Towards a Comprehensive Measure of the Ambient Population

Annabel Elizabeth Whipp

Submitted in accordance with the requirements for the degree of Doctor of Philosophy

The University of Leeds

School of Geography

April 2022

The candidate confirms that the work submitted is her own, except where work which has formed part of jointly-authored publications has been included. The contribution of the candidate and the other authors to this work has been explicitly indicated below. The candidate confirms that appropriate credit has been given within the thesis where reference has been made to the work of others.

The following chapters contain jointly authored manuscripts where Annabel Elizabeth Whipp is the lead author:

The work in Chapter 3 of the thesis has appeared in publication as:

Whipp, A., Malleson, N., Ward, J. and Heppenstall, A., 2021. Estimates of the Ambient Population: Assessing the Utility of Conventional and Novel Data Sources. ISPRS International Journal of Geo-Information, 10(3), p.131.

Annabel Elizabeth Whipp was the lead author and responsible for conceptualisation, writing, data analysis, visualisation, review and editing. Nicolas Malleson, Alison Heppenstall and Jonathan Ward supervised this study and provided editorial and advisory comments.

The work in Chapter 4 of the thesis has been submitted to a peer reviewed journal as:

Whipp, A., Malleson, N., Ward, J. and Heppenstall, A. Towards a comprehensive measure of the ambient population: Building estimates using geographically weighted regression.

Annabel Elizabeth Whipp was the lead author and responsible for conceptualisation, writing, data collection and analysis, visualisation, review and editing. Nicolas Malleson, Alison Heppenstall and Jonathan Ward supervised this study and provided editorial and advisory comments.

The work in Chapter 5 of the thesis has been submitted to a peer-reviewed journal as:

Whipp, A., Malleson, N., Ward, J. and Heppenstall, A. Alternative measures of the population at risk and their impact on the spatial distribution of crime.

Annabel Elizabeth Whipp was the lead author and responsible for conceptualisation, writing, data analysis, visualisation, review and editing. Nicolas Malleson, Alison Heppenstall and Jonathan Ward supervised this study and provided editorial and advisory comments.

Acknowledgements

I would like to thank my supervisors Nick Malleson, Jonathan Ward, and Alison Heppenstall for the extensive knowledge shared and the valuable guidance provided. To Professor Lex Comber, thank you for encouraging me to have the confidence to begin this journey.

I wish to acknowledge my appreciation of the financial support provided by the Economic and Social Research Council (ES/R501062/1) and my partnership with Leeds City Council. I would like to extend my thanks to Stephen Blackburn and John Eboe at Leeds City Council who provided data and offered support during my work placement.

In addition, I wish to express my gratitude to Keiran Suchak, Stella Spriggs, and Ellie Marfleet for their friendship and support throughout the PhD. I am indebted to Sedar Olmez and my parents, Annette and Nigel, who have inspired, encouraged, and tolerated me throughout this process.

Abstract

Traditional estimates of the population focus on residential populations and capture a single point in time. These estimates fail to account for the frequent fluctuations in human mobility, which significantly impacts the size and demographic composition of small area populations. Despite the impact and utility of estimates of the ambient population, they are not currently published as part of official population statistics. Estimates of the ambient population are a valuable asset in policymaking and can be utilised to inform emergency planning, retail analysis, epidemiological models, and crime analysis. In this thesis, estimates of the ambient population are produced and applied to a study of the spatial distribution of crime rates.

The thesis begins by identifying and critiquing data sources which may be useful for building estimates of the ambient population. This provides a framework of reference for researchers within urban analytics and other areas in which an accurate measurement of the ambient population is required. A method of statistical modelling is utilised in conjunction with novel data to produce daytime and night-time estimates of the ambient population in an urban area. These estimates are then employed to demonstrate any influence that the choice of denominator has on the spatial distribution of crime. This is the first time that a study of crime has utilised a comprehensive measure of the ambient population, drawing on high-resolution footfall counts and other novel sources of data.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

AIC: Akaike Information Criterion

API: Application Programming Interface

ATM: Automated Teller Machine

BC: British Columbia

CCTV: Closed-Circuit Television

CSEW: Crime Survey of England and Wales

GPS: Global Positioning System

GWR: Geography Weighted Regression

KDE: Kernel Density Estimation

LSOA: Lower Super Output Area

LISA: Local Indicator of Spatial Association

MAC: Media Access Control

MAUP: Modifiable Areal Unit Problem

OLS: Ordinary Least Squares

OSM: Open Street Map

SD: Standard Deviation

SMS: Short Messaging Service

UK: United Kingdom

US: United States

Chapter 1

Introduction

1.1 Introduction to the research

In recent decades, intensive urbanisation has resulted in over 55% of the global population living in cities or urban areas (United Nations, 2018). This figure is expected to continue to rise to 68% by 2050, with a total of 2.5 billion people estimated to live in urban areas by this date (United Nations, 2018). Increases in the size of the population living and working in urban areas pose significant challenges to the planning and management of numerous issues, including public safety, health, and infrastructure (Kuddus et al., 2020). A fundamental barrier to effective planning and management within urban areas is the lack of estimates of the size of these populations and the absence of a standardised methodology to produce these estimates. Most commonly, estimates of the size of the population enumerate the so-called resident population, which is a measure of the spatial distribution of the night-time population (Bhaduri et al., 2007). Therefore, these estimates fail to account for fluctuations in the size of the populations that occur due to human activities. Consequently, estimates of the resident population have limited value as they do not capture the entire population within an area. As levels of global urbanisation continue to rise, small area estimates of the size of the non-residential population will be a vital tool within research and policymaking and will enable the impacts of urbanisation on a range of services to be better understood.

A key challenge of producing small area estimates of the size of the ambient population is the selection of appropriate population data and suitable methodology. The use of open-source data and a reproducible methodology is key to ensuring that the accurate estimates of the size of the ambient population can be easily produced for applications across public sectors and within academic research. Despite the

utilisation of suitable population data being key to the production of accurate estimates, many previous studies have simply employed novel types of data, such as mobile phone or social media data, as a proxy of the size of the ambient population (Malleson and Andresen, 2016; Kounadi et al., 2018; Ristea et al., 2018; Tucker et al., 2021). When used in isolation, these novel types of data can be problematic as there is often socio-economic disparities in who they do and do not represent. Consequently, estimates employed in previous studies fail to capture all sectors of the ambient population and an approach that can enumerate the whole ambient population is required. The validation of these datasets has also posed a significant barrier due to the lack of ground truth data available. This thesis aims to address these challenges.

One application for which small area estimates of the size of the ambient population are crucial is within crime studies. Crime rates are considered the most meaningful statistic within crime studies and are valuable for communicating risk, informing resource allocation, and influencing policymaking and planning (Boggs, 1965; National Academy of Sciences, 2016). The importance of appropriate measures of the population at risk for calculating crime rates has been highlighted within the literature (Boggs, 1965; Harries, 1981). Existing studies have demonstrated that crime rates calculated using the resident population as a measure of the population at risk may not be accurate and, therefore, do not accurately communicate the real risk of crime (Boggs, 1965; Malleson and Andresen, 2015; Hanaoka, 2018; Ristea et al., 2018; He et al., 2020). However, existing studies have not yet explored the suitability of different measures of the non-residential population as measures of the population at risk. This gap in the literature provides an opportunity to assess the impact of different measures of the population at risk, including the ambient population, on the spatial distribution of crime rates. Identifying the most appropriate measure of the population at risk will enable the production of more accurate crime which can then inform policing.

The production of small area estimates of the size of the ambient population presents an opportunity to explore suitable data types and methodologies. These estimates can then be applied to aid the development of a better understanding of the spatial distributions of crime rates. This thesis explores types of population data and their suitability for producing estimates of the size of the ambient population. An approach is then developed which employs these data in conjunction with a method of statistical modelling to produce estimates of the size of the ambient population. A case study of West Yorkshire (UK) is employed to demonstrate both the utility and importance of these estimates. The variations in the spatial distributions of the crime rates of two crime types ('theft from the person' and 'violence and sexual offences'), calculated using different measures of the population (estimates of the resident, workday, and the ambient populations), are investigated.

1.2 Research aim and objectives

The overall aim of this research is to explore the development of small area estimates of the size of the ambient population in an urban area. To fulfil this aim, the following objectives have been established:

1. Review and discuss the literature relating to quantifying the size of the ambient population and comparable small area estimates of populations and their use within crime studies.
2. Assess and critique sources of population data that have the potential to be used to produce estimates of the size of the ambient population in urban areas, including those utilised in the existing literature.
3. Develop small area estimates of the size of the ambient population for an urban area.
4. Produce a validation dataset that captures footfall counts in an urban area.
5. Employ the validation dataset to assess the accuracy of the manual footfall counts, the footfall camera counts and the model estimates.

6. Utilise the estimates of the size of the ambient population to examine the impact of different measures of the population on the spatial distribution of the rates of two crime types; 'theft from the person' and 'violence and sexual offences'.

1.3 Thesis structure

This thesis is presented in the alternative format as described by the University of Leeds. The six chapters of the thesis are outlined below, and Table 1.1 highlights the structure of the thesis in relation to the research objectives. It should be noted that chapters 3, 4 and 5 of the thesis are self-contained manuscripts published in or submitted to peer-reviewed journals.

Table 1.1 The thesis structure in relation to the research objectives.

Research Objective	Chapter number
1. Review and discuss the literature relating to quantifying the size of the ambient population and comparable small area estimates of populations and their use within crime studies.	Chapter 2
2. Assess and critique sources of population data that have the potential to be used to produce estimates of the size of the ambient population in urban areas, including those utilised in the existing literature.	Chapter 2 Chapter 3
3. Develop small area estimates of the size of the ambient population for an urban area.	Chapter 4
4. Produce a validation dataset that captures footfall counts in an urban area.	Chapter 4
5. Employ the validation dataset to assess the accuracy of the manual footfall counts, the footfall camera counts and the model estimates.	Chapter 4

6. Utilise the estimates of the size of the ambient population to examine the impact of different measures of the population on the spatial distribution of the rates of two crime types; 'theft from the person' and 'violence and sexual offences'.	Chapter 5

Chapter 2 presents a critical review of literature relevant to producing estimates of the size of the ambient population. This chapter defines the term the 'ambient population' and highlights early studies of the topic. Then follows a critical discussion of both conventional and novel types of population data which may be useful in producing small area estimates of the size of the ambient population. The chapter then explores studies within environmental criminology that employ estimates of the ambient population to assess the relationships between the crime rates, calculated using different measures of the population at risk and their spatial distributions. The work in this chapter highlights the need to develop a better understanding of the spatio-temporal differences between population data that have the potential to be used to produce estimates of the size of the ambient population. This opportunity is explored within the subsequent chapter.

The work in Chapter 3 has been published in the ISPRS International Journal of Geo-Information as:

Whipp, A., Malleson, N., Ward, J. and Heppenstall, A., 2021. Estimates of the Ambient Population: Assessing the Utility of Conventional and Novel Data Sources. ISPRS International Journal of Geo-Information, **10**(3), p.131.

Chapter 3 aims to identify and assess the utility of sources of novel and conventional data for producing estimates of the size of the ambient population. The work begins with a critical discussion of types of population data and then explores and compares the spatio-temporal distribution of the available data. The results of

the work highlight that footfall camera count data have not yet been explored within the literature and are a potentially viable type of population data for producing estimates of the size of the ambient population. The work in this chapter also acknowledges that while footfall camera count data may be valuable, they have not yet been validated.

The work in Chapter 4 is ready to be submitted to a peer reviewed journal as:

Whipp, A., Malleson, N., Ward, J. and Heppenstall, A. Towards a comprehensive measure of the ambient population: Building estimates using geographically weighted regression.

Chapter 4 develops a novel approach for producing estimates of the size of the ambient population in an urban area. A method of statistical modelling, geographically weighted regression, is utilised in conjunction with indicators of the size of the population to estimate the number of footfall counts. The footfall camera count data and the model estimates are validated using a novel dataset enumerating manual footfall counts produced as part of this research. The work in this chapter highlights the utility of openly available, novel data for producing estimates of the size of the ambient population in urban areas.

The work in Chapter 5 has been accepted with revisions by PlosOne:

Whipp, A., Malleson, N., Ward, J. and Heppenstall, A. Alternative measures of the population at risk and their impact on the spatial distribution of crime.

Chapter 5 presents a study in which alternative measures of the population at risk are employed to investigate the impact on the spatial distributions of two crime types: 'theft from the person' and 'violence and sexual offences'. Three measures of the population at risk are used: the resident population, the workday population, and the ambient population. The estimates of the size of the ambient population utilised

in this study are produced using the approach detailed in Chapter 4. The study finds that for rates of 'theft from the person' and rates of 'violence and sexual offences', the use of both the resident and workday populations results in the risk of victimisation within urban centres being overestimated. In contrast, the risk in residential areas is underestimated. The findings of the study highlight the value of estimates of the ambient population for producing accurate crime rates and support the demand for geographically comprehensive estimates that can be utilised by police forces and policymakers.

The thesis is concluded in Chapter 6. The chapter begins by outlining the novelty of the thesis. The chapter then proceeds to provide a summary of the thesis and demonstrates the extent to which the research aim and objectives have been met. The limitations of the research are noted and recommendations for future work are made. An outlook on producing and utilising estimates of the size of the ambient population and concluding remarks are also documented.

Reference list

Bhaduri, B., Bright, E., Coleman, P. and Urban, M.L. 2007. LandScan USA: a high-resolution geospatial and temporal modeling approach for population distribution and dynamics. *GeoJournal*. **69**(1–2), pp.103–117.

Boggs, S.L. 1965. Urban crime patterns. *American sociological review*. **30**(6), pp.899–908.

Hanaoka, K. 2018. New insights on relationships between street crimes and ambient population: Use of hourly population data estimated from mobile phone users' locations. *Environment and Planning B: Urban Analytics and City Science*.

Harries, K.D. 1981. Alternative denominators in conventional crime rates. *Environmental criminology*., pp.147–165.

He, L., Páez, A., Jiao, J., An, P., Lu, C., Mao, W. and Long, D. 2020. Ambient Population

and Larceny-Theft: A Spatial Analysis Using Mobile Phone Data. *ISPRS International Journal of Geo-Information*. **9**(6), p.342.

Kounadi, O., Ristea, A., Leitner, M. and Langford, C. 2018. Population at risk: using areal interpolation and Twitter messages to create population models for burglaries and robberies. *Cartography and Geographic Information Science*. **45**(3), pp.205–220.

Kuddus, M.A., Tynan, E. and McBryde, E. 2020. Urbanization: A problem for the rich and the poor? *Public Health Reviews*. **41**(1).

Malleson, N. and Andresen, M.A. 2016. Exploring the impact of ambient population measures on London crime hotspots. *Journal of Criminal Justice*. **46**, pp.52–63.

Malleson, N. and Andresen, M.A. 2015. The impact of using social media data in crime rate calculations: Shifting hot spots and changing spatial patterns. *Cartography and Geographic Information Science*.

National Academy of Sciences 2016. *Modernizing Crime Statistics: Report 1: Defining and Classifying Crime*.

Ristea, A., Andresen, M.A. and Leitner, M. 2018. Using tweets to understand changes in the spatial crime distribution for hockey events in Vancouver. *Canadian Geographer*.

Tucker, R., O'Brien, D.T., Ciomek, A., Castro, E., Wang, Q. and Phillips, N.E. 2021. Who 'Tweets' Where and When, and How Does it Help Understand Crime Rates at Places? Measuring the Presence of Tourists and Commuters in Ambient Populations. *Journal of Quantitative Criminology*.

United Nations 2018. 68% of the world population projected to live in urban areas by 2050, says UN. *United Nations News*.

Chapter 2

Understanding the Ambient Population

2.1 Introduction

The key aim of this thesis is to explore the development of small area estimates of the size of the ambient population in an urban area. To fulfil this aim, a thorough understanding of the ambient population and the relevant literature is required. This chapter provides a comprehensive review of the existing literature related to the ambient population, types of population data, and the application of estimates of the ambient population within environmental criminology. Section 2.2 defines the term the 'ambient population' and highlights early studies of the topic. Section 2.3 follows with a critical discussion of traditional and novel types of population data which may be valuable for enumerating the size of the ambient population. In Section 2.4 studies within environmental criminology that have employed estimates of the ambient population to assess the relationships between the crime rates calculated using different measures of the population at risk and their spatial distributions are presented.

2.2 Defining the ambient population

Typically, estimates of the size of the population enumerate the resident population of an area at a single point in time, i.e., the number of people who have lived or intend to live at an address continuously for a twelve-month period (United Nations, 2008a). However, the utility of these estimates is limited as they do not capture the significant fluctuations in the size of the population which occur due to human activity patterns, i.e., the ambient population (Bell and Ward, 2000)Be. The ambient population can be defined as "the number of people within a given geographical area at a specific point in time, excluding individuals at their place of residence and those utilising modes of transport" (Whipp et al., 2021, p.131). This definition is used throughout this thesis, but it should be noted that there are

numerous terms utilised within the existing literature that are used synonymously with the term the 'ambient population'. These terms include 'real-time census' (Kontokosta and Johnson, 2017), 'temporary population' (Charles-Edwards et al., 2008), 'daytime population' (Schmitt, 1956; Boeing, 2018), 'non-residential population' (Berry et al., 2016), 'mobile population' (Malleson and Andresen, 2015b), 'service population' (Markham et al., 2013), 'functional population' (Nelson and Nicholas, 1992), and 'seasonal population' (Adamiak et al., 2017). This chapter reviews works that employ these terms, or the term the 'ambient population'.

There is a considerable need for estimates of the size of the ambient population across a range of applications, including emergency planning and management (Smith et al., 2005; Chen and McAneney, 2006; Bengtsson et al., 2011), retail analysis (Soundararaj et al., 2020), service planning (Markham et al., 2013), environmental criminology (Malleson and Andresen, 2015a; Malleson and Andresen, 2015b; Malleson and Andresen, 2016; Hipp et al., 2019; Ristea et al., 2020; Haleem et al., 2020; Jung et al., 2020; He et al., 2020; Tucker et al., 2021), and urban analytics (Ratti et al., 2006; Reades et al., 2007; Reades et al., 2009). The importance of both understanding and estimating the size of the ambient population was first highlighted in studies published in the 1950s by Foley (1954) and Schmitt (1956). Despite the need for estimates of the size of the ambient population being acknowledged within the literature over sixty years ago, they are not currently a part of the standard suite of official population statistics in any nation. Historically, this has been due to the lack of available data and the absence of a standardised methodology to produce estimates of the size of the ambient population (Malleson and Andresen, 2015a; Malleson and Andresen, 2016; Crols and Malleson, 2019; He et al., 2020). However, it was stated by Smith (1989) that the development of a single methodology that can be utilised across varying contexts and geographies was likely impossible. It should be noted that this assertion was made prior to the proliferation of spatio-temporal data produced by emergent technologies in recent decades. However, geographic variations in the drivers of the size of the ambient population have not yet been explored within the

literature. The following section critically reviews types of population data that may be valuable for producing estimates of the size of the ambient population.

2.3 Types of population data

The ability to estimate the size of the ambient population has previously been limited by the lack of suitable data. However, in the last two decades vast volumes of spatio-temporal data have been produced by emergent technologies. This section critically reviews types of population data that may be valuable for producing estimates of the ambient population. These include conventional data types, such as census data, official statistics, and travel survey data, and novel data types, including mobile phone data, geo-located social media data, Wi-Fi sensor data, and mobility data.

2.3.1 Census data and official population statistics

Estimates of the size of populations have conventionally been a product of national surveys, most notably censuses and other official population statistics. Unlike many other sources of data, census data are geographically comprehensive and enumerate close to the whole population, for example, the 2011 Census of England and Wales was completed by approximately 95% of households (Office for National Statistics, 2013b). Consequently, census data are considered to represent the gold standard of data collection (Rees et al., 2002).

There are, however, several significant limitations of national census data and other official population statistics in that they are only representative of a single point in time (Census Day) and the data are commonly only collected at five or ten-year intervals. Consequently, the data fail to capture diurnal, seasonal, and event-driven fluctuations, which contribute to the size of the ambient population and become outdated. Additionally, due to the length of data processing, data are often published over a year after data collection; thus, they become quickly outdated. To address this

issue and provide more up-to-date official estimates of the resident population, the statistical agencies in several countries produce mid-year population estimates, including England, Wales, and the US. Mid-year population estimates are calculated using the most recent estimates of the resident population from the national census, in addition to the numbers of births, deaths, and internal and external migration. This allows more current estimates of the resident population to be produced; however, mid-year estimates still do not account for any short-term visitors or diurnal movements.

As estimates of the resident population only enumerate the number of individuals living at a residential address, they are considered to have limited utility for enumerating the size of the ambient population, as they fail to capture human mobility due to work, leisure, and other activities. However, the 2011 the Census for England and Wales captured estimates of the workday population which provide the typical location of individuals during standard working hours, in addition to the number of people in the area who are not in employment (Office for National Statistics, 2013a). The US Census Bureau produces similar data in the form of Commuter-Adjusted Population Estimates, which are akin to workday population estimates. These estimates are calculated by adding the total resident population to the total workers working in an area, then subtracting the total workers who live in the area (United States Census Bureau, 2017). Although work-related movements are a key element of diurnal populations, estimates of the workday population do not capture fluctuations in the size of the ambient population, which result from other activities.

Despite the limitations of census data, as described above, they have nonetheless been employed in a number of studies to produce estimates of the ambient population in conjunction with other datasets (Gober and Mings, 1984; Smith et al., 2005; Martin et al., 2009a; Martin et al., 2015). A database of the ambient population was produced by Smith et al. (2005) for use in hazard modelling and provided geographically comprehensive coverage of England and Wales (UK). The

database drew on census estimates of the resident and workday populations and official population statistics enumerating care home, prison, hospital, and school populations, in addition to the locations of leisure facilities and retail centres. Martin et al. (2015) later proposed a framework for estimating the ambient population, which drew on similar datasets, including census data and other forms of official population statistics, such as higher education statistics and hospital episode statistics. Studies by Gao et al. (2014) and Qi et al. (2015) utilised demographic data and resident population estimates from the Chinese census with land use data to produce estimates of the daytime ambient population. In 2018, daytime ambient population densities were calculated in a study of San Francisco (US) using US census data and payroll statistics (Boeing, 2018). The aforementioned studies utilised traditional datasets, which provided comprehensive spatial information, but did not employ novel sources of data that are able to add temporal detail, such as seasonality.

Census data have also been utilised, in conjunction with ancillary data, to produce LandScan, a dataset of estimates of the ambient population. LandScan is a global dataset developed annually by Oak Ridge National Laboratory. The dataset is produced using census data, remotely sensed images and dasymetric modelling to disaggregate estimates of the ambient population, averaged over a one-year period, within a spatial boundary (Oak Ridge National Laboratory, n.d.). The data represent estimates of the ambient population for a 24-hour period at a spatial resolution of $1km^2$ (Oak Ridge National Laboratory, n.d.). LandScan data are freely available for academic use and have consequently been utilised in several studies (Dobson et al., 2000; Sutton et al., 2001; Chen, 2002; Dobson et al., 2003; Bhaduri et al., 2007); however, the data lack detailed temporal information. As the data are averaged over a 24-hour period, diurnal fluctuations, and seasonality, which are crucial to producing accurate estimates of the ambient population, are not captured.

As census data do not quantify temporary populations, which significantly impact the size of the ambient population, previous studies have successfully employed other forms of official statistics. In order to estimate the size of the service

population in the context of Indigenous persons in Australia, Markham et al. (2013) utilised health service data, while Adamiak et al. (2017) measured seasonal populations in Finland using the registries of second home ownership. These studies highlight the utility of official statistics to enumerate temporary populations which are not captured by census data and offer a low-cost alternative to novel sources of data, such as mobile phone data, which can be financially expensive to acquire. However, it should be noted that while these studies were able to estimate the size of seasonal populations not captured by the census, they did not enumerate more temporary visitors, such as workers and visitors.

While data from national censuses and official population statistics provide extensive geographical coverage and high levels of enumeration, the frequency of data collection and the temporal representation of the data significantly limit their utility. Despite these limitations, these data sources have been employed extensively within the existing literature, highlighting their utility for producing estimates of the size of the ambient population.

2.3.2 Travel survey data

Travel survey data capture information regarding human activity patterns that lead to fluctuations in the size of the ambient population. Travel survey data often provide information regarding journey purpose, destination, and duration, which can be used to estimate the size of the ambient population. Both national governments and private organisations conduct travel surveys at local and national levels, but commonly lack the geographical comprehensiveness of census data.

Travel survey data have been successfully employed to produce estimates of the size of the ambient population in both Australia and Japan. Lau (2009) employed travel survey data to estimate the size of the ambient population in Melbourne (Australia) for vehicle route-planning. However, at the time of publication, the travel survey data used were over a decade old; thus, the findings were unlikely

to be representative of the true spatio-temporal distribution of the ambient population. Kashiyama et al. (2017) employed agent-based modelling to produce an open dataset of estimates of the ambient population across urban areas of Japan using travel survey data and national census data. The estimates were then validated using commercial mobility and traffic census data. However, as travel survey data for the study area are collected once every five years, the data cannot reflect seasonal changes in human activity patterns, which are a crucial element of the ambient population (Bell, 2001; Kashiyama et al., 2017). Additionally, the complexity of the model developed is a barrier to the approach being widely utilised within research and policymaking.

Travel surveys can be utilised to quantify seasonal populations, which official statistics often fail to capture. Warchivker et al. (2000) utilised travel surveys to explore spatial changes in the residence of Aboriginal communities, a group characterised by high levels of mobility, in central Australia. The travel surveys used in this study were conducted during different seasons to account for seasonality and evidenced high levels of inter-community and intra-community mobility. Happel and Hogan (2002) adopted a similar approach in a study that examined seasonal movements of population in Arizona, Florida and Texas (US); however, the authors noted the significant limitations relating to the suitability of the data. These studies demonstrated that travel survey data can represent seasonal fluctuations in the ambient population if surveys are conducted at frequent intervals to capture different seasons and annual fluctuations. However, the frequent collection of data is often beyond the scope of surveys undertaken by national and local governments.

While travel survey data capture the activity patterns of individuals, there are several limitations to their use in producing estimates of the size of the ambient population. Travel survey data are most commonly self-reported, resulting in omitted journeys that significantly impact both data accuracy and quality (Sallis and Saelens, 2000; Stopher et al., 2005; Bricka et al., 2009; Lubans et al., 2011). Additionally, there are challenges associated with the development of statistically valid sampling

methods to represent seasonal populations (Happel and Hogan, 2002; Kashiyama et al., 2017). The issue of infrequent data collection, which is due to the financial and temporal expenses associated with conducting surveys, is acknowledged in the studies reviewed. The studies reviewed in this section highlight the value of travel survey data if collected at frequent intervals. However, some novel types of population data, such as mobile phone data, offer alternatives to travel surveys as they continuously capture human activity patterns; thus, offering high levels of spatio-temporal detail. The subsequent sections explore novel types of population data.

2.3.3 Mobile phone data

Mobile phone data have garnered significant interest within the research community due to the high penetration rates of mobile phones and the spatio-temporal data they produce. These data are collected by network operators and provide mobile positioning data that log a mobile phone's approximate location. Several methods can be employed to geo-locate mobile phones; however, due to higher levels of accuracy, the triangulation of signal strengths from surrounding cell towers is most commonly used (Toole et al., 2012). Consequently, the accuracy of mobile phone positioning data is higher in areas that contain higher densities of cell towers, such as cities and urban areas. As mobile phones are used globally, the utility of these data for estimating human activity patterns in both developed and developing countries has been acknowledged within the existing literature (Blondel et al., 2015; Wesolowski et al., 2015; Manley and Dennett, 2019). Therefore, mobile phone data have the potential to be utilised as a part of a standardised framework to produce estimates of the size of the ambient population.

Mobile phone data have been utilised to produce estimates of the ambient population in several studies. To aid understanding of urban environments in a case study of Milan (Italy), Ratti et al. (2006) employed data from a major European telecommunications provider. The work visualised cell phone call density to demonstrate activity patterns across the city. Reades et al. (2007) used mobile phone

data to characterise locations in Rome (Italy) and identify clusters, such as night-time leisure and early morning commuting, based on mobile phone usage. Data from Telecom Italia Mobile were utilised in conjunction with commercial premises data in order to demonstrate a measurable link between mobile phone usage and business activity in Rome (Italy) (Reades et al., 2009). A study by Deville et al. (2014) utilised over one billion mobile phone call records from Portugal and France to estimate population densities; however, the data represented only 20% of the market in Portugal and 30% of the market in France (Deville et al., 2014). The estimates of population density were then validated using remotely sensed images and ancillary data (Deville et al., 2014). Several studies have employed mobile phone data in conjunction with other data types to produce estimates of the ambient population (Lwin et al., 2016). Lwin et al. (2016) utilised mobile call records, person trip data, and geo-located Tweets in conjunction with a space-time multiple regression model to produce grid-based estimates of the population for the city of Kobe (Japan), at thirty-minute intervals. Mobile phone data have also been employed to study seasonal and temporary fluctuations in the population, which are crucial elements of the ambient population. Silm and Ahas (2010) attempted to quantify movements of seasonal populations in Estonia, while Bengtsson et al. (2011) explored shifts in the size of the population following the 2010 earthquake and consequent cholera outbreak in Haiti.

Despite mobile phone data providing high spatio-temporal resolution information regarding the movements of mobile phone users, the data have several limitations. While mobile phone data are produced at a high-spatial resolution, the data are often aggregated to a lower geographical level to protect user privacy. This significantly limits the utility of the data, as the size of the ambient population will vary considerably across space; thus, spatial detail is crucial for producing accurate estimates. Additionally, there are concerns regarding the spatial accuracy of mobile phone data. The positioning accuracy of 3G and 4G networks is 200 metres and above 150 metres, respectively; however, the introduction of 5G cellular networks will allow higher accuracy spatial data to be produced (up to 1 metre). However, increased spatial accuracy will have a limited impact on the utility of the data if they are

aggregated to a low geographical level. It is also important to note that the geography of studies that employ mobile phone data is reflective of data availability and accessibility. For example, several studies have focussed on Italian cities, particularly Milan and Rome, as aggregated and anonymised mobile phone data for these areas were made available by a national telecommunications provider. As mobile phone data are commonly commercial datasets, they can be extremely costly to acquire and may not be available with high levels of spatio-temporal detail nor for all geographic regions.

2.3.4 Geo-located social media data

The increasing popularity of social media platforms, such as Twitter, Instagram and Foursquare, has resulted in the availability of large volumes of spatio-temporal information. The aforementioned social media platforms allow users to share their location with a post, providing spatio-temporal data which can be utilised to quantify the size of the ambient population. Geo-located social media data provide more accurate spatial information than mobile phone data as they often utilise GPS coordinates (García-Palomares et al., 2018). The spatial accuracy of geo-located social media data, therefore, allows these data to be used in conjunction with other types of data, such as land use data, which can aid in building more comprehensive estimates of the ambient population. Geo-located social media data have been utilised as a proxy of the ambient population in several crime-based studies (Steiger et al., 2015; Malleson and Andresen, 2015b; Hipp et al., 2019) and to estimate the size of visitor populations (Hamstead et al., 2018).

Twitter data have been acknowledged as being particularly well suited for building estimates of the size of the ambient population as Tweet data are openly available via an Application Programming Interface (API); however, the data have several shortcomings. It should be noted that only a limited volume of data are available, as only a random sample of Tweets that accounts for 1% of the feed can be downloaded (Tucker et al., 2021). Furthermore, a study by Sloan et al. (2013) found

that only 0.85% of Twitter data are geotagged, resulting in a small sample of data. Rates of geotagging may be low as location services on Twitter are disabled by default, thus must be enabled by the user. Consequently, users who have privacy concerns relating to sharing their location or those who are unaware of the service may not use the service. Demographic factors, such as age and socio-economic status, have a significant impact on the volume and frequency of geotagged Tweets (Longley et al., 2015). While the demographics of Twitter users have been quantified (OFCOM, 2020), a study by Sloan et al. (2013) found that the characteristics of users who geotag Tweets were statistically significantly different from the wider Twitter users. Thus, geotagged Tweets are neither representative of all Twitter users nor of the ambient population as a whole. In 2019, Twitter announced the ability to geotag Tweets with precise geographic information would no longer be available, thus limiting the use of Twitter data in future attempts to estimate the size of the ambient population (Tucker et al., 2021).

2.3.5 Pedestrian counters

Pedestrian counters are a source of individual level movement data which are most commonly captured using either footfall cameras or Wi-Fi sensors. Footfall cameras capture data using a counting device, often mounted on walls or streetlights, which record high quality video. Image processing and target-specific tracking algorithms are employed to identify pedestrians and enumerate them as they pass a virtual line. Due to the spatio-temporal detail these data are able to capture and their ability to capture the whole population, and not only those who use particular services, such as Wi-Fi enabled devices or social media platforms, they have the potential to enumerate the ambient population. At the time of writing, there are no studies that the author is aware of that employ footfall camera data to estimate the size of the ambient populations, or produce other small area estimates of the population. Therefore, there is a marked opportunity to explore the utility of footfall camera count data for producing estimates of the size of the ambient population.

Wi-Fi sensors are a potentially valuable source of data for producing estimates of the ambient population due to their ability to enumerate individuals and provide spatio-temporally detailed data (Crols and Malleson, 2019; Soundararaj et al., 2020). Wi-Fi sensors log counts when a Wi-Fi probe request emitted from a Wi-Fi enabled device, such as a mobile phone or tablet, is received. Probe requests are emitted periodically; therefore, as an individual with a device moves through an urban area, the device is likely to connect to multiple access points (Oliveira et al., 2018). This produces highly detailed spatio-temporal data, which can be utilised to estimate the size of the ambient population.

Data access is a significant barrier to the use of Wi-Fi sensor datasets as they are often commercial products that can be costly to obtain. There are also concerns regarding the representativeness of Wi-Fi sensor data as despite the high penetration rates of Wi-Fi enabled mobile devices, the average number of devices carried by an individual is unknown. This may be particularly problematic in urban areas, where one individual may be carrying multiple Wi-Fi enabled devices and will, therefore, be logged multiple times at each location. There have also been ethical and privacy concerns regarding the use of Wi-Fi sensor data as individual movements can be captured, as probe requests contain the unique media access control (MAC) address of each device (Freudiger, 2015). To address these privacy issues, both Apple and Android devices periodically randomise MAC addresses; consequently, the same user and device may be logged by a sensor multiple times under different MAC addresses (Martin et al., 2017; Android, 2020). Due to this randomisation of the MAC addresses, it is not possible to detect multiple logs of the same user and device and omit them from the dataset; thus, the number of logs may be significantly higher than the number of people in the area.

Wi-Fi sensor data have not been widely used for producing estimates of the ambient population; however, there are two examples within the existing literature. Kontokosta and Johnson (2017) used over 20 million Wi-Fi probe requests in conjunction with traditional data sources to develop a real-time census for building

occupancy New York City (US). The estimates produced were then validated using survey data and were found to be within 5% of the survey estimates. However, this study failed to discuss the issue of MAC address randomisation which may result in the same device being logged multiple times. Wi-Fi probe requests were also used by Crols and Malleson (2019), together with official statistics and travel survey data, to quantify the ambient population in Otley (UK). A notable limitation of this work is that empirical validation of the estimates of the size of the ambient population was not conducted due to a lack of data (Crols and Malleson, 2019). While Wi-Fi sensor data provide detailed spatio-temporal information, data access poses a significant challenge, and consequently, the data have not been used extensively in the existing literature. However, these studies demonstrate the potential value of Wi-Fi sensor data which could be explored within future work.

2.3.6 Mobility data

To aid efforts in remediating the impacts of the COVID-19 pandemic, three companies, Google, Apple and CityMapper, made mobility data publicly available. While not designed to enumerate the size of the ambient population, these datasets may be useful for aiding the understanding of the human activity patterns, which result in fluctuations in its size.

