* 9.4, pp. 305–328.

Aoyama, Y., Murphy, J. T., and Hanson, S. (2010). *Key concepts in economic geography*. Sage.

Arribas-Bel, D., Kourtit, K., and Nijkamp, P. (2016). "The sociocultural sources of urban buzz". In: *Environment and Planning C: Government and Policy* 34.1, pp. 188–204.

Arribas-Bel, D., Kourtit, K., Nijkamp, P., and Steenbruggen, J. (2015). "Cyber cities: social media as a tool for understanding cities". In: *Applied Spatial Analysis and Policy* 8.3, pp. 231–247.

Arribas-Bel, D., Nijkamp, P., and Scholten, H. (2011). "Multidimensional urban sprawl in Europe: A self-organizing map approach". In: *Computers, environment and urban systems* 35.4, pp. 263–275.

Arribas-Bel, D., Patino, J. E., and Duque, J. C. (2017). "Remote sensing-based measurement of Living Environment Deprivation: Improving classical approaches with machine learning". In: *PloS one* 12.5, e0176684.

Arsanjani, J. J., Mooney, P., Zipf, A., and Schauss, A. (2015). "Quality assessment of the contributed land use information from OpenStreetMap versus authoritative datasets". In: *OpenStreetMap in GIScience*. Springer, pp. 37–58.

Ash, J., Kitchin, R., and Leszczynski, A. (2018a). *Digital Geographies*. SAGE Publications. ISBN: 9781526455369. URL: https://books.google.co.uk/books?id=bpFyDwAAQBAJ.

Ash, J., Kitchin, R., and Leszczynski, A. (2018b). "Digital turn, digital geographies?" In: *Progress in Human Geography* 42.1, pp. 25–43.

Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., and Ives, Z. (2007). "Dbpedia: A nucleus for a web of open data". In: *The semantic web*. Springer, pp. 722–735.

Awcock, H. (2018). "Contesting the Capital: Space, Place, and Protest in London, 1780-2010". English. PhD thesis. Royal Holloway, University of London.

Backstrom, L., Sun, E., and Marlow, C. (2010). "Find me if you can: improving geographical prediction with social and spatial proximity". In: *Proceedings of the 19th international conference on World wide web*. ACM, pp. 61–70.

Ballatore, A. (2016). "Prolegomena for an ontology of place". In: *Advancing geographic information science*, pp. 91–103.

Ballatore, A. and De Sabbata, S (2019). "Los Angeles as a digital place: the geographies of user-generated content". In: *Transactions in GIS*.

Ballatore, A. and De Sabbata, S. (2018). "Charting the Geographies of Crowdsourced Information in Greater London". In: *Geospatial Technologies for All*. Ed. by A. Mansourian, P. Pilesjö, L. Harrie, and R. van Lammeren. Cham: Springer International Publishing, pp. 149–168. ISBN: 978-3-319-78208-9.

Barocas, S. and Selbst, A. D. (2016). "Big data's disparate impact". In:

Barron, C., Neis, P., and Zipf, A. (2014). "A comprehensive framework for intrinsic OpenStreetMap quality analysis". In: *Transactions in GIS* 18.6, pp. 877–895.

Barton, M. (2016). "An exploration of the importance of the strategy used to identify gentrification". In: *Urban Studies* 53.1, pp. 92–111.

Batey, P. and Brown, P. (2007). "The spatial targeting of urban policy initiatives: a geodemographic assessment tool". In: *Environment and Planning A* 39.11, pp. 2774–2793.

Bawa-Cavia, A. (2011). “Sensing the urban: using location-based social network data in urban analysis”. In: *Pervasive PURBA Workshop*. Vol. 5.

Berners-Lee, T. (1998). *The World Wide Web: A very short personal history [online]*. Availabe at https://www.w3.org/People/Berners-Lee/ShortHistory.html (Accessed April 29th, 2021).

Berthon, P., Pitt, L., Kietzmann, J., and McCarthy, I. P. (2015). “CGIP: managing consumer-generated intellectual property”. In: *California Management Review* 57.4, pp. 43–62.

Bettencourt, L. M., Lobo, J., Helbing, D., Kühnert, C., and West, G. B. (2007). “Growth, innovation, scaling, and the pace of life in cities”. In: *Proceedings of the national academy of sciences* 104.17, pp. 7301–7306.

Blaschke, T., Merschdorf, H., Cabrera-Barona, P., Gao, S., Papadakis, E., and Kovacs-Györi, A. (2018). “Place versus space: From points, lines and polygons in gis to place-based representations reflecting language and culture”. In: *ISPRS International Journal of Geo-Information* 7.11, p. 452.

Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., and Kalai, A. T. (2016). “Man is to computer programmer as woman is to homemaker? debiasing word embeddings”. In: *Advances in Neural Information Processing Systems*.

Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J., and Yakhnenko, O. (2013). “Translating embeddings for modeling multi-relational data”. In: *Advances in neural information processing systems*, pp. 2787–2795.

Borth, D., Chen, T., Ji, R., and Chang, S.-F. (2013). “Sentibank: large-scale ontology and classifiers for detecting sentiment and emotions in visual content”. In: *Proceedings of the 21st ACM international conference on Multimedia*. ACM, pp. 459–460.

Bottou, L., Curtis, F. E., and Nocedal, J. (2018). “Optimization methods for large-scale machine learning”. In: *Siam Review* 60.2, pp. 223–311.

Boulos, M. N. K., Peng, G., and VoPham, T. (2019). *An overview of GeoAI applications in health and healthcare*.

Bowen, D. A., Mercer Kollar, L. M., Wu, D. T., Fraser, D. A., Flood, C. E., Moore, J. C., Mays, E. W., and Sumner, S. A. (2018). “Ability of crime, demographic and business data to forecast areas of increased violence”. In: *International journal of injury control and safety promotion* 25.4, pp. 443–448.

Boy, J. D. and Uitermark, J. (2016). “How to study the city on Instagram”. In: *PLoS one* 11.6, e0158161.

Boyd, D and Crawford, K (2012). “Critical Questions For Big Data. Information”. In: *Communication & Society* 15.5, pp. 662–679.

Brantner, C. and Rodriguez-Amat, J. R. (2016). “Constructing Public Space| New “Danger Zone” in Europe: Representations of Place in Social Media–Supported Protests”. In: *International Journal of Communication* 10, p. 22.

Brown, P. J. (1991). “Exploring geodemographics”. In: *Handling Geographical Information*, pp. 221–258.

Bruns, A. (2012). “HOW LONG IS A TWEET? MAPPING DYNAMIC CONVERSATION NETWORKS ON TWITTER USING GAWK AND GEPHI”. In: *Information, Communication & Society* 15.9, pp. 1323–1351. DOI: 10.1080/1369118X.2011.635214. eprint: https://doi.org/10.1080/1369118X.2011.635214. URL: https://doi.org/10.1080/1369118X.2011.635214.

Bruns, A. and Burgess, J. (2011). “# ausvotes: How Twitter covered the 2010 Australian federal election”. In: *Communication, Politics & Culture* 44.2, p. 37.

Buhrmester, V., Münch, D., and Arens, M. (2019). “Analysis of explainers of black box deep neural networks for computer vision: A survey”. In: *arXiv preprint arXiv:1911.12116*.

Butler, D. (2006). *The web-wide world*.

Cai, G. and Xia, B. (2015). "Convolutional neural networks for multimedia sentiment analysis". In: *Natural Language Processing and Chinese Computing*. Springer, pp. 159–167.

Callahan, E. S. and Herring, S. C. (2011). "Cultural bias in Wikipedia content on famous persons". In: *Journal of the American society for information science and technology* 62.10, pp. 1899–1915.

Cambridge, U. (2009). "Introduction to information retrieval". In:

Capineri, C. (2016a). "Kilburn high road revisited". In: *Urban Planning* 1.2, pp. 128–140.

Capineri, C. (2016b). "The nature of volunteered geographic information". In: *European Handbook of Crowdsourced Geographic Information*, p. 15.

Chandar, S, Khapra, M. M., Larochelle, H, and Ravindran, B (2016). "Correlational Neural Networks." In: *Neural Computation* 28.2, p. 257.

Chang, H.-w., Lee, D., Eltaher, M., and Lee, J. (2012). "@ Phillies tweeting from Philly? Predicting Twitter user locations with spatial word usage". In: *Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012)*. IEEE Computer Society, pp. 111–118.

Chang, J.-W., Bista, R., Kim, Y.-C., and Kim, Y.-K. (2007). "Spatio-temporal similarity measure algorithm for moving objects on spatial networks". In: *International Conference on Computational Science and Its Applications*. Springer, pp. 1165–1178.

Chaniotakis, E. and Antoniou, C. (2015). "Use of geotagged social media in urban settings: Empirical evidence on its potential from twitter". In: *2015 IEEE 18th International Conference on Intelligent Transportation Systems*. IEEE, pp. 214–219.

Chen, J., Deng, S., and Chen, H. (2017). "Crowdgeokg: Crowdsourced geo-knowledge graph". In: *China Conference on Knowledge Graph and Semantic Computing*. Springer, pp. 165–172.

Chen, K., Zhou, Y., and Dai, F. (2015). "A LSTM-based method for stock returns prediction: A case study of China stock market". In: *2015 IEEE international conference on big data (big data)*. IEEE, pp. 2823–2824.

Chen, Y., Lv, Y., Wang, X., and Wang, F. (2017). "A convolutional neural network for traffic information sensing from social media text". In: *2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC)*, pp. 1–6.

Chen, Y., Zhao, J., Hu, X., Zhang, X., Li, Z., and Chua, T.-S. (2013). "From interest to function: Location estimation in social media". In: *Twenty-Seventh AAAI Conference on Artificial Intelligence*.

Cheng, T. and Wicks, T. (2014). "Event detection using Twitter: a spatio-temporal approach". In: *PloS one* 9.6, e97807.

Cheshire, J., Batty, M., Reades, J., Longley, P., Manley, E., and Milton, R. (2019). "CyberGIS for Analyzing Urban Data". In: *CyberGIS for Geospatial Discovery and Innovation*. Springer, pp. 33–52.

Chinchor, N. A. (1998). *Overview of muc-7/met-2*. Tech. rep. SCIENCE APPLICATIONS INTERNATIONAL CORP SAN DIEGO CA.

Cho, K. (2013). "Simple sparsification improves sparse denoising autoencoders in denoising highly corrupted images". In: *International Conference on Machine Learning*, pp. 432–440.

Chollet, F. et al. (2015). *Keras*. https://github.com/fchollet/keras.

Chong, W.-H. and Lim, E.-P. (2017). "Exploiting contextual information for fine-grained tweet geolocation". In: *Eleventh International AAAI Conference on Web and Social Media*.