Google COVID-19 Community Mobility Reports are calculated using aggregated and anonymised data from Google account users who have the 'Location History' setting enabled, which is disabled by default. The data account for daily changes in the number of visits by Google account users to six place categories; grocery and pharmacy, parks, transit stations, retail and recreation, residential, and workplaces (Google, 2020). The number of daily visits is compared to a baseline figure for the corresponding day of the week during the 5-week period between January 3rd and February 6th, 2020 (Google, 2020). The data provide a percentage change in the number of visits to each of the six place categories and do not provide the actual number of visits. The location accuracy of the data is acknowledged to vary

geographically, and regions are excluded from the dataset if the levels of data are not statistically significant (Google, 2020). Google COVID-19 Community Mobility Reports were successfully utilised by Halford et al. (2020) to measure the mobility elasticity of four crime types in the UK during the pandemic. As the mobility elasticity of crime is calculated using the percentage change of the number of crimes and the percentage change of the number of visits, Google COVID-19 Community Mobility Reports can be used. However, the data are not suitable for studies in which raw numbers are required.

The technology company Apple publish daily Mobility Trends Reports which capture routing requests in the iOS application 'Apple Maps' (Apple, 2020). The data represent the percentage changes in the number of direction requests via public transport, walking and driving since January 13th, 2020. However, the number of routing requests is not necessarily indicative of the number of journeys that took place. The dataset provides coverage of 63 countries, and the data are aggregated to country, region, sub-region, and city levels (Apple, 2020). While the Apple Mobility Trends Reports provide data at a higher spatial resolution than the Google COVID-19 Community Mobility Reports, the data still lack sufficient detail for use in the production of small area estimates of the ambient population.

Citymapper, a GPS navigation application, publish the Citymapper Mobility Index, which is calculated by comparing the number of journeys planned via the Citymapper iOS and Android applications to a typical usage period (Citymapper, 2020). Users can plan journeys via public transport, walking, cycling, taxis and micro-mobility (such as e-bikes and electric scooters) (Citymapper, 2020). However, the application cannot be used to plan journeys by a private car. The typical usage period is the four weeks between January 6th and February 2nd, 2020 (Citymapper, 2020). In some geographic locations, an alternative typical usage period is employed to reflect typical usage more accurately in these areas. Examples include Hong Kong and Singapore, for which the typical usage period utilised was December 2nd to the 22nd, 2019 (Citymapper, 2020). Data are available for 36 cities; thus, the dataset is a smaller

geographic coverage than Google COVID-19 Community Mobility Reports and Apple Mobility Trends Reports.

Google COVID-19 Community Mobility Reports, Apple Mobility Trends Reports and the Citymapper Mobility Index offer novel, open mobility data that are indicative of fluctuations in the ambient population. However, it is important to note the shortcomings of these datasets as they may limit data utility. All three datasets compare the percentage change in the number of journeys made during the pandemic but do not quantify the number of journeys; thus, the data cannot be utilised as a proxy for the size of the ambient population in a given geographic location. The number of users for each of the applications is unknown; consequently, the size and representativeness of the samples cannot be determined. For both the Citymapper Mobility Index and Apple Mobility Trends Reports, the index is calculated using the number of trips planned, not the number of trips actually taken and, therefore, does not accurately represent of mobility. Google COVID-19 Community Mobility Reports, however, only contain data for those movements which have occurred. An additional limitation is that both Google COVID-19 Community Mobility Reports and Apple Mobility Trends Reports are only available for a limited period during the COVID-19 pandemic, which will limit their use in future research. At the time of writing, Apple Mobility Trends Reports and the Citymapper Mobility Index have not yet been employed in studies of the size of the ambient population.

2.4 Estimation methods

In his seminal work, Smith (1989) identified two approaches for estimating temporary populations, the direct and the indirect approach. The direct approach draws on information collected from the temporary population via censuses and large-scale surveys. The indirect approach draws upon variables which are symptomatic of changes in the size of temporary population. However, the difference between these two approaches has become increasingly unclear due to the emergence of novel

sources of data. Types of novel data, such as mobile phone data and Wi-Fi sensors, can be treated as either direct or indirect, depending on the approach utilised (Panczak et al., 2020). Previous studies have employed both direct and indirect approaches to generate estimates of temporary populations (Rigall-I-Torrent, 2010; Lwin et al., 2016). More recent work by Crols and Malleson (2019) used simulation and modelling methodologies to produce estimates of the temporary population.

2.4.1 Direct approaches

The earliest method of deriving estimates of temporary populations was to employ direct estimates from national censuses, survey data or through the combination of several different data sources. Studies that used these methods often combined information about workplaces (Office for National Statistics, 2013a), residential status (Gober and Mings, 1984) and place of enumeration (Bell and Ward, 2000) to identify two or more states of population distributions (for example, daytime and night-time). A similar approach involves the scaling of population estimates from single or multiple sources to the entire population based on expansion factors. This approach has been utilised in a range of studies, some employed survey data (Foley, 1954), while others used mobile phone data (Bengtsson et al., 2011), night-time census estimates (Deville et al., 2014) or ancillary data (Kontokosta and Johnson, 2017).

2.4.2 Component-based approaches

Estimates of the temporary population have been produced using component-based approaches. These methods utilise either generic or area-specific equations to calculate estimates of temporary populations by subtracting and adding the numbers of people arriving in or leaving an area across a given timeframe. A component-based approach was utilised in conjunction with journey to work data from the US Census was used in work by McKenzie et al., 2013, while Adamiak et al. (2017) used second home ownership data to capture seasonal variations in temporary populations. This

approach was also used by Swanson and Tayman (2011) to derive estimates of different subgroups of the temporary population.

2.4.3 Symptomatic data redistribution

Symptomatic data have been utilised with areal interpolation methods, such as dasymetric mapping, to redistribute estimates of temporary populations from larger (coarser) to smaller (more granular) geographical units (Tenerelli et al., 2015). The redistribution of symptomatic data often uses ancillary data such as building type (Greger, 2015) or land use data, in addition to data from other sources including censuses and other large scale surveys, mobile phones, transportation (Ma et al., 2017), and social media (Lwin et al., 2016). Dobson et al. (2000) utilised a multi-dimensional dasymetric model to produce estimates of the ambient population for grid cells at a global scale. The symptomatic data can be related to the size of the population through sampling techniques (Mennis, 2003; Mennis and Hultgren, 2006; Wu et al., 2008), regression analysis (Chen, 2002; Wu et al., 2006; Briggs et al., 2007; Silván-Cárdenas et al., 2010; Lu et al., 2010) or expert knowledge (Eicher and Brewer, 2001).

2.4.4 Modelling and simulation

Several more contemporary studies have used various methods of modelling or simulation to produce estimates of temporary populations, including agent-based modelling (Walker and Barros, 2012; Kashiyama et al., 2017; Crols and Malleson, 2019), cellular automation modelling (Khakpour and Rod 2016), and neural networks (Liu et al., 2018; Chen et al., 2018).

2.4.5 Section summary

It has been suggested within the existing literature that there have been limited attempts to estimate the size of the ambient population due to the lack of suitable data (Malleson and Andresen, 2015a; Malleson and Andresen, 2016; Crols and Malleson, 2019; He et al., 2020). Despite the increased availability of spatio-temporal data, such as mobile phone, social media, and mobility data, within the last two decades, there has been little progress in the development of estimates of the size of the ambient population. This section has highlighted that many of the data types reviewed offer fine-scale, spatio-temporal data which can be explored and potentially employed to produce estimates of the size of the ambient population, despite their limitations. However, it is noted that due to their ability to provide spatio-temporally detailed data, which is able to represent the whole ambient population, pedestrian counter data, most notably footfall camera counts, should be investigated further. Several of the data types reviewed in this section have previously been utilised within the context of environmental criminology and are discussed in Section 2.4.

2.5 Estimates of the ambient population and the geography of crime

Crime rates are a valuable statistic utilised to measure and communicate risk. They are commonly calculated using estimates of the resident population as a measure of the population at risk. However, as crime clusters within space and time, the size of the resident population may not always be a reflective measure. To account for fluctuations in the size of the population within crime rates, existing studies have employed estimates of the ambient population as a measure of the population at risk. This section begins by defining crime rates and highlighting their importance within crime studies and policymaking. A discussion of theories within environmental criminology that support the use of the ambient population within the calculation of crime rates is then present. This section then reviews the existing literature related to the use of these measures in the calculation of crime rates, the exploration of the spatial distribution of crimes, and the investigation of relationships between crime rates.

2.5.1 Calculation of crime rates

Crime rates are the most common measurement of crime and quantify the ratio of police-recorded crimes to the size of the population in a geographic area. They are considered to be the most meaningful statistic employed within crime studies (Boggs, 1965) and have a diverse range of applications. They are a valuable tool used to inform resource allocation, influence planning and policymaking by police forces and local governments, and convey messages regarding safety and risk to members of the public (National Academy of Sciences, 2016). Crime rates are also critical within environmental criminology as they enable researchers to better understand the spatial distribution of crime and analyse patterns (Boggs, 1965). Crime rates are currently calculated by dividing the number of crimes that occur (numerator) by the size of the population at risk within a given geographic area (denominator). The denominator most commonly employed in the calculation of crime rates is the resident population, which is captured by national censuses (Andresen and Jenion, 2010).

Despite the importance of appropriate denominators, estimates of the resident population are still widely utilised (Harries, 1981). The use of unsuitable denominators is financially and temporally expensive and result in the production of crime rates that do not accurately represent the risk of a crime occurring (Harries, 1981). As the denominator is the size of the population at risk, it would be expected that the denominator should vary depending on the level of risk exposure and on the crime type (Boggs, 1965). Barriers to the use of alternative measures of the population at risk, such as a lack of data and substantial cost, have been noted within the existing literature (Boggs, 1965; Andresen and Jenion, 2010). However, the proliferation of spatio-temporal data over the last two decades has resulted in the emergence of data types that may be useful for producing estimates of the size of the ambient population (Crols and Malleson, 2019).

The need for estimates of the ambient population for the use in the calculation of crime rates is supported by two of the most prominent theories within environmental criminology; routine activity theory and crime pattern theory (Cohen and Felson, 1979; Brantingham et al., 1981). Routine activity theory states that three elements must converge in space and time for a crime to occur; a target, an offender and the absence of a capable guardian (Cohen and Felson, 1979). This theory supports the need for estimates of the ambient population as the convergence of these three elements is related to both the size of and fluctuations within the population. Crime pattern theory suggests that crimes are more likely to occur in proximity to the nodes, paths, and edges which are familiar to the perpetrator (Brantingham et al., 1981). Brantingham and Brantingham (1995), expanded on crime pattern theory through the introduction of crime attractors and generators. Crime attractors are those locations, such as car parks, which attract criminals for the sole purpose of committing a crime. While crime generators, including shopping centres and train stations, lead to the convergence of victims and criminals in space and time. Both crime attractors and generators are commonly located in areas, such as urban centres, which attract large volumes of people and experience significant fluctuations in the size of the ambient population. Thus, it is evident that in accordance with both routine activity theory and crime pattern theory, estimates of the size of the ambient population are required to calculate accurate crime rates.

2.5.2 Exploration of the spatial distribution of crimes

The utility of estimates of the ambient population in crime studies was first noted by Boggs (1965) over fifty years ago. Boggs (1965) noted that the use of the resident population in crime rates may lead to high crime rates in central business districts in which there are few residents but high numbers of potential victims and targets. In her seminal work on patterns of urban crime, Boggs (1965) presents crime rates based on the environmental opportunities for different crime types. The results of a factor analysis test indicated that different crime types are concentrated by neighbourhood, while the impact of an alternative denominator varied between crime types. The study identified that car theft and business robbery would benefit

from the use of alternative denominators, while for homicide, aggravated assault and resident daytime burglary, crime specific rates did not add value. This study demonstrated that for some crime types, the use of the resident population as a denominator is inappropriate and, therefore, may be misleading.

Estimates of the ambient population, produced using traditional data types, have been utilised within environmental criminology to explore the spatial distribution of crimes. Andresen and Jenion (2008) demonstrate that estimates of the ambient population can be utilised at primary, secondary, and tertiary levels of crime prevention. The authors identify that LandScan data can be used to better understand areas that currently experience high rates of crime (tertiary), identify areas at risk of developing high crime rates (secondary), and quantify the relationship between crime and fluctuations in the ambient population to inform policy (primary). Stults and Hasbrouck (2015) examine the impact of commuting on crime rates in cities in the US using commuter adjusted daytime population estimates produced by the United States Census Bureau. The study found that for violent crime and three types of property crime, those cities which experienced an increase in the ambient population due to commuting trends have higher rates of crime (Stults and Hasbrouck, 2015). Work by Felson and Boivin (2015) employed travel survey data to explore the relationship between daily visitors in Eastern Canada for violent and property crimes. The study concludes that the size of the ambient population could potentially be more influential than the size of the resident population in determining the spatial distribution of crimes.

Novel types of data have also been employed to explore the spatial distribution of crimes in several studies. Work by Malleson and Andresen (2015b) used Twitter data as a proxy of the ambient population and found that different spatial patterns of crime emerge when using the ambient population and the resident population in Leeds, UK. An interesting finding from this work is that the city centre of the study area, where high numbers of crimes are recorded, does not contain any statistically significant clusters of violent crime. This finding is critical to the discussion

around the use of alternative measures of the population at risk, as it evidences the significant difference between the use of the ambient population and the resident population. Twitter data are also used with geographically weighted regression to forecast hotspots of street crime in Portland (US) (Ristea et al., 2018). The model was relatively successful and was able to predict hotspots containing 23% of future street crimes. Haleem et al. (2020) integrated mobile phone data and census data to estimate the size of the ambient and the exposed populations to assess hotspots of violent crime in public spaces on Saturday nights in Greater Manchester (UK). The study found a non-linear relationship between population size and occurrences of violent crimes, which established an expected link between violent crime and the night-time economy.

2.5.3 Investigation of relationships

Measures of the ambient population have been used within environmental criminology to examine the relationships between the size of the ambient population and the number of crime events. Work by Andresen and Jenion (2010) compared crime rates calculated using the ambient population, represented by LandScan data, to those calculated using the resident population. The study highlighted that the two crime rates have a very weak statistical relationship, which supports the argument to employ estimates of the size of the ambient population, particularly for the calculation of rates of violent crimes. However, this is the only study that employs a traditional data type to investigate the relationship between crime rates calculated using a measure of the ambient population.

More recent studies have employed mobile phone data as a measure of the ambient population to explore relationships. Hanaoka (2018a) examined the relationship between snatch and run offences and hourly estimates of the ambient population based on mobile phone data. The study found that the effects of the ambient population are substantial and that the effects differ between daytime and night-time hours. During daytime hours, when the ambient population is larger in size,

fewer snatch and run offences are committed. However, the same pattern was not evident during night-time hours as the number of snatch and run offences were weakly correlated with estimates of the ambient population. Work by Jung et al. (2020) compares the relationship between assault density and the ambient and the resident populations using a generalised linear model. The ambient population, which was represented using mobile phone data, is associated with the number of assaults throughout the four examined time periods, while the resident population fails to account for the spatio-temporal variation. He et al. (2020) also use spatially referenced mobile phone data to measure the size and activity patterns of the ambient population. They find that the size of the non-local (i.e. the non-resident) population is significantly correlated with the spatial variation of larceny-theft (He et al., 2020). This study is unique as the authors had access to mobile phone activity data from three state-owned and operated telecommunications providers in Xi'an in the Shaanxi province of China; consequently, location data for all mobile phone users within the study area were available. While these studies highlight the utility of mobile phone data for exploring the relationships between crime rates, acquiring mobile phone data with such extensive geographical coverage and high levels of enumeration would be unachievable in most countries.

Twitter data have also been used to investigate the relationship between crimes and the ambient population. Hipp et al. (2019) use Twitter data to investigate the relationship between the size of the ambient population and the number of reported crimes throughout a day. The work demonstrated that the number of violent crimes increases as the size of the ambient population increases and, interestingly, also evidenced that an increased ambient population has a strong negative correlation with rates of property-based burglary. Consequently, the study highlighted the value of the use of estimates of the size of the ambient population for calculating rates of crime that target properties rather than individuals. Twitter data are also employed by Tucker et al. (2021) to estimate the size of sectors of the ambient population; local residents, commuters and tourists and investigate any correlation with crime rates. In non-residential areas, commuters and tourists have a

positive correlation with public violence and private conflict. However, commuters and tourists have a relationship with violent crime and daytime hours during weekdays, while commuters and tourists only have a statistically significant relationship with private conflict during daytime hours on both weekdays and weekend days. It is important to note that the authors find that Twitter data has limited utility as the geographical coverage is sparse for residential areas (Tucker et al., 2021). These studies highlight that Twitter data can be employed in studies of the relationships between the size of the ambient population and crime; however, the data are limited in terms of geographical coverage.

2.5.4 Section summary

The existing literature has highlighted the importance of accurate crime rates and provided evidence that estimates of the resident population are not reflective of the population at risk for all crime types. However, there remains an opportunity to explore variations in both the spatial distribution of, and relationships between, crime rates calculated using different measures of the population at risk, including the ambient population. This opportunity is explored in Chapter 5 of the thesis, which investigates the impact of the use of three different measures of the population at risk on the rates of two crime types.

2.6 Chapter summary - Understanding the Ambient Population

This chapter provided a critical review of studies that developed or employed estimates of the size of the ambient population and highlighted the importance of these estimates within environmental criminology. Section 2.2 provided a brief introduction to the ambient population. The term the 'ambient population' was then defined, and the utility of estimates of the size of the ambient population was highlighted. The limitations to producing estimates of the size of the ambient population, primarily the lack of suitable data and a standardised methodology, were noted.

Section 2.3 critically discussed traditional and novel types of population data, which may be valuable for enumerating the size of the ambient population. The section highlighted relevant sources of population data and how they have been employed in existing studies as estimates of the size of the ambient population. The section also explored the advantages and disadvantages of both traditional and novel data types. It was noted that pedestrian counter data, produced by footfall cameras and Wi-Fi sensors, are a potentially viable source of population data for producing estimates of the size of the ambient population and their utility should be explored further. Pedestrian counter data are explored in more detail in Chapter 3 of this thesis and are employed to produce estimates of the size of the ambient population in chapters 4 and 5.

Section 2.4 provided an overview of methods and approaches utilised for producing estimates of population, both of the ambient population and of populations more generally.

Section 2.5 highlighted the importance of accurate crime rates for various applications and discussed early studies of the ambient population within environmental criminology. The section also outlined two of the most prominent theories within environmental criminology, routine activity theory and crime pattern theory, both of which support the use of alternative measures of the population at risk. The section then presented studies within environmental criminology that have employed estimates of the ambient population to assess the relationships between the crime rates calculated using different measures of the population at risk and their spatial distributions. The work in this chapter identifies an opportunity to explore variations in both the spatial distribution of, and relationships between, crime rates calculated using different measures of the population at risk, including the ambient population. This opportunity is explored in Chapter 5 of the thesis.

Reference list

Adamiak, C., Pitkänen, K. and Lehtonen, O. 2017. Seasonal residence and counterurbanization: the role of second homes in population redistribution in Finland. *GeoJournal*. **82**(5).

Andresen, M.A. and Jenion, G.W. 2010. Ambient populations and the calculation of crime rates and risk. *Security Journal*. **23**(2), pp.114–133.

Andresen, M.A. and Jenion, G.W. 2008. Crime prevention and the science of where people are. *Criminal Justice Policy Review*. **19**(2), pp.164–180.

Android 2020. Privacy: MAC Randomization. [Accessed 7 December 2020]. Available from: https://source.android.com/devices/tech/connect/wifi-mac-randomization.

Apple 2020. Mobility Trends Reports. *Mobility Trends Reports*. [Online]. [Accessed 8 February 2021]. Available from: https://covid19.apple.com/mobility.

Bell, M. 2001. Understanding circulation in Australia. *Journal of Population Research*. **18**(1).

Bengtsson, L., Lu, X., Thorson, A., Garfield, R. and von Schreeb, J. 2011. Improved response to disasters and outbreaks by tracking population movements with mobile phone network data: A post-earthquake geospatial study in haiti. *PLoS Medicine*. **8**(8).

Berry, T., Newing, A., Davies, D. and Branch, K. 2016. Using workplace population statistics to understand retail store performance. *International Review of Retail, Distribution and Consumer Research*. **26**(4), pp.375–395.

Bhaduri, B., Bright, E., Coleman, P. and Urban, M.L. 2007. LandScan USA: a high-resolution geospatial and temporal modeling approach for population distribution and dynamics. *GeoJournal*. **69**(1–2), pp.103–117.

Blondel, V.D., Decuyper, A. and Krings, G. 2015. A survey of results on mobile phone datasets analysis. *EPJ Data Science*. **4**(1).

Boeing, G. 2018. Estimating local daytime population density from census and payroll data. *Regional Studies, Regional Science*. **5**(1).

Boggs, S.L. 1965. Urban crime patterns. *American Sociological Review*., pp.899–908.

Brantingham, P.J., Brantingham, P.L. and others 1981. *Environmental criminology*. Sage Publications Beverly Hills, CA.

Brantingham, Patricia and Brantingham, Paul 1995. Criminality of place. *European journal on criminal policy and research*. **3**(3), pp.5–26.

Bricka, S., Zmud, J., Wolf, J. and Freedman, J. 2009. Household travel surveys with GPS an experiment. *Transportation Research Record*. (2105).

Charles-Edwards, E., Bell, M., Brown, D. and others 2008. Where people move and when: Temporary population mobility in Australia. *People and Place*. **16**(1), p.21.

Chen, K. 2002. An approach to linking remotely sensed data and areal census data. *International Journal of Remote Sensing*. **23**(1).

Chen, K. and McAneney, J. 2006. High-resolution estimates of Australia's coastal population. *Geophysical Research Letters*.

Citymapper 2020. Citymapper mobility Index: About the Data. [Accessed 23 August 2021]. Available from: https://citymapper.com/cmi.

Cohen, L.E. and Felson, M. 1979. Social change and crime rate trends: A routine activity approach. *American sociological review*., pp.588–608.

Crols, T. and Malleson, N. 2019. Quantifying the ambient population using hourly population footfall data and an agent-based model of daily mobility. *Geoinformatica*. **23**(2), pp.201–220.

Deville, P., Linard, C., Martin, S., Gilbert, M., Stevens, F.R., Gaughan, A.E., Blondel, V.D. and Tatem, A.J. 2014. Dynamic population mapping using mobile phone data. *Proceedings of the National Academy of Sciences of the United States of America*. **111**(45).

Dobson, J., Bright, E., Coleman, P. and Bhaduri, B. 2003. LandScan 2000: A new global population geography. *In*: *Remotely-sensed cities*.

Dobson, J.E., Bright, E.A., Coleman, P.R., Durfee, R.C. and Worley, B.A. 2000. LandScan: A global population database for estimating populations at risk. *Photogrammetric Engineering and Remote Sensing*.

Felson, M. and Boivin, R. 2015. Daily crime flows within a city. *Crime Science*. **4**(1), p.31.

Foley, D.L. 1954. Urban daytime population: A field for demographic-ecological analysis. *Social Forces*. **32**(4).

Freudiger, J. 2015. How talkative is your mobile device? An experimental study of Wi-Fi probe requests *In*: *Proceedings of the 8th ACM Conference on Security & Privacy in Wireless and Mobile Networks*., pp.1–6.

Gao, X., Yuan, H., Qi, W. and Liu, S. 2014. Assessing the social economic vulnerability of urban areas to disasters: A case study in Beijing, China. *International Review for Spatial Planning and Sustainable Development*. **2**(1).

García-Palomares, J.C., Salas-Olmedo, M.H., Moya-Gómez, B., Condeço-Melhorado, A. and Gutiérrez, J. 2018. City dynamics through Twitter: Relationships between land use and spatiotemporal demographics. *Cities*. **72**.

Gober, P. and Mings, R.C. 1984. A geography of nonpermanent residence in the U.S. *Professional Geographer*. **36**(2).

Google 2020. COVID19 Mobility Reports. [Accessed 8 February 2021]. Available from: https://www.google.com/covid19/mobility/?hl=en.

Haleem, M.S., Do Lee, W., Ellison, M. and Bannister, J. 2020. The 'Exposed' Population, Violent Crime in Public Space and the Night-time Economy in Manchester, UK. *European Journal on Criminal Policy and Research*.

Halford, E., Dixon, A., Farrell, G., Malleson, N. and Tilley, N. 2020. Crime and coronavirus: Social distancing, lockdown, and the mobility elasticity of crime. *Crime Science*. **9**(1).

Hamstead, Z.A., Fisher, D., Ilieva, R.T., Wood, S.A., McPhearson, T. and Kremer, P. 2018. Geolocated social media as a rapid indicator of park visitation and equitable park access. *Computers, Environment and Urban Systems*. **72**, pp.38–50.

Hanaoka, K. 2018. New insights on relationships between street crimes and ambient population: Use of hourly population data estimated from mobile phone users'

locations. *Environment and Planning B: Urban Analytics and City Science*.

Happel, S.K. and Hogan, T.D. 2002. Counting snowbirds: The importance of and the problems with estimating seasonal populations. *Population Research and Policy Review*. **21**(3).

Harries, K.D. 1981. Alternative denominators in conventional crime rates. *Environmental criminology*., pp.147–165.

He, L., Páez, A., Jiao, J., An, P., Lu, C., Mao, W. and Long, D. 2020. Ambient Population and Larceny-Theft: A Spatial Analysis Using Mobile Phone Data. *ISPRS International Journal of Geo-Information*. **9**(6), p.342.

Hipp, J.R., Bates, C., Lichman, M. and Smyth, P. 2019. Using social media to measure temporal ambient population: does it help explain local crime rates? *Justice Quarterly*. **36**(4), pp.718–748.

Jung, Y., Chun, Y. and Kim, K. 2020. Modeling Crime Density with Population Dynamics in Space and Time: An Application of Assault in Gangnam, South Korea. *Crime and Delinquency*.

Kashiyama, T., Pang, Y. and Sekimoto, Y. 2017. Open PFLOW: Creation and evaluation of an open dataset for typical people mass movement in urban areas. *Transportation Research Part C: Emerging Technologies*. **85**.

Kontokosta, C.E. and Johnson, N. 2017. Urban phenology: Toward a real-time census of the city using Wi-Fi data. *Computers, Environment and Urban Systems*. **64**, pp.144–153.

Lau, K.H. 2009. A GIS-based stochastic approach to generating daytime population distributions for vehicle route planning. *Transactions in GIS*. **13**(5–6).

Longley, P.A., Adnan, M. and Lansley, G. 2015. The geotemporal demographics of twitter usage. *Environment and Planning A*. **47**(2).

Lubans, D.R., Boreham, C.A., Kelly, P. and Foster, C.E. 2011. The relationship between active travel to school and health-related fitness in children and adolescents: A systematic review. *International Journal of Behavioral Nutrition and Physical Activity*. **8**.

Lwin, K.K., Sugiura, K. and Zettsu, K. 2016. Space–time multiple regression model for grid-based population estimation in urban areas. *International Journal of Geographical Information Science*. **30**(8).

Malleson, N. and Andresen, M.A. 2016. Exploring the impact of ambient population measures on London crime hotspots. *Journal of Criminal Justice*. **46**, pp.52–63.

Malleson, N. and Andresen, M.A. 2015a. Spatio-temporal crime hotspots and the ambient population. *Crime Science*.

Malleson, N. and Andresen, M.A. 2015b. The impact of using social media data in crime rate calculations: Shifting hot spots and changing spatial patterns. *Cartography and Geographic Information Science*.

Manley, E. and Dennett, A. 2019. New Forms of Data for Understanding Urban Activity in Developing Countries. *Applied Spatial Analysis and Policy*. **12**(1), pp.45–70.

Markham, F., Bath, J. and Taylor, J. 2013. *New directions in Indigenous service population estimation* [Online]. Available from: http://hdl.handle.net/1885/147837.

Martin, D., Cockings, S. and Leung, S. 2015. Developing a Flexible Framework for Spatiotemporal Population Modeling. *Annals of the Association of American Geographers*.

Martin, D., Cockings, S. and Leung, S. 2009. *Population 24/7: building time-specific population grid models*.

Martin, J., Mayberry, T., Donahue, C., Foppe, L., Brown, L., Riggins, C., Rye, E.C. and Brown, D. 2017. A study of MAC address randomization in mobile devices and when it fails. *arXiv*.

National Academy of Sciences 2016. *Modernizing Crime Statistics: Report 1: Defining and Classifying Crime*.

Nelson, A.C. and Nicholas, J.C. 1992. Estimating Functional Population for Facility Planning. *Journal of Urban Planning and Development*. **118**(2).

Oak Ridge National Laboratory, U.D. of E. n.d. Documentation. *LandScan*. [Online]. [Accessed 23 August 2021]. Available from:

https://landscan.ornl.gov/documentation.

OFCOM 2020. *Adults' Media Use & Attitudes report 2020*.

Office for National Statistics 2013a. 2011 Census: The workday population of England and Wales - An alternative 2011 Census output base. *Office for National Statistics*. [Online]. Available from: https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/articles/theworkdaypopulationofenglandandwales/2013-10-31.

Office for National Statistics 2013b. 2011 Census Statistics for England and Wales: March 2011 QMI. *Office for National Statistics*. [Online]. [Accessed 5 November 2021]. Available from: https://www.ons.gov.uk/.

Oliveira, L., Henrique, J., Schneider, D., de Souza, J., Rodriques, S. and Sherr, W. 2018. Sherlock: Capturing probe requests for automatic presence detection *In*: *2018 IEEE 22nd International Conference on Computer Supported Cooperative Work in Design ((CSCWD))*., pp.848–853.

Qi, W., Liu, S., Gao, X. and Zhao, M. 2015. Modeling the spatial distribution of urban population during the daytime and at night based on land use: A case study in Beijing, China. *Journal of Geographical Sciences*. **25**(6), pp.756–768.

Ratti, C., Frenchman, D., Pulselli, R.M. and Williams, S. 2006. Mobile landscapes: using location data from cell phones for urban analysis. *Environment and planning B: Planning and design*. **33**(5), pp.727–748.

Reades, J., Calabrese, F. and Ratti, C. 2009. Eigenplaces: Analysing cities using the space - Time structure of the mobile phone network. *Environment and Planning B: Planning and Design*. **36**(5).

Reades, J., Calabrese, F., Sevtsuk, A. and Ratti, C. 2007. Cellular census: Explorations in urban data collection. *IEEE Pervasive computing*. **6**(3), pp.30–38.

Rees, P., Martin, D. and Williamson, P. 2002. *The census data system*. Wiley.

Ristea, A., Al Boni, M., Resch, B., Gerber, M.S. and Leitner, M. 2020. Spatial crime distribution and prediction for sporting events using social media. *International*

Journal of Geographical Information Science.

Ristea, A., Kounadi, O. and Leitner, M. 2018. Geosocial Media Data as Predictors in a GWR Application to Forecast Crime Hotspots (Short Paper) *In*: *10th International Conference on Geographic Information Science (GIScience 2018)*.

Sallis, J.F. and Saelens, B.E. 2000. Assessment of physical activity by self-report: Status, limitations, and future directions. *Research Quarterly for Exercise and Sport*. **71**.

Schmitt, R.C. 1956. Estimating Daytime Populations. *Journal of the American Institute of Planners*. **22**(2), pp.83–85.

Silm, S. and Ahas, R. 2010. The seasonal variability of population in estonian municipalities. *Environment and Planning A*. **42**(10).

Sloan, L., Morgan, J., Housley, W., Williams, M., Edwards, A., Burnap, P. and Rana, O. 2013. Knowing the Tweeters: Deriving sociologically relevant demographics from Twitter. *Sociological Research Online*. **18**(3).

Smith, G., Arnot, C., Fairburn, J. and Walker, G. 2005. A National Population Data Base for Major Accident Hazard Modelling.

Smith, S.K. 1989. Toward a methodology for estimating temporary residents. *Journal of the American Statistical Association*. **84**(406).

Soundararaj, B., Cheshire, J. and Longley, P. 2020. Estimating real-time high-street footfall from Wi-Fi probe requests. *International Journal of Geographical Information Science*. **34**(2), pp.325–343.

Steiger, E., Westerholt, R., Resch, B. and Zipf, A. 2015. Twitter as an indicator for whereabouts of people? Correlating Twitter with UK census data. *Computers, environment and urban systems*. **54**, pp.255–265.

Stopher, P.R., Greaves, S.P. and FitzGerald, C. 2005. Developing and deploying a new wearable GPS device for transport applications *In*: *Second International Colloquium on the Behavioural Foundations of Integrated Land Use and Transportation Models: Frameworks, Models, and Applications, Toronto, June*., pp.13–15.

Stults, B.J. and Hasbrouck, M. 2015. The Effect of Commuting on City-Level Crime

Rates. *Journal of Quantitative Criminology*.

Sutton, P., Roberts, D., Elvidge, C. and Baugh, K. 2001. Census from Heaven: An estimate of the global human population using night-time satellite imagery. *International Journal of Remote Sensing*. **22**(16).

Toole, J.L., Ulm, M., González, M.C. and Bauer, D. 2012. Inferring land use from mobile phone activity *In*: *Proceedings of the ACM SIGKDD international workshop on urban computing*., pp.1–8.

Tucker, R., O'Brien, D.T., Ciomek, A., Castro, E., Wang, Q. and Phillips, N.E. 2021. Who 'Tweets' Where and When, and How Does it Help Understand Crime Rates at Places? Measuring the Presence of Tourists and Commuters in Ambient Populations. *Journal of Quantitative Criminology*.

United Nations 2008. Principles and recommendations for population and housing censuses - Revision 2. *Statistical Papers*. **3**(March).

United States Census Bureau 2017. Calculating Commuter-Adjusted Population Estimates. [Accessed 26 August 2021]. Available from: https://www.census.gov/.

Warchivker, I., Tjapangati, T. and Wakerman, J. 2000. The turmoil of Aboriginal enumeration: Mobility and service population analysis in a Central Australian community. *Australian and New Zealand Journal of Public Health*. **24**(4).

Wesolowski, A., Metcalf, C.J.E., Eagle, N., Kombich, J., Grenfell, B.T., Bjørnstad, O.N., Lessler, J., Tatem, A.J. and Buckee, C.O. 2015. Quantifying seasonal population fluxes driving rubella transmission dynamics using mobile phone data. *Proceedings of the National Academy of Sciences of the United States of America*. **112**(35).

Zandvliet, R. and Dijst, M. 2005. Research Note—The Ebb and Flow of Temporary Populations: The Dimensions of Spatial-Temporal Distributions of Daytime Visitors in the Netherlands. *Urban Geography*. **26**(4), pp.353–364.

Chapter 3

Estimates of the Ambient Population: Assessing the Utility of Novel Data Sources

This chapter has been published as:

Whipp, A., Malleson, N., Ward, J. and Heppenstall, A., 2021. Estimates of the Ambient Population: Assessing the Utility of Conventional and Novel Data Sources. ISPRS International Journal of Geo-Information, **10**(3), p.131.

The aim of this chapter, which addresses Research Objective 2 is to assess and critique sources of population data that have the potential to be used to produce estimates of the size of the ambient population in urban areas, including those utilised in the existing literature.

This chapter builds upon the work in Chapter 2 which identified the need to assess the utility of and differences between both traditional and novel types of population data, particularly pedestrian counter data.

Abstract

This paper will critically assess the utility of conventional and novel data sources for building fine-scale spatio-temporal estimates of the ambient population. It begins with a review of data sources employed in existing studies of the ambient population, followed by preliminary analysis to further explore the utility of each data set. The identification and critiquing of data sources which may be useful for building estimates of the ambient population is a novel contribution to the literature. This paper will provide a framework of reference for researchers within urban analytics and other areas where an accurate measurement of the ambient population is required. The work has implications for national and international applications where accurate small area estimates of the ambient population are crucial in the planning and management of urban areas, the development of realistic models and informing

policy. This research highlights workday population estimates, in conjunction with footfall camera and Wi-Fi sensors data as potentially valuable for building estimates of the ambient population.

3.1 Introduction

The United Nations (2018) estimates that 68% of the global population will be living in cities or other urban centres by 2050. This predicted rise in the size of urban populations highlights the urgent need to be able to quantify the ambient population. The ability to produce estimates of the ambient population is integral to the management and planning of urban areas and allows the development of insights into socioeconomic and environmental issues that impact cities (Batty, 2013). In this paper, the ambient population is defined as the number of people within a given geographical area at a specific point in time, excluding individuals at their place of residence and those utilising modes of transport.

This paper assesses the utility of conventional and novel data sources for producing estimates of the ambient population and identifies appropriate data sources recommended for use in future work. A UK-based case study in the city of Leeds, West Yorkshire is utilised to demonstrate spatio-temporal patterns produced by different data sources. The study is widely generalisable as similar data are available worldwide. The work addresses an omission in the existing literature by producing an assessment of potential data sources and recommends the utilisation of a combination of conventional and novel data sources to produce estimates of the ambient population.

There is a clear need to develop estimates of the ambient population in order to better understand urban dynamics and the needs of growing urban populations. Existing studies regarding the ambient population have employed a range of data sources, both conventional and novel; however, there is a lack of research assessing

the viability of these data sources. While the systematic literature review by Panczak, Charles-Edwards and Corcoran (2020) identifies potential data sources, it does not assess their suitability for building estimates of the ambient population. This paper assesses the viability of datasets previously employed and identifies those which may be useful and therefore should be validated. This is a necessary step in order to ensure the development of appropriate estimates of the ambient population in future work. The next section of the paper will evaluate conventional and novel data sources identified as potentially useful for quantifying the ambient population.

3.2 Data types

Despite estimates of the ambient population being highlighted as beneficial by Boggs (1965) over 50 years ago, there has been limited research within this area. Andresen et al. (2012) suggest that the lack of research is due to temporal and financial constraints. Often novel data were privately owned, thus unavailable or expensive (Andresen et al., 2012). However, these constraints are no longer as significant due to advances in technology resulting in high resolution population data being more widely available (Andresen, 2011).

This paper examines what will be referred to as conventional and novel data sources. Conventional data are those typically acquired from surveys, interviews and questionnaires and are available from national statistical agencies. Novel data are those collected from novel sources such as sensors, mobile phones, social media platforms and footfall cameras. Table 3.1 provides a summary of the data sources reviewed in this paper. These sources were selected as they are able to provide estimates of population which are relevant to the ambient population. The primary focus is on data available in the United Kingdom, but similar datasets exist in many other countries so the review will generalise widely.

Table 3.1 A summary of data sources reviewed

Category	Data type	Data source(s)	Description	Frequency of data collection	Open access	Ability to represent daytime population	Ability to provide detailed spatio-temporal information
Census data	Conventional	Usual resident population	The number of residents at each household on census day.	Decennial	Yes	No	No
		Mid-year population	A combination of various administrative datasets which aim to provide more up to date estimates of the usual resident population.	Annual	Yes	No	No

		Workday population	Workday population is the number of individuals in a geographical area who are in employment and whose workplace is within the specified area, in addition to those who are not in employment and are usual residents. These data are not collected in all countries.	Decennial	Yes	Yes	No
Other administrative datasets	Conventional	Travel surveys	Data on the movement of individuals. They are conducted at national and local levels, by government agencies	Annually	Yes	Yes	Yes
Mobile data	Novel	Mobile phone activity data	Produced either when a mobile phone receives or makes a call or SMS message or when a	Each time the phone	No	Yes	Yes

			device moves between cell towers. The data are highly granular, thus are spatially detailed. All records are timestamped.	communicates with a mask			
		Smartphone location data/mobility reports	Gathered by a variety of smartphone applications that track the location of a user.	Variable.	No	Yes	Yes
		Cell tower locations (OpenCellID)	OpenCellID is an open dataset of cell tower locations. The data are contributed by commercial organisations and by individuals. The dataset is not comprehensive and does not	When a user uploads data.	Yes	Yes	No

			provide full geographical coverage of area.				
Geo-located social media data	Novel	Twitter, Flickr, Foursquare, Facebook, etc.	These data are produced when users upload social- media posts with an attached geographical location.	When a user uploads a post.	Data is subject to restrictions such access to a limited sample and limited spatio-temporal detail.	Yes	Yes

Pedestrian counters	Novel	Footfall cameras	Counts of individuals passing a specific geographic point. These data are usually captured by local governments and private organisations operating in spaces such as shopping centres and city centres.	When a person passes a camera	Private organisations do not publicly release the data, but it is often available on agreement.	Yes	Yes
		Wi-Fi sensors	Wi-Fi sensors capture the MAC addresses of nearby Wi-Fi enabled mobile devices as they attempt to connect to a hub. The data are spatio-temporally detailed.	When a Wi-Fi enabled device passes a sensor.	Privately owned.	Yes	Yes

3.2.1 Conventional data sources

This section reviews 'conventional' sources that have been used to estimate the ambient population. The utility of conventional data is assessed in order to determine whether data lacking fine spatio-temporal detail have value for building estimates of the ambient population.

3.2.1.1 Census data

Estimates of populations have traditionally been derived from household surveys and government data sources, most notably population censuses. Data from the 2011 UK census includes estimates of the usual resident population, mid-year population and workday population. These measures of the population are currently widely used for academic research and industrial purposes (Kobayashi et al., 2011; Wardrop et al., 2018). The usual resident population is the count of the number of individuals usually resident at a given address (Nomis, 2013). Mid-year population estimates are calculated using the most recent census in addition to data regarding internal and external migration, births, deaths, etc (Office for National Statistics, 2021). Workday population data were introduced in the 2011 UK Census to quantify individuals at their place of work during typical working hours, in addition to those who are unemployed residents (Office for National Statistics, 2013a). Workday population data can provide an overview of the usual daytime population, unlike the residential population and mid-year population estimates. However, estimates of the workday population are not able to capture fluctuations in the size of the ambient population which occur throughout time. Additionally, similarly to estimates of the resident population, estimates of the workday population are only representative of a single day of a year (census day). Estimates of the workday population are not universally available and are only currently available for England, Wales, and the US (Mckenzie et al., 2013; Office for National Statistics, 2013b; Panczak et al., 2020).

Censuses held by national statistical offices represent the gold standard of data collection and are geographically comprehensive (Rees et al., 2002). There are examples of estimates of the ambient population being constructed from multiple data sources, typically including census data. Bhaduri et al. (2007) used census data as a primary input, combining it with remote sensing images to capture the average ambient population over a 24-hour period at a resolution of $1km^2$. Smith et al. (2005) produced a population database for hazard modelling that combined a variety of data sources, including measures from the UK census, leisure facilities and retail data. The limitations of this work included data accuracy and the rapidity with which census data become outdated. Martin, Cockings and Leung (2009b) proposed a framework that uses a range of administrative datasets including the census, Higher Education and Hospital Episode Statistics to produce a grid model of the average ambient population (Martin et al., 2009b). A weakness of the framework is that it relies on annual data and fails to include data which are produced by novel sources and contain high-levels of spatio-temporal detail. However, the authors acknowledge the potential value of novel data which supports the rationale of this work. Highlighting the utility of data from other national censuses, data from the Chinese census were used by Qi et al. (2015) to build daytime population estimates through the addition of tourism, school registration, hospital patient, and land use data. However, this research did little to expand work by Martin, Cockings and Leung (2009b), despite the availability of novel data, such as geo-located social media data, in 2015.

Despite these examples, the data are impacted by several issues, including under-enumeration and respondent errors (Sullivan, 2020). In the UK, it can take over 12 months for census data to be processed and released (Office for National Statistics, n.d.); thus, censuses conducted decennially are quickly outdated (Wardrop et al., 2018). Urban areas are continually in a state of flux, with changes in the residential population and workday population varying significantly within a short period of time (Nemeškal et al., 2020). These changes which occur at relatively fine temporal scales cannot be captured by a decennial census. However, conventional data sources may have been heavily utilised due to their accessibility.

3.2.1.2 Travel survey data

Data from travel surveys are able to provide detailed information regarding the movements of individuals. Travel surveys are conducted by a number of national and local governments across the world, but there is no general framework, resulting in inconsistent data. The frequency at which these surveys are conducted varies greatly, and many countries do not collect any travel data.

Travel surveys were primarily introduced to inform policymaking regarding transport planning and land-use, but recently they have also been used to examine the ambient population (Nitsche et al., 2014). Zandvliet and Dijst (2005) use the Netherlands National Travel Survey to examine temporary, visitor populations and determine the demographic characteristics of this temporary population. Similarly, Charles-Edwards et al. (2008) employed the National Australian Visitor Survey to gain insight into the temporary movements of the population, for purposes such as leisure activities. The surveys collect information regarding the typical journeys that people make, including journey length and the purpose of the trip. They are also able to capture valuable socio-demographic information about individuals which many novel data sources are not able to capture.

Many of the studies which utilise travel survey data are now becoming outdated, principally due to the availability of alternative, novel data sources. A primary advantage offered by travel survey data, especially when compared to novel data, is the ability to provide information regarding demographics, reason for travel and mode of transport. These features are not required to quantify the ambient population but may be valuable to future work examining the demographic characteristics of the ambient population. However, it should be noted that travel surveys are limited by their sample size. Due to the expense associated with collecting travel survey data, the samples are often very limited and may prevent findings from being extrapolated to the wider population (Faber and Fonseca, 2014).

Estimates of the ambient population can benefit from the use of conventional data sources, primarily due to their extensive geographical coverage. Workday population estimates are able to reveal more detail regarding the geographic location of individuals during a typical working day, deeming them valuable in attempting to quantify the ambient population. Despite the utility of workday population data, these data are still plagued by infrequent data collection and lack estimates of the numbers of people in an urban area for activities, such as shopping, socialising and tourism which are required to produce estimates of the ambient population. Although this limits their use as a sole measure of the ambient population, there may be value in combining these data with others (as Section 2.2 will discuss in detail). Consequently, the following sections discuss additional data from novel sources that may be useful in building estimates of the ambient population and fill the gaps in the more traditional sources.

3.2.2 Novel data sources

Several data sources have emerged in recent years that provide detailed spatio-temporal data that can be useful for building estimates of the ambient population. Due to the secondary nature of the data, i.e., many of the data sources were not designed to capture the ambient population, many have limitations, and few have been extensively explored (Steiger et al., 2015). The utility of novel data sources will be assessed in the remainder of this section.

3.2.2.1 Mobile phone data

Mobile phone activity data have been utilised by several studies that explore the ambient population. Ratti et al. (2006) demonstrate the benefits of mobile phone data for use within urban analytics and city planning through the production of a visual representation of urban activities in Milan, Italy. In a similar study, Reades et al. (2007) employ mobile phone data to build visualisations of mobile phone usage across Rome; however, due to the demographic characteristics of mobile phone users - i.e.

a small proportion of the elderly population use a mobile phone - the data fail to reflect the entire ambient population (Reades et al., 2007). Work by Terada, Nagata and Koboyashi (2013) accounts for socio-demographic characteristics such as age and gender and employed mobile phone activity data to produce spatial estimates of the population of Japan. Crucially, Reades et al. (2007) acknowledge that while traditional datasets have limited temporal detail, data access and ethical issues are barriers to the use of mobile phone activity data. He et al. (2020) used geo-referenced mobile phone data as a measure of the ambient population to assess the relationship between larceny (theft) in Xi'an, China. The dataset utilised provided full coverage of all mobile phone users within the study area and includes information such as gender and date of birth. The authors state that the work highlights the utility of mobile phone data for estimating the ambient population; however, they do not acknowledge that access to such a comprehensive dataset is not possible in many countries. Smartphone location data were utilised by (Hanaoka, 2018b) as an estimate of the ambient population. It is unknown whether these data are publicly available or if similar data are available for other countries. The work fails to assess the representativeness of the data and doesn't indicate whether the smartphone location data are able to reflect the size of the ambient population. Mobile phone activity data are not analysed in further detail due to the associated ethical concerns and the lack of data available at a sufficiently small geographical scale.

Since the outbreak of COVID-19, several technology companies, such as Apple and Google, have made mobility data available. Apple produce daily mobility reports which demonstrate the changes in routing requests via the Apple Maps application (Apple, 2020). The data are able to indicate changes in the percentage of requests for walking, driving and public transport routes (Apple, 2020). While the data are able to depict temporal trends in the percentage change of route requests, the representativeness of the data is a significant concern. Firstly, there is no information regarding whether people take the journeys they requested directions for using the Apple Maps application. Secondly, journeys which are not planned using the application are not captured. It can be assumed that routing requests for journeys

made more regularly, such as commuting to work and travelling to the supermarket, are less common. Additionally, the spatial detail of the data is limited and does not provide any indication to the number of journeys made into or out of an area, limiting the use of the data for quantifying the ambient population.

Google mobility reports indicate the percentage change in the visits to different location categories including retail and recreation, supermarket and pharmacy, parks, workplaces and public transport (Google, 2020). The level of spatial detail varies significantly between countries. The data are gathered from Google Account users who have devices that are able to track their movements (typically smart phones) and enable 'Location History'. Google state that the data may or may not be representative of the wider population (Google, 2020). Unlike Apple mobility data, Google mobility reports indicate journeys which have taken place; however, the number of journeys made and information regarding the representativeness of the data are unknown. Google also state the reports will only be available for a limited period of time, thus may not be available for use in future research (Google, 2020).

Although the data used to generate mobility reports (particularly the traces of individual peoples' movements that are used in the Google reports) may provide a valuable source of high-resolution information about the ambient population, at present the products are not released at a sufficient spatial granularity to be of direct use here. Typically, a single mobility estimate covers an entire city or borough. While these estimates may provide a useful picture of regional behaviour change, they are not sufficiently detailed to estimate the dynamics of the ambient population and will not be reviewed.

3.2.2.2 Geo-located social media data

Social media platforms are a novel source of vast quantities of real-time volunteered geographic data (Goodchild, 2007). Many social media platforms allow

users to share geographic data, including; Facebook, FourSquare and Twitter (Hecht and Stephens, 2014).

Volunteered geographic data generated on Twitter are noted as being exceptionally well suited to building estimates of the ambient population (Stefanidis et al., 2013; Malleson and Andresen, 2015; Steiger et al., 2015; Hamstead et al., 2018; Hipp et al., 2019). This is due to the open and accessible API and the detailed spatio-temporal information provided. However, if a request through the API exceeds 1% of total Tweets, the data are then limited to a random sample of 1% of all Tweets (Tucker et al., 2021). This sample is extremely limited and, consequently, may not be generalisable to the wider population (Faber and Fonseca, 2014). However, techniques such as geoparsing can be used to derive geographic information from Tweets that are not geotagged. Geoparsing can be utilised to convert free text descriptions of a location (toponyms) into geographic locations in the form of coordinates. Geoparsing can be conducted through a range of applications, such as the Python library Mordecai (Halterman, 2017; Gritta et al., 2020).

Despite the limitations of geo-located social media data have been utilised in diverse applications, from measuring tourism attractiveness (Giglio et al., 2019) to quantifying human mobility (Roy et al., 2019) and predictive crime modelling (Ristea et al., 2020). However, there are concerns regarding the generalisability of the data. Socio-economic characteristics, such as age and socioeconomic group, have a significant influence on the volume and temporal frequency of geo-located social media data (Liu et al., 2015). For example, in the UK 95% of 16-24 year olds have at least one social media profile, however this decreases to 39% of people aged between 65 and 74 (OFCOM, 2020). Twitter data have been used in existing work to quantify the size of the ambient population, to estimate the size of the population at risk from specific crimes and to test criminological theory (Malleson and Andresen, 2015a; Hipp et al., 2019; Liu et al., 2020).

Geo-located social media data are able to provide insight at fine spatio-temporal scales but are limited by their lack of generalisability. Further research into the representativeness of geo-located social media data would allow these types of data to be utilised within studies of the ambient population. However, the future of Twitter data in academic research may be limited as in 2019 Twitter announced that the option to geo-tag Tweets was going to be removed as most users do not use the feature (Tucker et al., 2021).

3.2.2.3 Wi-Fi sensor data

Wi-Fi sensors are a potentially viable tool for counting the number of individuals in an area and providing real-time data (Crols and Malleson, 2019; Soundararaj et al., 2020). Wi-Fi sensors record a count every time a Wi-Fi probe request is received from a Wi-Fi enabled device (Freudiger, 2015), such as a mobile phone. As a device moves through an urban area, it will attempt to connect to multiple access points, thus is counted at multiple geographical locations, providing detailed spatio-temporal data (Oliveira et al., 2018). When the sensor data are calibrated and validated, there can be certainty in the numbers of devices counted, but it is not yet evident how many people carry no Wi-Fi enabled device, or even multiple devices. Given the proliferation of the use of Wi-Fi enabled smart phones, Wi-Fi sensors are a cheap and feasible method of collecting data regarding the ambient population.

Ethical concerns regarding the use of Wi-Fi sensor data have recently become less significant due to technological developments. Wi-Fi sensors are able to capture the movements of individuals as probe requests contain a device's unique media access control (MAC) address (Freudiger, 2015). Many mobile device users will be unaware that their device emits probe requests, nor that probe requests would allow them to be tracked (Vanhoef et al., 2016). However, both Apple and Android devices now periodically change MAC addresses to prevent device users from being tracked (Martin et al., 2017) (Android, 2020). An additional barrier to the use of Wi-Fi sensors

data is accessibility. Often the data are privately owned, thus can only be acquired through an agreement, often financial.

Wi-Fi sensor data have not yet been used extensively, however the small number of studies that have employed them have demonstrated their value. Kontokosta and Johnson (2017) developed a real-time census with hourly estimates of the ambient population for Lower Manhattan, New York City using over 20 million Wi-Fi probe data points, in conjunction with data from conventional sources. User groups, such as daily, weekly, first-timers, or occasional visitors, were identified based upon hourly connections to the Wi-Fi sensors. This enabled the extraction of population estimates for workers, residents and visitors (Kontokosta and Johnson, 2017). The work provided an excellent foundation in using modelling techniques and Wi-Fi data to produce estimates of the population. Highlighting the value of using Wi-Fi sensor data in conjunction with other sources, Crols and Malleson (2019) used a combination of administrative datasets and footfall counts from Wi-Fi sensors to build an agent-based model of demographic characteristics of commuters. A significant limitation of this study was the lack of empirical data; thus, a validation process was not carried out.

While Wi-Fi sensor data may be a useful source of detailed spatio-temporal information for building estimates of the ambient population, the lack of accessible, open Wi-Fi sensor data may be a barrier to its use.

3.2.2.4 Footfall camera data

Footfall cameras are another source of individual movement data and are typically operated by private companies, thus, there is limited information regarding data accuracy.

Footfall cameras rely on a physical device to capture data. Therefore, it is crucial that the cameras are situated in appropriate locations and that there are sufficient devices to capture footfall in different geographical areas. Ensuring the equitable distribution of footfall cameras is a crucial issue, as highlighted by Robinson and Franklin (2020). While sensors are able to produce new data and subsequently new knowledge about urban population, where there is a lack of coverage gaps emerge, resulting in so-called 'sensor deserts' (Robinson and Franklin, 2020).

The most commonly used footfall camera technology is target-specific tracking. Target-specific tracking utilises counting devices mounted on the sides of buildings and CCTV columns. High-definition video is used with image processing algorithms to produce counts of pedestrians as they cross a virtual line. The cameras can be employed outdoors to measure footfall in urban centres. There are ethical concern regarding the use of these cameras as they have the potential to identify and track individuals (Righetti et al., 2018). However, it should be noted that not all cameras that utilise this technology record or store data.

Counts from footfall cameras have not been employed extensively within academic research, thus it is challenging to assess the potential benefits of the use of footfall camera data. Footfall cameras do offer spatio-temporally detailed data and are an unobtrusive way of quantifying the population. Further exploration of these data is needed to assess their utility and their accuracy.

3.3 Data assessment: A case study in a large UK city

This section will assess the suitability of conventional and novel datasets for building estimates of the ambient population by examining their spatio-temporal characteristics. Recall that the aim of the paper is not to create a new estimate of the ambient population, but to assess the viability of the datasets discussed previously and identify those which may be useful.

The study area for this section will be the city centre of Leeds, United Kingdom (UK), shown in Figure 3.1. The area which is encapsulated by the term 'city centre' was determined and agreed upon with Leeds City Council. While the following analysis benefits from the use of a case study, the findings regarding the efficacy of the datasets are globally generalisable. Leeds is the third-largest city in the UK with a population of 751,485 (Office for National Statistics, 2011a), while the usual resident population of the study area is 16,022. Leeds is the biggest commercial centre in the region, thus experiences high volumes of workers commuting into the city and is a popular destination for shopping and other leisure activities. Leeds is a major urban centre which experiences fluctuations in the ambient population, making it an ideal testbed for this work. Where data have been aggregated, workplace zones (the lowest level of UK geography), an administrative boundary is used and LSOAs have been utilised. Workplace zones were designed to represent consistent numbers of workers across England and Wales, thus vary in size based on the workday population in each workplace zone. LSOA are an administrative boundary which, on average, contain approximately 1500 residents or 650 households (Office for National Statistics, 2016).

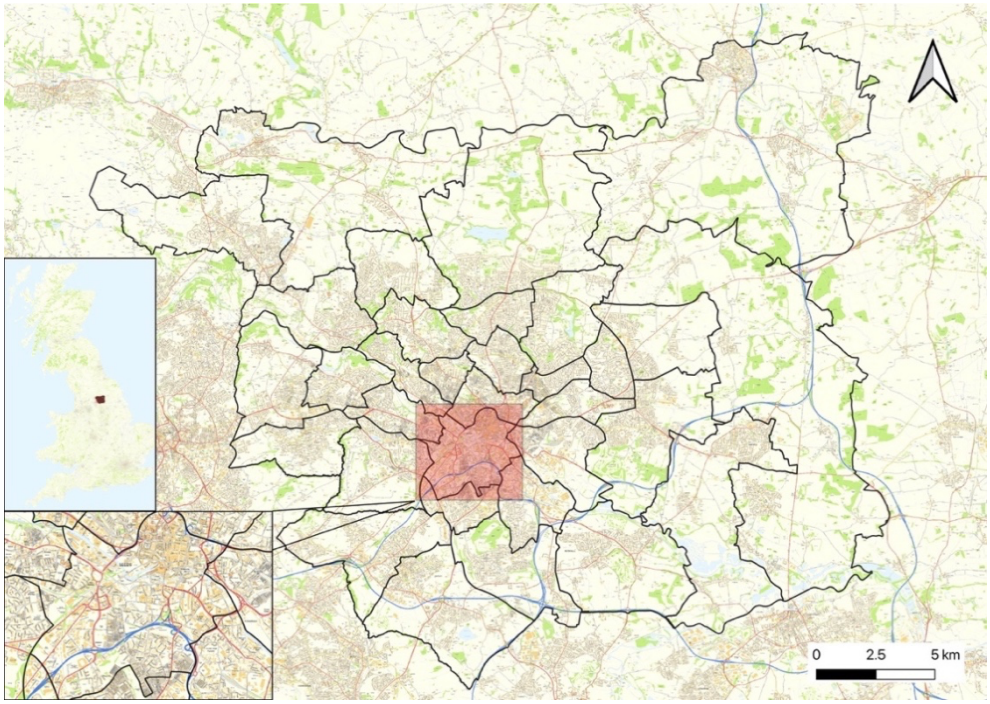


Figure 3.1 The study area, Leeds, United Kingdom. The inset maps highlight the focus area, which is the city centre of Leeds, in addition to the location of Leeds within the UK. The city centre covers an area of 4 km^2.

3.3.1. Census data

Estimates of the usual resident population and workday population are commonly used in small area estimates of the population. Within Leeds, there are vast differences between the two estimates due to the city centre attracting visitors, shoppers and workers, with the ward of City and Holbeck, which features the city centre, experiencing a 346% increase between estimates of the usual resident and workday populations (Office for National Statistics, 2011b).

Figure 3.2 demonstrates the workday population per workplace zone and the resident population per LSOA. The usual resident population is very low across much of the West of the study area, while it higher in the East. The maximum workday population is much higher than the maximum usual resident population. It should be noted that the workday population and the usual resident population are not available at the same geographic scale (workplace zone and LSOA level) and cannot be directly compared.

Figure 3.2 The workday population per workplace zone and the usual resident population per LSOA in Leeds city centre (Office for National Statistics, 2011b).

3.3.2. OpenCellID data

OpenCellID data, highlighted in Figure 3.3 represent the density of cell towers within the study area. Kernel Density Estimation (KDE) calculates the density of features, in this instance for point data, around each output raster cell (esri, n.d.). The KDE was produced using ArcGIS Pro 2.8.0. The KDE works by fitting a smoothly curved surface over each point (cell tower). The value of the surface is highest at the geographic location of the point and decreases as the distance from the feature increases (esri, n.d.). The bandwidth was calculated using an algorithm in ArcGIS Pro 2.8.0 which calculates the default search radius. The data are a cumulative record of cell towers and were downloaded on the 3rd December 2020 and, therefore capture all cell towers within the area. The cell towers are located primarily in the areas around the Leeds train station and the Trinity shopping centre. OpenCellID data are useful in helping to identify areas which are likely to experience high volumes of people, however there are no data regarding the number of individuals using a mobile device in each location.

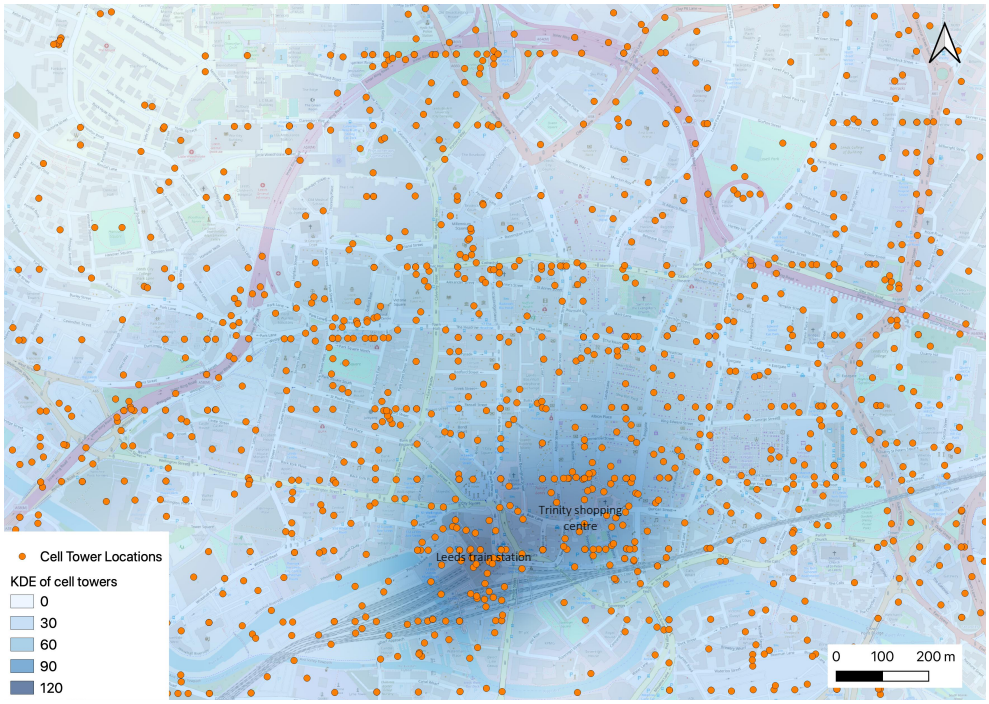


Figure 3.3 KDE of cell towers in Leeds city centre using a radius of 200m and a cell size of 2.79m^2. There are 1261 cell towers within the study area according to the OpenCellID database.

3.3.3 Geo-located social media data

Social media platforms have recently emerged as a possible source of data for building estimates of the ambient population, with Twitter being the most commonly used source of geo-located data. Figure 3.4 highlights hotspots of a random sample of 10,000 geo-located Tweets in the Leeds local authority district collected from 4th December 2015 to 14th February 2017. The data were collected using the Twitter Streaming API, listening for all tweets within the UK and filtering those with precise coordinates. The KDE of geo-located Tweets has a very different distribution to the KDE of cell towers, seen in Figure 3.3. The areas with the highest density of geo-located tweets are primarily located around the main shopping and leisure areas of the city centre. Towards the west of the city centre, in the business district, the density of geo-located Tweets is lower which suggests that people in the city for leisure purposes are the producers of geo-located Tweets. While the number of

Twitter users who send geo-located Tweets can be quantified, it is more challenging to determine their proportion within the ambient population.

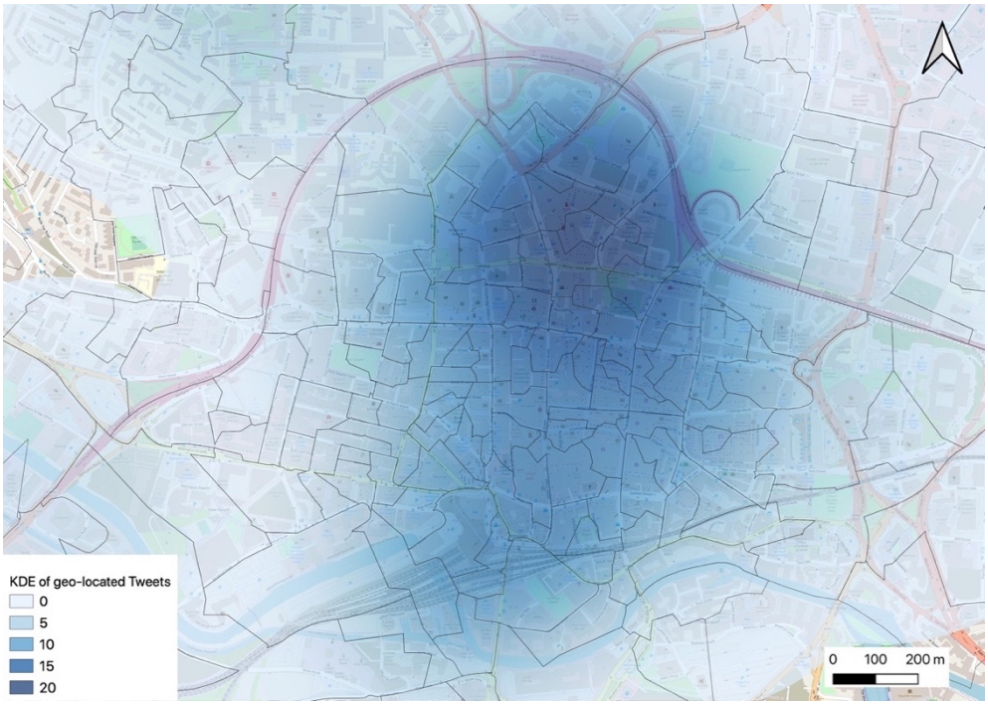


Figure 3.4 KDE of geo-located Tweets in Leeds city centre using a radius of 200m and a cell size of 2.79m^2.

3.3.4 Wi-Fi sensor data

Wi-Fi sensor counts are logs of Wi-Fi probe requests which occur when a Wi-Fi enabled device passes a Wi-Fi sensor. The Wi-Fi data used in this study were produced by the Local Data Company in partnership with the Consumer Data Research Centre. The sensor data were downloaded at 5-minute intervals for the 12 months of 2017. While this example is based in Leeds, there are other examples in cities such as London (UK) and Singapore. As can be seen in Figure 3.5, Wi-Fi sensors produce pedestrian counts at specific geographic points and enable the detection of patterns and fluctuations at different temporal levels such as daily or hourly. Figure 3.5 also highlights the importance of enumerating individuals who visit the city centre for leisure purposes as Saturday experiences the highest pedestrian count in all locations except St. Pauls Street, which is located in the business district.

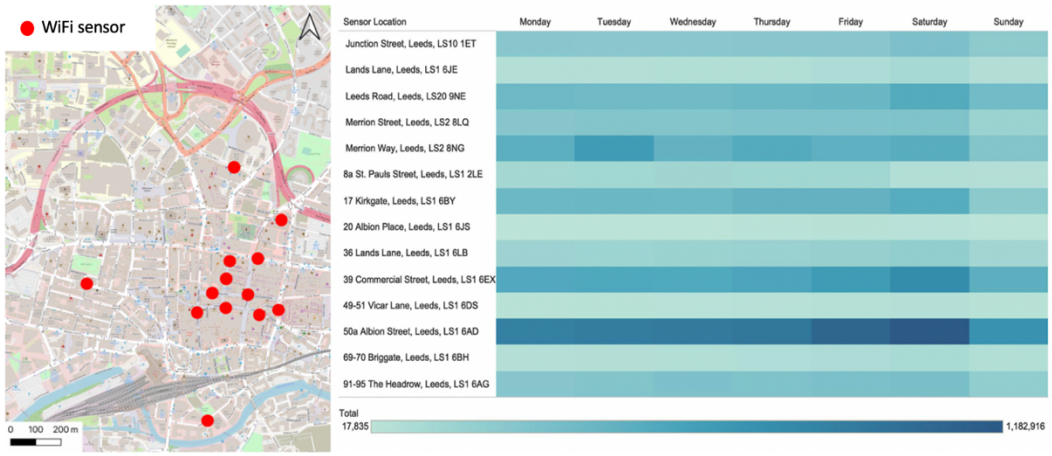


Figure 3.5 Counts from Wi-Fi sensor data capturing daily fluctuations (sum for a 24-hour period, averaged over 12 months) by location.

3.3.5 Footfall camera counts

Footfall cameras are a novel data source which are able to capture fluctuations in hourly and daily counts of pedestrians. The data examined in this section were aggregated from 8 footfall cameras located in Leeds in May 2018. Figure 3.6 highlights the hourly changes in footfall recorded by the footfall cameras. Mondays, Tuesdays and Wednesdays exhibit similar trends in hourly counts, likely to be due to workers in the city. Thursdays and Fridays experience higher footfall than days earlier in the week and footfall in the early evening, between 17:00 and 19:00 is evident. This could be linked to later shop closing times on these days and an increase in people socialising in the city centre towards the end of the typical working week. On both Saturday and Sunday, footfall reaches a peak later in the day, around 14:00, when compared to the working week.

The data offer fine spatio-temporal detail and enumerate non-workers in the city centre, both of which benefit estimates the ambient population. Footfall camera data could be used with workday population data in order to enumerate the ambient population by capturing workers and non-workers in an area.

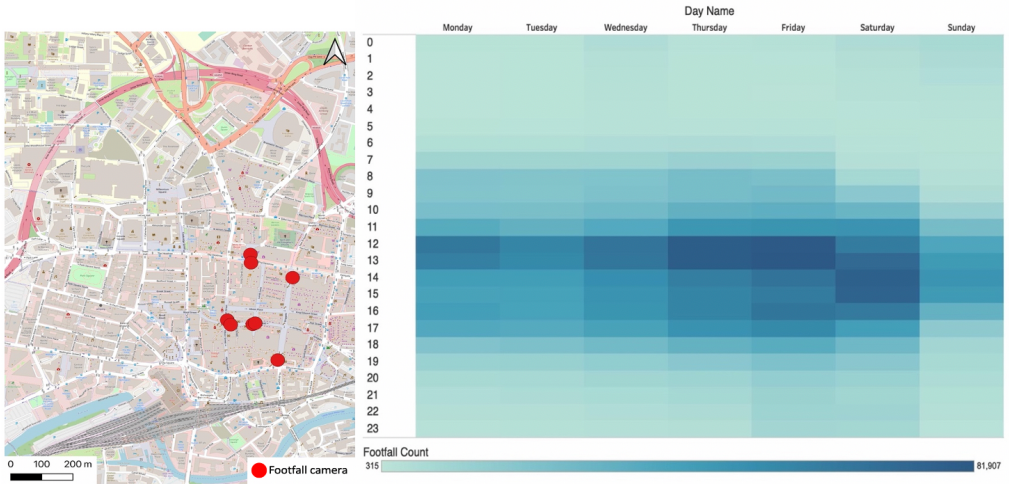


Figure 3.6 Hourly fluctuations in pedestrian counts from eight footfall cameras located in Leeds city centre and a map showing the locations of the cameras.

In Figure 3.7 the temporal trends of Wi-Fi sensor counts and footfall camera counts are highlighted. Both data sources demonstrated decreases in counts on the 7th May, 14th May, 21st May and 29th May. The peaks in Wi-Fi sensor counts occur more frequently than in the footfall camera data and do not share any overlaps temporally. Reasons for this may include that the Wi-Fi sensors and footfall cameras are located in different parts of the city centre, thus are not enumerating the same spatial locations. Additionally, the counts are captured in different ways; footfall cameras count the number of passing pedestrians, while Wi-Fi sensors count the number of Wi-Fi enabled devices that emit a probe request. Thus, the counts from the two different data sources would not be expected to be identical. Conducting a validation process will enable better understanding of the accuracy of counts from each data source.