Chu, G., Potetz, B., Wang, W., Howard, A., Song, Y., Brucher, F., Leung, T., and Adam, H. (2019). "Geo-aware networks for fine-grained recognition". In: *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pp. 0–0.

CIA (2011). *The World Factbook 2011*. Central Intelligence Agency.

Cioffi-Revilla, C. (2010). "Computational social science". In: *WIREs Computational Statistics* 2.3, pp. 259–271. DOI: 10.1002/wics.95. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/wics.95. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/wics.95.

Ciuccarelli, P., Lupi, G., and Simeone, L. (2014). *Visualizing the data city: social media as a source of knowledge for urban planning and management*. Springer Science & Business Media.

Clelland, D. and Hill, C. (2019). "Deprivation, policy and rurality: The limitations and applications of area-based deprivation indices in Scotland". In: *Local Economy* 34.1, pp. 33–50.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., and Kuksa, P. (2011). "Natural language processing (almost) from scratch". In: *Journal of Machine Learning Research* 12.Aug, pp. 2493–2537.

Cormode, G. and Duffield, N. (2014). "Sampling for big data: a tutorial". In: *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 1975–1975.

Cortes, C. and Vapnik, V. (1995). "Support-vector networks". In: *Machine learning* 20.3, pp. 273–297.

Cox, D. T., Shanahan, D. F., Hudson, H. L., Fuller, R. A., and Gaston, K. J. (2018). "The impact of urbanisation on nature dose and the implications for human health". In: *Landscape and urban planning* 179, pp. 72–80.

Crampton, J. W. (2011). *Mapping: A critical introduction to cartography and GIS*. Vol. 11. John Wiley & Sons.

Crampton, J. W., Graham, M., Poorthuis, A., Shelton, T., Stephens, M., Wilson, M. W., and Zook, M. (2013). "Beyond the geotag: situating 'big data'and leveraging the potential of the geoweb". In: *Cartography and geographic information science* 40.2, pp. 130–139.

Crang, M., Crang, P., and May, J. (1999). *Virtual geographies: bodies, space and relations*. Psychology Press.

Crawford, K. and Schultz, J. (2014). "Big data and due process: Toward a framework to redress predictive privacy harms". In: *BCL Rev.* 55, p. 93.

Crossley, M. (1999). "A guide to coordinate systems in Great Britain". In: *Ordnance Survey*.

Cumbria Intelligence Observatory, C. C. C. (2019). *The English Indices of Deprivation (IoD) Index of Multiple Deprivation (IMD) Published September 2019 [online]*. Available at https://cumbria.gov.uk/elibrary/Content/Internet/536/671/4674/17217/17223/422771749.PDF (Accessed May 18th, 2021).

Cumbria Vision, C. C. C. (2009). *Cumbria Economic Strategy [online]*. Available at https://www.cumbria.gov.uk/elibrary/content/internet/534/576/6304/407851554.pdf (Accessed May 18th, 2021).

Curtis, F. E. and Scheinberg, K. (2017). "Optimization methods for supervised machine learning: From linear models to deep learning". In: *Leading Developments from INFORMS Communities*. INFORMS, pp. 89–114.

Dacey, M. F. (1965). *A review on measures of contiguity for two and k-color maps*. Tech. rep. NORTHWESTERN UNIV EVANSTON ILL.

Danyllo, W., Alisson, V., Alexandre, N., Moacir, L., Jansepetrus, B., and Oliveira, R. F. (2013). "Identifying relevant users and groups in the context of credit analysis based on data from Twitter". In: *Cloud and Green Computing (CGC), 2013 Third International Conference on*. IEEE, pp. 587–592.

Davis Jr, C. A., Pappa, G. L., Oliveira, D. R. R. de, and L. Arcanjo, F. de (2011). "Inferring the location of twitter messages based on user relationships". In: *Transactions in GIS* 15.6, pp. 735–751.

De Sabbata, S. and Liu, P. (2019). "Deep learning geodemographics with autoencoders and geographic convolution". In: *Proceedings of the 22nd AGILE conference on Geographic Information Science, Limassol, Greece*.

Dedman, D., Hennell, T., Hooper, J., Tocque, K., and Bellis, M (2006). "Using geodemographics to illustrate health inequalities". In: *Liverpool: North West Public Health Observatory, Liverpool John Moores University*.

Defferrard, M., Bresson, X., and Vandergheynst, P. (2016). "Convolutional neural networks on graphs with fast localized spectral filtering". In: *Advances in neural information processing systems*, pp. 3844–3852.

DeLyser, D. and Sui, D. (2013). "Crossing the qualitative-quantitative divide II: Inventive approaches to big data, mobile methods, and rhythmanalysis". In: *Progress in human geography* 37.2, pp. 293–305.

Demšar, U., Harris, P., Brunsdon, C., Fotheringham, A. S., and McLoone, S. (2013). "Principal component analysis on spatial data: an overview". In: *Annals of the Association of American Geographers* 103.1, pp. 106–128.

Dickson, A and Young, R (1985). "Social Policy and Local Government: Urban Deprivation in Strathclyde Region". In: *International Journal of Sociology and Social Policy* 5.1, pp. 79–94.

Ding, C. (1998). "The GIS-based human-interactive TAZ design algorithm: examining the impacts of data aggregation on transportation-planning analysis". In: *Environment and Planning B: Planning and Design* 25.4, pp. 601–616.

Ding, Y., Korotkiy, M, Omelayenko, B., Kartseva, V, Zykov, V, Klein, M., Schulten, E., and Fensel, D. (2002). "Goldenbullet: Automated classification of product data in e-commerce". In: *Proceedings of the 5th international conference on business information systems*.

DiNucci, D. (1999). "Fragmented future." In: *Print* 53.4, pp. 32–33.

Dodge, M. (1998). "The geographies of Cyberspace. A research note". In: *NETCOM: Réseaux, communication et territoires/Networks and communication studies* 12.4, pp. 383–396.

Dodge, M. and Kitchin, R. (2004). "Flying through code/space: the real virtuality of air travel". In: *Environment and planning A* 36.2, pp. 195–211.

Dodge, M. and Kitchin, R. (2005). "Codes of life: Identification codes and the machine-readable world". In: *Environment and Planning D: Society and Space* 23.6, pp. 851–881.

Doersch, C., Singh, S., Gupta, A., Sivic, J., and Efros, A. A. (2015). "What makes Paris look like Paris?" In: *Communications of the ACM* 58.12, pp. 103–110.

Economics, G. (2016). "Economic Evidence Base for London 2016". In: *Water supply and drainage*, p295–296.

Efthimion, P. G., Payne, S., and Proferes, N. (2018). "Supervised machine learning bot detection techniques to identify social twitter bots". In: *SMU Data Science Review* 1.2, p. 5.

Eisenstein, J., Ahmed, A., and Xing, E. P. (2011). "Sparse additive generative models of text". In:

Eisenstein, J., O'Connor, B., Smith, N. A., and Xing, E. P. (2010). "A latent variable model for geographic lexical variation". In: *Proceedings of the 2010 conference on* *empirical methods in natural language processing*. Association for Computational Linguistics, pp. 1277–1287.

Elwood, S. (2010). “Geographic information science: Emerging research on the societal implications of the geospatial web”. In: *Progress in human geography* 34.3, pp. 349–357.

Elwood, S. and Leszczynski, A. (2013). “New spatial media, new knowledge politics”. In: *Transactions of the Institute of British Geographers* 38.4, pp. 544–559.

Felt, M. (2016). “Social media and the social sciences: How researchers employ Big Data analytics”. In: *Big Data & Society* 3.1, p. 2053951716645828. DOI: 10.1177/2053951716645828. eprint: https://doi.org/10.1177/2053951716645828. URL: https://doi.org/10.1177/2053951716645828.

Fisher, P. and Tate, N. J. (2015). “Modelling class uncertainty in the geodemographic Output Area Classification”. In: *Environment and Planning B: Planning and Design* 42.3, pp. 541–563.

Fisher, P. and Unwin, D. (2001). “Virtual reality in geography: an introduction”. In: *Virtual reality in geography*. CRC Press, pp. 10–12.

Flanagin, A. J. and Metzger, M. J. (2008). “The credibility of volunteered geographic information”. In: *GeoJournal* 72.3-4, pp. 137–148.

Flatow, D., Naaman, M., Xie, K. E., Volkovich, Y., and Kanza, Y. (2015). “On the accuracy of hyper-local geotagging of social media content”. In: *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*. ACM, pp. 127–136.

Floridi, L. (2014). *The fourth revolution: How the infosphere is reshaping human reality*. OUP Oxford.

Fotheringham, A. S. (1997). “Trends in quantitative methods I: stressing the local”. In: *Progress in Human Geography* 21.1, pp. 88–96.

Fout, A., Byrd, J., Shariat, B., and Ben-Hur, A. (2017). “Protein interface prediction using graph convolutional networks”. In: *Advances in neural information processing systems*, pp. 6530–6539.

Fox, K. (2012). *OpenStreetMap: 'It's the Wikipedia of maps' [online]*. Available at https://www.theguardian.com/theobserver/2012/feb/18/openstreetmap-world-map-radicals (Accessed August 12, 2020).

Frias-Martinez, V. and Frias-Martinez, E. (2014). “Spectral clustering for sensing urban land use using Twitter activity”. In: *Engineering Applications of Artificial Intelligence* 35, pp. 237–245.

Fuchs, C. (2008). “Wikinomics: How mass collaboration changes everything”. In: *International Journal of Communication* 2, pp. 1–11.

Gajarla, V. and Gupta, A. (2015). “Emotion Detection and Sentiment Analysis of Images”. In: *Georgia Institute of Technology*.

Gale, C. G. and Longley, P. (2013). “Temporal uncertainty in a small area open geodemographic classification”. In: *Transactions in GIS* 17.4, pp. 563–588.

Gale, C. G., Singleton, A. D., Bates, A. G., and Longley, P. A. (2016). “Creating the 2011 area classification for output areas (2011 OAC)”. In: *Journal of Spatial Information Science* 2016.12, pp. 1–27.

Gale, C. G. (2014). “Creating an open geodemographic classification using the UK Census of the Population”. PhD thesis. UCL (University College London).

Galster, G. (2001). “On the nature of neighbourhood”. In: *Urban studies* 38.12, pp. 2111–2124.

Gao, S., Janowicz, K., McKenzie, G., and Li, L. (2013). “Towards Platial Joins and Buffers in Place-Based GIS.” In: *Comp@ Sigspatial*, pp. 42–49.