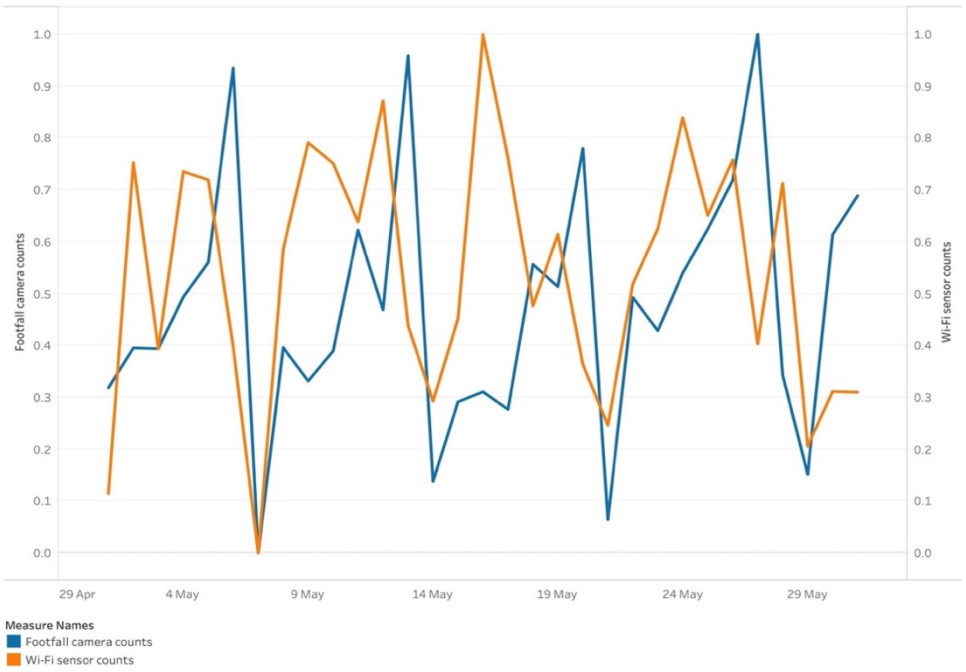


Figure 3.7 The number of pedestrians/Wi-Fi enabled devices captured in Leeds during May 2017. Hourly counts by location have been aggregated to daily counts and have been normalised.

3.4 Discussion and conclusion

This paper assesses the utility of conventional and novel data sources that have previously been identified as potential sources of data regarding the ambient population. It provides an assessment of the advantages and disadvantages of data previously employed to quantify the ambient population and identifies potentially useful data for use in future research. Future research may include data validation and the development of a methodological framework to quantify the ambient population. At the time of writing the authors are unaware of any other study that assesses the viability of data sources for producing estimates of the ambient population and identifies those which may be useful in future work.

The work notes the limited utility of conventional data sources to estimate the ambient population in cities, due to the infrequency of data collection and the lack of

spatio-temporal detail provided. However, these data have extensive geographic coverage and enumerate the majority of the population, encapsulating most, if not all, demographic groups (Rees et al., 2002). Workday population data were highlighted as a potentially useful measure for estimating the ambient population if used in conjunction with novel data which capture fluctuations throughout the day (Malleson and Andresen, 2016).

Novel sources of data, previously utilised in existing studies of the ambient population, have been acknowledged to have several significant limitations. OpenCellID data are able to indicate where people are likely to be located, but they are limited by the inability to enumerate the mobile devices connecting to a cell tower (Ulm et al., 2015). Consequently, the data have limited utility in producing estimates of the ambient population. Geo-located social media data have been identified as being able to provide detailed spatio-temporal logs of the locations of individuals; however, the data only represent a small-proportion of social media users and not the entire population of an area (Tucker et al., 2021). Finally, mobile phone data provide temporally frequent data but are expensive to purchase from network providers, which is a significant research barrier. Additionally, there are significant ethical issues surrounding the consent of mobile service users.

Footfall camera data have limited ethical concerns and are able to capture all individuals who pass the camera, thus can be representative of the whole population. However, it is possible that individuals may be counted by the same camera multiple times or be counted by multiple cameras. As with Wi-Fi sensors, a physical device has to be installed to capture data, therefore ensuring there are sufficient devices within a geographical area is crucial to ensure that there enough data and that the data are representative (Robinson and Franklin, 2020). Footfall camera data are able to capture non-working and atypical working populations at fine spatio-temporal scales. Wi-Fi sensors also offer spatio-temporal detail, but do not capture the entire population as they only capture the number of Wi-Fi enabled devices (Freudiger, 2015). Footfall camera data are able to enumerate the whole population without the

bias of the digital divide. The availability of Wi-Fi sensor and footfall camera data remains a significant issue, however this work has provided evidence that they are potentially valuable sources of data for building estimates of the ambient population. Following the direct comparison of footfall camera and Wi-Fi sensor counts, it is clear that a validation process must be undertaken to assess the accuracy of the data.

Assessing the utility of data sources for quantifying the ambient population is a crucial step in producing accurate estimates. While no single dataset is able to capture the ambient population, this paper has highlighted data sources which may be valuable for estimating the ambient population. Estimates of the ambient population would benefit from data which are geographically comprehensive and spatio-temporally detailed. Conventional data sources, such as the census are able to provide data which are geographically comprehensive, but they lack temporal detail (Office for National Statistics, 2013b). However, workday population estimates are able to provide an indication of work-related temporal fluctuations, in addition to providing an extensive geographical coverage (Office for National Statistics, 2013a). Footfall cameras and Wi-Fi sensors are able to pr ovide spatio-temporally detailed data which do not have associated ethical concerns, unlike mobile phone activity data (Reades et al., 2007). While geo-located social media are also able to provide data at a high spatio-temporal resolution, there is insufficient information regarding the representativeness of the data (Goodchild, 2007; Tucker et al., 2021). Additionally, Twitter data will no longer be geo-located which limits it's use in future research (Tucker et al., 2021). Consequently, Wi-Fi sensor and footfall camera data have been recommended as potentially valuable for estimating the ambient population. Issues such as data access and counting individuals multiple times remain, but validating and exploring these datasets further would enable the development of a framework for building estimates of the ambient population (Bernardin and Stiefelhagen, 2008). Future work should include the validation of counts from footfall cameras and Wi-Fi sensors and the production of a comprehensive framework to estimate the ambient population.

Reference list

Andresen, M.A. 2011. The ambient population and crime analysis. The Professional Geographer. **63**(2), pp.193–212.

Andresen, M.A., Jenion, G.W. and Reid, A.A. 2012. An evaluation of ambient population estimates for use in crime analysis. Crime Mapping: A Journal of Research and Practice. **4**(1), pp.7–30.

Android 2020. Privacy: MAC Randomization. [Accessed 7 December 2020]. Available from: https://source.android.com/devices/tech/connect/wifi-mac-randomization.

Apple 2020. Mobility Trends Reports. ***Mobility Trends Reports***. [Online]. [Accessed 8 February 2021]. Available from: https://covid19.apple.com/mobility.

Batty, M. 2013. ***The new science of cities***. MIT press.

Bhaduri, B., Bright, E., Coleman, P. and Urban, M.L. 2007. LandScan USA: a high-resolution geospatial and temporal modeling approach for population distribution and dynamics. ***GeoJournal***. **69**(1–2), pp.103–117.

Boggs, S.L. 1965. Urban crime patterns. ***American sociological review***. **30**(6), pp.899–908.

Charles-Edwards, E., Bell, M., Brown, D. and others 2008. Where people move and when: Temporary population mobility in Australia. ***People and Place***. **16**(1), p.21.

Crols, T. and Malleson, N. 2019. Quantifying the ambient population using hourly population footfall data and an agent-based model of daily mobility. ***Geoinformatica***. **23**(2), pp.201–220.

Echtner, C.M., Ritchie, J.R.B. and Ritchie, A.B. 1993. The Measurement of Destination Image: An Empirical Assessment.

Freudiger, J. 2015. How talkative is your mobile device? An experimental study of Wi-Fi probe requests ***In***: Proceedings of the 8th ACM Conference on Security & Privacy in Wireless and Mobile Networks., pp.1–6.

Giglio, S., Bertacchini, F., Bilotta, E. and Pantano, P. 2019. Using social media to identify tourism attractiveness in six Italian cities. Tourism management. 72, pp.306–312.

Goodchild, M.F. 2007. Citizens as sensors: the world of volunteered geography. GeoJournal. **69**(4), pp.211–221.

Google 2020. COVID19 Mobility Reports. [Accessed 8 February 2021]. Available from: https://www.google.com/covid19/mobility/?hl=en.

Hamstead, Z.A., Fisher, D., Ilieva, R.T., Wood, S.A., McPhearson, T. and Kremer, P. 2018. Geolocated social media as a rapid indicator of park visitation and equitable park access. Computers, Environment and Urban Systems. **72**, pp.38–50.

Hanaoka, K. 2018. New insights on relationships between street crimes and ambient population: Use of hourly population data estimated from mobile phone users' locations. Environment and Planning B: Urban Analytics and City Science. **45**(2), pp.295–311.

He, L., Páez, A., Jiao, J., An, P., Lu, C., Mao, W. and Long, D. 2020. Ambient Population and Larceny-Theft: A Spatial Analysis Using Mobile Phone Data. ISPRS International Journal of Geo-Information. **9**(6), p.342.

Hecht, B. and Stephens, M. 2014. A tale of cities: Urban biases in volunteered geographic information In: Eighth International AAAI Conference on Weblogs and Social Media.

Hipp, J.R., Bates, C., Lichman, M. and Smyth, P. 2019. Using social media to measure temporal ambient population: does it help explain local crime rates? Justice Quarterly. **36**(4), pp.718–748.

Kobayashi, T., Medina, R.M. and Cova, T.J. 2011. Visualizing diurnal population change in urban areas for emergency management. ***Professional Geographer***.

Kontokosta, C.E. and Johnson, N. 2017. Urban phenology: Toward a real-time census of the city using Wi-Fi data. Computers, Environment and Urban Systems. **64**, pp.144–153.

Kounadi, O., Ristea, A., Leitner, M. and Langford, C. 2018. Population at risk: using areal interpolation and Twitter messages to create population models for burglaries and robberies. Cartography and Geographic Information Science. **45**(3), pp.205–220.

Liu, L., Lan, M., Eck, J.E., Yang, B. and Zhou, H. 2020. Assessing the Intraday Variation of the Spillover Effect of Tweets-Derived Ambient Population on Crime. Social Science Computer Review.

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. Annals of the Association of American Geographers. **105**(3), pp.512–530.

Malleson, N. and Andresen, M.A. 2015a. Spatio-temporal crime hotspots and the ambient population. Crime Science.

Malleson, N. and Andresen, M.A. 2015b. The impact of using social media data in crime rate calculations: Shifting hot spots and changing spatial patterns. Cartography and Geographic Information Science.

Martin, D., Cockings, S. and Leung, S. 2015. Developing a Flexible Framework for Spatiotemporal Population Modeling. *Annals of the Association of American Geographers*.

Martin, D., Cockings, S. and Leung, S. 2009. *Population 24/7: building time-specific population grid models*.

Martin, J., Mayberry, T., Donahue, C., Foppe, L., Brown, L., Riggins, C., Rye, E.C. and Brown, D. 2017. A study of MAC address randomization in mobile devices and when it fails. *arXiv*.

Nitsche, P., Widhalm, P., Breuss, S., Brändle, N. and Maurer, P. 2014. Supporting large-scale travel surveys with smartphones - A practical approach. *Transportation Research Part C: Emerging Technologies*.

OFCOM 2020. *Adults' Media Use & Attitudes report 2020*.

Office for National Statistics 2011. 2011 Census aggregate data. *UK Data Service*. [Online]. [Accessed 20 September 2021]. Available from: infuse2011gf.ukdataservice.ac.uk/.

Oliveira, L., Henrique, J., Schneider, D., de Souza, J., Rodriques, S. and Sherr, W. 2018. Sherlock: Capturing probe requests for automatic presence detection *In*: *2018 IEEE 22nd International Conference on Computer Supported Cooperative Work in Design ((CSCWD))*., pp.848–853.

Panczak, R., Charles-Edwards, E. and Corcoran, J. 2020. Estimating temporary populations: a systematic review of the empirical literature. *Palgrave Communications*.

Qi, W., Liu, S., Gao, X. and Zhao, M. 2015. Modeling the spatial distribution of urban population during the daytime and at night based on land use: A case study in Beijing, China. *Journal of Geographical Sciences*. **25**(6), pp.756–768.

Ratti, C., Frenchman, D., Pulselli, R.M. and Williams, S. 2006. Mobile landscapes: using location data from cell phones for urban analysis. *Environment and planning B: Planning and design*. **33**(5), pp.727–748.

Reades, J., Calabrese, F., Sevtsuk, A. and Ratti, C. 2007. Cellular census: Explorations in urban data collection. *IEEE Pervasive computing*. **6**(3), pp.30–38.

Rees, P., Martin, D. and Williamson, P. 2002. *The census data system*. Wiley.

Robinson, C. and Franklin, R.S. 2020. The sensor desert quandary: What does it mean (not) to count in the smart city? *Transactions of the Institute of British Geographers*.

Roy, K.C., Cebrian, M. and Hasan, S. 2019. Quantifying human mobility resilience to extreme events using geo-located social media data. *EPJ Data Science*. **8**(1), p.18.

Smith, G., Arnot, C., Fairburn, J. and Walker, G. 2005. A National Population Data Base for Major Accident Hazard Modelling.

Soundararaj, B., Cheshire, J. and Longley, P. 2020. Estimating real-time high-street footfall from Wi-Fi probe requests. *International Journal of Geographical Information Science*. **34**(2), pp.325–343.

Stefanidis, A., Crooks, A. and Radzikowski, J. 2013. Harvesting ambient geospatial information from social media feeds. *GeoJournal*. **78**(2), pp.319–338.

Steiger, E., Westerholt, R., Resch, B. and Zipf, A. 2015. Twitter as an indicator for whereabouts of people? Correlating Twitter with UK census data. *Computers, environment and urban systems*. **54**, pp.255–265.

Sullivan, T. 2020. *Census 2020: Understanding the Issues* 1st ed. (T. Sullivan, ed.). Springer.

Terada, M., Nagata, T. and Kobayashi, M. 2013. Population estimation technology for mobile spatial statistics. *NTT DOCOMO Techn. J*. **14**, pp.10–15.

Tucker, R., O'Brien, D.T., Ciomek, A., Castro, E., Wang, Q. and Phillips, N.E. 2021. Who 'Tweets' Where and When, and How Does it Help Understand Crime Rates at Places? Measuring the Presence of Tourists and Commuters in Ambient Populations. *Journal of Quantitative Criminology*.

United Nations 2018. 68% of the world population projected to live in urban areas by 2050, says UN. *United Nations News*.

Vanhoef, M., Matte, C., Cunche, M., Cardoso, L.S. and Piessens, F. 2016. Why MAC address randomization is not enough: An analysis of Wi-Fi network discovery mechanisms *In*: *Proceedings of the 11th ACM on Asia Conference on Computer and Communications Security*., pp.413–424.

Wardrop, N.A., Jochem, W.C., Bird, T.J., Chamberlain, H.R., Clarke, D., Kerr, D., Bengtsson, L., Juran, S., Seaman, V. and Tatem, A.J. 2018. Spatially disaggregated population estimates in the absence of national population and housing census data. *Proceedings of the National Academy of Sciences of the United States of America*.

Zandvliet, R. and Dijst, M. 2005. Research Note—The Ebb and Flow of Temporary Populations: The Dimensions of Spatial-Temporal Distributions of Daytime Visitors in the Netherlands. *Urban Geography*. **26**(4), pp.353–364.

Zarsky, T.Z. 2016. Incompatible: the GDPR in the age of big data. *Seton Hall L. Rev.* **47**, p.995.

Chapter 4

Towards a Comprehensive Measure of the Ambient Population: Building Estimates Using Geographically Weighted Regression

This chapter is ready to be submitted to a peer reviewed journal as:

Whipp, A., Malleson, N., Ward, J. and Heppenstall, A. Towards a comprehensive measure of the ambient population: Building estimates using geographically weighted regression.

This chapter has three aims:

1. To develop small area estimates of the size of the ambient population for an urban area.
2. To produce a validation dataset that captures footfall counts in an urban area.
3. To employ the validation dataset to assess the accuracy of the manual footfall counts, the footfall camera counts and the model estimates.

The work in this chapter fulfils research objectives 3, 4, and 5. This chapter builds on the work in Chapter 3 which identified the need to produce small-area estimates of the ambient population using both traditional and novel data types.

The paper was submitted with supplementary materials which included an R script containing all of the code needed to replicate the model.

Abstract

Estimates of the resident population fail to account for human mobility, which significantly impacts the numbers of people in urban areas. Employing the ambient population provides a more nuanced approach to small-area population estimation. This paper utilises a method of statistical modelling, geographically weighted

regression, and novel data to estimate the size of the ambient population in an urban area. Models of the daytime and night-time ambient populations are produced for the city of Leeds, West Yorkshire, UK. Interestingly, the presence of cash machines and hospitality venues were found to be statistically significant and were identified as the most important predictors of the ambient population. In contrast to the literature, the number of retail hubs, transport hubs, and the density of mobile phone cell towers were not found to have statistically significant relationships with footfall camera counts. Footfall camera data and the results of the predictive model were validated through comparison with manually collected pedestrian counts. The results of this validation process demonstrated that at five out of the six locations in Leeds city centre, the model produced expected estimates of the size of the ambient population. The results suggest that the approach of this study can be used as a tool to inform decision-making within local government and studies in which small area estimates of ambient populations are required.

4.1 Introduction

Estimates of the ambient population quantify fluctuations in the non-residential population. They are a valuable asset in policymaking and are an essential tool within social science research. The ambient population can be defined as "the number of people within a given geographical area at a specific point in time, excluding individuals at their place of residence and those utilising modes of transport" (Whipp, Malleson, et al., 2021, p.131). Estimates of the ambient population can be used to gain a better understanding of human mobility (Ratti et al., 2006; Reades et al., 2007; Kontokosta and Johnson, 2017), to inform hazard management (Smith et al., 2005; Chen and McAneney, 2006; Løvholt et al., 2012) and to represent crime rates with greater accuracy (Andresen and Jenion, 2010; Andresen, 2011; Mburu and Helbich, 2016). Despite the utility of estimates of the ambient population, previously there have been limited attempts to produce and employ them.

Much of the recent literature has focussed on the use of single, novel data sources as a proxy of the ambient population, rather than combining multiple sources to produce more comprehensive estimates. This study aims to estimate the size of the ambient population in an urban area using numerous traditional and novel datasets. This aim is fulfilled by employing geographically weighted regression to estimate the size of the ambient population using high-resolution spatio-temporal footfall camera counts and contextual factors, such as estimates of the workday population, the locations of mobile phone cell towers, and land use data. The estimates of the ambient population are then externally validated using manual footfall counts to assess the accuracy of both the model estimates and of the footfall camera data.

The article is structured as follows. Section 4.2 reviews existing research relating to quantifying and exploring the ambient population. The data and methods utilised are presented in Section 4.3, while Section 4.4 outlines and discusses the results from the predictive models. In Section 4.5, the validation process is outlined, and the results are presented. Section 4.6 provides a conclusion and recommendations for future work.

4.2 Background

Estimates of the ambient population can aid the development of a more detailed understanding of the fluctuations in human activity patterns. There are currently no widely accepted methods for estimating the size of the ambient population, despite the many applications for its use and the value of these estimates being acknowledged within the literature over 60 years ago (Schmitt, 1956; Martin et al., 2009b).

Existing work has attempted to use traditional datasets, such as censuses, in conjunction with areal interpolation methods to produce spatially detailed estimates

of the ambient population. Various areal interpolation methods have been employed to produce estimates of populations, including dasymetric mapping (Mennis and Hultgren, 2006; Bhaduri et al., 2007; Sims et al., 2017), grid-based modelling (Martin, 1989; Martin, 1996; Martin et al., 2009b; Martin et al., 2015) and pycnophylactic interpolation (Tobler, 1979; Tobler et al., 1995; Tobler et al., 1997). These methods commonly utilise land use data in conjunction with auxiliary data to produce small area estimates of the population. The primary advantage of areal interpolation methods is the ability to disaggregate coarse data to a finer spatial resolution to produce small-area estimates of the ambient population (Mennis and Hultgren, 2006). However, due to the increased volume and availability of spatio-temporally detailed data, the utility of areal interpolation methods has diminished.

There are several limitations of areal-based interpolation methods. Dasymetric mapping assumes that all features are equitably distributed within a geographical area and the accuracy of the results is dependent on the resolution of the auxiliary data utilised. The auxiliary data are most commonly remotely sensed images; however, using these images to enumerate high-rise buildings, which are typically located in urban areas, is a sizable challenge and can lead to reduced data accuracy (Li et al., 2020). Consequently, there are concerns regarding the accuracy of remotely sensed images for urban areas, as high-rise buildings have a significant impact on the number of people within an area. Grid-based modelling techniques, first developed in work by Martin (1989), interpolate values from a control point, often the centroid of a polygon. The control point has a significant impact on the results of the interpolation, as, if the geometric centroid is outside of the polygon boundary it can impact the reliability of the interpolated values (Comber and Zeng, 2019). Pycnophylactic interpolation was used in a study by Tobler (1979) to produce interpolated values of the residential population in Michigan, U.S. However, as pycnophylactic interpolation produces a smooth interpolated surface, the method is only suitable for use with continuously observed phenomena; thus, use with population data is not appropriate (Comber and Zeng, 2019). Due to the limitations of areal interpolation-based methods and the reduced need to disaggregate coarse

data, this study focusses on the use of both traditional and novel sources of data with a method of statistical modelling.

As high-resolution spatio-temporal data became more readily available there was a shift towards the use of novel data as a proxy of the ambient population. Spatio-temporal data, including geo-located social media data (Kounadi et al., 2018; Giglio et al., 2019; Roy et al., 2019), mobile phone activity data (Reades et al., 2007; Terada et al., 2013) and Wi-Fi sensor counts (Kontokosta and Johnson, 2017; Crols and Malleson, 2019), have been utilised both as a proxy of the ambient population and to provide insight into its spatial distribution. However, there have been limited attempts to use these types of data in conjunction with quantitative methods to produce estimates of the ambient population. To address this gap in the literature, this paper aims to use high-resolution spatio-temporal data and statistical modelling to produce estimates of the ambient population in an urban area.

To produce comprehensive estimates of the ambient population, geographically weighted regression (GWR) is employed in this study. GWR is a local spatial regression technique that can capture spatial heterogeneity. GWR has been used in a range of applications, such as crime studies (Cahill and Mulligan, 2007; Stein et al., 2016; Xu et al., 2019; Maldonado-Guzmán, 2020), land use (Tu and Xia, 2008; Wang et al., 2011; Liu et al., 2015; Chen et al., 2016; Munira and Sener, 2020), transport (Cardozo et al., 2012; Selby and Kockelman, 2013; Pirdavani et al., 2014; Chiou et al., 2015; Yang et al., 2017), pollution (Robinson et al., 2013) and health (Comber et al., 2011; Yang and Matthews, 2012; Kauhl et al., 2016). GWR has been used for population estimation to predict population size based upon satellite imagery data (Lo, 2008; Chu et al., 2019; Roni and Jia, 2020). However, the method has not yet been employed to build estimates of the ambient population using non-image data. Given this gap in the literature, this paper aims to utilise novel sources of data and GWR to produce estimates of the ambient population in an urban area.

4.3 Data and methodology

4.3.1 Study area and geography

The selected study area is the city centre of Leeds, UK (see Figure 4.1). The area has a large retail offering, a significant financial and legal district and is a popular tourist destination which result in fluctuations in the ambient population, making it an ideal test-bed for this work. The study area has a resident population of 16,022 and a workday population of 134,244 (Office for National Statistics, 2011b), highlighting the diurnal fluctuations which occur in the centre of the city due to the dynamics of the ambient population. The geographic area encapsulated as the city centre of Leeds was developed through discussions with Leeds City Council and is shown in Figure 4.1. The approach used in this study could be generalised to other city or urban centres where the necessary data are available. Other candidate cities which could be suitable for the use of this approach for estimating the ambient population include Liverpool and Manchester, due to the similarities between the retail offerings, geographic concentrations of workplaces, and the night-time economy. The data sources are described in detail in sections 4.2 and 4.3.

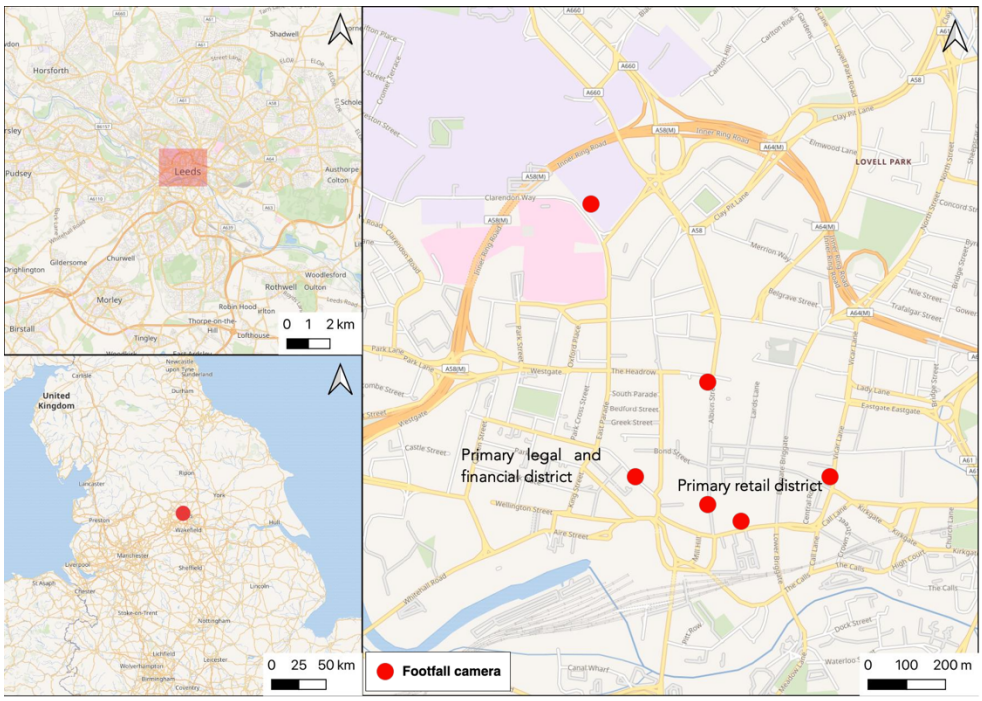


Figure 4.1 The study area of the city centre of Leeds, UK. The inset maps represent the location of Leeds within the UK. The study area covers an area of 4km2. Basemap data copyrighted OpenStreetMap contributors and available from https://www.openstreetmap.org.

The data utilised have been aggregated to workplace zones. Workplace zones are a set of output geographies from the 2011 UK Census and are the smallest level of UK geography designed specifically to represent the workday (i.e. non-residential) population (Office for National Statistics, 2014). Workplace zones were selected as the unit of geography as using the smallest level of geography enables small-area estimates to be produced which can provide a more representative indication of how the size of the ambient population varies across space. The study area contains 116 workplace zones, and each workplace zone was designed to contain an average of 500 people (House of Commons, 2014).

4.3.2 Dependent variables

The dependent variables utilised in the statistical models are daytime and night-time counts of pedestrians from eight footfall cameras located within the study area from the 1st to the 31st of May 2019. The cameras, operated by Leeds City Council, capture hourly counts of pedestrians passing the cameras and the data are openly available from Data Mill North (https://datamillnorth.org/), a source of open data created by Leeds City Council. The counts from the eight cameras were divided into daytime counts, between the hours of 07:00 and 19:00, and night-time counts, between 19:00 and 07:00. The time periods that represent the daytime and the night-time were selected based on discussions held with Leeds City Council. As commuters typically enter the city from 07:00 onwards, this was selected as the start of the daytime period, while 19:00 was selected as the end of the daytime period as this is when there is an increase in activity associated with the night-time economy.

The data were then aggregated by workplace zone, producing daytime and night-time counts for each of the zones. The accuracy of the footfall camera counts is

unknown; thus, the data are validated using manual counts in Section 4.5. The footfall cameras cover a limited geographical area, spanning six workplace zones; consequently, in two zones the mean of two separate camera counts is used. All eight cameras are in the primary shopping district; therefore, the observed dependent variables may not capture all human activity patterns that produce fluctuations in the ambient population and may impact the model fit. Footfall camera data are becoming increasingly accessible and are currently openly available for several cities including Melbourne (Australia), Auckland (New Zealand) and Dublin (Ireland). Creating similar models for such cities and comparing them to the model produced here presents an interesting opportunity for future work.

4.3.3 Independent variables

The independent variables used in this study were selected based upon factors identified in the existing literature as being associated with levels of footfall or the size of the ambient population. The selected variables are discussed in detail below and summarised in Table 4.1. The independent variables were assessed for multicollinearity using a variance inflation factor which identifies correlation between independent variables. The variance inflation factor was below 10 in all instances, thus there was not multicollinearity present between variables (O'Brien, 2007).

Table 4.1 The candidate independent variables used in the daytime and night-time OLS and GWR models of the ambient population.

Variable	Data source and year	Description
OpenCelliD	OpenCelliD, 2020	The number of cell towers and corresponding cells per workplace zone. The data are gathered using software which captures the GPS position of the cell towers that users are connected to. The data are available from OpenCelliD

		(https://opencellid.org/). The data used in this paper were downloaded on December 3rd, 2020. Expected correlation with the dependent variable: Positive
ATMs	OpenStreetMap, 2021	The number of ATMs per workplace zone according to OpenStreetMap (https://www.openstreetmap.org/). The data were filtered using the tag 'amenity=atm'. Expected correlation with the dependent variable: Positive
Higher and further education	OpenStreetMap, 2021	The number of college and university buildings per workplace zone according to OpenStreetMap. The data were filtered using the tags 'amenity=college' and 'amenity=university'. Expected correlation with the dependent variable: Positive
Hospitality	OpenStreetMap, 2021	The number of bars, cafes, restaurants, and pubs per workplace zone according to OpenStreetMap. The data were filtered using the following tags: 'amenity=bar', 'amenity=cafe', 'amenity=restaurant' and 'amenity=pub'. Expected correlation with the dependent variable: Positive
Retail	OpenStreetMap, 2021	The number of shops per workplace zone according to OpenStreetMap. The retail data were filtered using the tag 'building=retail'. Expected correlation with the dependent variable: Positive

Transport hubs	OpenStreetMap, 2021	The number of transport hubs per workplace zone according to OpenStreetMap. The number of bicycle parking facilities, bus stops, car parks, and train stations were downloaded using the following tags: 'amenity=bicycle_parking', 'highway=bus_stop', 'amenity=parking', and 'building=train_station'. Expected correlation with the dependent variable: Positive
Workday population	UK 2011 Census (Office for National Statistics, 2011b)	The workday population can be defined as the number of people in employment in an area, in addition to those who are residents in the area and are not in employment (Office for National Statistics, 2013a). These data are available at workplace zone level from the UK Census. Expected correlation with the dependent variable: Positive

4.3.3.1 Mobile phone cell tower density: OpenCelliD

OpenCelliD is a cumulative database of cell tower locations and their network coverage areas (known as cells). The data are collected in a crowd-sourced manner via a smartphone application that captures the location of the cell that a user is connected to. The size of a cell is dependent on environmental factors, such as typography, and the number of expected mobile phone users in the area. Telecommunication providers install cell towers where there are likely to be higher numbers of people; thus, mobile phone cell tower density data may be a good proxy of the ambient population. OpenCelliD data are considered to be indicative of the ambient population and were utilised as a proxy of the ambient population to produce crime rates in Vancouver, BC (Johnson et al., 2020). OpenCelliD is a global database, thus the data could be used to estimate the size of the ambient population in other

urban areas (OpenCelliD, 2018). The data utilised in this study were downloaded on the 3rd December 2020.

4.3.3.2 Points of Interest: OpenStreetMap

OpenStreetMap (OSM) is a collaborative mapping service and corresponding database. The data, which are geographic features, are available to download from the OSM database. Geographic features are represented as points (nodes), lines (ways) and polygons (relations) which are commonly tagged with attribute data. The tagging of features is encouraged to create a common basemap and to ensure that the data are of high quality, despite the activity being time intensive (Liu et al., 2011; Mooney and Corcoran, 2012). Despite tagging being encouraged, there are missing and incorrect tags; thus, it is probable that the true number of features in each workplace zone will be higher than stated. However, existing work suggests that OSM data tend to be more comprehensive in urban areas (Hagenauer and Helbich, 2012).

There are no formal standards for tagging features, but it is expected that users should use both a key, which describes a category, and a value, which is a feature (OpenStreetMap, 2020). The key and value should be separated by an equals sign, for example a shop should be tagged as 'amenity=shop'. The tags used to download data utilised in this study can be seen in Table 1. The OSM data were downloaded using the 'osmdata' R package (Padgham et al., 2017). Four of the variables used in this study were produced using OSM data, these are the number of ATMs, higher and further education buildings, retail premises and transportation hubs. To validate the OpenStreetMap data points, the geographic locations of the features of interest (ATMs, institutes of higher and further education, hospitality venues, retailers, and transport hubs) were cross-referenced with Google Maps. In instances where the feature was not present in both sources, it was removed from the downloaded OSM point data.

ATMs (automated teller machines) are generally located in areas which experience high levels of footfall to increase the number of transactions (Introna and Whittaker, 2007; Kisore and Koteswaraiah, 2017; Ashtikar et al., 2019). It would therefore be expected that higher numbers of ATMs are indicative of a higher ambient population in an area, thus were selected as a candidate predictor variable. Although beyond the scope of this paper, it will be interesting to verify whether ATMs continue to be a useful predictor of footfall following the increase in cashless payment options employed during COVID-19 restrictions.

Student populations can have a significant impact on the ambient population in many urban areas, with regards to both daytime and night-time activity. There are numerous ways in which the student population can be quantified; in this study the number of university and college buildings has been utilised. These data have been selected as they are indicative of where students are likely to be located, particularly during daytime hours, rather than providing information regarding their residential location. Estimates of the numbers of students in further and higher education were also included in a spatio-temporal model of the population for flood risk assessment in urban areas by Smith et al. (2016).

Footfall counts have been utilised in retail geography and planning as an indicator of potential spend (Genecon, 2011), in addition to being employed as a method of site selection for new retail offerings (Brown, 2006; Wood and Browne, 2007). Smith et al. (2016) used retail destination data, as retail areas attract large numbers of non-residential populations and shopping is a significant element within human activity patterns. Consequently, it would be expected that areas with a high number of retail premises also experience high levels of footfall and the number of retail venues has been selected for use as a dependent variable. The same trend would be expected with the hospitality sector, including bars, cafes, restaurants, and pubs. There are several ways in which retail and hospitality could be quantified; in this study the aggregated number of shops, bars, cafes, restaurants, and pubs in each

workplace zone has been used. Alternative methods include using feature density or areal coverage.

Transport hubs and transportation more generally have been highlighted as important attractors in urban environments in a number of studies (Echtner et al., 1993; Choi et al., 2007; Mazanec et al., 2007; Tang et al., 2009). Consequently, the aggregated counts of bicycle parking facilities, bus stops, car parks, and train stations per workplace zone were selected as candidate predictor variables.

4.3.3.3 Workday population: 2011 UK Census

Workplaces are a key element of human activity patterns, thus it is expected that the workday population and the ambient population are intrinsically linked (Martin et al., 2015; Berry et al., 2016). Estimates of the workday population are able to capture those in employment within an area, in addition to those are residents but are not in employment (Office for National Statistics, 2013a); thus, these data were selected as a candidate independent variable. In a study by Smith et al. (2016), the working age and retired populations were included in a model of the population; however, since this research was conducted, workday population data captured by the UK Census have been made available (Smith et al., 2016). As UK Census data are only collected decennially, the data used in this study are ten years old. This is a clear limitation of the dataset; however, it is the most geographically comprehensive estimate of the workday population currently available.

4.3.4 Geographically weighted regression

GWR is a form of local analysis which captures non-stationarity and allows spatial heterogeneity to be explored (Brunsdon et al., 1996; Fotheringham et al., 1997; Charlton and Fotheringham, 2002). In global regression models, such as an Ordinary Least Squares (OLS) regression, the relationship between y and x is assumed

to be unchanging across the study area. The equation for a simple linear regression is as follows,

$$y_i = \beta_0 + \sum_{k=1}^{m} \beta_k X_{ik} + e_i$$

where, for observations i=1..n, y_i is the dependent variable, X_{ik} is the value of the k[th] predictor variable, m is the number of independent variables, β_0 is the intercept term, β_k is the regression coefficient for the k[th] predictor variable, and e_i is the random error term.

However, spatial data are often not compatible for use with an OLS model as they are spatially autocorrelated and often represent spatial patterns that are challenging to explore using global models and statistics. Spatial autocorrelation is a measure of similarity between values that are close in space. If data are spatially autocorrelated, the strength of the relationship between variables will vary across space. GWR enables this spatial heterogeneity to be captured. GWR is similar to simple linear regression, but a GWR model is fitted at geographic coordinate locations. The GWR model is,

$$y_i = \beta_{i0} + \sum_{k=1}^{p} \beta_{ik} X_{ik} + \varepsilon_i$$

where, y_i is the dependent variable at location i, x is the value of the kth covariate at location i, β_{i0} is the intercept, β_{ik} is the regression coefficient of the kth covariate, p is the number of independent variables, and ε_i is the random error at location i.

GWR was selected for use in this study as it has not yet been used in the literature to produce small-area estimates of the population. Other methods of areal interpolation have been employed to produce estimates of populations, including dasymetric mapping (Mennis and Hultgren, 2006; Bhaduri et al., 2007; Sims et al., 2017), grid-based modelling (Martin, 1989; Martin, 1996; Martin et al., 2009b; Martin et al., 2015) and pycnophylactic interpolation (Tobler, 1979; Tobler et al., 1995; Tobler et al.,

1997). These methods commonly utilise land use data in conjunction with auxiliary data to produce small area estimates of the population. The primary advantage of areal interpolation methods is the ability to disaggregate coarse data to a finer spatial resolution to produce small-area estimates of the ambient population (Mennis and Hultgren, 2006). However, due to the increased volume and availability of spatio-temporally detailed data, the utility of areal interpolation methods has diminished. Consequently, this provided an opportunity to explore the use of GWR as a method of modelling the ambient population.

Daytime and night-time OLS and GWR models of the ambient population were produced with the seven candidate independent variables, summarised in Table 4.1. These models will be referred to as the 'full models'. In the final models, the variables that were not statistically significant in the full models were removed and the models were re-run. The R^2 values are used to compare the predictive capacity of the OLS and GWR models and the AIC values are used as a relative measure of the goodness of fit. The OLS models were fitted in R using the 'lm' function which is part of the built-in 'stats' package. The GWR models were produced in R using the 'spgwr' package (Bivand et al., 2020). In instances where the GWR models predicted a negative estimate of the ambient population, the minimum value was capped at 0. The values were capped to more accurately reflect the ambient population which cannot have a negative value in the real-world. The estimates of the GWR model were visualised using QGIS. All basemap data are copyrighted OpenStreetMap contributors and are available from https://www.openstreetmap.org.

4.4 Results

4.4.1 The daytime ambient population: Model results

In the full OLS and GWR models (Table 4.2), two of the seven variables have statistically significant relationships with the daytime ambient population. The number of ATMs was identified as the most significant predictor of the daytime

ambient population with a beta coefficient of 2.657 (p< 0.001). The number of hospitality venues is negatively associated with the daytime ambient population (beta coefficient= -1.759, p<0.001), which is inconsistent with the expected relationship based on the existing literature. This unexpected finding will be revisited in Section 6. The number of retail premises are positively associated with the daytime ambient population, but the relationship is not statistically significant. The variables OpenCelliD, hospitality, higher and further education, transport hubs, and the workday population were all negatively associated with the daytime ambient population, but the relationships were not significant.

Table 4.2 Full OLS and GWR models of the daytime ambient population.

	OLS model		GWR model		
Variable	Coefficient	Standard error	t-value	Minimum	Maximum
Intercept	1.046	6.488	1.612	9.968	1.145
OpenCelliD	-4.242	3.017	-1.406	-4.505	-4.261
ATMs	2.657 ***	3.373	7.877	2.651	2.727
Higher and further education	-6.100	3.330	-0.018	-1.032	-3.359
Hospitality	-1.759 **	6.134	-2.867	-1.829	-1.765
Retail	3.883	7.137	0.544	3.705	3.989
Transport hubs	-2.250	3.138	-0.717	-2.563	-2.124
Workday population	-8.892	1.779	-0.500	-9.847	-7.487
OLS diagnostics: Adjusted R^2: 0.301			GWR diagnostics: Adjusted R^2: 0.339		

AIC: 2760.190 Significance codes: 0 '***', 0.001 '**'	AIC: 2750.114

In the final OLS and GWR models (Table 3), the adjusted R^2 values for the full daytime OLS and GWR models are 0.312 and 0.332 respectively. The AIC value of the GWR model of the daytime ambient population is lower than that of the OLS (by 6.344), which along with the R^2 value indicates that the GWR model has increased predictive capacity. The GWR model accounts for around 33% of the variation in the daytime ambient population. In the final model, ATMs remain the most statistically significant independent variable with a beta coefficient of 2676.46 and a p-value less than 0.001. The hospitality variable in the final model has a beta coefficient of -160.11 and a p-value less than 0.001.

Table 4.3 Final OLS and GWR models of the daytime ambient population.

	OLS model		GWR model		
Variable	Coefficient	Standard error	t-value	Minimum	Maximum
Intercept	361.53	410.38	0.881	315.11	575.95
ATMs	2676.46 ***	330.26	8.104	2647.48	2927.77
Hospitality	-160.11 **	59.75	-2.680	-184.53	-158.94
OLS diagnostics: Adjusted R^2: 0.312 AIC: 2753.222 Significance codes: 0 '***', 0.001 '**'			GWR diagnostics: Adjusted R^2: 0.332 AIC: 2746.878		

The estimates of the daytime ambient population can be seen in Figure 4.2. Estimates are highest in the primary retail area of the study area, and in the central areas which feature high numbers of attractors, such as bars and restaurants. In the

financial district, estimates are significantly lower than in the retail areas. Despite high numbers of workers in this area, it is unlikely to experience fluctuations in the ambient population due to visitors or tourists. The spatial distribution of the estimates appears to be expected across the study area and the estimates will be validated in Section 5.

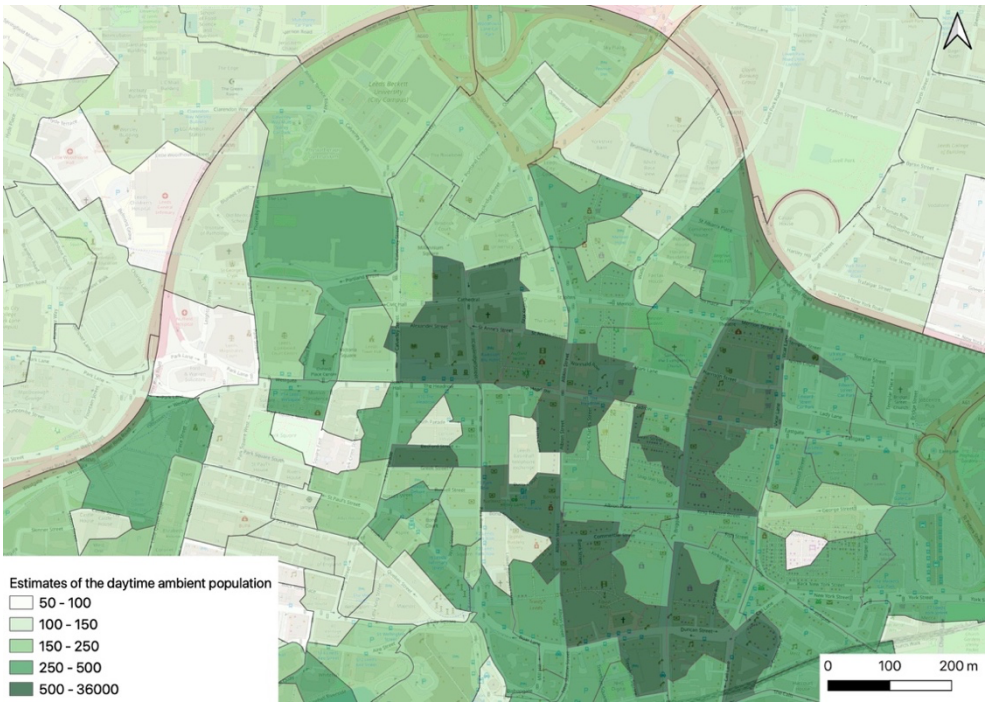


Figure 4.2 The spatial distribution of the estimates of the daytime ambient population.

4.4.2 The night-time ambient population: Model results

In the full OLS and GWR models (Table 4.4), two of the seven variables have a statistically significant relationship with the night-time ambient population. As in the models of the daytime ambient population, the ATM variable was identified as the most significant predictor of the night-time ambient population, with a beta coefficient of 285.897 and a p-value of less than 0.01. The hospitality variable is negatively associated with the night-time ambient population with beta coefficient of -21.000 and is statistically significant ($p<0.001$). The variables OpenCelliD, higher and further education, transport hubs, and the workday population all have a negative association with the night-time ambient population, but these relationships are not statistically significant. The retail variable has a positive association with the night-

time ambient population, which is to be expected based on the literature; however, the relationship is not statistically significant.

Table 4.4 Full OLS and GWR models of the night-time ambient population.

	OLS model		GWR model		
Variable	Coefficient	Standard error	t-value	Minimum	Maximum
Intercept	178.692	104.840	1.704	170.722	194.024
OpenCelliD	-6.334	4.875	-1.299	-6.726	-6.324
ATMs	285.897 ***	54.512	5.245	283.987	292.044
Higher and further education	-0.304	5.380	-0.057	-0.388	-0.258
Hospitality	-21.000 *	9.911	-2.119	-21.885	-20.975
Retail	1.993	11.532	0.173	1.653	2.191
Transport hubs	-2.877	5.071	-0.567	-3.159	-2.721
Workday population	-0.001	0.002	-0.463	-0.001	-2.721
OLS diagnostics: Adjusted R^2: 0.166 AIC: 2234.796 Significance codes: 0 '***', 0.01 '*'			GWR diagnostics: Adjusted R^2: 0.193 AIC: 2232.612		

In the final OLS and GWR models (Table 4.5), the adjusted R^2 values for the full daytime OLS and GWR models are 0.166 and 0.182 respectively. The AIC value of the GWR model of the night-time population is 2229.427, which is 5.369 lower than that

of the OLS model. The AIC and R^2 values indicate that the GWR model has increased predictive capacity. The GWR model is able to account for around 19% of the variation in the night-time ambient population. In the final model, the ATM variables is the most statistically significant variable, with a beta coefficient of 291.026 and a p-value of less than 0.01. The hospitality variable has a beta coefficient of -19.035 and a p-value of less than 0.05.

Table 4.5 Final OLS and GWR models of the night-time ambient population.

	OLS model		GWR model		
Variable	Coefficient	Standard error	t-value	Minimum	Maximum
Intercept	72.530	66.131	1.097	66.324	93.016
ATMs	291.026 ***	53.220	5.468	287.011	305.496
Hospitality	-19.035 .	9.628	-1.977	-20.704	-18.766
OLS diagnostics: Adjusted R2: 0.166 AIC: 2234.796 Significance codes: 0 '***', 0.05 '.'			GWR diagnostics: Adjusted R2: 0.182 AIC: 2229.427		

The spatial distribution of the estimates of the night-time ambient population can be seen in Figure 4.3. Estimates of the night-time ambient population are highest in the central and eastern areas of the study area where there are high numbers of attractors. Estimates of the ambient population are lowest in workplace zones in the West and North-West of the study area, which primarily contain workplaces and low numbers of attractors. Thus, the spatial distribution of the night-time population is reflective of the expected patterns.

There are clear differences between the estimates of the daytime and the night-time ambient population. The daytime ambient population is much higher across the whole of the study area, particularly in the retail areas of Leeds city centre. The night-time population is much lower across the study, with higher levels of the ambient population in areas with high concentrations of features associated with the night-time economy, such as the arena in the North East of the study area.

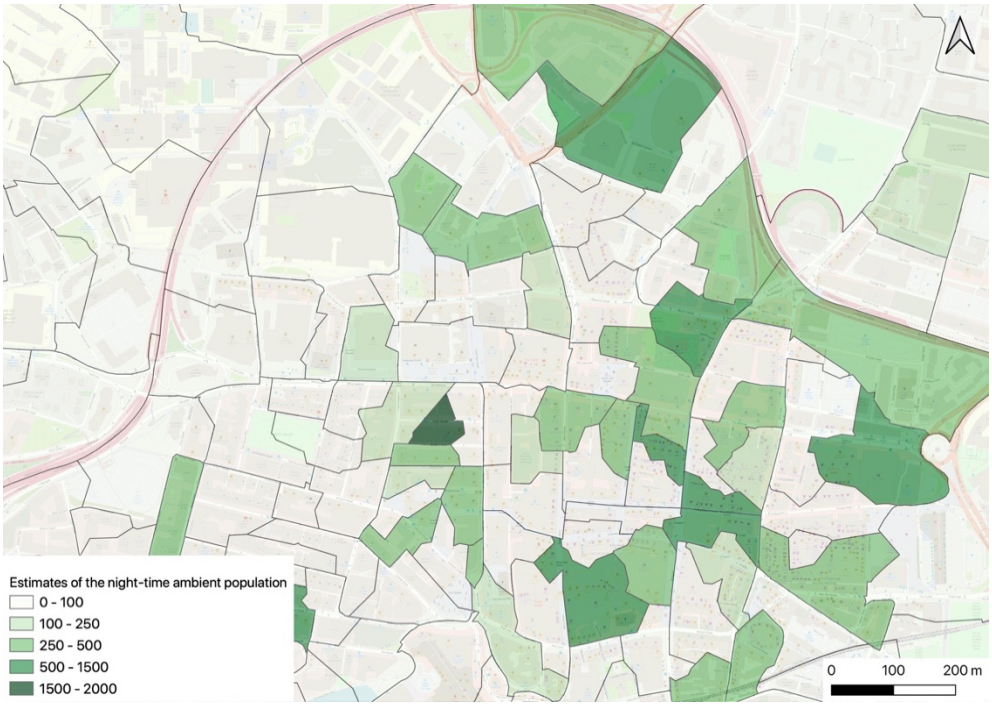


Figure 4.3 The spatial distribution of the estimates of the night-time ambient population

4.4.3 Model testing

The model was tested in two other locations within the metropolitan borough of Leeds: Headingley and Wetherby, using the coefficients of the final daytime model of Leeds city centre. These case study locations were selected to assess the validity of the model in areas of a similar geographic size with different demographic characteristics, hence different human activity patterns. Age was the demographic selected for use as it is expected to impact the type of activities people engage in and the frequency at which these activities occur. Headingley is a suburb of Leeds, 3km Northwest of Leeds city centre and a popular student area, with 78% of the usual

residents aged between 18 and 30 (Office for National Statistics, 2011b). Wetherby is a market town 20km Northwest of Leeds city centre. The residential population of Wetherby is 10% older than the national average, with 30.6% of households over the age of 65 (Office for National Statistics, 2011b).

Estimates of the daytime ambient populations in Headingley and Wetherby were produced using the coefficient determined in the final daytime model. The same dependent variables were utilised as those used in the final model of Leeds city centre (the number of ATMs and the number of hospitality venues). However, the lack of ATMs in the centre of the town of Wetherby may be a limitation of the variables selected for use in the model. This issue may be relevant in other locations which are less urban and, therefore, have fewer ATMs. Consequently, alternative variables may need to be identified and utilised when producing estimates of the ambient population in smaller towns and less urbanised areas.

Estimates of the daytime ambient population in Headingley, which can be seen in Figure 4.4, shows an expected spatial distribution. Estimates of the ambient population are highest in the workplace zones in proximity to the train station, sports stadium, university student accommodation, and along the primary high street. Estimates are lowest in workplace zones in which building density is low, thus the spatial distribution of the estimates is rational.

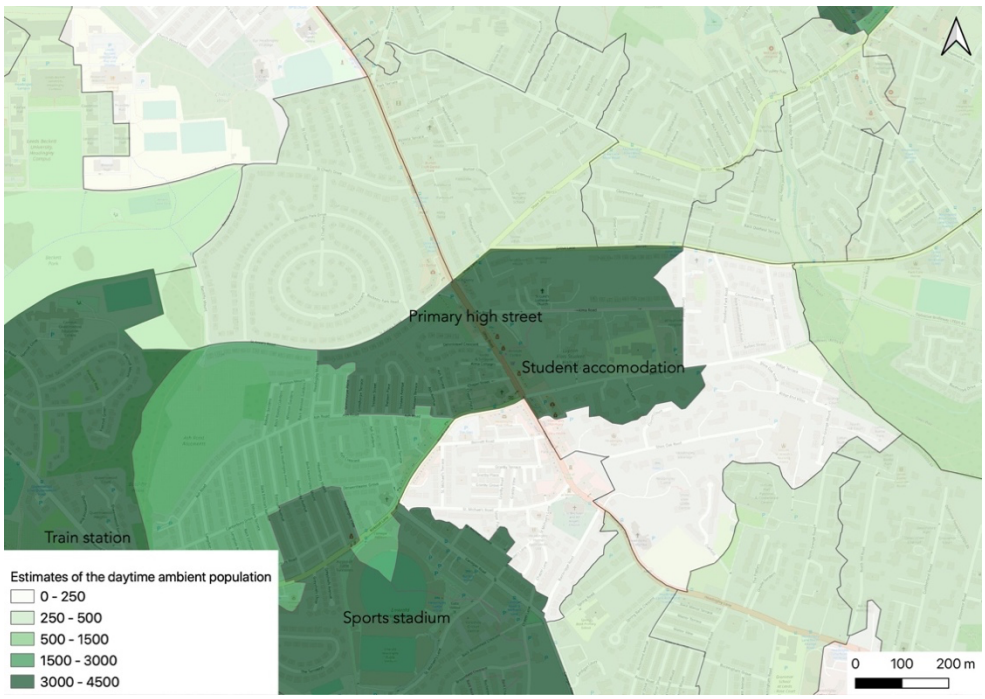


Figure 4.4 Estimates of the daytime ambient population in Headingley.

Estimates of the daytime ambient population in Wetherby, shown in Figure 4.5, are highest in the workplace zones near the centre of town and along part of the main high street, where a major supermarket and a high density of shops and attractors are located. Estimates of the daytime ambient population are capped at zero in workplace zones towards the southeast of the study area where there are fewer attractors, and the area is primarily residential. However, the workplace zone in the centre of the study area has a high number of attractors, primarily shops and cafes, but the model estimates in this area are capped at 0; thus, do not fit the expected spatial distribution of the ambient population. This reinforces the need to develop a more robust understanding of the potential indicators of the size of the ambient population.

Figure 4.5 Estimates of the daytime ambient population in Wetherby.

4.5 Validation

Due to the limited number of locations at which footfall counts are collected and the consequent impact on assessing the accuracy of estimates produced by the GWR models, the results were validated using manual counts. As there is a lack of openly available data quantifying the size of the ambient population, it was necessary to collect manual pedestrian counts. There are three aims of the validation process. Firstly, to compare the manual counts collected by the three different data collectors at each location to determine whether the manual counts can themselves be trusted (Section 4.5.1). Secondly, to assess the accuracy of the counts produced by the footfall cameras used in this study (Section 4.5.2) and lastly to evaluate the estimates produced by the GWR models (Section 4.5.3). Manual pedestrian counts allow observations to be taken at specific geographical locations over a limited period of time and are a relatively cost-effective way of gathering data (Bauer et al., 2011). Three data collectors were located at each of the ten sites chosen as validation locations. Three data collectors were stationed at each site to enable a mean number of counts to be taken for each time period, at each location. The ten sites were selected to capture locations which typically experience significant fluctuations in the

ambient population due to features such as prominent retail areas, areas close to a large university, and those with high concentrations of offices. Six of the ten sites were selected as footfall cameras are installed in these locations and taking measurements at these locations allowed the footfall camera counts to be validated using manual counts.

In sections 4.5.1 and 4.5.2, the two-sample Kolmogorov-Smirnov test is utilised to assess whether two independent samples come from the same distribution (Smirnov, 1948; Kolmogorov, 1992; Berger and Zhou, 2014). The null hypothesis is that the two samples come from the same distribution. If the Kolmogorov-Smirnov statistic is high and the p-value is below 0.05, the null hypothesis can be rejected. If the Kolmogorov-Smirnov statistic is low and the p-value is above 0.05, the null hypothesis cannot be rejected.

4.5.1 Validation of manual counts

The two-sample Kolmogorov-Smirnov test is used to assess whether the samples of manual counts recorded by the data collectors are likely to come from the same distribution. The test is conducted for eight locations and the results of the tests can be seen in Table 6. For seven of the eight locations, in each of the three tests the null hypothesis could not be rejected; thus, the three samples likely come from the same distribution. For the manual counts collected at North Street (Wetherby), the Kolmogorov-Smirnov statistic and p-value indicate that comparing the samples taken by data collectors 1 and 2, and 2 and 3, the null hypothesis, that the counts from each pair of data collectors are sampled from the same distribution, can be rejected. It is unclear why the samples differ in distribution at this location; possible reasons include counting errors or mistakes. Therefore, the results of the Kolmogorov-Smirnov statistic indicates that at seven of the eight locations the data collectors likely captured similar samples of the population; thus, are suitable for use in the validation of the footfall camera counts.

Table 4.6 The results of Kolmogorov-Smirnov tests on manual count samples at eight locations.

Count location	Data collectors	Kolmogorov Smirnov statistic	p-value	Null hypothesis
Otley Road/B617, Headingley	1 and 2	0.333	0.930	Fail to reject
	1 and 3	0.333	0.930	Fail to reject
	2 and 3	0.166	0.999	Fail to reject
High Street/A661, Wetherby	1 and 2	0.333	0.930	Fail to reject
	1 and 3	0.333	0.930	Fail to reject
	2 and 3	0.333	0.930	Fail to reject
North Street, Wetherby	1 and 2	0.833	0.025	Reject
	1 and 3	0.666	0.142	Fail to reject
	2 and 3	0.833	0.025	Reject
Otley Road/Wood Lane, Headingley	1 and 2	0.166	0.999	Fail to reject
	1 and 3	0.333	0.930	Fail to reject
	2 and 3	0.333	0.930	Fail to reject
The Headrow, Leeds city centre	1 and 2	0.166	0.999	Fail to reject
	1 and 3	0.333	0.930	Fail to reject
	2 and 3	0.333	0.930	Fail to reject
Calverley Street, Leeds city centre	1 and 2	0.285	0.871	Fail to reject
	1 and 3	0.214	0.990	Fail to reject
	2 and 3	0.333	0.930	Fail to reject

Commercial Street, Leeds city centre	1 and 2 1 and 3 2 and 3	0.500 0.333 0.500	0.474 0.940 0.474	Fail to reject Fail to reject Fail to reject
Briggate, Leeds city centre	1 and 2 1 and 3 2 and 3	0.333 0.333 0.333	0.930 0.930 0.930	Fail to reject Fail to reject Fail to reject

4.5.2 Validation of footfall camera counts

There is limited information available regarding the accuracy of pedestrian counting devices such as footfall cameras, especially those installed in outdoor environments, as most devices were developed for indoor environments, such as shopping centres (Greene-Roesel et al., 2008). There are several physical factors that may impact the accuracy of the data, including the weather, lighting and occlusion (when individuals are not visually isolated, leading to under-enumeration) (Greene-Roesel et al., 2008; Lindsey, 2015). Data from three footfall cameras, located at Briggate, Commercial Street and The Headrow (locations 2, 3 and 5 in Figure 4.6) in Leeds city centre are validated using manual counts. Both the footfall camera counts and manual counts were recorded on the 8th and 9th July 2021 between the hours of 10:00 and 16:00 and the Kolmogorov-Smirnov test is utilised to determine whether the counts are likely to have come from the same distribution.

The results of the Kolmogorov-Smirnov test (Table 4.7) indicate that for data sampled at Commercial Street and The Headrow Square, the null hypothesis cannot be rejected; thus, the manual counts and the footfall camera counts likely come from the same distribution. However, at Briggate it is unlikely that the two samples came from the same distribution as the high Kolmogorov-Smirnov statistic and low p-value indicate that the null hypothesis can be rejected. This is supported by the percentage differences between the hourly counts from the average manual counts and the

footfall camera counts (Table 4.8). The average hourly percentage difference between the footfall camera counts and the manual counts at Briggate is 55.20%, which is substantially higher than the values for Commercial Street (1.32%) and The Headrow (16.39%). At Briggate a counting discrepancy may have occurred, such as the manual counts capturing a different geographical area to the footfall camera. It is not always clear precisely which part of the street the footfall cameras are covering, and this is exacerbated at Briggate as the camera is located at a busy intersection between multiple roads. However, as the counts at Commercial Street and The Headrow are similar, we are confident that the cameras are recording sufficiently accurate numbers and the failure of the Briggate camera to match the manual counts is the result of a discrepancy in where, precisely, the population is being recorded.

Table 4.7 Results of the Kolmogorov-Smirnov test on the average manual counts and the footfall camera counts at three locations.

Count location	Kolmogorov-Smirnov statistic	p-value	Null hypothesis
Commercial Street	0.333	0.931	Fail to reject
The Headrow	0.333	0.780	Fail to reject
Briggate	1	0.01	Reject

Table 4.8 A summary of the footfall camera counts and the manual counts.

Footfall camera location	Time	Footfall camera count	Mean manual count
	10:00	1512	610
	11:00	2120	941
	12:00	2551	1140

Briggate	13:00	2204	1134
	14:00	2598	1138
	15:00	2501	1103
Commercial Street	10:00	1106	1115
	11:00	1145	1539
	12:00	2197	1962
	13:00	2214	1673
	14:00	1900	1838
	15:00	1726	1644
The Headrow	10:00	275	445
	11:00	511	548
	12:00	632	858
	13:00	774	907
	14:00	800	728
	15:00	849	725

4.5.3 Validation of the models of the ambient population

In this section, footfall camera counts spanning the six-hour period between 10:00 and 16:00 were utilised to compare the estimates of the model with the manual counts. The final daytime model, outlined in Section 4.4, was re-run using six-hour estimates of footfall. This allows direct comparisons between the two samples recorded at six locations in Leeds city centre which can be seen in Figure 4.6, in addition to two sites in Headingley (Figure 4.7) and two sites in Wetherby (Figure 4.8).

At all six locations in Leeds city centre the model under-predicts the ambient population; however, given the R^2 of the model is 0.332, differences between the model estimates and the mean manual counts are to be expected. The mean manual

counts at the six locations within Leeds city centre and the model estimates for the surrounding workplace zone can be seen in Table 4.9, in addition to the model estimate as a percentage of the mean manual count. At locations two and four the mean manual counts, and the model estimates are relatively similar, while at locations three, five, and six the model estimates as a percentage of the mean manual counts range between 20% and 37% which is expected given the R^2. Location one is likely to be an outlier as there is a significant difference between the mean manual counts and the model estimates, this is likely to be due to the location having few attractors but being in proximity to a university campus and a large hospital; this will result in large numbers of people passing through the area which are not captured by the model. While the 'workday population' and the 'higher and further education' variables were not statistically significant, alternative measures of the groups represented could be explored in future work.

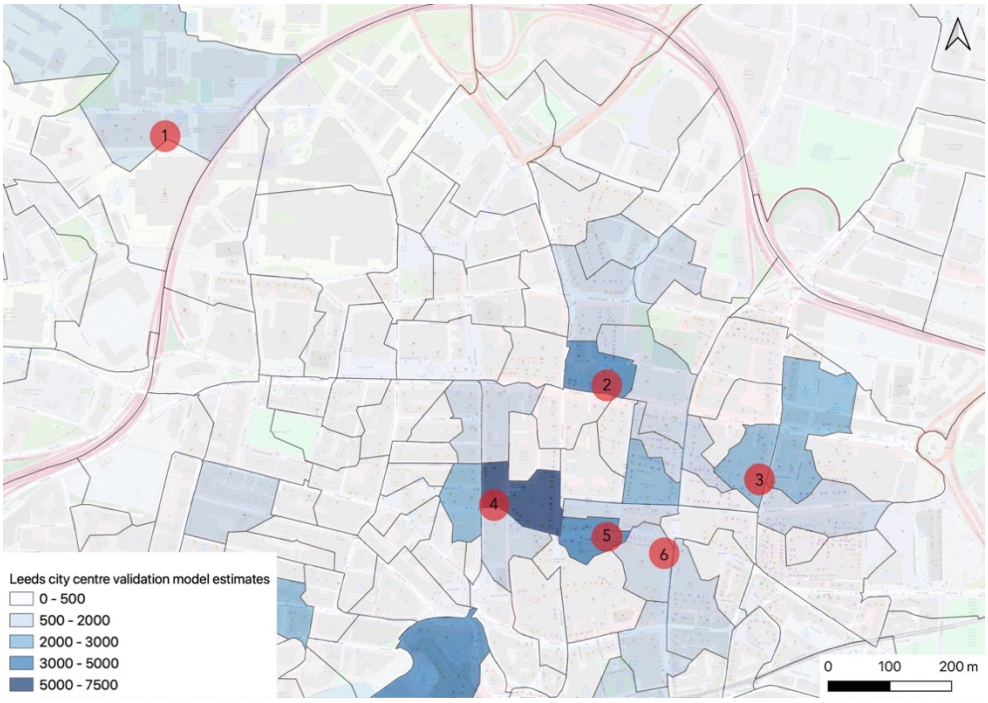


Figure 4.6 Average estimates of the size of the ambient population between the hours of 10:00 and 16:00 in Leeds city centre.

Table 4.9 The mean manual counts at six locations within Leeds city centre and the model estimates for the surrounding workplace zone

Manual count location	Mean manual count	Model estimate	The model estimate as a percentage of the mean manual count
1	2136	110	5
2	4203	3369	80
3	3252	1223	37
4	8251	6641	80
5	7834	2298	29
6	6066	1232	20

In the validation model for the study location Headingley, at location one (Figure 4.7) the average manual count is 1708 while the model estimate is 1825. At location two (Figure 4.7) the average manual count was 697, while the model estimate within the workplace zone is 1078. At both locations the mean manual counts are similar in value to the model estimates, the counts would not be expected to be identical as the model estimate represents the mean counts for a one-month period. The spatial distribution of the ambient population is expected, with higher counts along the main high street and in locations with high numbers of attractors. There are three workplace zones within the study area which have model estimates capped at 0, this is likely to be due to these areas being primarily residential and consequently lacking ATMs and hospitality venues. However, the model was developed to predict the ambient population in urban areas, such as Leeds city centre, which are likely to have high numbers of ATMs and hospitality venues.

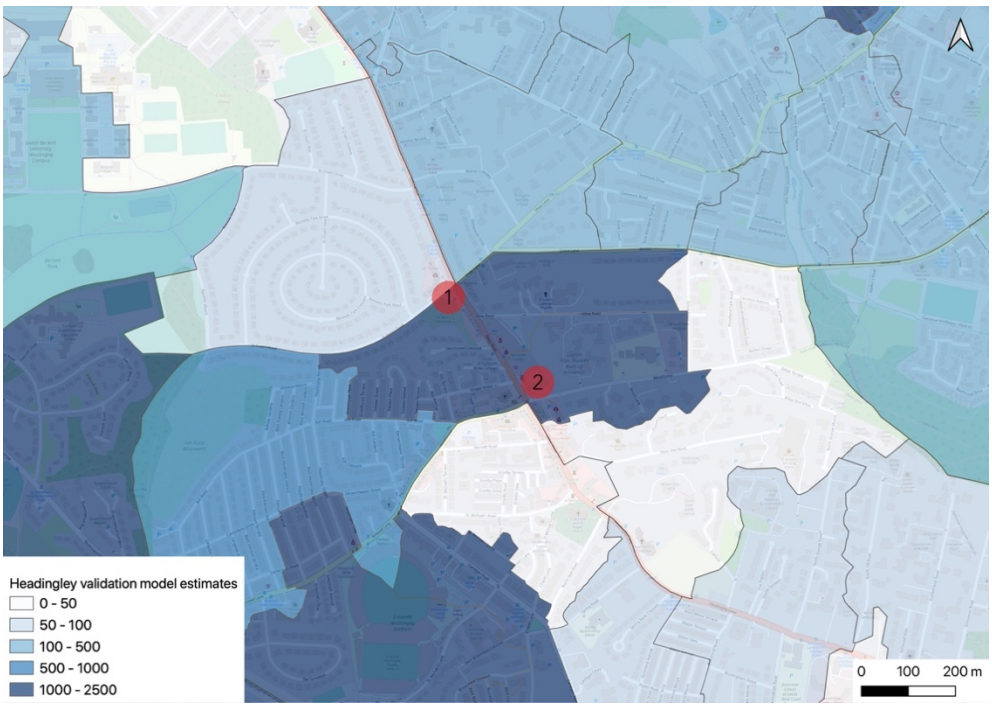


Figure 4.7 Average estimates of the size of the ambient population between the hours of 10:00 and 16:00 in Headingley.

In the validation model for the study location of Wetherby, the manual count location one (Figure 4.8), the average manual counts are 739, while the model estimate across the six-hour period between 10:00 and 16:00 was 854. At location two (see Figure 4.8) in Wetherby, the average manual count is 865, while model estimate was capped at 0 for this location. The model estimate at location two does not accurately represent the ambient population of the area, this is likely due to the workplace zone covering a geographically small area which does not feature any ATMs. This issue may limit the accuracy of the model for producing estimates of the ambient population in small-geographical areas.

Figure 4.8 Average estimates of the size of the ambient population between the hours of 10:00 and 16:00 in Wetherby.

4.6 Discussion

The aim of this work is to estimate the size of the ambient population in an urban area using both traditional and novel datasets. In estimating both the daytime and the night-time ambient populations, the GWR models have a higher predictive capacity than the OLS models, thus account for a larger proportion of the variation in the ambient population. This is to be expected as the ambient population is unevenly distributed across space and GWR can account for this spatial heterogeneity (Oshan et al., 2019; Ćwiakowski, 2020). The number of retail premises and ATMs are positively associated with the daytime and night-time ambient populations, but the relationship is only statistically significant for the ATM variable. The variables OpenCelliD, hospitality, higher and further education, transport hubs, and the workday population were all negatively associated with both the daytime and night-time ambient populations; this contradicts information from the existing literature, but the relationships were not statistically significant (Echtner et al., 1993; Choi et al., 2007; Mazanec et al., 2007; Tang et al., 2009; Smith et al., 2016). The results of this study suggest that those factors identified in the literature as having a relationship

with the ambient population need to be explored in more detail. This is increasingly important post the COVID-19 pandemic as there may be shifts in the ways individuals utilise public space, particularly within urban areas, which will impact both the size of the ambient population and the spatio-temporal fluctuations within it (Sharifi and Khavarian-Garmsir, 2020; Dubois and Dimanche, 2021; Ramani and Bloom, 2021; Florida et al., 2021). The exploration of the relationships between physical factors and the ambient population in other urban areas would be a valuable contribution to the literature, as these relationships may be place specific. The night-time model has a lower R^2 than the daytime model which is to be expected as fluctuations in the night-time ambient population are likely to be related to factors which were not included in the model, such as cultural and social events (Hanaoka, 2018). This highlights an opportunity to investigate factors which influence the ambient population during the night-time, as this has received little focus within the existing literature.

The validation process demonstrated that in Leeds city centre at five of the six locations, the model produced expected values given the R^2. The model estimates at location one in the city centre appear to be an outlier; this may be due to the large numbers of people passing through the area to access the nearby university campus and large hospital. This highlights an opportunity for future research to explore the relationship between pedestrian flow and the ambient population (Trasberg et al., 2021). At both locations in Headingley and at location one in Wetherby, the model estimates were similar to the manual footfall counts. At location two in Wetherby the model estimates were significantly lower than the manual footfall counts; however, this is likely due to a lack of ATMs in the workplace zone as it covers a relatively small geographical area. This highlighted the limitations of employing the model to produce estimates of the ambient population in locations which comprise of a small geographical area.