Gao, S., Janowicz, K., Montello, D. R., Hu, Y., Yang, J.-A., McKenzie, G., Ju, Y., Gong, L., Adams, B., and Yan, B. (2017). "A data-synthesis-driven method for detecting and extracting vague cognitive regions". In: *International Journal of Geographical Information Science* 31.6, pp. 1245–1271.

Gao, Y., Zhao, S., Yang, Y., and Chua, T.-S. (2015). "Multimedia social event detection in microblog". In: *International Conference on Multimedia Modeling*. Springer, pp. 269–281.

García-Palomares, J. C., Salas-Olmedo, M. H., Moya-Gomez, B., Condeco-Melhorado, A., and Gutierrez, J. (2018). "City dynamics through Twitter: Relationships between land use and spatiotemporal demographics". In: *Cities* 72, pp. 310–319.

Gardner, Z, Mooney, P., De Sabbata, S, and Dowthwaite, L (2020). "Quantifying gendered participation in OpenStreetMap: responding to theories of female (under) representation in crowdsourced mapping". In: *GeoJournal* 85.6, pp. 1603–1620.

Gers, F. A., Eck, D., and Schmidhuber, J. (2002). "Applying LSTM to time series predictable through time-window approaches". In: *Neural Nets WIRN Vietri-01*. Springer, pp. 193–200.

Ghani, N. A., Hamid, S., Hashem, I. A. T., and Ahmed, E. (2019). "Social media big data analytics: A survey". In: *Computers in Human Behavior* 101, pp. 417–428.

Gibbons, J., Nara, A., and Appleyard, B. (2018). "Exploring the imprint of social media networks on neighborhood community through the lens of gentrification". In: *Environment and Planning B: Urban Analytics and City Science* 45.3, pp. 470–488.

Gleason, C., Carrington, P., Cassidy, C., Morris, M. R., Kitani, K. M., and Bigham, J. P. (2019). ""It's almost like they're trying to hide it": How User-Provided Image Descriptions Have Failed to Make Twitter Accessible". In: *The World Wide Web Conference*, pp. 549–559.

Glenn, D. (2012). *Is the Status Update Dead? 36% of Tweets Are Photos [Infographic] [online]*. Availabe at https://www.adweek.com/performance-marketing/is-the-status-update-dead-36-of-tweets-are-photos-infographic/ (Accessed April 29th, 2021).

Glorot, X. and Bengio, Y. (2010). "Understanding the difficulty of training deep feedforward neural networks". In: *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pp. 249–256.

Goby, V. (2003). "Physical space and cyberspace: how do they interrelate? A study of offline and online social interaction choice in Singapore". In: *CyberPsychology & Behavior* 6.6, pp. 639–644.

Gomide, J., Veloso, A., Meira Jr, W., Almeida, V., Benevenuto, F., Ferraz, F., and Teixeira, M. (2011). "Dengue surveillance based on a computational model of spatio-temporal locality of Twitter". In: *Proceedings of the 3rd international web science conference*. ACM, p. 3.

González-Bailón, S. (2013). "Social science in the era of big data". In: *Policy & internet* 5.2, pp. 147–160.

Goodchild, M. F. (1992). "Geographical information science". In: *International journal of geographical information systems* 6.1, pp. 31–45.

Goodchild, M. F. (2007). "Citizens as sensors: the world of volunteered geography". In: *GeoJournal* 69.4, pp. 211–221.

Goodchild, M. F. (2008). "patial accuracy 2.0". In: *Proceedings of the 8th symposium on spatial accuracy assessment in natural resources and environmental sciences*.

Goodchild, M. F. (2011). "Formalizing place in geographic information systems". In: *Communities, neighborhoods, and health*. Springer, pp. 21–33.

Goodchild, M. F. and Janelle, D. G. (2004). *Spatially integrated social science*. Oxford University Press.

Goodfellow, I, Bengio, Y, and Courville, A (2016). "Deep Learning | The MIT Press". In: *Cambridge, Massachusetts*.

Goodman, B. and Flaxman, S. (2017). "European Union regulations on algorithmic decision-making and a "right to explanation"". In: *AI magazine* 38.3, pp. 50–57.

Graells-Garrido, E., Lalmas, M., and Menczer, F. (2015). "First women, second sex: Gender bias in Wikipedia". In: *Proceedings of the 26th ACM Conference on Hypertext & Social Media*, pp. 165–174.

Graham, M. (2012). "Big data and the end of theory". In: *The Guardian* 9.

Graham, M., De Sabbata, S., and Zook, M. A. (2015a). "Towards a study of information geographies:(im) mutable augmentations and a mapping of the geographies of information". In: *Geo: Geography and environment* 2.1, pp. 88–105.

Graham, M., Hale, S. A., and Gaffney, D. (2014a). "Where in the world are you? Geolocation and language identification in Twitter". In: *The Professional Geographer* 66.4, pp. 568–578.

Graham, M., Hogan, B., Straumann, R. K., and Medhat, A. (2014b). "Uneven geographies of user-generated information: Patterns of increasing informational poverty". In: *Annals of the Association of American Geographers* 104.4, pp. 746–764.

Graham, M., Shelton, T., and Zook, M. (2013a). "Mapping Zombies: A Guide for Digital Pre-Apocalyptic Analysis and Post-Apocalyptic Survival". In:

Graham, M., Straumann, R. K., and Hogan, B. (2015b). "Digital divisions of labor and informational magnetism: Mapping participation in Wikipedia". In: *Annals of the Association of American Geographers* 105.6, pp. 1158–1178.

Graham, M., Zook, M., and Boulton, A. (2013b). "Augmented reality in urban places: contested content and the duplicity of code". In: *Transactions of the Institute of British Geographers* 38.3, pp. 464–479.

Gray, J., Buckner, L., Policy, S., and Comber, A. (2018). "Exploring social dynamics: predictive geodemographics". In:

Gross, J. L. and Yellen, J. (1999). *Graph theory and its applications*.

Grover, A. and Leskovec, J. (2016). "node2vec: Scalable feature learning for networks". In: *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, pp. 855–864.

Guo, D. and Chen, C. (2014). "Detecting non-personal and spam users on geo-tagged Twitter network". In: *Transactions in GIS* 18.3, pp. 370–384.

Guo, X., Liu, X., Zhu, E., and Yin, J. (2017). "Deep clustering with convolutional autoencoders". In: *International Conference on Neural Information Processing*. Springer, pp. 373–382.

Gupta, A., Joshi, A., and Kumaraguru, P. (2012). "Identifying and characterizing user communities on twitter during crisis events". In: *Proceedings of the 2012 workshop on Data-driven user behavioral modelling and mining from social media*, pp. 23–26.

Gurevich, O. and Ghosh, R. A. (2014). *Systems and methods for identifying geographic locations of social media content collected over social networks*. US Patent App. 13/853,687.

Hahmann, S., Purves, R., and Burghardt, D. (2014). "Twitter location (sometimes) matters: Exploring the relationship between georeferenced tweet content and nearby feature classes". In: *Journal of Spatial Information Science* 2014.9, pp. 1–36.

Haklay, M., Singleton, A., and Parker, C. (2008). "Web mapping 2.0: The neogeography of the GeoWeb". In: *Geography Compass* 2.6, pp. 2011–2039.

Hamilton, W., Ying, Z., and Leskovec, J. (2017). "Inductive representation learning on large graphs". In: *Advances in neural information processing systems*, pp. 1024–1034.

Han, B., Cook, P., and Baldwin, T. (2014). "Text-based twitter user geolocation prediction". In: *Journal of Artificial Intelligence Research* 49, pp. 451–500.

Hao, M., Jiang, D., Ding, F., Fu, J., and Chen, S. (2019). "Simulating spatio-temporal patterns of terrorism incidents on the indochina peninsula with GIS and the random forest method". In: *ISPRS International Journal of Geo-Information* 8.3, p. 133.

Harris, R. (2003). "10 Population mapping by geodemographics and digital imagery". In: *Remotely-sensed cities*, p. 223.

Harris, R., O'Sullivan, D., Gahegan, M., Charlton, M., Comber, L., Longley, P., Brunsdon, C., Malleson, N., Heppenstall, A., Singleton, A., et al. (2017). "More bark than bytes? Reflections on 21+ years of geocomputation". In: *Environment and Planning B: Urban Analytics and City Science* 44.4, pp. 598–617.

Hartigan, J. A. and Wong, M. A. (1979). "Algorithm AS 136: A k-means clustering algorithm". In: *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 28.1, pp. 100–108.

Hartshorne, R. (1939). "H. THE CHARACTER OF REGIONAL GEOGRAPHY". In: *Annals of the Association of American Geographers* 29.4, pp. 436–456.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). "Deep residual learning for image recognition". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778.

Henaff, M., Bruna, J., and LeCun, Y. (2015). "Deep convolutional networks on graph-structured data". In: *arXiv preprint arXiv:1506.05163*.

Herring, C. (1994). "An architecture of cyberspace: Spatialization of the Internet". In: *US Army Construction Engineering Research Laboratory: Champaign, IL, USA*.

Hinchliffe, S. J., Crang, M., Reimer, S. M., and Hudson, A. (1997). "Software for qualitative research: 2. Some thoughts on 'aiding'analysis". In: *Environment and Planning A* 29.6, pp. 1109–1124.

Hine, C. (2008). "Virtual ethnography: Modes, varieties, affordances". In: *The SAGE handbook of online research methods*, pp. 257–270.

Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A.-r., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. N., et al. (2012). "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups". In: *IEEE Signal Processing Magazine* 29.6, pp. 82–97.

Hinton, G. E. and Salakhutdinov, R. R. (2006). "Reducing the dimensionality of data with neural networks". In: *science* 313.5786, pp. 504–507.

Hoffart, J., Suchanek, F. M., Berberich, K., and Weikum, G. (2013). "YAGO2: A spatially and temporally enhanced knowledge base from Wikipedia". In: *Artificial Intelligence* 194, pp. 28–61.

Hollenstein, L. and Purves, R. (2010). "Exploring place through user-generated content: Using Flickr tags to describe city cores". In: *Journal of Spatial Information Science* 2010.1, pp. 21–48.

Hotelling, H. (1992). "Relations between two sets of variates". In: *Breakthroughs in statistics*. Springer, pp. 162–190.

Hu, Y. and Wang, R.-Q. (2020). "Understanding the removal of precise geotagging in tweets". In: *Nature Human Behaviour*, pp. 1–3.

Hu, Y. and Zhang, Y. (2020). "Spatial–temporal dynamics and driving factor analysis of urban ecological land in Zhuhai city, China". In: *Scientific reports* 10.1, pp. 1–15.