A limitation of the study is that the footfall camera data were only captured at six locations within Leeds city centre and validation data were only captured at ten sites across the three locations: Leeds city centre, Headingley, and Wetherby. This

highlights the importance of the equitable distribution of footfall cameras across geographical areas to ensure that different types of locations are captured and to ensure the data are valuable (Shelton et al., 2015; Hoffmann, 2019; Robinson and Franklin, 2020). The selection of the manual count locations aimed to capture locations which would be expected to experience a relatively large ambient population, i.e., primary retail locations in each of the areas. The study may benefit from additional data collected from a larger number of sites across all three study areas; however, this was not within the scope of the project. Additional data could provide insight regarding the size of the ambient population in different weather conditions, during weekends, and may allow the detection of further unexpected spatial distributions of the estimates of the ambient population.

4.7 Conclusion

This paper has estimated the size of the ambient population in an urban area, using novel sources of data and a method of statistical modelling, thus contributing to the existing literature. Models of the ambient population were developed using GWR and OpenStreetMap data capturing indicators of the size of the ambient population, the numbers of ATMs and hospitality venues. Interestingly, the numbers of higher and further education centres, retail hubs, the density of cell towers, the number of transport hubs, and estimates of the workday population did not have statistically significant relationships with footfall camera counts. The validation process assessed the similarity of the samples collected by the three data collectors, quantified the accuracy of the footfall cameras, and allowed model estimates to be compared to manual counts at ten sites using a novel, empirical dataset. This work has potential to provide insight regarding the size and spatial distribution of the ambient population in urban areas and consequently can inform studies such as those exploring exposure to air pollution, crime risk, and public safety.

Reference list

Andresen, M.A. 2011. The ambient population and crime analysis. The Professional Geographer. 63(2), pp.193–212.

Andresen, M.A. and Jenion, G.W. 2010. Ambient populations and the calculation of crime rates and risk. Security Journal. 23(2), pp.114–133.

Ashtikar, O., Kendurkar, C., Basangar, A. and Marachakkanavar, M. 2019. Smart Automated Teller Machine Applications Using Artificial Intelligence. SSRN Electronic Journal.

Bauer, D., Ray, M. and Seer, S. 2011. Simple Sensors Used for Measuring Service Times and Counting Pedestrians Strengths and Weaknesses. Transportation Research Record: Journal of the Transportation Research. 2214, pp.77–84.

Berry, T., Newing, A., Davies, D. and Branch, K. 2016. Using workplace population statistics to understand retail store performance. International Review of Retail, Distribution and Consumer Research. 26(4), pp.375–395.

Bhaduri, B., Bright, E., Coleman, P. and Urban, M.L. 2007. LandScan USA: a high-resolution geospatial and temporal modeling approach for population distribution and dynamics. GeoJournal. 69(1–2), pp.103–117.

Bivand, R., Yu, D., Nakaya, T. and Garcia-Lopez, M.A. 2020. 'spgwr' package.

Brown, S. 2006. International Review of Retail, Distribution and Consumer Research Retail location theory: evolution and evaluation. International Review of Retail, Distribution and Consumer Research. 3(2), pp.185–299.

Brunsdon, C., Fotheringham, A.S. and Charlton, M.E. 1996. Geographically weighted regression: a method for exploring spatial nonstationarity. Geographical Analysis.

Cahill, M. and Mulligan, G. 2007. Using geographically weighted regression to explore local crime patterns. Social Science Computer Review.

Cardozo, O.D., García-Palomares, J.C. and Gutiérrez, J. 2012. Application of geographically weighted regression to the direct forecasting of transit ridership at station-level. Applied Geography.

Charlton, M. and A Stewart Fotheringham 2002. Introduction to Geographically Weighted Regression. Science Foundation Ireland under the National Development Plan.

Chen, K. and McAneney, J. 2006. High-resolution estimates of Australia's coastal population. Geophysical Research Letters.

Chen, Q., Mei, K., Dahlgren, R.A., Wang, T., Gong, J. and Zhang, M. 2016. Impacts of land use and population density on seasonal surface water quality using a modified geographically weighted regression. Science of the Total Environment.

Chiou, Y.C., Jou, R.C. and Yang, C.H. 2015. Factors affecting public transportation usage rate: Geographically weighted regression. Transportation Research Part A: Policy and Practice.

Choi, S., Lehto, X.Y. and Morrison, A.M. 2007. Destination image representation on the web: Content analysis of Macau travel related websites. Tourism Management. 28(1).

Chu, H.J., Yang, C.H. and Chou, C.C. 2019. Adaptive non-negative geographically weighted regression for population density estimation based on nighttime light. ISPRS International Journal of Geo-Information.

Comber, A. and Zeng, W. 2019. Spatial interpolation using areal features: A review of methods and opportunities using new forms of data with coded illustrations. Geography Compass. 13(10).

Comber, A.J., Brunsdon, C. and Radburn, R. 2011. A spatial analysis of variations in health access: Linking geography, socio-economic status and access perceptions. International Journal of Health Geographics.

Crols, T. and Malleson, N. 2019. Quantifying the ambient population using hourly population footfall data and an agent-based model of daily mobility. Geoinformatica. 23(2), pp.201–220.

Echtner, C.M., Ritchie, J.R.B. and Ritchie, A.B. 1993. The Measurement of Destination Image: An Empirical Assessment.

Fotheringham, A.S., Charlton, M. and Brunsdon, C. 1997. Two techniques for exploring non-stationarity in geographical data. Geographical Systems.

Genecon 2011. Understanding High Street Performance [Online]. [Accessed 18 May 2021]. Available from: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/31823/11-1402-understanding-high-street-performance.pdf.

Giglio, S., Bertacchini, F., Bilotta, E. and Pantano, P. 2019. Using social media to identify tourism attractiveness in six Italian cities. Tourism management. 72, pp.306–312.

Greene-Roesel, R., Diogenes, M.C., Ragland, D.R. and Lindau, L. a 2008. Effectiveness of a Commercially Available Automated Pedestrian Counting Device in Urban Environments: Comparison with Manual Counts. TRB 2008 Annual Meeting. c.

Hagenauer, J. and Helbich, M. 2012. Mining urban land-use patterns from volunteered geographic information by means of genetic algorithms and artificial neural networks. International Journal of Geographical Information Science. 26(6), pp.963–982.

House of Commons 2014. Who works where? What the 2011 Census tells us about co-workers and commutes. [Accessed 16 March 2021]. Available from: https://commonslibrary.parliament.uk/.

Introna, L.D. and Whittaker, L. 2007. The Information Society Power, Cash, and Convenience: Translations in the Political Site of the ATM.

Johnson, P., Andresen, M.A. and Malleson, N. 2020. Cell Towers and the Ambient Population: a Spatial Analysis of Disaggregated Property Crime. European Journal on Criminal Policy and Research.

Kauhl, B., Schweikart, J., Krafft, T., Keste, A. and Moskwyn, M. 2016. Do the risk factors for type 2 diabetes mellitus vary by location? A spatial analysis of health insurance claims in Northeastern Germany using kernel density estimation and geographically weighted regression. International Journal of Health Geographics.

Kisore, N. and Koteswaraiah, C.H. 2017. Improving ATM coverage area using density based clustering algorithm and voronoi diagrams. Information Sciences. 376, pp.1–20.

Kontokosta, C.E. and Johnson, N. 2017. Urban phenology: Toward a real-time census of the city using Wi-Fi data. Computers, Environment and Urban Systems. 64, pp.144–153.

Kounadi, O., Ristea, A., Leitner, M. and Langford, C. 2018. Population at risk: using areal interpolation and Twitter messages to create population models for burglaries and robberies. Cartography and Geographic Information Science. 45(3), pp.205–220.

Li, X., Zhou, Y., Gong, P., Seto, K.C. and Clinton, N. 2020. Developing a method to estimate building height from Sentinel-1 data.

Lindsey, G. 2015. The Minnesota Bicycle and Pedestrian Counting Initiative: Implementation Study [Online]. [Accessed 30 March 2021]. Available from: http://www.lrrb.org/pdf/201534.pdf.

Liu, D., Wang, M., Hua, X.S. and Zhang, H.J. 2011. Semi-automatic tagging of photo albums via exemplar selection and tag inference. IEEE Transactions on Multimedia. 13(1), pp.82–91.

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. Annals of the Association of American Geographers. 105(3), pp.512–530.

Lo, C.P. 2008. Population estimation using geographically weighted regression. GIScience and Remote Sensing.

Løvholt, F., Glimsdal, S., Harbitz, C.B., Zamora, N., Nadim, F., Peduzzi, P., Dao, H. and Smebye, H. 2012. Tsunami hazard and exposure on the global scale. Earth-Science Reviews.

Maldonado-Guzmán, D.J. 2020. Airbnb and crime in Barcelona (Spain): testing the relationship using a geographically weighted regression. Annals of GIS., pp.1–14.

Martin, D. 1996. An assessment of surface and zonal models of population. International Journal of Geographical Information Systems. 10(8).

Martin, D. 1989. Mapping population data from zone centroid locations. Transactions - Institute of British Geographers. 14(1).

Martin, D., Cockings, S. and Leung, S. 2015. Developing a Flexible Framework for Spatiotemporal Population Modeling. Annals of the Association of American Geographers.

Martin, D., Cockings, S. and Leung, S. 2009. Population 24/7: building time-specific population grid models.

Mazanec, J.A., Wöber, K. and Zins, A.H. 2007. Tourism destination competitiveness: From definition to explanation? Journal of Travel Research. 46(1).

Mburu, L.W. and Helbich, M. 2016. Crime risk estimation with a commuter-harmonized ambient population. Annals of the American Association of Geographers. 106(4), pp.804–818.

Mennis, J. and Hultgren, T. 2006. Cartography and Geographic Information Science Intelligent Dasymetric Mapping and Its Application to Areal Interpolation Intelligent Dasymetric Mapping and Its Application to Areal Interpolation. Cartography and Geographic Information Science. 33(3), pp.179–194.

Mooney, P. and Corcoran, P. 2012. The Annotation Process in OpenStreetMap. Transactions in GIS. 16(4), pp.561–579.

Munira, S. and Sener, I.N. 2020. A geographically weighted regression model to examine the spatial variation of the socioeconomic and land-use factors associated with Strava bike activity in Austin, Texas. Journal of Transport Geography.

Office for National Statistics 2013. 2011 Census: The workday population of England and Wales - An alternative 2011 Census output base. Office for National Statistics. [Online]. Available from: https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/articles/theworkdaypopulationofenglandandwales/2013-10-31.

Office for National Statistics 2011. 2011 Census - Office for National Statistics. local authorities in the United Kingdom.

Office for National Statistics 2014. Workplace Zones: A new geography for workplace statistics.

OpenCelliD 2018. What is OpenCelliD? [Accessed 17 February 2021]. Available from: http://wiki.opencellid.org/wiki/What_is_OpenCellID.

OpenStreetMap 2020. Tags - OpenStreetMap Wiki. [Accessed 6 May 2021]. Available from: https://wiki.openstreetmap.org/wiki/Tags.

Padgham, M., Lovelace, R., Salmon, M. and Rudis, B. 2017. osmdata. The Journal of Open Source Software. 2(14), p.305.

Pirdavani, A., Bellemans, T., Brijs, T. and Wets, G. 2014. Application of geographically weighted regression technique in spatial analysis of fatal and injury crashes. Journal of Transportation Engineering.

Ratti, C., Frenchman, D., Pulselli, R.M. and Williams, S. 2006. Mobile landscapes: using location data from cell phones for urban analysis. Environment and planning B: Planning and design. 33(5), pp.727–748.

Reades, J., Calabrese, F., Sevtsuk, A. and Ratti, C. 2007. Cellular census: Explorations in urban data collection. IEEE Pervasive computing. 6(3), pp.30–38.

Robinson, C. and Franklin, R.S. 2020. The sensor desert quandary: What does it mean (not) to count in the smart city? Transactions of the Institute of British Geographers.

Robinson, D.P., Lloyd, C.D. and Mckinley, J.M. 2013. Increasing the accuracy of nitrogen dioxide (NO 2) pollution mapping using geographically weighted regression (GWR) and geostatistics. International Journal of Applied Earth Observation and Geoinformation. 21, pp.374–383.

Roni, R. and Jia, P. 2020. An optimal population modeling approach using geographically weighted regression based on high-resolution remote sensing data: A case study in Dhaka City, Bangladesh. Remote Sensing.

Roy, K.C., Cebrian, M. and Hasan, S. 2019. Quantifying human mobility resilience to extreme events using geo-located social media data. EPJ Data Science. 8(1), p.18.

Schmitt, R.C. 1956. Estimating Daytime Populations. Journal of the American Institute of Planners. 22(2), pp.83–85.

Selby, B. and Kockelman, K.M. 2013. Spatial prediction of traffic levels in unmeasured locations: Applications of universal kriging and geographically weighted regression. Journal of Transport Geography.

Sims, K.M., Weber, E.M., Bhaduri, B.L., Thakur, G.S. and Resseguie, D.R. 2017. Application of social media data to high-resolution mapping of a special event population In: Advances in Geographic Information Science.

Smith, A., Martin, D. and Cockings, S. 2016. Spatio-Temporal Population Modelling for Enhanced Assessment of Urban Exposure to Flood Risk. Applied Spatial Analysis and Policy.

Smith, G., Arnot, C., Fairburn, J. and Walker, G. 2005. A National Population Data Base for Major Accident Hazard Modelling.

Stein, R., Conley, J.F. and Davis, C. 2016. The differential impact of physical disorder and collective efficacy: a geographically weighted regression on violent crime. GeoJournal.

Tang, L., Choi, S., Morrison, A.M. and Lehto, X.Y. 2009. The many faces of Macau: A correspondence analysis of the images communicated by online tourism information sources in English and Chinese. Journal of Vacation Marketing. 15(1).

Terada, M., Nagata, T. and Kobayashi, M. 2013. Population estimation technology for mobile spatial statistics. NTT DOCOMO Techn. J. 14, pp.10–15.

Tobler, W., Deichmann, U. and Gottsegen, J. 1995. UC Santa Barbara NCGIA Technical Reports Title The Global Demography Project (95-6) Publication Date [Online]. [Accessed 16 March 2021]. Available from: https://escholarship.org/uc/item/0kt69058.

Tobler, W., Deichmann, U., Gottsegen, J. and Maloy, K. 1997. World population in a grid of spherical quadrilaterals. International Journal of Population Geography. 3(3).

Tobler, W.R. 1979. Smooth pycnophylactic interpolation for geographical regions. Journal of the American Statistical Association. 74(367).

Tu, J. and Xia, Z.G. 2008. Examining spatially varying relationships between land use and water quality using geographically weighted regression I: Model design and evaluation. Science of the Total Environment.

Wang, Y., Kockelman, K.M. and Wang, X. 2011. Anticipation of land use change through use of geographically weighted regression models for discrete response. Transportation Research Record.

Whipp, A., Malleson, N., Ward, J. and Heppenstall, A. 2021. Estimates of the Ambient Population: Assessing the Utility of Conventional and Novel Data Sources. ISPRS International Journal of Geo-Information. 10(3), p.131.

Wood, S. and Browne, S. 2007. Convenience store location planning and forecasting — a practical research agenda. International Journal of Retail & Distribution Management. 35(4), pp.233–255.

Xu, Y.H., Pennington-Gray, L. and Kim, J. 2019. The Sharing Economy: A Geographically Weighted Regression Approach to Examine Crime and the Shared Lodging Sector. Journal of Travel Research.

Yang, H., Lu, X., Cherry, C., Liu, X. and Li, Y. 2017. Spatial variations in active mode trip volume at intersections: a local analysis utilizing geographically weighted regression. Journal of Transport Geography.

Yang, T.C. and Matthews, S.A. 2012. Understanding the non-stationary associations between distrust of the health care system, health conditions, and self-rated health in the elderly: A geographically weighted regression approach. Health and Place.

Chapter 5

Alternative Measures of the Population at Risk and their Impact on the Spatial Distribution of Crime

The work in Chapter 5 has been accepted with revisions by PlosOne:

Whipp, A., Malleson, N., Ward, J. and Heppenstall, A. Alternative measures of the population at risk and their impact on the spatial distribution of crime.

The aim of this chapter is to investigate the impact of different measures of the population at risk (the residential, workday, and ambient populations) on the spatial distribution of crime rates for two crime types: 'theft from the person' and 'violence and sexual offences'. The work addresses Research Objective 5, as set out in Chapter 1 of the thesis. This work utilises the approach developed in Chapter 4 of the thesis to produce estimates of the size of the ambient population.

Abstract

Traditionally, crime rates are calculated using a measure of the resident population. However, as crimes are not only committed against residents of an area, but also against temporary populations, such as workers and visitors, the resident population may not reliably represent the population at risk. This study explores the impact that three measures of the population at risk have on the spatial distribution of crime rates for two crime types: 'theft from the person' and 'violence and sexual offences'. The rates of both crime types are calculated using measures of the resident, workday, and ambient populations and are then explored using correlation analysis, global and local indicators of spatial autocorrelation, and hot spot analysis. The results of the study evidence that for rates of 'theft from the person' and rates of 'violence and sexual offences', the use of the both the resident and workday populations overestimate the risk of victimisation within urban centres and underestimate the risk

in residential areas. The findings of the study highlight the value of estimates of the ambient population for producing accurate crime rates and support the demand for geographically comprehensive estimates that can be utilised by police forces and policymakers.

5.1 Introduction

Crime rates are the most common measurement of crime and are considered to be the most meaningful statistic employed within crime studies (Boggs, 1965). Crime rates communicate the risk of an individual becoming a victim of a specific crime type. They are a valuable tool used to inform resource allocation, influence planning and policymaking by police forces and local governments, and to convey messages regarding safety to members of the public (National Academy of Sciences, 2016). Crime rates are calculated by dividing the number of recorded crimes by the size of the population at risk, most commonly the resident population, within a geographic area (Andresen and Jenion, 2010). However, due to the significant fluctuations in population size which occur due to human activity patterns, measures of the resident population do not effectively communicate the risk associated with crime types which target individuals (Boggs, 1965; Harries, 1981; Andresen and Jenion, 2010). The existing literature lacks conclusive evidence regarding the measure of the population at risk that should be employed in the calculation of crime rates instead of the resident population (Cohen et al., 1985; Andresen and Jenion, 2010).

As alternative sources of non-resident population data have emerged, it is now possible to produce alternative measures of the population at risk for use in the calculation of crime rates. This research investigates the impact of different measures of the population at risk on the spatial distribution of crime rates for two crime types. The three measures of the population used are the resident population, the workday population, and the ambient population. In this study the ambient population is defined as “the number of people within a given geographical area at a specific point in time, excluding individuals at their place of residence and those utilising modes of transport” (Whipp et al., 2021). The two crime types ‘theft from the person’ and

'violence and sexual offences'. were chosen because the victim is not certain to be located at a residence (as would be the case with crime types such as burglary) and data for recorded incidents are readily available in the UK. Four methods were used to investigate the relationship between measures of the population at risk and the numbers of crimes across the study area and analyse the spatial distribution of the crime rates. A correlation analysis, measures of both local and global spatial autocorrelation, and hot spot analysis are employed.

The results of this study demonstrate that the spatial distributions of rates of theft from the person and rates of violence and sexual offences vary considerably when alternative measures of the population at risk are utilised. Interestingly, the study finds that there are larger, statistically significant variations in the spatial distributions of the crime rates between the use of the workday and the ambient populations (i.e., the number of individuals present in an area for work and/or leisure purposes), than between the resident and workday populations for both crime types. The results of this study consolidate findings from the existing literature in that they support the use of a non-residential measure of the population at risk. Crucially, the findings support the use of the ambient population as the measure of the population at risk within crime rates, due to their ability to enumerate both the work and non-work-related populations.

5.2 Background

The importance of using appropriate denominators for the calculation of crime rates has been explored within the existing literature, although it was first highlighted in work by Boggs (1965). The work noted that changes in human activities throughout a given timeframe produce fluctuations in the number of targets available to offenders, for example, the numbers of people or vehicles present. This led Boggs (1965) to hypothesise that differences in opportunities for crimes to occur should be reflected by the denominators of crime rates. The work demonstrated that the use of alternative measures of the population at risk had significant impacts on the rates of

car theft, non-residential daytime burglary, non-residential night-time burglary, and grand larceny (theft over the value of 1000 US dollars) (Boggs, 1965).

Traditional sources of population data, such as national censuses and travel surveys, have previously been employed to improve understanding of the spatial distribution of crime rates (Stults and Hasbrouck, 2015; Felson and Boivin, 2015; Malleson and Andresen, 2016). Felson and Boivin (2015) employed transport survey data to test criminological theory and found that the use of the numbers of daily visitors may be more influential than the size of the resident population in the spatial distribution of crimes. Similarly, Stults and Hasbrouck (2015) used travel survey data to explore the impact of the commuting population on the spatial distribution of crimes and found that the risk of crime in cities is over-estimated when the resident population is used as a measure of the population at risk. Malleson and Andresen (2016) identified the workday population, an output of the 2011 Census of England and Wales, as an appropriate measure of the population at risk. However, the authors noted that the data fail to capture non-work-related fluctuations in the size of the population. Additionally, Malleson and Andresen (2016) only explored the impact of using an alternative measure of the population at risk on the spatial distribution of one type of crime, theft from the person. LandScan data, which are average estimates of the population calculated using data from national censuses and remotely sensed images and provides global coverage, have been used as a measure of the population at risk in several studies. Andresen (2007) used LandScan data to test criminological theory, while work by Andresen and Jenion (2008) employed LandScan data to assess the use of the ambient population at different levels of crime prevention. Later work by Andresen and Jenion (2010) quantified the value of the ambient populations in the calculation of rates of violent crime, using LandScan data. However, LandScan data are limited in their utility, as the data are produced using estimates of the resident population. While studies that employ traditional sources of data acknowledge the potential value of using alternative measures of the population at risk, further investigation using non-residential measures to explore a range of crime types in different locations is required (Andresen and Jenion, 2008; Andresen and Jenion,

2010; Felson and Boivin, 2015; He et al., 2020). The workday population in particular is highlighted in the literature as a potentially valuable source of estimates of the population at risk and should be explored further (Malleson and Andresen, 2016). As a result, the most recent, available estimates of the size of the workday population from the 2011 Census of England and Wales were selected for use in this study.

More recently, geo-located data from the social media platform Twitter have been employed as a proxy for the ambient population in several studies (Malleson and Andresen, 2015a; Malleson and Andresen, 2015b; Ristea et al., 2018; Tucker et al., 2021). Twitter data were used by Malleson and Andresen (2015a) to identify spatio-temporal clusters of robbery and theft from the person events. They have also been used in conjunction with hot spot analysis to examine the emergence of hot spots of violent crimes (Malleson and Andresen, 2015b) and with geographically weighted regression (GWR) to forecast hotspots of street crime (Ristea et al., 2018). Tucker et al. (2021) employed Twitter data to investigate the relationship between ambient, resident, commuter, and tourist populations on the rates of public violence and private conflict across the city of Boston (US) using a machine learning clustering algorithm. The findings from these studies all supported the use of the ambient population as a potential alternative measure of the population at risk. However, the sparsity of Twitter data within residential areas was identified as a limitation (Tucker et al., 2021). Additionally, in 2019 Twitter announced that users would no longer be able to share Tweets with a precise geographical location, i.e., geographic coordinates, limiting their use as a measure of population at risk for small areas in any future studies (Benton, 2019). However, techniques such as geoparsing can be used to derive geographic information from Tweets that are not geotagged. Geoparsing can be utilised to convert free text descriptions of a location (toponyms) into geographic locations in the form of coordinates. Geoparsing can be conducted through a range of applications, such as the Python library Mordecai (Halterman, 2017; Gritta et al., 2020).

Call data records from mobile phones that provide the locations of mobile phone users, have also been used as a proxy of the ambient population and employed in studies of the spatial distributions of crime (Hanaoka, 2018; He et al., 2020; Jung et al., 2020). These data have been used to produce hourly estimates of the size of the ambient population in order to quantify the relationship between these estimates and the number of snatch and run offences in Osaka (Japan) (Hanaoka, 2018). The relationship between the size of the ambient population and the spatial variation of larceny-theft in Xi'an (China) was investigated by He et al. (2020). In this study, data which enumerated all mobile phone users in the area were employed; however, as mobile phone data are expensive to acquire and are often unavailable at a fine spatial scale due to data privacy restrictions, the enumeration of all users is not possible in most countries. In another recent study, Jung et al. (2020) compared the relationship between assault density and the ambient and resident populations using a generalised linear model. Similarly to the studies that employed geo-located Twitter data, all the aforementioned studies that employed call data records support the use of alternative measures of the population at risk in the calculation of crime rates; however, data access poses a significant barrier to its use. As a result, in this study only which is openly available is utilised. This allows the work to be easily reproducible, which is critical if alternative measures of the population at risk are to be routinely employed by local governments and police forces for policymaking and resource allocation.

The need for an appropriate measure of the population at risk is also supported by two of the most prominent theories within environmental criminology: routine activity theory and crime pattern theory. Routine activity theory states that for a crime to occur, three elements must come together in space and time: a target, an offender, and the absence of a capable guardian (Cohen and Felson, 1979). According to this theory, the ambient population, due to the hourly, daily, and seasonal fluctuations in its size, influences where and when these elements will converge and consequently when and where crimes are committed. Crime pattern theory focuses on the ways in which the environmental backcloth, i.e., the physical

environment, impacts crime. Brantingham et al. (1981) suggest that the locations in which crimes are committed are not selected randomly, but are concentrated around nodes, paths and edges that the perpetrator is familiar with. Brantingham et al., (1995) later introduced the idea of crime generators (areas in which offenders and victims converge, e.g., offices and train stations) and crime attractors (areas known for opportunities to commit crimes, e.g., shopping centres and car parks) to crime pattern theory. As urban centres typically have higher concentrations of both crime attractors and crime generators, higher crime rates would be expected in these types of location. This suggests that the use of a non-residential measure of the population at risk would impact the spatial distribution of crime rates within urban centres in particular, as these areas attract high numbers of people. The changes in the spatial distribution of crime rates within urban centres and cities within West Yorkshire are explored in this study.

To summarise, existing studies have begun to explore the impact of different measures of the population at risk on the spatial distribution of different crime types. This study addresses this gap within the literature as it explores differences between the spatial distributions of crime rates calculated using both residential and non-residential measures of the population at risk, in addition to exploring the differences between two alternative measures of the non-residential population. Given the importance of accurate crime rates, for uses such as the communication of risk and the implementation of effective crime prevention strategies, further research into alternative measures of the population at risk is required.

5.3 Data and methodology

5.3.1 Study area and geography

The chosen study area is the county of West Yorkshire, UK, which covers an area of approximately 2000 km^2 (see Figure 5.1). West Yorkshire has a resident population of 2,226,058 people (Office for National Statistics, 2011) and significant

central business districts in the cities of Leeds, Bradford, and Wakefield which attract large numbers of people for both work and leisure purposes. Central business districts are typically associated with high levels of crime as they generally attract large numbers of people and contain high numbers of crime generators and attractors (Brantingham and Brantingham, 1995). In 2021, West Yorkshire had the second highest total crime rate of all police force areas in England and Wales (Office for National Statistics, 2021). Consequently, West Yorkshire is a suitable study area for this work and was selected for use in this study.

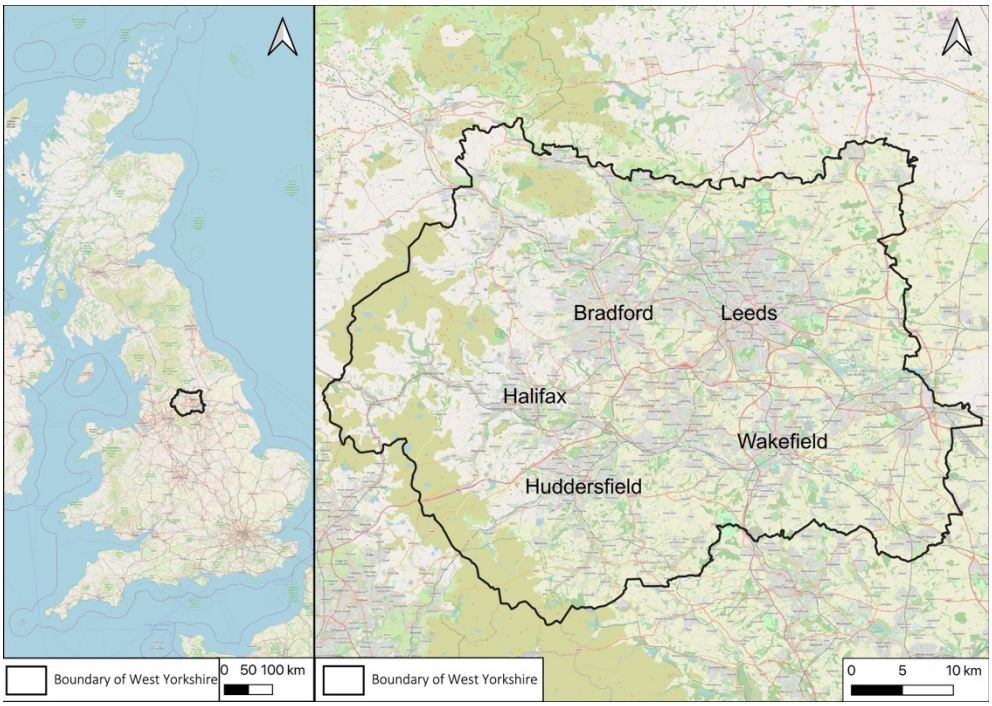


Figure 5.1 The study area of West Yorkshire, with the cities of Bradford, Leeds, and Wakefield labelled. Two other large towns, Halifax and Huddersfield, are also labelled. The inset map demonstrates the position of West Yorkshire within the UK.

In this study, data are aggregated to Lower Super Output Areas (LSOA) level. The LSOA boundaries were produced in 2011 and were downloaded from the UK Data Service (borders.ukdataservice.ac.uk). Within the study area there are 1389 LSOAs and, on average, each LSOA contains 1500 residents or 650 households (Office for National Statistics, 2016). LSOAs have been selected for use due to their compatibility

with police recorded crime data and are the level of geography in a number of UK-based studies of crime (Tompson et al., 2015; Malleson and Andresen, 2016).

5.3.2 Data

This study employs crime data, enumerating recorded incidences of theft from the person and violence and sexual offences and three measures of the population: the resident population, the workday population, and the ambient population (i.e., the number of individuals present in an area for work and/or leisure purposes). The crime data and the three measures of the population are used to produce crime rates for each LSOA, which are calculated using Equation 5.1.

Equation 5.1 The equation for calculating the rate of crime per 1000 people, per LSOA.

$$\text{Crime rate per 1000 people in each LSOA} = \left(\frac{\text{Number of crime events}}{\text{Size of the population at risk per LSOA}}\right) \times 1000$$

This section provides a detailed description of each of the data sources used in the study. Descriptive statistics for the data utilised to produce the rates of both crime types can be found in the Appendix A Table 1, while the summary statistics for the crime rates explored in this study and the variable names used in the remainder of the paper are presented in the Appendix A Table 2.

5.3.2.1 Crime data

The crime data used in this study enumerate the number of police recorded incidences of theft from the person and violence and sexual offences per LSOA in West Yorkshire between 1st January and 31st December 2019. The police recorded crime data were downloaded from data.police.uk (https://data.police.uk/) at LSOA level. Visualisations of the counts of the two crime types are available in the

Appendix A Figure 1 and Appendix A Figure 2, and visualisations of the rates are in the Appendix A Figures 6-8. 'Theft from the person' is defined as property stolen while "being held or carried by the victim" (Office for National Statistics, 2017a). Crimes of 'violence' include harassment, abuse, wounding, and homicide and 'sexual offences' include rape, sexual assault, and indecent exposure (Office for National Statistics, 2017b). While violent crimes and sexual offences are defined separately in raw police records, it is not possible to distinguish between the crime types within the publicly available data. Recorded crimes that did not have an assigned location cannot be used and were removed prior to analysis (8.6% of theft from the person incidents and 3.7% of violence and sexual offences incidents). Incidences of violence and sexual offences were selected for use in this study as it was the most commonly committed crime type in West Yorkshire in 2019 and can significantly impact both survivors and their families. Due to their severity, it is imperative to reduce both the numbers of these offences, and their impact, through effectively directed policies. Theft from the person was selected as incidences of this crime, by definition, occur outside of the home; thus, it would be expected that the use of a non-residential measure of the size of the population would be appropriate for this crime type. In contrast, incidences of violence and sexual offences are likely to occur in both residential and non-residential areas. Thus, it would be expected that the impacts of the measure of the population at risk will vary between the two crime types. Therefore, the use of theft from the person and violence and sexual offences allows the impact of alternative measures of the size of the population at risk to be explored.

The geographic coordinates associated with each crime event in the police.uk data are an approximation of where a crime actually occurred (College of Policing, 2021). In order to ensure anonymity, however, each data point must be located in the centre of a street, over a public place, or over a commercial premise and the area around each point must contain either more than eight postal addresses or none (College of Policing, 2021). This anonymisation of the crime data produces spatial inaccuracies as the spatial points do not reflect the exact location at which a crime was committed. However, as the crime data used in this study are aggregated to LSOA

level, the location accuracy will have a minimal impact on the analysis (Tompson et al., 2015).

The crime data used in this study only enumerate recorded crimes, i.e., those reported to the police, thus the crime rates produced are not fully representative of the true number of offences committed. According to the 2016 Crime Survey of England and Wales (CSEW), which attempts to enumerate all crimes committed and not only those which are reported to the police, only 52% of violent crimes were reported (Office for National Statistics, 2017b). The CSEW does not attempt to quantify the under-reporting of sexual offences to the police due to low levels of reporting of these crimes to both police forces and in the CSEW (Office for National Statistics, 2017b). Evidence from the CSEW suggests that incidences of theft from the person are under-reported by between 40% and 50% (Office for National Statistics, 2017a). Consequently, for both theft from the person and violence and sexual offences the numbers of crimes recorded will be substantially lower than the true number of offences committed. Thus, it should be noted that the crime rates produced and analysed in this study are only representative of recorded crimes and may not accurately reflect true crime rates.

5.3.2.2 Estimates of the resident population

Crime rates are commonly calculated using the resident population as a measure of the population at risk. In this study, estimates of the usual resident population from the 2011 Census of England and Wales are employed as the measure of the resident population. The Office for National Statistics, the producers of census data for England and Wales, define the usual resident population as "anyone, who on census day, was in the UK and had stayed or intended to stay in the UK for a period of 12 months or more, or had a permanent UK address and was outside the UK and intended to be outside the UK for less than 12 months" (Nomis, 2013). The data are open access and were downloaded at LSOA level from the UK Data Service (infuse2011gf.ukdataservice.ac.uk/). As census data for England and Wales are

collected decennially, and the results from the 2021 census have not yet been published, these data are the most recent estimates of the resident population available.

5.3.2.3 Estimates of the workday population

As employment can result in significant changes in the number of people in an area, when compared to the usual resident population, estimates of the workday population may be valuable as a measure of the population at risk. Workday population estimates for England and Wales were first captured in the 2011 census. The Office for National Statistics define the workday population as "where the usually resident population is re-distributed to their places of work, while those not in work are recorded at their usual residence" (Office for National Statistics, 2013). These estimates represent the workday population on Census Day (27th March) 2011. As with estimates of the resident population, the data are open access and were downloaded at LSOA level from the UK Data Service (infuse2011gf.ukdataservice.ac.uk/).

5.3.2.4 Estimates of the ambient population

Within this study, the ambient population is defined as "the number of people within a given geographical area at a specific point in time, excluding individuals at their place of residence and those utilising modes of transport" (Whipp et al., 2021). As estimates of the size of the ambient population are currently not part of any standard suite of population statistics, this study employs estimates produced using an approach developed by (Whipp, Malleson, et al., 2021).The estimates of footfall counts are produced using indicators of footfall, in conjunction with a method of statistical modelling developed in (Whipp, Malleson, et al., 2021). Two independent variables are utilised; the numbers of ATMs and the number of hospitality venues, which includes bars, cafes, restaurants, and pubs. The data for both variables was downloaded from OpenStreetMap and aggregated to LSOA level. These two variables

are then used in conjunction with geographically weighted regression to produce estimates of the number of footfall counts per LSOA (the dependent variable). These estimates are then added to the estimates of the workday population which are defined in Section 5.3.2.3. The amalgamation of the footfall counts and the estimates of the workday population aims to ensure that the estimates of the ambient population capture the number of people within an LSOA for both work and non-work-related purposes. The model currently accounts for 33.2% of the variation of the dependent variable, i.e., the footfall camera counts; however, the predictive capacity of the model could be improved by developing a more in depth understanding of the drivers of the size of the ambient population in urban centres. Despite this limitation, this measure of the ambient population offers a significant advantage as a measure of the population at risk, when compared to the resident and ambient populations, as it captures, to some extent, both work and non-work-related populations. Visualisations of the spatial distribution of the ambient, resident, and workday populations in West Yorkshire are available in the Appendix A Figure 3-5.

5.3.3 Methodology

The aim of this study is to investigate the impact of different measures of the population at risk (the denominator in Equation One) on the spatial distribution of crime rates. To fulfil this aim, four empirical methods are employed: descriptive global analysis (correlation), a global measure of spatial autocorrelation (Moran's I), a local measure of spatial autocorrelation (Local Indicators of Spatial Analysis (LISA)) and hotspot analysis (Getis Ord GI*).

5.3.3.1 Descriptive global analysis - Correlation

Correlation analysis is used to test the relationship between two variables and has been used in a number of crime studies to assess the relationships between crimes rates and different measures of the population (Malleson and Andresen, 2016; Hanaoka, 2018; Hipp et al., 2019; He et al., 2020; Tucker et al., 2021). In this study,

correlation analysis is used to investigate the relationship between crime rates calculated using three different measures of the population at risk (the resident, workday, and ambient populations). The analysis is conducted using the '.corr' function within the 'pandas' package in Python. The Spearman's rank correlation coefficient is specified as the method as all crime rates have a negatively skewed distribution.

5.3.3.2 Global measure of spatial autocorrelation - Moran's I

Moran's I is an inferential statistic which measures spatial autocorrelation and has been employed in existing studies investigating the spatial distribution of crime (Andresen, 2011; Kadar et al., 2017; Kounadi et al., 2018; Lan et al., 2019). The statistic evaluates whether the pattern produced is clustered, dispersed, or random in nature based on the value and location of a feature. The significance of the statistic can be evaluated using the p-value and z-score. The values produced by the tool range between 1 and -1. A Moran's I value which is higher than the Expected I Moran's I indicate a positive correlation, while values lower than the Expected Moran's I indicate negative spatial autocorrelation. A value of 0 suggests a random pattern with no correlation. The analyses are conducted using ArcGIS Pro 2.8.0 using the 'Spatial Autocorrelation' tool.

5.3.3.3 Local measure of spatial autocorrelation – LISA

The LISA statistic is utilised to indicate the spatial clustering of low or high values (the crime rate per LSOA) within a dataset and to highlight outliers. Work by Andresen (2011) demonstrated the value of the LISA statistic to analyse the spatial distribution of violent crimes in Vancouver, Canada. The sum of all LISAs for a dataset is equivalent to a global indicator of spatial autocorrelation, such as Moran's I (Anselin, 1995). Thus, using the LISA statistic provides a more in depth understanding of spatial autocorrelation as each area of analysis receives its own measure of spatial autocorrelation. Each area is then compared to its neighbouring area, in this instance LSOAs. A positive LISA suggests that the feature has similar values to the neighbouring features; thus, the data are clustered. The clusters can be groups of high value

features or low value features. A negative LISA suggests that the feature is a spatial outlier and is dissimilar to the neighbouring features. The spatial outliers can be low value features surrounded by high value neighbours, or high value features surrounded by low value neighbours. A summary of the cluster and outlier definitions can be seen in Table 5.1. The analyses are conducted using ArcGIS Pro 2.8.0 using the 'Cluster and Outlier Analysis' tool. As spatial statistics require a feature to have a minimum of one neighbour for the analysis to be reliable, an appropriate distance band must be selected. Within ArcGIS Pro 2.8.0 the 'Calculate Distance Band from Neighbour Count' tool was utilised to evaluate the minimum, average, and maximum distances for a specified number of neighbours (a minimum of eight neighbours recommended).The recommended distance band, which determines the number of neighbours each feature has, was calculated to be 4.415km. The visualisations of the clusters and outliers for crime rates calculated using the resident population are not included due to the large volume of results generated. However, the visualisations can be found in the Supplementary Information (Appendix A Figure 12 and Appendix A Figure 13).

Table 5.1 Definitions of cluster and outlier types which are identified using the LISA statistic.

Cluster/outlier name	Description
Low-Low cluster	A low value in a low-value neighbourhood
High-High cluster	A high value in a low-value neighbourhood
Low-High outlier	A low value in a high-value neighbourhood
High-Low outlier	A high value in a low-value neighbourhood

5.3.3.4 Hot spot analysis - Getis Ord GI*

Hot spot analysis enables the detection of statistically significant clusters of high or low values within a dataset, referred to as hot spots and cold spots respectively. Hot spot analysis is a commonly used technique within crime studies and has been used to inform resource allocation, facilitate problem-solving, and measure and analyse crime patterns (Eck et al., 2005; Chainey et al., 2008; Chainey and Ratcliffe, 2013; Malleson and Andresen, 2016). The use of hot spot analysis adds further insight, as it identifies where the sum values within a neighbourhood (number of neighbouring areas) is high or low relative to the global average, while the local measure of spatial autocorrelation measures the degree to which the value of an area is similar to the values of neighbouring areas.

There are a number of statistical methods available to calculate hotspots (Chainey et al., 2008); in this study the Getis Ord GI* statistic was selected as it is an indicator of local, rather than global, spatial autocorrelation (Getis and Ord, 1992; Ord and Getis, 1995). A Getis Ord GI* statistic is produced for each feature (the crime rate per LSOA) within the dataset, and each feature is analysed within the context of the neighbouring features. A z-score (a GI* statistic) and a p-value are generated for each feature. A hot spot is determined by a high, positive z-score and a small p-value, while a low, negative z-score and small p-value determine a cold spot. As spatial statistics require a feature to have a minimum of one neighbour for the analysis to be reliable, an appropriate distance band must be selected. Within ArcGIS Pro 2.8.0 the 'Calculate Distance Band from Neighbour Count' tool was utilised to evaluate the minimum, average, and maximum distances for a specified number of neighbours (a minimum of eight neighbours recommended). The size of the distance band is 4.415km and was measured using Euclidean distance. The analyses are conducted using ArcGIS Pro 2.8.0 using the 'Hot Spot Analysis' tool. The visualisations of the hot spot analysis for crime rates calculated using the resident population are not included due to the large volume of results generated but are available in the Supplementary Information (S14 and S15 Figs).

5.4 Results

5.4.1 Spatial distribution of the populations and the numbers of crime events

The results of this study show that there is little variation in the spatial distribution of the usual resident population across West Yorkshire, which ranges between 1011 and 4156 people. This pattern is expected as census boundaries are designed to distribute the population evenly across a geographic area. The spatial distribution of the workday population is more varied than that of the usual resident population. Estimates of the workday population are high (between 20,000 and 32,261 people) in LSOAs in proximity to Leeds, Wakefield, and Bradford and the Northeast of the study area. High estimates of the workday population are expected in large towns and cities due to people travelling to these areas for work. With regards to the spatial distribution of the ambient population, there are pockets of high estimates in and around the cities of Bradford, Leeds, and Wakefield, in addition to the towns of Halifax and Huddersfield. The fact that there are notable differences in the size and the spatial distributions of the three measures of the population illustrate the impact these different measures will have on crime rates.

The spatial distributions of the counts of theft from the person events and violence and sexual offences are low (ranging between 0 and 300 offences) across most of the study area. However, there are distinct pockets of high counts of theft from the person in the city of Leeds. While for violence and sexual offences, areas in both Leeds and Halifax contain high numbers of crimes, ranging between 900 and 1781 events in 2019.

5.4.2 Spatial distribution of the rates of 'theft from the person'

Rates of theft from the person calculated using the resident population are generally low across West Yorkshire, with small pockets of higher rates of these crimes in Leeds, Bradford, and Halifax. This spatial distribution occurs because although the resident populations of these urban centres are low, there are high

numbers of crime committed, as these types of centres typically contain high numbers of crime attractors and crime generators. When the workday population is used as a measure of the population at risk, there are no pockets of high rates present in the major cities and towns. For ambient theft (i.e., rates of theft from the person calculated using the ambient population), the spatial distribution is very similar to that of the workday theft variable (i.e., rates of theft from the person calculated using the workday population) (see Appendix A Figure 7 and Appendix A Figure 8) which suggests that there may be similarities between spatial distributions of the workday and the ambient populations.

5.4.3 Spatial distribution of the rates of 'violence and sexual offences'

The rates of violence and sexual offences calculated using the usual resident population are consistent across the study area, with concentrations of high rates in LSOAs in the central areas of Bradford and Leeds. When estimates of the workday population are employed, the spatial distribution of rates of violence and sexual offences changes considerably. Rates across much of the study area range between 4.509 and 250 per 1000 people per LSOA, with small pockets of higher rates dispersed across the centre and Southeast of the study area. It should be noted that rates in the LSOAs within urban centres fall within the lowest class utilised (0-250 crimes per 1000 people). For ambient violence (i.e., rates of violence and sexual offences calculated using the ambient population), the spatial distribution is very similar to that of workday violence. This highlights that while there are significant differences between the spatial distribution of resident violence and workday violence, there is little spatial variation between workday violence and ambient violence. Visualisations of the rates are available in the Appendix A Figures 9-11. As temporal information regarding the occurrence of violence and sexual offences is not available, it is not possible to specify

whether these crimes are more commonly committed at specific times of day. However, the Office for National Statistics (2015) noted that over half of all violent crime in the UK in the financial year 2013-2014 was identified as alcohol related. Due to the relationship between alcohol consumption and the night-time economy (Hadfield et al., 2009) it could be suggested that violent crimes may be more likely to be committed at night. This is supported by the lack of a statistically significant correlation between resident violence (rates of violence and sexual offences calculated using the resident population) and ambient violence (rates of violence and sexual offences calculated using the ambient population). It should be noted that there is no openly available information regarding the times of day that sexual offences are most commonly committed in the UK.

5.4.4 Correlation analysis

The relationships between rates of theft from the person and violence and sexual offences calculated using the three measures of the population at risk are examined using a correlation analysis. As all rates have a negatively skewed distribution, Spearman's rank correlation coefficient (ρ) was utilised; the results of the correlation analysis can be seen in Figure 5.2.

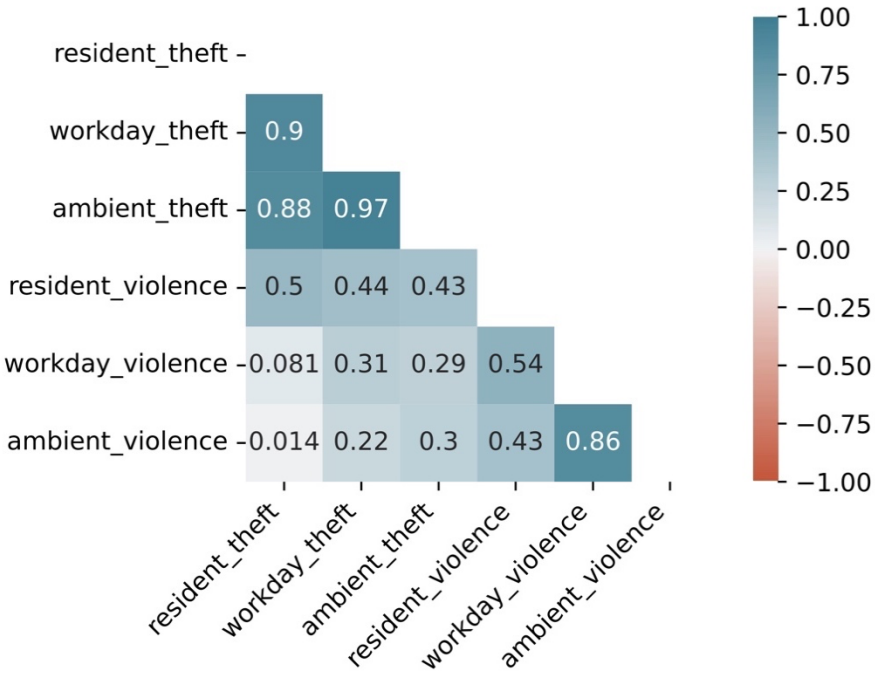


Figure 5.2 Correlation matrix highlighting the relationship between rates of theft from the person and violence and sexual offences, per 1000 people within West Yorkshire, calculated using three different measures of the population (the resident, workday, and ambient populations).

Rates of theft from the person calculated using the resident population (resident theft) and those calculated using the workday population (workday theft) are strongly, positively correlated ($\rho=0.903$, $p<0.001$). Ambient theft (rates of theft from the person calculated using the ambient population) is strongly, positively correlated with both resident theft ($\rho=0.880$, $p<0.001$) and workday theft ($\rho=0.969$, $p<0.001$). These results suggest that the measures are capturing similar rates, which is supported by the similarities between their spatial distributions.

Rates of resident violence (rates of violence and sexual offences calculated using the resident population) and workday violence (rates of violence and sexual offences calculated using the workday population) are moderately correlated, and the relationship is statistically significant ($\rho=0.540$, $p<0.001$). Given the variation between

the spatial distributions of resident violence and workday violence, a weak relationship between the rates would have been expected. The ambient violence variable (rates of violence and sexual offences calculated using the ambient population) is not significantly correlated with both resident violence (ρ=0.428, $p<0.001$) and workday violence (ρ=0.863, $p<0.001$). However, the relationship between ambient violence and workday violence is stronger than the relationship between resident violence and ambient violence. This is expected given the similarities between the spatial distribution of the two rates.

5.4.5 Global spatial autocorrelation - Moran's I

The results of the Moran's I statistical analyses, which were performed on the total crime rates for West Yorkshire as it is a global statistic, can be seen in Table 5.2. The z-scores for the Moran's I statistics for resident theft (rates of theft calculated using the residential population), workday theft (rates of theft calculated using the workday population), and ambient theft (rates of theft calculated using the ambient population) are positive and statistically significant ($p<0.001$). The z-score for the Moran's I statistics for all three measures of the rate of theft from the person are both positive and are statistically significant ($p<0.001$). Thus, the null hypothesis that the values demonstrate complete spatial randomness can be rejected for both crime types and the three different measures of the population at risk, as the distribution of high or low values in the dataset is more clustered than would be expected if the underlying spatial process were random.

Table 5.2 Outputs of the Global Moran's I statistic for the two crime types, calculated using three different measures of the population at risk.

Variable	Expected Moran's I	Moran's I	z-score	p-value
Resident theft (rates of theft calculated using the	-0.001	0.046	7.743	<0.001

residential population)				
Workday theft (rates of theft calculated using the workday population)	-0.001	0.132	18.904	<0.001
Ambient theft (rates of theft calculated using the ambient population)	-0.001	0.088	13.227	<0.001
Resident violence (rates of violence and sexual offences calculated using the residential population)	-0.001	0.161	23.356	<0.001
Workday violence (rates of violence and sexual offences calculated using the workday population)	-0.001	0.109	15.503	<0.001
Ambient violence (rates of violence and sexual offences calculated using the	-0.001	0.076	11.383	<0.001

ambient population)				

The results of the Moran's I statistic indicate that spatial clustering is present both in rates of theft from the person and in rates violence and sexual offences for all three measures of the population at risk. As spatial autocorrelation is present, it can be further investigated using a local measure of spatial autocorrelation in Section 5.4.6.

5.4.6 Local spatial autocorrelation – Local Indicators of Spatial Analysis

The locations of High-High clusters (an LSOA with a high crime rate in a neighbourhood of LSOAs with high crime rates) for both resident theft (rates of theft calculated using the resident population) and workday theft (rates of theft calculated using the workday population) (Figure 5.3) are largely focussed around the city of Leeds, with a smaller group of clusters around Bradford. However, there are differences in the locations of High-Low (an LSOA with a high crime rate in a neighbourhood of LSOAs with low crime rates) outliers, which are concentrated in the Northeast of West Yorkshire for resident theft and dispersed throughout the study area for workday theft. These findings are supported by the results of the Global Moran's I (Table 5.2), which indicated higher levels of spatial clustering for workday theft when compared to resident theft. With regards to ambient theft (Figure 5.4), High-High clusters are dispersed throughout the study area, with Low-High clusters located on the periphery. Unlike resident theft and workday theft, for ambient theft (rates of theft calculated using the ambient population) there are no large groups of clusters or outliers around large towns or cities.

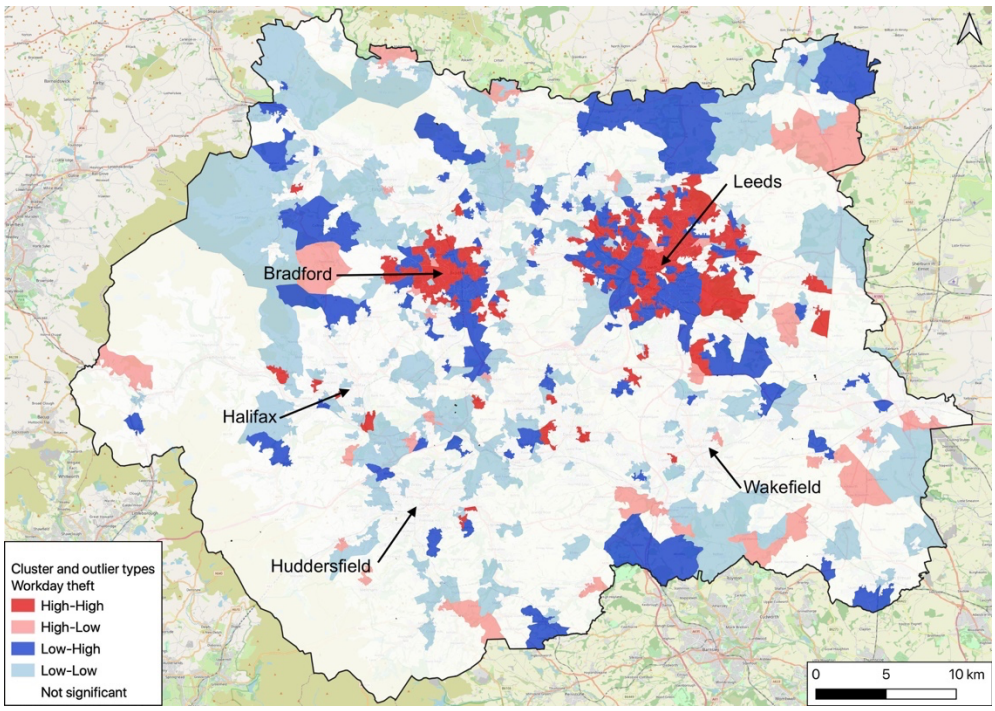


Figure 5.3 The spatial distribution of clusters and outliers for rates of theft calculated using the workday population (Basemap: © OpenStreetMap contributors, 2021 and Ordnance Survey data © Crown copyright and database right 2010-19).

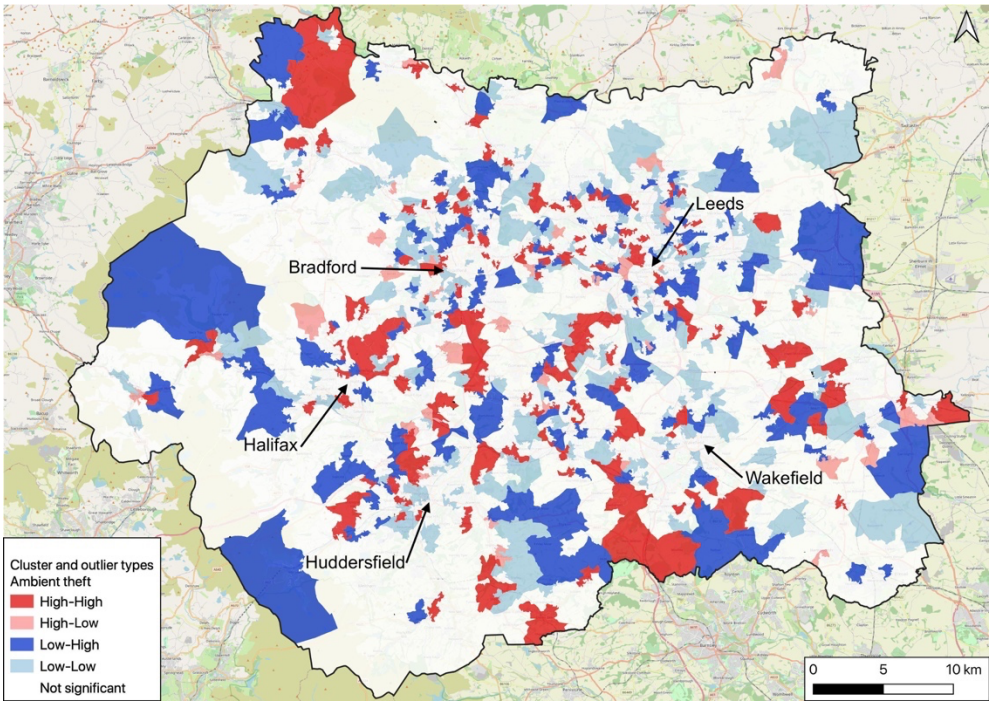


Figure 5.4 The spatial distribution of clusters and outliers for rates of theft calculated using the ambient population (Basemap: © OpenStreetMap contributors, 2021 and Ordnance Survey data © Crown copyright and database right 2010-19).

For resident violence (rates of violence and sexual offences calculated using the resident population) and workday violence (rates of violence and sexual offences calculated using the workday population), the spatial distributions of Low-Low clusters (an LSOA with a low crime rate in a neighbourhood of LSOAs with high crime rates) and High-Low outliers are similar across the study area. For workday violence (Figure 5.5), the High-High clusters present around the cities of Leeds and Bradford, which are also present for resident violence, contain large pockets of Low-High outliers. For both resident violence and workday violence, there is a group of High-High clusters and Low-High outliers South-West of Bradford. Low-Low clusters are primarily located along the North and South-West periphery of the study area for both resident violence and workday violence. In contrast, for the ambient violence (rates of violence and sexual offences calculated using the ambient population) variable (Figure 5.6), both clusters and outliers are distributed throughout the study area and are not present around any of cities in West Yorkshire. The spatial distribution of all four cluster types for ambient violence are significantly different to those for the resident violence and workday violence variables, highlighting the value of the ambient population as a measure of the population at risk.

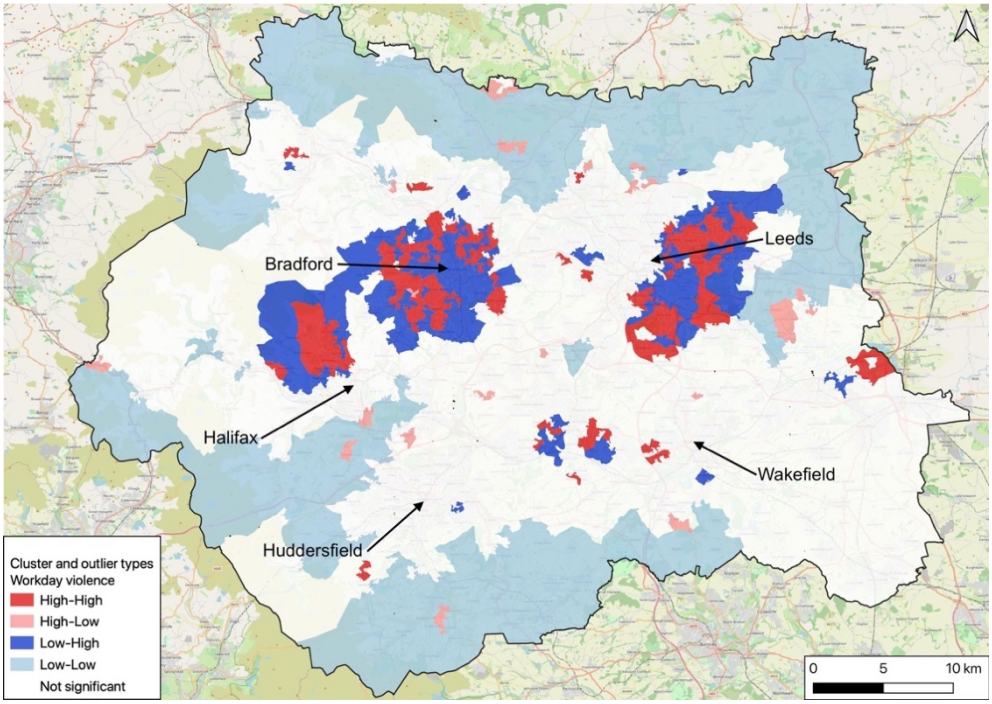


Figure 5.5 The spatial distribution of clusters and outliers for rates of violence and sexual offences calculated using the workday population (Basemap: © OpenStreetMap contributors, 2021 and Ordnance Survey data © Crown copyright and database right 2010-19).

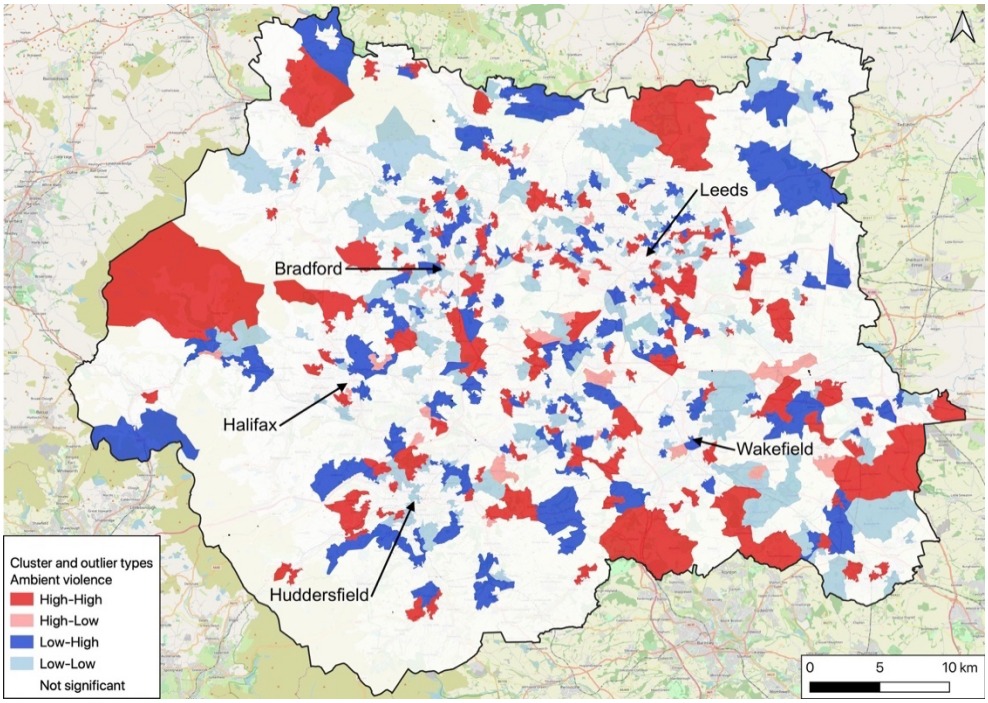


Figure 5.6 The spatial distribution of clusters and outliers for rates of violence and sexual offences calculated using the ambient population (Basemap: © OpenStreetMap contributors, 2021 and Ordnance Survey data © Crown copyright and database right 2010-19).

5.4.7 Hot spot analysis – Getis Ord GI*

For both rates of resident theft (rates of theft calculated using the resident population) and workday theft (rates of theft calculated using the workday population), there are prominent groups of hot spots mainly located around the city of Leeds. For workday theft (Figure 5.7) there are additional groups of hot spots present around Bradford. Furthermore, for workday theft the hot spots located around Leeds span a larger geographical area and there is a large group of hot spots around the city of Bradford. However, a notable difference between the spatial distributions is that there are no cold spots present for resident theft, whereas for

workday theft and ambient theft cold spots are dispersed throughout the study area. For ambient theft (rates of theft calculated using the ambient population) (Figure 5.8), the hot spots are dispersed across the study area and there are no clusters present around the cities of Leeds and Bradford, or any of the other urban centres. The spatial variations between the hot spots for workday theft and ambient theft highlight the impact of the measure of the population at risk, particularly within urban centres.

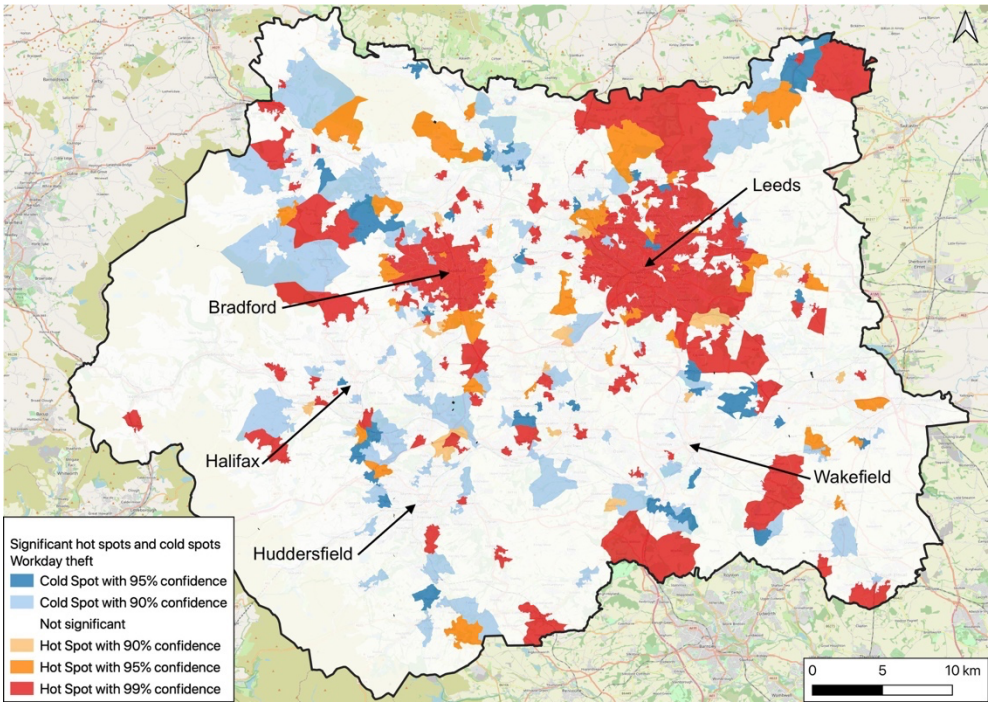


Figure 5.7 The spatial distribution of hot spots and cold spots for rates of theft calculated using the workday population (Basemap: © OpenStreetMap contributors, 2021 and Ordnance Survey data © Crown copyright and database right 2010-19).

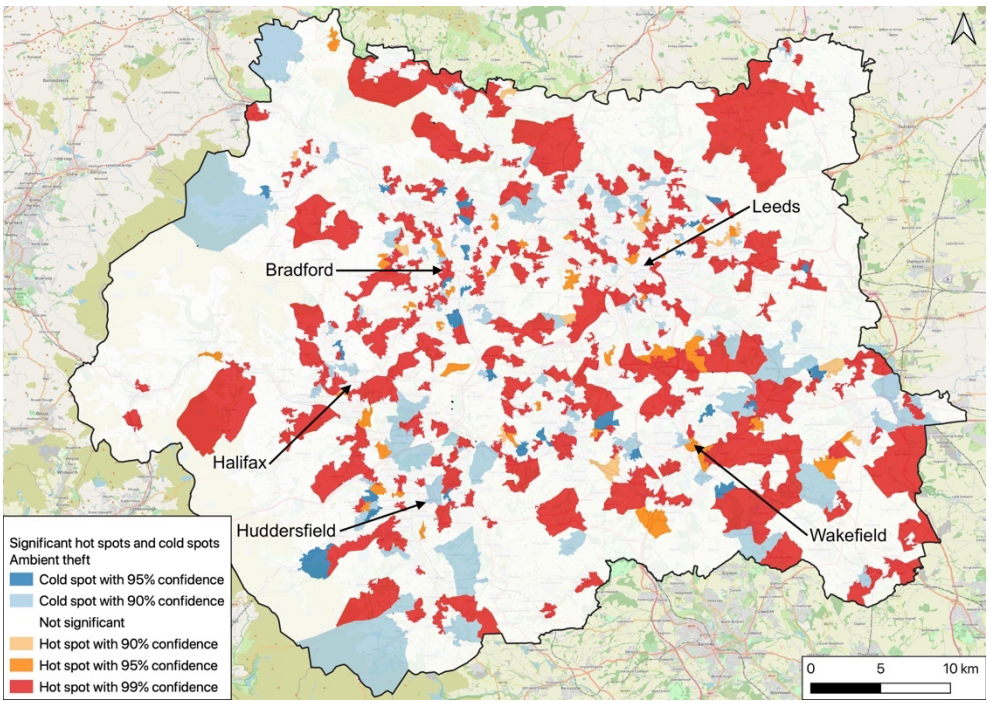


Figure 5.8 The spatial distribution of hot spots and cold spots for rates of theft calculated using the ambient population (Basemap: © OpenStreetMap contributors, 2021 and Ordnance Survey data © Crown copyright and database right 2010-19).

The patterns of hot spots and cold spots for resident violence (rates of violence and sexual offences calculated using the resident population) and workday violence (rates of theft calculated using the workday population) (Figure 5.9) are very similar. When these measures of the population at risk are employed, there are large clusters of hot spots shifts around the cities of Leeds and Bradford. The spatial distribution of cold spots is also similar, with groups of cold spots located on the North and South borders of the study area. The spatial distribution of hot spots and cold spots for ambient violence is significantly different to those for resident violence and workday violence, as demonstrated in Figure 5.10. In contrast, when the ambient population is employed (Figure 5.10), hot spots and cold spots are dispersed throughout the study area and, most notably, there are no pronounced clusters of hot spots around any urban centres.

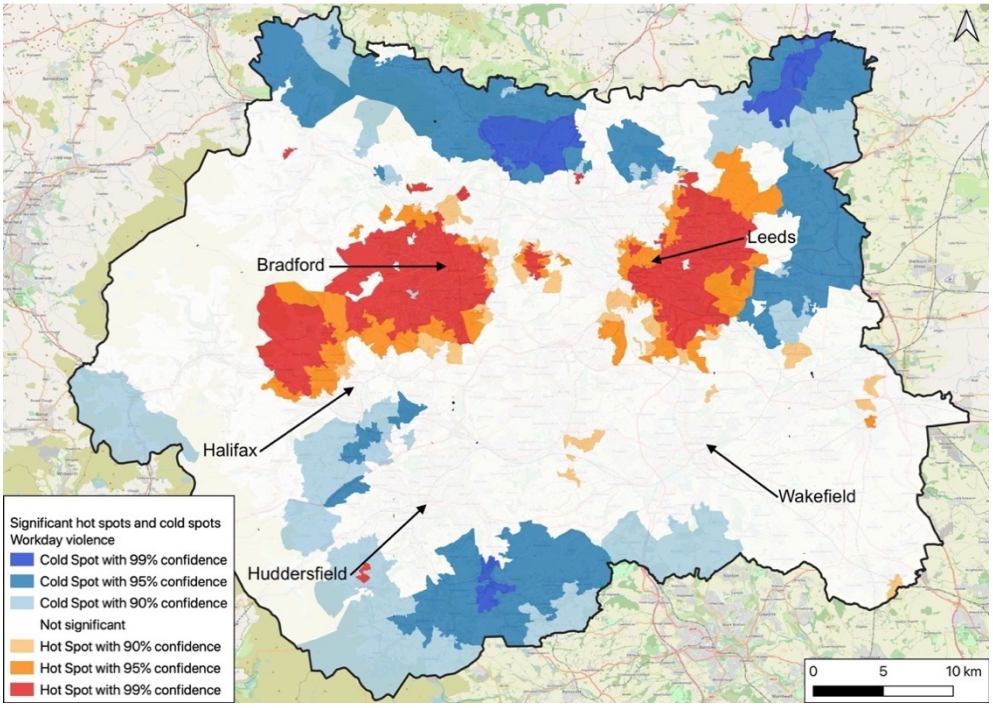


Figure 5.9 The spatial distribution of hot spots and cold spots for rates of violence and sexual offences calculated using the workday population (Basemap: © OpenStreetMap contributors, 2021 and Ordnance Survey data © Crown copyright and database right 2010-19).