Huang, B. and Carley, K. M. (2017). "On predicting geolocation of tweets using convolutional neural networks". In: *International conference on social computing, behavioral-cultural modeling and prediction and behavior representation in modeling and simulation*. Springer, pp. 281–291.

Huang, X., Wang, C., Li, Z., and Ning, H. (2018). "A visual–textual fused approach to automated tagging of flood-related tweets during a flood event". In: *International Journal of Digital Earth*, pp. 1–17.

Hubbard, P., Bartley, B., Fuller, D., and Kitchin, R. (2002). *Thinking geographically: Space, theory and contemporary human geography*. A&C Black.

Ifrim, G., Shi, B., and Brigadir, I. (2014). "Event Detection in Twitter using Aggressive Filtering and Hierarchical Tweet Clustering." In: *SNOW-DC@ WWW*, pp. 33–40.

Information Policy Team, D. f. E. (2017). *Leicester and Leicestershire Area Review: Final Report [online]*. Available at https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/634666/Leicester_and_Leicestershire_Area_Review_Report_FINAL.pdf (Accessed May 10th, 2021).

Ingold, D. and Soper, S. (2016). *Amazon doesn't consider the race of its customers. Should it? Bloomberg, April 21, 2016*.

Jaccard, P. (1901). "Distribution de la flore alpine dans le bassin des Dranses et dans quelques régions voisines". In: *Bull Soc Vaudoise Sci Nat* 37, pp. 241–272.

Jacobs, A. (2009). "The pathologies of big data". In: *Communications of the ACM* 52.8, pp. 36–44.

Janowicz, K., Gao, S., McKenzie, G., Hu, Y., and Bhaduri, B. (2020). *GeoAI: spatially explicit artificial intelligence techniques for geographic knowledge discovery and beyond*.

Jenkins, A., Croitoru, A., Crooks, A. T., and Stefanidis, A. (2016). "Crowdsourcing a collective sense of place". In: *PloS one* 11.4.

Jenks, G. F. (1967). "The data model concept in statistical mapping". In: *International yearbook of cartography* 7, pp. 186–190.

Jiang, Z. (2016). "Spatial Big Data Analytics: Classification Techniques for Earth Observation Imagery". In:

Jiang, Z. and Shekhar, S. (2017). "Spatial big data science". In: *Schweiz: Springer International Publishing AG*.

Jivraj, S and Finney, J (2013). *Geographies of Diversity in Leicestershire*.

Kam, H. T. (1995). "Random decision forest". In: *Proceedings of the 3rd International Conference on Document Analysis and Recognition*. Vol. 1416. Montreal, Canada, August, p. 278282.

Kellerman, A. (2016). "Image spaces and the geography of Internet screen-space". In: *GeoJournal* 81.4, pp. 503–517.

Kietzmann, J. H., Hermkens, K., McCarthy, I. P., and Silvestre, B. S. (2011). "Social media? Get serious! Understanding the functional building blocks of social media". In: *Business horizons* 54.3, pp. 241–251.

Kingma, D. P. and Ba, J. (2014). "Adam: A method for stochastic optimization". In: *arXiv preprint arXiv:1412.6980*.

Kingma, D. P. and Welling, M. (2013). "Auto-encoding variational bayes". In: *arXiv preprint arXiv:1312.6114*.

Kipf, T. N. and Welling, M. (2016a). "Semi-Supervised Classification with Graph Convolutional Networks". In:

Kipf, T. N. and Welling, M. (2016b). "Variational graph auto-encoders". In: *arXiv preprint arXiv:1611.07308*.

Kitchin, R. (2013). "Big data and human geography: Opportunities, challenges and risks". In: *Dialogues in human geography* 3.3, pp. 262–267.

Kitchin, R. (2014). "Big Data, new epistemologies and paradigm shifts". In: *Big Data & Society* 1.1, p. 2053951714528481.

Knaap, E., Wolf, L., Rey, S., Kang, W., and Han, S. (2019). "The Dynamics of Urban Neighborhoods: A Survey of Approaches for Modeling Socio-Spatial Structure". In:

Kong, L., Liu, Z., and Huang, Y. (2014). “Spot: Locating social media users based on social network context”. In: *Proceedings of the VLDB Endowment* 7.13, pp. 1681–1684.

Kononenko, I., Bratko, I., and Kukar, M. (1997). “Application of machine learning to medical diagnosis”. In: *Machine Learning and Data Mining: Methods and Applications* 389, p. 408.

Kontopantelis, E., Mamas, M. A., Marwijk, H. van, Buchan, I., Ryan, A. M., and Doran, T. (2018). “Increasing socioeconomic gap between the young and old: temporal trends in health and overall deprivation in England by age, sex, urbanity and ethnicity, 2004–2015”. In: *J Epidemiol Community Health* 72.7, pp. 636–644.

Kramer, M. A. (1991). “Nonlinear principal component analysis using autoassociative neural networks”. In: *AIChE journal* 37.2, pp. 233–243.

Krishnan, M. (2019). “Against Interpretability: a Critical Examination of the Interpretability Problem in Machine Learning”. In: *Philosophy & Technology*, pp. 1–16.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). “Imagenet classification with deep convolutional neural networks”. In: *Advances in neural information processing systems*, pp. 1097–1105.

Krumm, J., Davies, N., and Narayanaswami, C. (2008). “User-generated content”. In: *IEEE Pervasive Computing* 7.4, pp. 10–11.

Kullback, S. and Leibler, R. A. (1951). “On information and sufficiency”. In: *The annals of mathematical statistics* 22.1, pp. 79–86.

Kwalitan (2018). *Kwalitan [online]*. Available at https://www.kwalitan.nl/ (Accessed April 29th, 2021).

Landwehr, P. M., Wei, W., Kowalchuck, M., and Carley, K. M. (2016). “Using tweets to support disaster planning, warning and response”. In: *Safety science* 90, pp. 33–47.

Laney, D. (2001). “3D data management: Controlling data volume, velocity and variety”. In: *META Group Research Note* 6, p. 70.

Lansley, G. and Longley, P. A. (2016). “The geography of Twitter topics in London”. In: *Computers, Environment and Urban Systems* 58, pp. 85–96.

Lau, J. H., Chi, L., Tran, K.-N., and Cohn, T. (2017). “End-to-end network for twitter geolocation prediction and hashing”. In: *arXiv preprint arXiv:1710.04802*.

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A.-L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., et al. (2009). “Computational social science”. In: *Science* 323.5915, pp. 721–723.

Le, Q. and Mikolov, T. (2014). “Distributed representations of sentences and documents”. In: *International conference on machine learning*, pp. 1188–1196.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). “Deep learning”. In: *Nature* 521.7553, pp. 436–444.

Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., et al. (2017). “Photo-realistic single image super-resolution using a generative adversarial network”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4681–4690.

Lee, C.-H., Yang, H.-C., Chien, T.-F., and Wen, W.-S. (2011). “A novel approach for event detection by mining spatio-temporal information on microblogs”. In: *2011 International Conference on Advances in Social Networks Analysis and Mining*. IEEE, pp. 254–259.

Leicester-Shire Rutland Statistics Research, L. C. C. B. I. S. (2010). *Leicester Leicestershire Economic Assessment [online]*. Available at https://www.lsr-online.org/uploads/summary-may-20102.pdf (Accessed May 10th, 2021).

Leidner, J. L. and Lieberman, M. D. (2011). “Detecting geographical references in the form of place names and associated spatial natural language”. In: *SIGSPATIAL Special* 3.2, pp. 5–11.

Lessig, L. (2003). “An Information Society: Free or Feudal?” In: *World Summit on the Information Society*.

Leszczynski, A. and Crampton, J. (2016). “Introduction: Spatial big data and everyday life”. In: *Big Data & Society* 3.2, p. 2053951716661366.

Leszczynski, A. and Elwood, S. (2015). “Feminist geographies of new spatial media”. In: *The Canadian Geographer/Le Géographe canadien* 59.1, pp. 12–28.

Li, L., Goodchild, M. F., and Xu, B. (2013). “Spatial, temporal, and socioeconomic patterns in the use of Twitter and Flickr”. In: *Cartography and geographic information science* 40.2, pp. 61–77.

Li, R., Wang, S., and Chang, K. C.-C. (2012a). “Multiple location profiling for users and relationships from social network and content”. In: *Proceedings of the VLDB Endowment* 5.11, pp. 1603–1614.

Li, R., Wang, S., Deng, H., Wang, R., and Chang, K. C.-C. (2012b). “Towards social user profiling: unified and discriminative influence model for inferring home locations”. In: *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, pp. 1023–1031.

Li, W. (2020). “GeoAI: Where machine learning and big data converge in GIScience”. In: *Journal of Spatial Information Science*.

Liaropoulos, A (2013). “Exercising State Sovereignty in Cyberspace: An International Cyber-Order under Construction?” In: *Journal of Information Warfare* 12.2, pp. 19–26. ISSN: 14453312, 14453347. URL: https://www.jstor.org/stable/26486852.

Liu, Y., Liu, X., Gao, S., Gong, L., Kang, C., Zhi, Y., Chi, G., and Shi, L. (2015). “Social sensing: A new approach to understanding our socioeconomic environments”. In: *Annals of the Association of American Geographers* 105.3, pp. 512–530.

Liu, Y., Yuan, Y., and Zhang, F. (2020). “Mining urban perceptions from social media data”. In: *Journal of Spatial Information Science* 2020.20, pp. 51–55.

Liu, Y. and Cheng, T. (2018). “Understanding public transit patterns with open geodemographics to facilitate public transport planning”. In: *Transportmetrica A: Transport Science*, pp. 1–28.

Liu, Z. and Huang, Y. (2016). “Closeness and structure of friends help to estimate user locations”. In: *International Conference on Database Systems for Advanced Applications*. Springer, pp. 33–48.

Liuying, Z. and Sichun, Y. (2018). “Automatic classification of geographical problems based on SVM”. In: *Application Research of Computers* 9, p. 36.

Lloyd, C. D. (2016). “Spatial scale and small area population statistics for England and Wales”. In: *International Journal of Geographical Information Science* 30.6, pp. 1187–1206.

Local Government Act, L. (1972). *Local Government Act 1972 [online]*. Available at https://www.legislation.gov.uk/ukpga/1972/70/contents (Accessed May 10th, 2021).

Logan, J. R. (2012). “Making a place for space: Spatial thinking in social science”. In: *Annual review of sociology* 38, pp. 507–524.

Longley, P. (2005). “Geographical information systems: A renaissance of geodemographics for public service delivery”. In: *Progress in Human Geography* 29.1, pp. 57–63.

Longley, P. (Nov. 2007). “Some challenges to geodemographic analysis and their wider implications for the practice of GIScience”. In: *Computers, Environment and Urban Systems* 31, pp. 617–622. DOI: 10.1016/j.compenvurbsys.2007.10.002.

Longley, P. and Singleton, A. (2014). *London Output Area Classification (LOAC): Final Report*. Available at https://data.london.gov.uk/dataset/london-area-classification (Accessed 1st March 2020).

Lueg, C. and Fisher, D. (2012). *From Usenet to CoWebs: interacting with social information spaces*. Springer Science & Business Media.

Luo, F., Cao, G., Mulligan, K., and Li, X. (2016). "Explore spatiotemporal and demographic characteristics of human mobility via Twitter: A case study of Chicago". In: *Applied Geography* 70, pp. 11–25.

Lynch, K. (1960). *The image of the city*. Vol. 11. MIT press.

Ma, X., Tao, Z., Wang, Y., Yu, H., and Wang, Y. (2015). "Long short-term memory neural network for traffic speed prediction using remote microwave sensor data". In: *Transportation Research Part C: Emerging Technologies* 54, pp. 187–197.

Mac Aodha, O., Cole, E., and Perona, P. (2019). "Presence-only geographical priors for fine-grained image classification". In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 9596–9606.

MacEachren, A. M., Jaiswal, A., Robinson, A. C., Pezanowski, S., Savelyev, A., Mitra, P., Zhang, X., and Blanford, J. (2011). "Senseplace2: Geotwitter analytics support for situational awareness". In: *2011 IEEE conference on visual analytics science and technology (VAST)*. IEEE, pp. 181–190.

MacLeod, G. (2018). "The Grenfell Tower atrocity: Exposing urban worlds of inequality, injustice, and an impaired democracy". In: *City* 22.4, pp. 460–489.

MacQueen, J. et al. (1967). "Some methods for classification and analysis of multivariate observations". In: *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*. Vol. 1. 14. Oakland, CA, USA, pp. 281–297.

Mai, G., Janowicz, K., Cai, L., Zhu, R., Regalia, B., Yan, B., Shi, M., and Lao, N. (2020). "SE-KGE: A location-aware Knowledge Graph Embedding model for Geographic Question Answering and Spatial Semantic Lifting". In: *Transactions in GIS*.

Mai, G., Yan, B., Janowicz, K., and Zhu, R. (2019). "Relaxing unanswerable geographic questions using a spatially explicit knowledge graph embedding model". In: *The Annual International Conference on Geographic Information Science*. Springer, pp. 21–39.

Manley, E. and Dennett, A. (2019). "New Forms of Data for Understanding Urban Activity in Developing Countries". In: *Applied Spatial Analysis and Policy* 12.1, pp. 45–70.

Mao, X. J., Shen, C., and Yang, Y. B. (2016). "Image Restoration Using Convolutional Auto-encoders with Symmetric Skip Connections". In:

Mark, D. M. (2003). "Geographic information science: Defining the field". In: *Foundations of geographic information science* 1, pp. 3–18.

Markham, A. and Buchanan, E. (2012). "Ethical decision-making and internet research: Version 2.0. recommendations from the AoIR ethics working committee". In: *Available online: aoir. org/reports/ethics2. pdf*.

Martí, P., Serrano-Estrada, L., and Nolasco-Cirugeda, A. (2017). "Using locative social media and urban cartographies to identify and locate successful urban plazas". In: *Cities* 64, pp. 66–78.

Martín, Y., Li, Z., and Cutter, S. L. (2017). "Leveraging Twitter to gauge evacuation compliance: Spatiotemporal analysis of Hurricane Matthew". In: *PLoS one* 12.7, e0181701.

Martinčić-Ipšić, S., Močibob, E., and Perc, M. (2017). "Link prediction on Twitter". In: *PloS one* 12.7, e0181079.

Marz, N. and Warren, J. (2015). *Big Data: Principles and best practices of scalable realtime data systems*. Manning Publications Co.

Mason, A (2001). "4. Unresolved Issues". In: *Journal of Economic Perspeftives* 2.2, pp. 41–58.

Mayer, M., Heck, D. W., and Mocnik, F.-B. (2020). "Shared mental models as a psychological explanation for converging mental representations of place–the example of OpenStreetMap". In: *Proceedings of the 2nd International Symposium on Platial Information Science (PLATIAL'19)*, pp. 43–50.

Mayer-Schonberger, V. and Cukier, K. (2013). "Big data: A revolution that will change how we live, work and think". In: *London: John Murray. ISBN* 978, p. 0544227750.

McLennan, D., Noble, S., Noble, M., Plunkett, E., Wright, G., and Gutacker, N. (2019). "The English Indices of Deprivation 2019: technical report". In: *Ministry of Housing, Communities and Local Government*.

McLuhan, M. (1975). "McLuhan's laws of the media". In: *Technology and Culture* 16.1, pp. 74–78.

Memon, I., Chen, L., Majid, A., Lv, M., Hussain, I., and Chen, G. (2015). "Travel recommendation using geo-tagged photos in social media for tourist". In: *Wireless Personal Communications* 80.4, pp. 1347–1362.

Metcalf, J. and Crawford, K. (2016). "Where are human subjects in big data research? The emerging ethics divide". In: *Big Data & Society* 3.1, p. 2053951716650211.

Mikheev, A., Moens, M., and Grover, C. (1999). "Named entity recognition without gazetteers". In: *Ninth Conference of the European Chapter of the Association for Computational Linguistics*.

Mikolov, T., Deoras, A., Povey, D., Burget, L., and Černockỳ, J. (2011). "Strategies for training large scale neural network language models". In: *Automatic Speech Recognition and Understanding (ASRU), 2011 IEEE Workshop on*. IEEE, pp. 196–201.

Miller, H. J. (2010). "The data avalanche is here. Shouldn't we be digging?" In: *Journal of Regional Science* 50.1, pp. 181–201.

Miller, H. J. and Goodchild, M. F. (2015a). "Data-driven geography". In: *GeoJournal* 80.4, pp. 449–461. ISSN: 1572-9893. DOI: 10.1007/s10708-014-9602-6. URL: https://doi.org/10.1007/s10708-014-9602-6.

Miller, H. J. and Goodchild, M. F. (2015b). "Data-driven geography". In: *GeoJournal* 80.4, pp. 449–461.

Mishra, S., Pappu, A., and Bhamidipati, N. (2019). "Inferring advertiser sentiment in online articles using wikipedia footnotes". In: *Companion Proceedings of The 2019 World Wide Web Conference*, pp. 1224–1231.

Mocnik, F.-B. (Jan. 2016). "A Scale-Invariant Spatial Graph Model". In:

Modai-Snir, T. and Ham, M. van (2018). "Neighbourhood change and spatial polarization: The roles of increasing inequality and divergent urban development". In: *Cities* 82, pp. 108–118.

Moncla, L., Gaio, M., and Mustiere, S. (2014a). "Automatic itinerary reconstruction from texts". In: *International Conference on Geographic Information Science*. Springer, pp. 253–267.

Moncla, L., Renteria-Agualimpia, W., Nogueras-Iso, J., and Gaio, M. (2014b). "Geocoding for texts with fine-grain toponyms: an experiment on a geoparsed hiking descriptions corpus". In: *Proceedings of the 22nd acm sigspatial international conference on advances in geographic information systems*, pp. 183–192.

Mora, H., Pérez-delHoyo, R., Paredes-Pérez, J. F., and Mollá-Sirvent, R. A. (2018). "Analysis of social networking service data for smart urban planning". In: *Sustainability* 10.12, p. 4732.

Nadeem, M. S., Franqueira, V. N., Zhai, X., and Kurugollu, F. (2019). “A survey of deep learning solutions for multimedia visual content analysis”. In: *IEEE Access* 7, pp. 84003–84019.

Naik, N., Kominers, S. D., Raskar, R., Glaeser, E. L., and Hidalgo, C. A. (2017). “Computer vision uncovers predictors of physical urban change”. In: *Proceedings of the National Academy of Sciences* 114.29, pp. 7571–7576.

Narayanan, A. and Shmatikov, V. (2009). “De-anonymizing social networks”. In: *2009 30th IEEE symposium on security and privacy*. IEEE, pp. 173–187.

Newsam, S. and Leung, D. (2019). “Georeferenced Social Multimedia as Volunteered Geographic Information”. In: *CyberGIS for Geospatial Discovery and Innovation*. Springer, pp. 225–246.

Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., and Ng, A. Y. (2011). “Multimodal deep learning”. In: *Proceedings of the 28th international conference on machine learning (ICML-11)*, pp. 689–696.

Nickel, M., Murphy, K., Tresp, V., and Gabrilovich, E. (2015). “A review of relational machine learning for knowledge graphs”. In: *Proceedings of the IEEE* 104.1, pp. 11–33.

Niu, T., Chen, Y., and Yuan, Y. (2020). “Measuring urban poverty using multi-source data and a random forest algorithm: A case study in Guangzhou”. In: *Sustainable Cities and Society* 54, p. 102014.

Noon, C. E. and Hankins, C. T. (2001). “Spatial data visualization in healthcare: supporting a facility location decision via GIS-based market analysis”. In: *Proceedings of the 34th Annual Hawaii International Conference on System Sciences*. IEEE, 10–pp.

Noronha, V. T. and Goodchild, M. F. (1992). “Modeling interregional interaction: Implications for defining functional regions”. In: *Annals of the Association of American Geographers* 82.1, pp. 86–102.

Nowak, J., Taspinar, A., and Scherer, R. (2017). “LSTM recurrent neural networks for short text and sentiment classification”. In: *International Conference on Artificial Intelligence and Soft Computing*. Springer, pp. 553–562.

NVIDIA, Vingelmann, P., and Fitzek, F. H. (2020). *CUDA, release: 10.2.89*. URL: https://developer.nvidia.com/cuda-toolkit.

Nystuen, J. D. and Dacey, M. F. (1961). “A graph theory interpretation of nodal regions”. In: *Papers of the Regional Science Association*. Vol. 7. 1. Springer, pp. 29–42.

O’neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown.

ONS (2011). *Census geography: An overview of the various geographies used in the production of statistics collected via the UK census. [online]*. Available at https://www.ons.gov.uk/methodology/geography/ukgeographies/censusgeography (Accessed April 30th, 2021).

ONS (2016). *Introduction to Output Areas - the building block of Census geography. [online]*. Available at https://www.ons.gov.uk/census/2001censusandearlier/dataandproducts/outputgeography/outputareas (Accessed April 30th, 2021).

Ooi, B. C. (1987). “Spatial kd-tree: A data structure for geographic database”. In: *Datenbanksysteme in Büro, Technik und Wissenschaft*. Springer, pp. 247–258.

OpenStreetMap (2004). *OpenStreetMap [online]*. Availabe at https://www.openstreetmap.org (Accessed April 29th, 2021).

O’Reilly, T. and Dougherty, D. (2004). *What Is Web 2.0: Design Patterns and Business Models for the Next Generation of Software [online]*. Availabe at https://www.oreilly.com/pub/a/web2/archive/what-is-web-20.html (Accessed April 29th, 2021).

Ostermann, F. O., Huang, H., Andrienko, G., Andrienko, N., Capineri, C., Farkas, K., and Purves, R. S. (2015). "Extracting and comparing places using geo-social media". In: *ISPRS Geospatial week 2015* 2.W5.

O'Sullivan, D. and Unwin, D. (2010). *Geographic information analysis*. John Wiley & Sons.

Pacione, M. (1989). "The urban crisis: poverty and deprivation in the Scottish city". In: *Scottish Geographical Magazine* 105.2, pp. 101–115.

Papamichail, G. P. and Papamichail, D. P. (2007). "The k-means range algorithm for personalized data clustering in e-commerce". In: *European Journal of Operational Research* 177.3, pp. 1400–1408.

Pereira, G. C., Rocha, M. C. F., and Florentino, P. V. (2013). "Spatial Representation: City and Digital Spaces". In: *International Conference on Computational Science and Its Applications*. Springer, pp. 524–537.

Peter, G. and Rodney, W. (1974). *Mental maps*.

Petersen, J., Gibin, M., Longley, P., Mateos, P., Atkinson, P., and Ashby, D. (2011). "Geodemographics as a tool for targeting neighbourhoods in public health campaigns". In: *Journal of Geographical Systems* 13.2, pp. 173–192.

Pinheiro, M. B. and Davis, C. A. (2018). "ThemeRise: a theme-oriented framework for volunteered geographic information applications". In: *Open Geospatial Data, Software and Standards* 3.1, p. 9.

Poorthuis, A. (2015). "Social space and social media: Analyzing urban space with big data". In:

Power, M. J., Neville, P., Devereux, E., Haynes, A., and Barnes, C. (2013). "'Why bother seeing the world for real?': Google Street View and the representation of a stigmatised neighbourhood". In: *new media & society* 15.7, pp. 1022–1040.

Prensky, M. (2009). "H. sapiens digital: From digital immigrants and digital natives to digital wisdom". In: *Innovate: journal of online education* 5.3.

Purves, R. S., Clough, P., Jones, C. B., Hall, M. H., and Murdock, V. (2018). "Geographic information retrieval: progress and challenges in spatial search of text". In: *Foundations and Trends in Information Retrieval* 12.2-3, pp. 164–318.

Purves, R. S., Winter, S., and Kuhn, W. (2019). "Places in information science". In: *Journal of the Association for Information Science and Technology* 70.11, pp. 1173–1182.

Putz, S. (1994). "Interactive information services using World-Wide Web hypertext". In: *Computer Networks and ISDN Systems* 27.2, pp. 273–280.

Qi, W., Procter, R., Zhang, J., and Guo, W. (2019). "Mapping consumer sentiment toward wireless services using geospatial twitter data". In: *IEEE Access* 7, pp. 113726–113739.

Qiu, P., Gao, J., Yu, L., and Lu, F. (2019). "Knowledge embedding with geospatial distance restriction for geographic knowledge graph completion". In: *ISPRS International Journal of Geo-Information* 8.6, p. 254.

Quirkos (2014). *Quirkos [online]*. Available at `https://www.quirkos.com/learn-qualitative/features.html` (Accessed April 29th, 2021).

Rawashdeh, A. and Ralescu, A. L. (2015). "Similarity Measure for Social Networks-A Brief Survey." In: *Maics*, pp. 153–159.

Reades, J., De Souza, J., and Hubbard, P. (2019). "Understanding urban gentrification through machine learning". In: *Urban Studies* 56.5, pp. 922–942.

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., et al. (2019). "Deep learning and process understanding for data-driven Earth system science". In: *Nature* 566.7743, pp. 195–204.

Rey, S. J., Anselin, L., Folch, D. C., Arribas-Bel, D., Sastré Gutiérrez, M. L., and Interlante, L. (2011). "Measuring spatial dynamics in metropolitan areas". In: *Economic Development Quarterly* 25.1, pp. 54–64.

Rezende, D. J., Mohamed, S., and Wierstra, D. (2014). "Stochastic backpropagation and approximate inference in deep generative models". In: *arXiv preprint arXiv:1401.4082*.

Ritchie, H. and Roser, M. (2018). "Urbanization". In: *Our World in Data*.

Roche, S. (2016). "Geographic information science II: Less space, more places in smart cities". In: *Progress in Human Geography* 40.4, pp. 565–573.

Rosenblatt, F. (1958). "The perceptron: a probabilistic model for information storage and organization in the brain." In: *Psychological review* 65.6, p. 386.

Rout, D., Bontcheva, K., Preoţiuc-Pietro, D., and Cohn, T. (2013). "Where's@ wally?: a classification approach to geolocating users based on their social ties". In: *Proceedings of the 24th ACM Conference on Hypertext and Social Media*. ACM, pp. 11–20.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. (2015). "Imagenet large scale visual recognition challenge". In: *International journal of computer vision* 115.3, pp. 211–252.

Sadilek, A., Kautz, H., and Bigham, J. P. (2012). "Finding your friends and following them to where you are". In: *Proceedings of the fifth ACM international conference on Web search and data mining*. ACM, pp. 723–732.

Salakhutdinov, R. and Hinton, G. (2009). "Semantic hashing". In: *International Journal of Approximate Reasoning* 50.7, pp. 969–978.

Samuel, A. L. (1959). "Some studies in machine learning using the game of checkers". In: *IBM Journal of research and development* 3.3, pp. 210–229.

Saunders, P. (2004). *Towards a credible poverty framework: from income poverty to deprivation*. Vol. 131. Social Policy Research Centre.

Scarselli, F., Gori, M., Tsoi, A. C., Hagenbuchner, M., and Monfardini, G. (2008). "The graph neural network model". In: *IEEE Transactions on Neural Networks* 20.1, pp. 61–80.

Schneider-Sliwa, R. (2001). "Urban Geography". In: *International Encyclopedia of the Social Behavioral Sciences* 24, pp. 16008–16015.

Scholten, H., Fruijter, S., Dilo, A., and Van Borkulo, E. (2008). "Spatial Data Infrastructure for emergency response in Netherlands". In: *Remote sensing and GIS technologies for monitoring and prediction of disasters*. Springer, pp. 179–197.

Sechelea, A., Do Huu, T., Zimos, E., and Deligiannis, N. (2016). "Twitter data clustering and visualization". In: *2016 23rd International Conference on Telecommunications (ICT)*. IEEE, pp. 1–5.

See, L., Mooney, P., Foody, G., Bastin, L., Comber, A., Estima, J., Fritz, S., Kerle, N., Jiang, B., Laakso, M., et al. (2016). "Crowdsourcing, citizen science or volunteered geographic information? The current state of crowdsourced geographic information". In: *ISPRS International Journal of Geo-Information* 5.5, p. 55.

Senaratne, H., Mobasheri, A., Ali, A. L., Capineri, C., and Haklay, M. (2016). "A review of volunteered geographic information quality assessment methods". In: *International Journal of Geographical Information Science*, pp. 1–29.

Seydi, S. T., Hasanlou, M., and Amani, M. (2020). "A New End-to-End Multi-Dimensional CNN Framework for Land Cover/Land Use Change Detection in Multi-Source Remote Sensing Datasets". In: *Remote Sensing* 12.12, p. 2010.

Shaw, E. (2017). "Parsing Perceptions of Place: Locative and Textual Representations of Place Émilie-Gamelin on Twitter". PhD thesis. Concordia University.

Shaw, J. and Graham, M. (2017). "An informational right to the city? Code, content, control, and the urbanization of information". In: *Antipode* 49.4, pp. 907–927.

Shaw, S.-L. and Sui, D. (2020). "Understanding the new human dynamics in smart spaces and places: Toward a Splatial framework". In: *Annals of the American Association of Geographers* 110.2, pp. 339–348.

Shelton, T. (2017). "Spatialities of data: mapping social media 'beyond the geotag'". In: *GeoJournal* 82.4, pp. 721–734.

Shelton, T., Poorthuis, A., Graham, M., and Zook, M. (2014). "Mapping the data shadows of Hurricane Sandy: Uncovering the sociospatial dimensions of 'big data'". In: *Geoforum* 52, pp. 167–179.

Shelton, T., Poorthuis, A., and Zook, M. (2015). "Social media and the city: Rethinking urban socio-spatial inequality using user-generated geographic information". In: *Landscape and urban planning* 142, pp. 198–211.

Shelton, T., Zook, M., and Graham, M. (2012). "The technology of religion: Mapping religious cyberscapes". In: *The Professional Geographer* 64.4, pp. 602–617.

Shen, Y. and Karimi, K. (2016). "Urban function connectivity: Characterisation of functional urban streets with social media check-in data". In: *Cities* 55, pp. 9–21.

Shildrick, T. (2018). "Lessons from Grenfell: Poverty propaganda, stigma and class power". In: *The Sociological Review* 66.4, pp. 783–798.

Shirky, C. (Jan. 2010). *Cognitive Surplus: Creativity and Generosity in a Connected Age*, p. 242.

Singleton, A., Pavlis, M., and Longley, P. A. (2016). "The stability of geodemographic cluster assignments over an intercensal period". In: *Journal of Geographical Systems* 18.2, pp. 97–123.

Skoumas, G., Pfoser, D., Kyrillidis, A., and Sellis, T. (2016). "Location estimation using crowdsourced spatial relations". In: *ACM Transactions on Spatial Algorithms and Systems (TSAS)* 2.2, pp. 1–23.

Slingsby, A., Dykes, J., and Wood, J. (2011). "Exploring uncertainty in geodemographics with interactive graphics". In: *IEEE Transactions on Visualization and Computer Graphics* 17.12, pp. 2545–2554.

Sloan, L. and Morgan, J. (2015). "Who tweets with their location? Understanding the relationship between demographic characteristics and the use of geoservices and geotagging on Twitter". In: *PloS one* 10.11, e0142209.

Sloggett, A. and Joshi, H. (1998). "Deprivation indicators as predictors of life events 1981-1992 based on the UK ONS Longitudinal Study." In: *Journal of Epidemiology & Community Health* 52.4, pp. 228–233.

Smelser, N. J., Baltes, P. B., et al. (2001). *International encyclopedia of the social & behavioral sciences*. Vol. 11. Elsevier Amsterdam.

Smith, T., Noble, M., Noble, S., Wright, G., McLennan, D., Plunkett, E., et al. (2015). "The English indices of deprivation 2015". In: *London: Department for Communities and Local Government*.

Soliman, A., Soltani, K., Yin, J., Padmanabhan, A., and Wang, S. (2017). "Social sensing of urban land use based on analysis of Twitter users' mobility patterns". In: *PloS one* 12.7, e0181657.

Steadman, I. (2013). "Big data and the death of the theorist". In: *wired. co. uk* 25, pp. 2013–01.

Stedman, R. C. (2002). "Toward a social psychology of place: Predicting behavior from place-based cognitions, attitude, and identity". In: *Environment and behavior* 34.5, pp. 561–581.

Steiger, E., Westerholt, R., and Zipf, A. (2016). "Research on social media feeds–A GIScience perspective". In: *European Handbook of Crowdsourced Geographic Information*, p. 237.

Sui, D. and DeLyser, D. (2012). “Crossing the qualitative-quantitative chasm I: Hybrid geographies, the spatial turn, and volunteered geographic information (VGI)”. In: *Progress in Human Geography* 36.1, pp. 111–124.

Sui, D. and Goodchild, M. (2011). “The convergence of GIS and social media: challenges for GIScience”. In: *International Journal of Geographical Information Science* 25.11, pp. 1737–1748.

Sui, D. Z. (2008). “The wikification of GIS and its consequences: Or Angelina Jolie’s new tattoo and the future of GIS”. In: *Computers, environment and urban systems* 1.32, pp. 1–5.

Sui, D. Z. and Goodchild, M. F. (2001). “GIS as media?” In: *International Journal of Geographical Information Science* 15.5, pp. 387–390.

Sulis, P., Manley, E., Zhong, C., and Batty, M. (2018). “Using mobility data as proxy for measuring urban vitality”. In: *Journal of Spatial Information Science* 2018.16, pp. 137–162.

Sun, S., Sarukkai, R., Kwok, J., and Shet, V. (2018). “Accurate deep direct geo-localization from ground imagery and phone-grade gps”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1016–1023.

Suresh, S., Chodosh, N., and Abello, M. (2018). “DeepGeo: Photo Localization with Deep Neural Network”. In: *arXiv preprint arXiv:1810.03077*.

Sutskever, I., Vinyals, O., and Le, Q. V. (2014). “Sequence to sequence learning with neural networks”. In: *Advances in neural information processing systems*, pp. 3104–3112.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2015). “Going deeper with convolutions”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–9.

Tait, A (2012). “Changes to output areas and super output areas in England and Wales, 2001 to 2011”. In: *Office for National Statistics*.

Talbot, R. J. (1991). “Underprivileged areas and health care planning: implications of use of Jarman indicators of urban deprivation.” In: *BMJ* 302.6773, pp. 383–386.

Tate, H. (2018). *Economic Impact of Tourism – Visitor Volume and Value 2018 [online]*. Available at https://www.cumbriatourism.org/what-we-do/research/economic-impact-of-tourism/ (Accessed May 18th, 2021).

The Scrutiny Commission Information Policy Team, R. C. C. (2018). *Supporting Evidence - Poverty in Rutland [online]*. Available at https://rutlandcounty.moderngov.co.uk/documents/s14016/Report%20No.%20237-2018%20Appendix%20B%20Supporting%20Data%20-%20Poverty%20in%20Rutland%20v.2.pdf (Accessed May 10th, 2021).

Thrift, N. J. (1983). “On the determination of social action in space and time”. In: *Environment and planning D: Society and space* 1.1, pp. 23–57.

Tobler, W. R. (1970). “A computer movie simulating urban growth in the Detroit region”. In: *Economic geography* 46.sup1, pp. 234–240.

Townsend, L. and Wallace, C. (2016). *Social media research: A guide to ethics [online]*. Available at http://www.dotrural.ac.uk/socialmediaresearchethics.pdf (Accessed 5th June, 2021).

Trisedya, B. D., Qi, J., and Zhang, R. (2019). “Entity alignment between knowledge graphs using attribute embeddings”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 33, pp. 297–304.

Trouillon, T., Welbl, J., Riedel, S., Gaussier, É., and Bouchard, G. (2016). “Complex embeddings for simple link prediction”. In: *International Conference on Machine Learning*, pp. 2071–2080.

Tsou, M.-H. (2015). “Research challenges and opportunities in mapping social media and Big Data”. In: *Cartography and Geographic Information Science* 42.sup1, pp. 70–74.

Tuan, Y.-F. (1977). *Space and place: The perspective of experience*. U of Minnesota Press.

Tuan, Y.-F. (1979). “Space and place: humanistic perspective”. In: *Philosophy in geography*. Springer, pp. 387–427.

Turner, A. (2021). *How Many Smartphones Are In The World? [online]*. Availabe at https://www.bankmycell.com/blog/how-many-phones-are-in-the-world (Accessed April 28th, 2021).

Twitter (2021a). *About Twitter's APIs [online]*. Available at https://help.twitter.com/en/rules-and-policies/twitter-api (Accessed April 30th, 2021).

Twitter (2021b). *Tweet location FAQs [online]*. Available at https://help.twitter.com/en/safety-and-security/tweet-location-settings (Accessed May 6th, 2021).

TwitterSupport (2019). *Most people don't tag their precise location in Tweets, so we're removing this ability to simplify your Tweeting experience. You'll still be able to tag your precise location in Tweets through our updated camera. It's helpful when sharing on-the-ground moments. [online]*. Available at https://twitter.com/TwitterSupport/status/1141039841993355264 (Accessed April 30th, 2021).

Uçar, A., Demir, Y., and Güzeliş, C. (2017). “Object recognition and detection with deep learning for autonomous driving applications”. In: *Simulation* 93.9, pp. 759–769.

Van Diggelen, F. and Enge, P. (2015). “The world's first GPS MOOC and worldwide laboratory using smartphones”. In: *Proceedings of the 28th international technical meeting of the satellite division of the institute of navigation (ION GNSS+ 2015)*, pp. 361–369.

Venerandi, A., Quattrone, G., Capra, L., Quercia, D., and Saez-Trumper, D. (2015). “Measuring urban deprivation from user generated content”. In: *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, pp. 254–264.

VoPham, T., Hart, J. E., Laden, F., and Chiang, Y.-Y. (2018). “Emerging trends in geospatial artificial intelligence (geoAI): potential applications for environmental epidemiology”. In: *Environmental Health* 17.1, p. 40.

Vrandečić, D. and Krötzsch, M. (2014). “Wikidata: a free collaborative knowledgebase”. In: *Communications of the ACM* 57.10, pp. 78–85.

Wadawadagi, R. S. and Pagi, V. B. (2020). “Sentiment Analysis on Social Media: Recent Trends in Machine Learning”. In: *Handbook of Research on Emerging Trends and Applications of Machine Learning*. IGI Global, pp. 508–527.

Wales, J. (2014). *At 15, Wikipedia is finally finding its way to the truth [online]*. Availabe at https://web.archive.org/web/20180505163037/https://www.wired.com/2016/01/at-15-wikipedia-is-finally-finding-its-way-to-the-truth (Accessed April 29th, 2021).

Wall, M. and Kirdnark, T. (2012). “Online maps and minorities: Geotagging Thailand's Muslims”. In: *New Media & Society* 14.4, pp. 701–716.

Wallgrün, J. O., Karimzadeh, M., MacEachren, A. M., and Pezanowski, S. (2018). “GeoCorpora: building a corpus to test and train microblog geoparsers”. In: *International Journal of Geographical Information Science* 32.1, pp. 1–29.

Wami, W. M., Dundas, R., Molaodi, O. R., Tranter, M., Leyland, A. H., and Katikireddi, S. V. (2019). “Assessing the potential utility of commercial ‘big data’for health research: Enhancing small-area deprivation measures with Experian^{TM} Mosaic groups”. In: *Health & place* 57, pp. 238–246.

Wang, H., Hu, Y., Tang, L., and Zhuo, Q. (2020a). "Distribution of Urban Blue and Green Space in Beijing and Its Influence Factors". In: *Sustainability* 12.6, p. 2252.

Wang, J., Hu, Y., and Joseph, K. (2020b). "NeuroTPR: A neuro-net toponym recognition model for extracting locations from social media messages". In: *Transactions in GIS*.

Wang, Q., Yuan, Z., Du, Q., and Li, X. (2018a). "GETNET: A general end-to-end 2-D CNN framework for hyperspectral image change detection". In: *IEEE Transactions on Geoscience and Remote Sensing* 57.1, pp. 3–13.

Wang, Q., Zhang, X., Chen, G., Dai, F., Gong, Y., and Zhu, K. (2018b). "Change detection based on Faster R-CNN for high-resolution remote sensing images". In: *Remote sensing letters* 9.10, pp. 923–932.

Wang, R., Liu, Y., Lu, Y., Zhang, J., Liu, P., Yao, Y., and Grekousis, G. (2019a). "Perceptions of built environment and health outcomes for older Chinese in Beijing: A big data approach with street view images and deep learning technique". In: *Computers, Environment and Urban Systems* 78, p. 101386.

Wang, S., Zhang, X., Ye, P., Du, M., Lu, Y., and Xue, H. (2019b). "Geographic knowledge graph (GeoKG): A formalized geographic knowledge representation". In: *ISPRS International Journal of Geo-Information* 8.4, p. 184.

Wang, Y. and Li, B. (2015). "Sentiment analysis for social media images". In: *Data Mining Workshop (ICDMW), 2015 IEEE International Conference on*. IEEE, pp. 1584–1591.

Wang, Z. and Xue, X. (2014). "Multi-class support vector machine". In: *Support Vector Machines Applications*. Springer, pp. 23–48.

Wang, Z., Ye, X., and Tsou, M.-H. (2016a). "Spatial, temporal, and content analysis of Twitter for wildfire hazards". In: *Natural Hazards* 83.1, pp. 523–540.

Wang, Z., Ye, X., and Tsou, M.-H. (2016b). "Spatial, temporal, and content analysis of Twitter for wildfire hazards". In: *Natural Hazards* 83.1, pp. 523–540. ISSN: 1573-0840. DOI: 10.1007/s11069-016-2329-6. URL: https://doi.org/10.1007/s11069-016-2329-6.

Watkins, D. A. (2012). "Digital facets of place: Flickr's mappings of the US-Mexico borderlands". PhD thesis. University of Oregon.

Webb, H., Jirotka, M., Stahl, B. C., Housley, W., Edwards, A., Williams, M., Procter, R., Rana, O., and Burnap, P. (2017). "The ethical challenges of publishing Twitter data for research dissemination". In: *Proceedings of the 2017 ACM on Web Science Conference*, pp. 339–348.

Wei, Y. D. and Ye, X. (2014). "Urbanization, urban land expansion and environmental change in China". In: *Stochastic environmental research and risk assessment* 28.4, pp. 757–765.

Weller, K., Bruns, A., Burgess, J., Mahrt, M., and Puschmann, C. (2014). *Twitter and society*. Vol. 89. Peter Lang.

Wikimapia (2006). *Wikimapia [online]*. Availabe at https://wikimapia.org (Accessed April 29th, 2021).

Wikimedia (2003). *Wikimedia Toolforge [online]*. Available at https://wikitech.wikimedia.org/wiki/Portal:Toolforge (Accessed May 18th, 2021).

Wilken, R. (2012). *Online territories: Globalization, mediated practice and social space*.

Wilson, M. W. (2015). "Morgan Freeman is dead and other big data stories". In: *Cultural geographies* 22.2, pp. 345–349.

Wohn, D. Y. and Na, E.-K. (2011). "Tweeting about TV: Sharing television viewing experiences via social media message streams". In: *First Monday*.

Wold, S., Esbensen, K., and Geladi, P. (1987). "Principal component analysis". In: *Chemometrics and intelligent laboratory systems* 2.1-3, pp. 37–52.

Wong, D. W. (2004). "The modifiable areal unit problem (MAUP)". In: *WorldMinds: geographical perspectives on 100 problems*. Springer, pp. 571–575.

Wu, W., Jiang, S., Liu, R., Jin, W., and Ma, C. (2020a). "Economic development, demographic characteristics, road network and traffic accidents in Zhongshan, China: gradient boosting decision tree model". In: *Transportmetrica A: transport science* 16.3, pp. 359–387.

Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., and Philip, S. Y. (2020b). "A comprehensive survey on graph neural networks". In: *IEEE Transactions on Neural Networks and Learning Systems*.

Wyly, E. (2014). "The new quantitative revolution". In: *Dialogues in Human Geography* 4.1, pp. 26–38.

Xing, J. and Sieber, R. E. (2018). "Propagation of Uncertainty for Volunteered Geographic Information in Machine Learning". In: *10th International Conference on Geographic Information Science*. Newcastle University.

Xu, C., Cetintas, S., Lee, K.-C., and Li, L.-J. (2014). "Visual sentiment prediction with deep convolutional neural networks". In: *arXiv preprint arXiv:1411.5731*.

Xu, J., Rahmatizadeh, R., Bölöni, L., and Turgut, D. (2017a). "Real-time prediction of taxi demand using recurrent neural networks". In: *IEEE Transactions on Intelligent Transportation Systems* 19.8, pp. 2572–2581.

Xu, Z., Liu, Y., Zhang, H., Luo, X., Mei, L., and Hu, C. (2017b). "Building the multimodal storytelling of urban emergency events based on crowdsensing of social media analytics". In: *Mobile Networks and Applications* 22.2, pp. 218–227.

Yan, B. (2019). *Geographic knowledge graph summarization*. Vol. 39. IOS Press.

Yan, B., Janowicz, K., Mai, G., and Zhu, R. (2019). "A spatially explicit reinforcement learning model for geographic knowledge graph summarization". In: *Transactions in GIS* 23.3, pp. 620–640.

Yang, H.-W., Pan, Z.-G., Wang, X.-Z., and Xu, B. (2004). "A personalized products selection assistance based on e-commerce machine learning". In: *Proceedings of 2004 International Conference on Machine Learning and Cybernetics (IEEE Cat. No. 04EX826)*. Vol. 4. IEEE, pp. 2629–2633.

Yang, J. and Leskovec, J. (2011). "Patterns of temporal variation in online media". In: *Proceedings of the fourth ACM international conference on Web search and data mining*. ACM, pp. 177–186.

Yang, Y., Lichtenwalter, R. N., and Chawla, N. V. (2015). "Evaluating link prediction methods". In: *Knowledge and Information Systems* 45.3, pp. 751–782.

Yardi, S. and Boyd, D. (2010). "Tweeting from the town square: Measuring geographic local networks". In: *Fourth International AAAI Conference on Weblogs and Social Media*.

Ye, J., Zhu, Z., and Cheng, H. (2013). "What's your next move: User activity prediction in location-based social networks". In: *Proceedings of the 2013 SIAM International Conference on Data Mining*. SIAM, pp. 171–179.

You, Q., Luo, J., Jin, H., and Yang, J. (2015). "Robust Image Sentiment Analysis Using Progressively Trained and Domain Transferred Deep Networks." In: *AAAI*, pp. 381–388.

Yu, R., Li, Y., Shahabi, C., Demiryurek, U., and Liu, Y. (2017). "Deep Learning: A Generic Approach for Extreme Condition Traffic Forecasting". In: *Proceedings of the 2017 SIAM International Conference on Data Mining*, pp. 777–785. DOI: 10.1137/1.9781611974973.87. eprint: https://epubs.siam.org/doi/pdf/10.1137/1.9781611974973.87. URL: https://epubs.siam.org/doi/abs/10.1137/1.9781611974973.87.

Zahra, K., Ostermann, F. O., and Purves, R. S. (2017). “Geographic variability of Twitter usage characteristics during disaster events”. In: *Geo-spatial information science* 20.3, pp. 231–240.

Zhan, X., Ukkusuri, S. V., and Zhu, F. (2014). “Inferring urban land use using large-scale social media check-in data”. In: *Networks and Spatial Economics* 14.3, pp. 647–667.

Zhang, K., Jin, Q., Pelechrinis, K., and Lappas, T. (2013). “On the importance of temporal dynamics in modeling urban activity”. In: *Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing*, pp. 1–8.

Zhang, M., He, C., and Liu, G. (2010). “Application of GIS in electronic commerce”. In: *2010 Second IITA International Conference on Geoscience and Remote Sensing*. Vol. 1. IEEE, pp. 483–486.

Zhang, W., Tang, P., and Zhao, L. (2019a). “Remote sensing image scene classification using CNN-CapsNet”. In: *Remote Sensing* 11.5, p. 494.

Zhang, W., Yu, Y., Qi, Y., Shu, F., and Wang, Y. (2019b). “Short-term traffic flow prediction based on spatio-temporal analysis and CNN deep learning”. In: *Transportmetrica A: Transport Science* 15.2, pp. 1688–1711.

Zhang, X., Chen, G., Wang, W., Wang, Q., and Dai, F. (2017). “Object-based land-cover supervised classification for very-high-resolution UAV images using stacked denoising autoencoders”. In: *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 10.7, pp. 3373–3385.

Zhang, Y., Cheng, T., and Aslam, N. S. (2019c). “Exploring the relationship between travel pattern and social-demographics using smart card data and household survey”. In: *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences-ISPRS Archives*. ISPRS, pp. 1375–1382.

Zhang, Y., Gao, Y., Xue, L., Shen, S., and Chen, K. (2008). “A common sense geographic knowledge base for GIR”. In: *Science in China Series E: Technological Sciences* 51.1, pp. 26–37.

Zhang, Y., Wang, S., Chen, B., and Cao, J. (2019d). “GCGAN: Generative adversarial nets with graph CNN for network-scale traffic prediction”. In: *2019 International Joint Conference on Neural Networks (IJCNN)*. IEEE, pp. 1–8.

Zheng, M., Li, T., Zhu, R., Chen, J., Ma, Z., Tang, M., Cui, Z., and Wang, Z. (2019). “Traffic accident’s severity prediction: A deep-learning approach-based CNN network”. In: *IEEE Access* 7, pp. 39897–39910.

Zheng, X., Han, J., and Sun, A. (2018). “A survey of location prediction on twitter”. In: *IEEE Transactions on Knowledge and Data Engineering* 30.9, pp. 1652–1671.

Zhou, J., Cui, G., Zhang, Z., Yang, C., Liu, Z., Wang, L., Li, C., and Sun, M. (2018). “Graph neural networks: A review of methods and applications”. In: *arXiv preprint arXiv:1812.08434*.

Zhou, W., Ming, D., Lv, X., Zhou, K., Bao, H., and Hong, Z. (2020). “SO–CNN based urban functional zone fine division with VHR remote sensing image”. In: *Remote Sensing of Environment* 236, p. 111458.

Zhou, X., Hristova, D., Noulas, A., Mascolo, C., and Sklar, M. (2017). “Cultural investment and urban socio-economic development: a geosocial network approach”. In: *Royal Society open science* 4.9, p. 170413.

Zhu, D. and Liu, Y. (2018). “Modelling Spatial Patterns Using Graph Convolutional Networks (Short Paper)”. In: *10th International Conference on Geographic Information Science (GIScience 2018)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

Zhu, R., Hu, Y., Janowicz, K., and McKenzie, G. (2016). “Spatial signatures for geographic feature types: Examining gazetteer ontologies using spatial statistics”. In: *Transactions in GIS* 20.3, pp. 333–355.

Zhu, X. and Ghahramani, Z. (2002). *Learning from labeled and unlabeled data with label propagation*. Tech. rep. Citeseer.

Zikopoulos, P., Eaton, C., et al. (2011). *Understanding big data: Analytics for enterprise class hadoop and streaming data*. McGraw-Hill Osborne Media.

Zook, M., Barocas, S., Crawford, K., Keller, E., Gangadharan, S. P., Goodman, A., Hollander, R., Koenig, B. A., Metcalf, J., Narayanan, A., et al. (2017). "Ten simple rules for responsible big data research". In: *PLoS computational biology* 13.3, e1005399.

Zook, M. and Breen, J. (2017). "Volunteered Geographic Information". In: *Encyclopedia of GIS*. Ed. by S. Shekhar, H. Xiong, and X. Zhou. Cham: Springer International Publishing, pp. 2434–2438. ISBN: 978-3-319-17885-1. DOI: 10.1007/978-3-319-17885-1_1656. URL: https://doi.org/10.1007/978-3-319-17885-1_1656.

Zook, M. and Graham, M. (2010). "Featured graphic: The virtual 'bible belt'". In: *Environment and Planning A* 42.4, pp. 763–764.

Zook, M., Graham, M., Shelton, T., and Gorman, S. (2010). "Volunteered geographic information and crowdsourcing disaster relief: a case study of the Haitian earthquake". In: *World Medical & Health Policy* 2.2, pp. 7–33.

Zook, M. A. and Graham, M. (2007). "Mapping DigiPlace: geocoded Internet data and the representation of place". In: *Environment and Planning B: Planning and Design* 34.3, pp. 466–482.