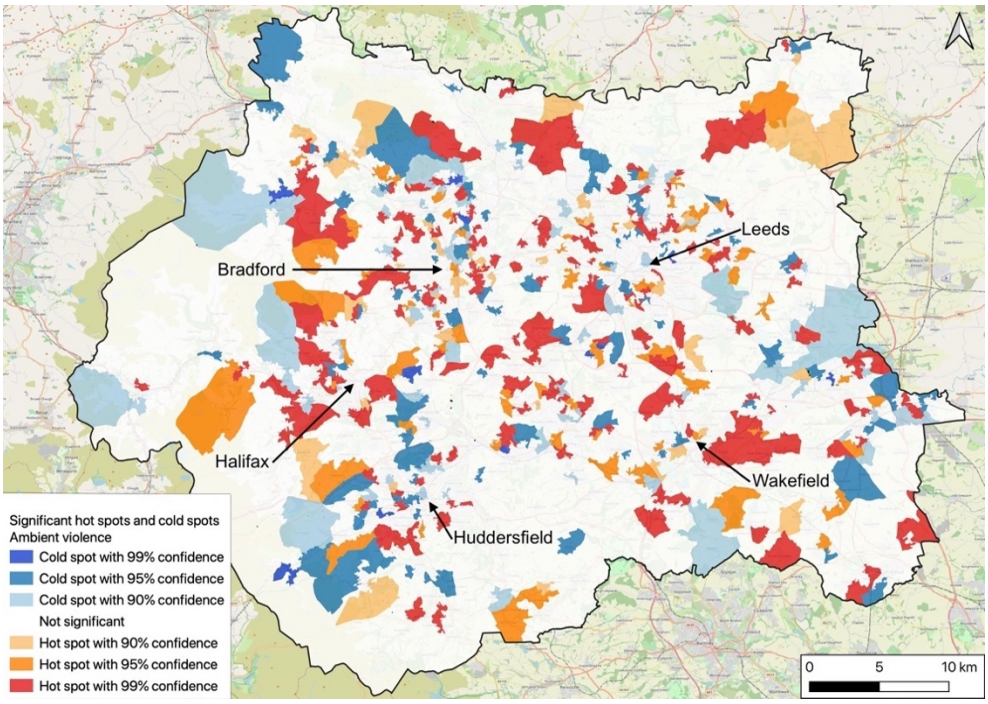


Figure 5.10 The spatial distribution of hot spots and cold spots for rates of violence and sexual offences calculated using the ambient population (Basemap: © OpenStreetMapContributors, 2021 and Ordnance Survey data © Crown copyright and database right 2010-19).

5.5 Discussion

5.5.1 Discussion of results

5.5.1.1 Theft from the person

When the ambient population is used as the measure of the population at risk the results of the cluster and outlier analysis in this study demonstrate a substantial reduction in the number of statistically significant High-High clusters of incidents of theft from a person (a high value in a high-value neighbourhood) in urban centres, when compared to use of both the resident and workday populations. The most striking reductions are around the city of Leeds. This difference is expected as crime pattern theory suggests that while urban centres have a high number of crime attractors and crime generators, they also experience significant increases in the size of the ambient population due to human activity (Brantingham et al., 1981; Brantingham and Brantingham, 1995, Kinney et al., 2008). Consequently, as the size of the ambient population is larger than that of the resident and workday populations, due to it capturing work and non-work-related human activity, crime rates calculated using the ambient population are lower than those calculated using the resident and workday populations. This means that the actual risk of victimisation is lower in urban centres than currently communicated. As the perception of risk may influence how some individuals utilise and move through space and affect any precautions taken to reduce their risk of victimisation, the accurate communication of risk is essential to public safety (Curtis, 2015; Prieto Curiel and Bishop, 2018). The use of the ambient population resulted in some High-High clusters of rates of theft from the person in areas with a primarily residential area. Given that the size of the ambient population

in residential areas would be expected to be lower than in urban centres and, therefore, contribute to an increased risk of victimisation, this finding is expected. However, it does contradict the low level of risk in residential areas communicated by rates of this crime type calculated using resident and workday populations. Therefore, the level of the risk of victimisation of theft from the person is under-estimated by the measure i.e., the resident population, currently employed by police forces. The use of the workday population similarly fails to communicate risk accurately.

The hot spot analysis produced similar results to those discussed above, with substantial reductions in the number of hot spots in urban centres when the ambient population was employed. This supports crime patterns theory (Brantingham et al., 1981; Brantingham et al., 2016) as although there are high numbers of crime attractors and generators in both Leeds and Bradford, these areas also attract large ambient populations. There were also increases in the number of hot spots present across the study area in LSOAs which would be described as primarily residential.

The results of this study evidence that the use of the ambient population as a measure of the population at risk from theft from the person is crucial. In addition to more accurately representing the size of the population at risk, the statistically significant differences in the spatial distribution of rates, evidence that the selection of the measure of the population risk should not be arbitrary.

5.5.1.2 Violence and sexual offences

When examining rates of violence and sexual offences, the results of the cluster and outlier analysis evidence that when the workday population is employed, groups of High-High clusters and Low-High outliers (a low value in a high-value neighbourhood) are heavily concentrated around the cities of Leeds and Bradford. Low-Low clusters (a low value in a low-value neighbourhood) are present on the periphery of the study area in the North and Southwest, in LSOAs which have a residential backcloth. Therefore, these results suggest that high rates of violence and

sexual offences tend to cluster in cities, while suburban, residential areas experience clusters of low rates, and are, therefore, safer for individuals. In contrast, the use of the ambient population produced substantially different distributions of statistically significant clusters and outliers. There were no concentrations of High-High clusters or Low-High outliers present around Leeds or Bradford. Instead, clusters and outliers appear to be distributed randomly throughout the study area, indicating that rates of violence and sexual offences do not cluster disproportionately in urban centres. As with rates of theft from the person, when the ambient population is employed as a measure of the population at risk, there are High-High clusters present in LSOAs that are primarily residential. This pattern is expected in residential areas, as some incidences of violence and sexual offences are committed at residential locations. The use of the workday and resident populations suggests that rates of violence and sexual offences are low in residential areas and, therefore, do not accurately represent the spatial distribution of rates of this crime type.

The results of the hot spot analysis illustrate marked differences between the use of the workday population and the ambient population as a measure of the population at risk in the calculation of crime rates. When the workday population is employed, hot spots are concentrated around the cities of Leeds and Bradford, indicating high rates of crime in these areas. There are also groups of cold spots across the North and Southwest periphery of the study area. However, when the ambient population is used as the measure of the population at risk, while some LSOAs around Leeds and Bradford do contain hot spots, there are no large groups of hot spots present around any urban centres. This illustrates that the use of the both the resident population and workday population as a measure of the population at risk suggests that only urban centres contain crime hot spots, while the use of the ambient population evidences that hot spots are in fact dispersed throughout West Yorkshire. Furthermore, when the ambient population is used within the calculation of rates of violence and sexual offences, there are additional hot spots present in the Northeast and West of the study area located in LSOAs, which are mainly residential. Hot spots are not present in residential areas when the resident and workday populations are

employed, which suggests that this measure may not accurately represent the level of risk of victimisation in residential areas. The use of the ambient population also resulted in a distinct reduction in the number of statistically significant cold spots on the periphery of West Yorkshire, which has a primarily residential backcloth, when compared to the use of the resident and workday populations. Therefore, rates of violence and sexual offences calculated using these measures inaccurately indicate higher levels of safety in residential areas.

The findings demonstrate the use of the resident and the workday populations as a measure of the population at risk, fail to accurately communicate the risk of victimisation from violence and sexual offences. The use of the ambient population demonstrates that levels of risk are not disproportionality higher in urban centres than in residential areas. This reinforces the need to employ the ambient population as a measure of the population at risk, to enable the calculation of accurate crime rates that can be utilised to inform effective policymaking.

5.5.1.3 Implications of the findings

The differences in the spatial distributions of rates of theft from the person and violence and sexual offences when different measures of the non-resident population are employed have significant implications for the use of crime rates. The spatial distributions of crime rates will influence resource allocation and the implementation of crime prevention policies. Therefore, if crime rates do not accurately represent the risk of a crime being committed, policies are likely to be ineffective. Consequently, those geographic areas that actually experience high crime rates and would benefit from the deployment of crime reduction policies will not be allocated appropriate resources. The use of both the resident and the workday populations in the calculation of crime rates suggest that cities of West Yorkshire experience concentrations of crime hot spots and clusters of crime. In contrast, the use of the ambient population demonstrates that these features are dispersed throughout the study area and large numbers of hot spots and clusters are not

present in urban centres. As measures of the resident and workday population do not capture the total population of an area, as the resident population fails to account for activities outside of the home and the workday population does not capture non-work-related activities, they cannot, therefore, accurately represent the size of the population at risk. Consequently, crime rates calculated using these measures are not able to communicate the actual risk of crime to members of the public. As the perception of risk may influence how some individuals utilise and move through space and affect any precautions taken to reduce their risk of victimisation, the accurate communication of risk is essential to public safety.

5.5.2 Limitations and opportunities for future work

5.5.2.1 The modifiable areal unit problem

The Modifiable Areal Unit Problem (MAUP) is a potential source of error within spatial analysis which occurs when the geographical boundaries imposed on data can impact the spatial patterns of aggregated data, as the variance structure of the data is altered (Charlton, 2009; Openshaw and Taylor, 1979). It is important to note that all data used in this study, as outlined in the Appendix A Table 1, are aggregated to LSOA level; thus, the results of the analyses employed are likely to be impacted by the effects of MAUP. Wong (2009) notes the importance of acknowledging the presence of MAUP; however, there are limited ways in which the effects can be managed. One approach is to conduct analyses at multiple geographical scales to investigate any variations in the results (Wong, 2009). However, only LSOAs are employed within this study as it is a commonly used geographic scale within crime analysis and is the smallest geography that can be used across all three measures of the population at risk. It should also be noted that the use of spatial units smaller than LSOAs, such as output areas or postcode areas, are not suitable for use as work by Tompson et al. (2015) illustrated that there is spatial error in police recorded crime data at these levels. While an alternative solution is to employ scale-independent analysis, this is not feasible within this study as crime rates must be representative of a geographical area (Su et al., 2011).

5.5.2.2 Spatio-temporal estimates of the population

Given that certain types of crimes will be more likely to be committed at specific times of day, it can be argued that crime rates should take into account the number of people in the area at the times at which these crimes are commonly committed (Newton and Felson, 2015). For example, the rates of alcohol-related offences may be more accurately communicated by employing a measure of the population at risk which enumerates night-time, non-residential populations. This is supported by data produced by the Metropolitan Police (2018) which highlights that between April 2017 and April 2018 in London (UK), 72% of alcohol related offences were committed between the hours of 18:00 and 06:00 and 38% occurred between 24:00 and 06:00. This study has not explored the use of spatio-temporal crime rates, as time-stamped crime data are not currently openly available.

5.5.2.3 Limitations of the data

In this section, limitations of the datasets employed in this study are noted. Data captured by the 2011 Census of England and Wales, i.e., the estimates of the size of the resident and workday populations, are limited by the frequency of data collection. While these data are geographically comprehensive and are considered to be the gold standard of population data (Rees et al., 2002), at the time of writing, the data utilised are over ten years old. It should be noted that data from the most recent census of England and Wales, conducted in 2021, have not yet been released. Consequently, these data may not accurately represent the resident and workday populations, and therefore will impact the accuracy of the crime rates produced. The estimates of the size of the workday population are also used to produce the estimates of the ambient population; thus, the estimates of the ambient population may also be outdated.

The accuracy of the footfall camera estimates used to produce the ambient population variable is impacted by the predictive capacity of the model used to

estimate footfall counts. The model currently accounts for 33.2% of the variation of the dependent variable, i.e., the footfall camera counts; however, the predictive capacity of the model could be improved by developing a more in depth understanding of the drivers of the size of the ambient population in urban centres. As the estimates of the ambient population are determined by the presence of ATMs, which may be limited in rural areas, and hospitality, which is a fluctuating market following the COVID-19 pandemic.

5.5.2.4 Exploration of other crime types

This study investigates the impact of different measures of the population at risk on the spatial distribution of the rates of two types of crime committed against the individual; theft from the person and violence and sexual offences. However, there may be value in the exploration of the use of the ambient population as a measure of the population at risk for the calculation of rates of crimes that do not target individuals, such as vehicle theft or burglary. One factor which supports this is that the size of the ambient population in residential areas will impact the number of capable guardians present and, therefore, in accordance with routine activity theory, will affect the numbers of crimes committed.

5.5.2.5 Exploration of other locations

While this study has explored patterns of crime rates across West Yorkshire, there remains an opportunity to investigate the impact of different measures of the population in different geographical locations. This would enable researchers to determine the generalisability of the work and determine instances in which the use of a resident (i.e. non-ambient) population may be more appropriate. In order to effectively assess the generalisability of the work, the impact in other regions, including urban, suburban, and rural areas, both in the UK and overseas, should be explored and assessed.

5.6 Conclusions

This study illustrates significant differences in the spatial distributions of rates of theft from the person and of rates of violence and sexual offences, calculated using the residential population as the measure of the population at risk, compared to measures of the non-residential population. Findings from this study are consistent with the literature, supporting that the use of the resident population as a measure of the population at risk is inappropriate and that there is value in employing estimates of the non-residential population. However, this study also expands on the existing literature by evidencing the value of using the ambient population, as opposed to the workday population, as a measure of the population at risk. The results of the cluster and outlier analysis and the hot spot analysis evidence that for both crime types, the use of the resident and workday population overestimate the risk of victimisation within urban centres and underestimate the risk in residential areas. There are, of course, limitations of the estimates of the ambient population used in this study, i.e., they may not capture the whole ambient population due to the model's predictive capacity. However, as these estimates capture both work and non-work-related increases in the size of the population, which the resident and workday populations do not, they offer a significant advantage over the current approach. This study demonstrates that estimates of the ambient population may be beneficial for the calculation of accurate crime rates and supports the demand for geographically comprehensive estimates of the size of the ambient population that can be utilised by police forces and policymakers resulting in more effective crime reduction strategies.

References

Andresen, M.A. 2007. Location quotients, ambient populations, and the spatial analysis of crime in Vancouver, Canada. *Environment and Planning A*. 39(10), pp.2423–2444.

Andresen, M.A. 2011. The ambient population and crime analysis. *The Professional Geographer*. 63(2), pp.193–212.

Andresen, M.A. and Jenion, G.W. 2010. Ambient populations and the calculation of crime rates and risk. *Security Journal*. **23**(2), pp.114–133.

Andresen, M.A. and Jenion, G.W. 2008. Crime prevention and the science of where people are. *Criminal Justice Policy Review*. **19**(2), pp.164–180.

Anselin, L. 1995. Local Indicators of Spatial Association—LISA. *Geographical Analysis*. **27**(2).

Benton, J. 2019. No TitleTwitter is removing precise-location tagging on tweets — a small win for privacy but a small loss for journalists and researchers. [Accessed 18 February 2021]. Available from: https://www.niemanlab.org/.

Boggs, S.L. 1965. Urban crime patterns. *American Sociological Review*., pp.899–908.

Brantingham, P.J., Brantingham, P.L. and others 1981. *Environmental criminology*. Sage Publications Beverly Hills, CA.

Brantingham, Patricia and Brantingham, Paul 1995. Criminality of place. *European journal on criminal policy and research*. **3**(3), pp.5–26.

Chainey, S. and Ratcliffe, J. 2013. *GIS and Crime Mapping*.

Chainey, S., Tompson, L. and Uhlig, S. 2008. The Utility of Hotspot Mapping for Predicting Spatial Patterns of Crime. *Security Journal*. **21**(1–2).

Charlton, M. 2009. Quantitative Data. *International Encyclopedia of Human Geography*., pp.19–26.

Cohen, L.E. and Felson, M. 1979. Social change and crime rate trends: A routine activity approach. *American sociological review*., pp.588–608.

College of Policing 2021. About data.police.uk. [Accessed 3 March 2021]. Available from: https://data.police.uk/about/.

Eck, J.E., Chainey, S., Cameron, J.G., Leitner, M. and Wilson, R.E. 2005. *Mapping Crime : Understanding Hot Spots*.

Felson, M. and Boivin, R. 2015. Daily crime flows within a city. *Crime Science*. **4**(1), p.31.

Getis, A. and Ord, J.K. 1992. The Analysis of Spatial Association by Use of Distance

Statistics. *Geographical Analysis*. **24**(3).

Hanaoka, K. 2018. New insights on relationships between street crimes and ambient population: Use of hourly population data estimated from mobile phone users' locations. *Environment and Planning B: Urban Analytics and City Science*. **45**(2), pp.295–311.

He, L., Páez, A., Jiao, J., An, P., Lu, C., Mao, W. and Long, D. 2020. Ambient Population and Larceny-Theft: A Spatial Analysis Using Mobile Phone Data. *ISPRS International Journal of Geo-Information*. **9**(6), p.342.

Hipp, J.R., Bates, C., Lichman, M. and Smyth, P. 2019. Using social media to measure temporal ambient population: does it help explain local crime rates? *Justice Quarterly*. **36**(4), pp.718–748.

Jung, Y., Chun, Y. and Kim, K. 2020. Modeling Crime Density with Population Dynamics in Space and Time: An Application of Assault in Gangnam, South Korea. *Crime and Delinquency*.

Kadar, C., Rosés Brüngger, R. and Pletikosa, I. 2017. Measuring ambient population from location-based social networks to describe urban crime *In*: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*.

Kounadi, O., Ristea, A., Leitner, M. and Langford, C. 2018. Population at risk: using areal interpolation and Twitter messages to create population models for burglaries and robberies. *Cartography and Geographic Information Science*. **45**(3), pp.205–220.

Lan, M., Liu, L., Hernandez, A., Liu, W., Zhou, H. and Wang, Z. 2019. The spillover effect of geotagged tweets as a measure of ambient population for theft crime. *Sustainability (Switzerland)*.

Malleson, N. and Andresen, M.A. 2016. Exploring the impact of ambient population measures on London crime hotspots. *Journal of Criminal Justice*. **46**, pp.52–63.

Malleson, N. and Andresen, M.A. 2015a. Spatio-temporal crime hotspots and the ambient population. *Crime Science*.

Malleson, N. and Andresen, M.A. 2015b. The impact of using social media data in crime rate calculations: Shifting hot spots and changing spatial patterns. *Cartography and Geographic Information Science*.

Metropolitan Police 2018. No Metropolitan Police Notifiable Offences by Time of Day. *London Datastore*.

National Academy of Sciences 2016. *Modernizing Crime Statistics: Report 1: Defining and Classifying Crime*.

Nomis 2013. Usual resident population. [Accessed 4 March 2021]. Available from: https://www.nomisweb.co.uk/census/2011/ks101uk.

Office for National Statistics 2013. 2011 Census: The workday population of England and Wales - An alternative 2011 Census output base. *Office for National Statistics*. [Online]. Available from: https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/articles/theworkdaypopulationofenglandandwales/2013-10-31.

Office for National Statistics 2011. 2011 Census aggregate data. *UK Data Service*. [Online]. [Accessed 20 September 2021]. Available from: infuse2011gf.ukdataservice.ac.uk/.

Office for National Statistics 2016. Census geography. [Accessed 4 March 2021]. Available from: https://www.ons.gov.uk/census/2001censusandearlier/dataandproducts/outputgeography/outputareas.

Office for National Statistics 2021. Crime in England and Wales: Police Force Area data tables.

Office for National Statistics 2017a. Overview of robbery and theft from the person: England and Wales. [Accessed 4 March 2021]. Available from: https://www.ons.gov.uk/peoplepopulationandcommunity/crimeandjustice/articles/overviewofrobberyandtheftfromtheperson/2017-07-20.

Office for National Statistics 2017b. Overview of violent crime and sexual offences.

[Accessed 9 September 2021]. Available from: https://www.ons.gov.uk/peoplepopulationandcommunity/crimeandjustice/compendium/focusonviolentcrimeandsexualoffences/yearendingmarch2016/overviewofviolentcrimeandsexualoffences.

OpenStreetMapContributors 2021. OpenStreetMap data. [Accessed 1 September 2021]. Available from: https://www.openstreetmap.org/.

Ord, J.K. and Getis, A. 1995. Local Spatial Autocorrelation Statistics: Distributional Issues and an Application. *Geographical Analysis*. **27**(4).

Rees, P., Martin, D. and Williamson, P. 2002. *The census data system*. Wiley.

Ristea, A., Kounadi, O. and Leitner, M. 2018. Geosocial Media Data as Predictors in a GWR Application to Forecast Crime Hotspots (Short Paper) *In*: *10th International Conference on Geographic Information Science (GIScience 2018)*.

Stults, B.J. and Hasbrouck, M. 2015. The Effect of Commuting on City-Level Crime Rates. *Journal of Quantitative Criminology*.

Tompson, L., Johnson, S., Ashby, M., Perkins, C. and Edwards, P. 2015. UK open source crime data: Accuracy and possibilities for research. *Cartography and Geographic Information Science*. **42**(2).

Tucker, R., O'Brien, D.T., Ciomek, A., Castro, E., Wang, Q. and Phillips, N.E. 2021. Who 'Tweets' Where and When, and How Does it Help Understand Crime Rates at Places? Measuring the Presence of Tourists and Commuters in Ambient Populations. *Journal of Quantitative Criminology*.

Whipp, A., Malleson, N., Ward, J. and Heppenstall, A. 2021. Estimates of the Ambient Population: Assessing the Utility of Conventional and Novel Data Sources. *ISPRS International Journal of Geo-Information*. **10**(3), p.131.

Wong, D.W. 2009. Modifiable Areal Unit Problem *In*: *International Encyclopedia of Human Geography*.

Chapter 6

Conclusions

The work within this thesis has fulfilled the research aim, which was to explore the development of small area estimates of the size of the ambient population in an urban area. In addition to achieving this aim, the work makes four significant novel contributions to the literature. Firstly, the thesis contributes to the literature through the critical assessment of the suitability of different data types for producing estimates of the size of the ambient population. This is a valuable contribution as the identification of suitable data underpins attempts to produce estimates of the size of the ambient population. Secondly, the approach developed in this thesis to produce estimates of the size of the ambient population using solely open data is a novel contribution to the literature. Thirdly, the dataset produced for the purpose of this research and utilised to validate the estimates of the size of the ambient population is currently the only openly available dataset that captures manual footfall counts. Consequently, this work is the first to validate estimates of the size of the ambient population using ground truth data. Lastly, the work identified that the use of the resident and workday populations overestimate the risk of victimisation within urban centres, while underestimating the risk in residential areas. This finding demonstrates that estimates of the size of the ambient population are critical for the accurate calculation of crime rates.

This chapter concludes the thesis and provides a summary of the research undertaken. Section 6.1 summarises the thesis and demonstrates the extent to which the aim and objectives, detailed in Chapter 1, have been met. The limitations of the thesis are discussed in Section 6.2, while Section 6.3 notes recommendations for future work. An outlook on both producing and utilising estimates of the size of the ambient population and concluding remarks are presented in Section 6.4.

6.1 Thesis summary and contribution to the literature

The aim of this thesis, as stated in Chapter 1, is to explore the development of small area estimates of the size of the ambient population in an urban area. To fulfil this aim, six research objectives were established. This section revisits these objectives and assesses the extent to which they have been met by the work in this thesis.

Objective One: Review and discuss the literature relating to quantifying the size of the ambient population and comparable small area estimates of populations and their use within crime studies.

Objective One was fulfilled through a review of the literature in Chapter 2. Chapter 2 initially defined the 'ambient population' and identified terms used synonymously within the existing literature. The demand for estimates of the size of the ambient population across a range of applications within research and policymaking was then acknowledged. The chapter noted that despite the utility of estimates of the size of the ambient population, they are not yet part of the standard suite of population statistics in any nation. Methodologies and data types utilised in previous studies of the ambient population were then explored. Early studies of the ambient population used methods of areal interpolation, including dasymetric mapping, grid-based modelling, and pycnophylactic interpolation. These methods were employed as they are able to disaggregate coarse data to a fine spatial scale; however, given the granularity of data now available, the utility of these methods for estimating the size of the ambient population has diminished. It was noted that many recent studies had employed novel data, such as geo-located Twitter data, Wi-Fi sensor counts, and mobile phone data, as a proxy of the size of the ambient population. However, this approach has significant limitations, as simply employing novel data as a measure of the ambient population fails to represent the entire population, resulting in inaccurate estimates. Many types of novel data are not representative of the entire population as they are subject to sampling bias. Thus,

individuals with certain demographic characteristics, i.e., those within specific age or socio-economic categories, may not be captured. Thus, Chapter 2 identified an opportunity to expand on the existing literature by exploring the use of high-resolution spatial data with a method of modelling to produce estimates of the size of the ambient population.

The work in Chapter 2 explored the use of estimates of the size of the ambient population within crime studies. The importance of accurate crime rates and their value within research and policymaking was illustrated. The literature highlighted that crime rates calculated using estimates of the resident population are both inappropriate and inaccurate, as these estimates fail to capture the size of the population at risk. Previous studies have employed alternative, non-residential measures of the population at risk. The results of these studies evidenced significant qualitative and quantitative differences in the spatial distributions of crime rates compared to the use of the resident population. While some of these studies use measures of the ambient, very few use comprehensive estimates of the ambient population and instead employ Twitter data as a proxy. Therefore, the use of comprehensive estimates of the size of the ambient population as a measure of the population at risk within crime rates needs to be further investigated.

Chapter 2 identified an opportunity to provide a novel contribution to the literature by comparing the differences between the use of two measures of the non-resident population, i.e., the workday population and the ambient population, and their impact on the spatial distribution of crimes rates. Work by Malleson and Andresen (2016) previously highlighted the size of the workday population as the most appropriate measure of the population at risk. However, the study only explored the use of the workday population on one crime type within London (UK). Therefore, there remained an opportunity to explore differences between the use of the workday population and a measure of the ambient population on the spatial distribution of two crime types within an alternative study area. This exploration would be a valuable contribution to both crime studies and policymaking. While

estimates of the workday population are readily available and are more appropriate than the measure currently used by police forces (i.e., the resident population), they fail to capture non-work-related fluctuations in the size of the population. Consequently, identifying any statistically significant differences in the spatial distributions of crime rates calculated using the workday and ambient populations allows an appropriate measure of the population at risk to be determined.

Objective Two: Assess and critique sources of population data that have the potential to be used to produce estimates of the size of the ambient population in urban areas, including those utilised in the existing literature.

The work presented in Chapter 3 critically reviewed and assessed the utility of sources of population data that have the potential to be utilised in the production of estimates of the size of the ambient population. This chapter highlighted that conventional types of data, such as data from national censuses, are able to enumerate a high proportion of the whole population. However, the utility of conventional data is often limited by the granularity of the data available and may limit the spatial scale of studies undertaken. Infrequent data collection is a further limitation of the data and is due to the temporal and financial cost of conducting large scale surveys. Despite these limitations, conventional data types, in particular census data, are a valuable asset within studies of the ambient population. This is due to the geographic comprehensiveness of the data, which is seldom available with novel data. While beyond the scope of this thesis, it should also be noted that conventional data often provide additional information about the populations they enumerate. For example, census data capture detailed data regarding the demographic characteristics of the population, while travel surveys provide information regarding journey purpose and duration. This additional information could enable a more in-depth understanding of the indicators and demographic composition of the ambient population to be developed. With many types of novel data, information regarding the individuals enumerated is often limited. Often, if individual-level information is captured, it is unlikely to be available for use due to privacy restrictions. Thus, the use

of conventional sources may be advantageous in studies that focus on the characteristics of the ambient population.

Following this, Chapter 3 explored novel data types, including mobile phone data (mobile phone activity data, smartphone location data, and cell tower location data), geo-located social media data, and pedestrian counters (footfall cameras and Wi-Fi sensors).

Mobile phone data are the most commonly utilised novel form of data within the existing literature and have been employed in a range of studies. However, mobile phone data are limited by significant ethical concerns and issues relating to data access. Both mobile phone activity and smartphone location data are produced by private organisations and can be expensive to acquire. Cell tower location data, however, is available free of charge from OpenCelliD and has been used in several studies to quantify the size of the ambient population. Despite this advantage, the data quantify cell tower density and do not provide information regarding the number of people in an area. Consequently, the utility of OpenCelliD data is limited, as they cannot be used to quantify the size of the ambient population.

The work in Chapter 3 then illustrated that while geo-located social media data can provide spatio-temporally detailed information about the population, the utility of the data is significantly limited by the size and representativeness of the samples. As geo-located social media data are only representative of a small proportion of individuals who utilise social media platforms, the demographic characteristics and the activity patterns of these individuals are unlikely to be representative of the whole ambient population.

Data from two types of pedestrian counters, Wi-Fi sensors and footfall cameras, are explored in Chapter 3. Wi-Fi sensors log the number of Wi-Fi enabled

devices passing a geographic point, while footfall cameras capture all individuals who pass a camera. Consequently, it can be argued that footfall camera data are more suitable for enumerating the whole population, despite high penetration levels of Wi-Fi enabled devices across different demographic groups. Data captured by pedestrian counters are becoming increasingly openly available and are currently accessible for cities in Australia, the UK, and the US. The work in Chapter 3 highlighted an opportunity to further explore the utility of pedestrian counter data in estimating the size of the ambient population.

Objective Three: Develop small area estimates of the size of the ambient population for an urban area.

The work presented in Chapter 4 builds a model that produces small area estimates of the size of the ambient population using openly available data. A preliminary predictive model was developed in which footfall camera counts were employed as the dependent variable, and seven independent variables were selected based on the existing literature. Footfall camera counts were chosen as the dependent variable, as they were highlighted by the work in Chapter 3 as a valuable data source and have not been explored extensively within the existing research. The seven independent variables (cell tower density, the size of the workday population, and the numbers of higher and further education buildings, retail premises, transport hubs, hospitality values, and ATMs) had been successfully employed in other studies reviewed in Chapter Two or highlighted as potentially valuable by the work in Chapter Three. Surprisingly, only two out of the seven variables tested, the number of ATMs and hospitality venues, were statistically significant. Based on the existing literature, it was expected that all seven independent variables would have a significant relationship with the size of the ambient population. This finding highlights that the understanding of the indicators of the size of the ambient population remains relatively limited, which illustrates that a more in-depth understanding is crucial to producing accurate estimates of the size of the ambient population. A final model was then developed in which the two variables that had a statistically significant

relationship with the footfall camera counts, the number of ATMs and hospitality venues, were employed as the independent variables. All other model conditions remained the same as in the preliminary model. The estimates of the size of the ambient population produced by the model were then validated to assess their accuracy. The dataset utilised to perform the validation, and the results of the process are discussed in the following sections.

Objective Four: Produce a validation dataset that captures footfall counts in an urban area.

To validate the footfall camera counts and estimates produced by the model developed in Chapter 4, footfall data from an alternative source were required. As no data were available, a novel dataset was produced. As outlined in Chapter 4, these manual footfall count data were captured by a team of data collectors at ten sites across the Metropolitan Borough of Leeds across a six-hour period. Three data collectors were stationed at each site, which allowed the mean total count to be taken and then utilised in the validation process. The validation data are openly available from the Consumer Data Research Centre (Whipp, 2021). These data also allow the accuracy of the manual footfall counts, the footfall camera counts, and the model estimates to be assessed (see Objective Five).

Objective Five: Employ the validation dataset to assess the accuracy of the footfall camera counts and the model estimates.

Validation is key to ensuring data accuracy. The validation process conducted in Chapter 4 assessed the similarity of the samples collected by the three data collectors, quantified the accuracy of the footfall cameras, and allowed model estimates to be compared to manual counts using a novel, empirical dataset. The results presented in Chapter 4 indicate that the samples collected by three data collectors at seven out of eight sites likely come from the same distribution. While the validation of the footfall camera counts demonstrated that the data were accurate

and had a similar distribution to the manual counts at two of three locations. The validation of the model estimates evidenced that at four of six locations, the accuracy of the estimates was as expected or higher than expected based on the model fit. No other studies within the existing literature have attempted to validate footfall camera data; thus, the findings are a novel contribution to the literature and highlight the value of the data for use in future studies.

Objective Six: Utilise the estimates of the size of the ambient population to examine the impact of different measures of the population on the spatial distribution of the rates of two crime types; 'theft from the person' and 'violence and sexual offences'.

Objective Six was fulfilled through a literature review presented in Chapter 2 and a case study in Chapter 5. Chapter 2 demonstrated that as crime clusters within space and time, crime rates calculated using the size of the resident population may not accurately represent the size of the population at risk. Existing studies within environmental criminology have employed alternative measures of the population at risk; however, the utility of a comprehensive estimates of the size of the ambient population on the spatial distribution of crime rates had not yet been explored. To assess the value of estimates of the size of the ambient population within crime studies and compare the use of both the resident population and different measures of the non-residential population at risk, a case study of West Yorkshire is employed in Chapter 5. This case study explored the impact of alternative measures of the population to assess the spatial distributions of the rates of two types of crime ('theft from the person' and 'violence and sexual offences'). The results of this study consolidate findings from the existing literature in that they support the use of a non-residential measure of the population at risk. The study also evidences that for rates of 'theft from the person' and rates of 'violence and sexual offences', the use of the both the resident and workday populations overestimate the risk of victimisation within urban centres and underestimate the risk in residential areas. The findings of the study highlight the value of estimates of the ambient population for producing

accurate crime rates and support the demand for geographically comprehensive estimates that can be utilised by police forces and policymakers.

6.2 Limitations of the research

This thesis has explored the development of small area estimates of the size of the ambient population in an urban area. The research has successfully developed a model which estimates the size of the ambient population. These estimates were utilised to investigate the impact of alternative measures of the population at risk on the spatial distributions of crime rates. However, there are several limitations of the research which are discussed in this section.

6.2.1 Indicators of the size of the ambient population

The work in Chapter 4 evidenced that indicators of the size of the population, which had been utilised in previous studies, were, surprisingly, not found to have a statistically significant relationship with the number of footfall counts. This finding suggests that the current understanding of indicators of the size of the ambient population is limited and needs to be developed. Furthermore, the indicators of the size of the ambient population will likely have changed within the last eighteen months due to the impacts of the COVID-19 pandemic. Changes in working patterns, which include working from home or hybrid forms of working, have resulted in changes in the size of the ambient population in urban areas. Government restrictions regarding social distancing and venue closures also resulted in fewer people visiting urban areas for leisure purposes. The lasting impact of both changes in working patterns and restrictions to social contact, and the subsequent effect on the size of the ambient population, is yet to be seen.

6.2.2 Limitations of the footfall camera count data

Despite the benefits of footfall camera count data, as discussed in chapters 2 and 3, these data also have some limitations. The primary limitation of footfall camera counts is that their utility is restricted by their representativeness. As footfall cameras only record the numbers of individuals who pass a specific geographic point, they do not enumerate the number of individuals within a wider geographical area. The representativeness is unknown as the ability of these data to enumerate large proportions of the population, particularly within urban areas, is heavily dependent on both the density of cameras and their location.

Additionally, as footfall cameras do not acquire any individual-level information, it is not possible to determine whether individuals pass a camera multiple times and are, therefore, double-counted, i.e., enumerated more than once. Double-counting will lead to overestimating the numbers of people in an area; however, quantifying this poses a significant challenge. Facial recognition algorithms offer a way to enumerate the issue; however, given the significant ethical issues associated with this approach, it is not a viable solution.

6.2.3 Limitations of the manually collected footfall data

The study in Chapter 3 produced estimates of the size of the ambient population using footfall camera counts and OpenStreetMap data with geographically weighted regression. These estimates were validated using manually collected footfall count data recorded at ten sites across the study area. However, the limitations of these data and the data collection process should be acknowledged.

The manually collected footfall data have both spatial and temporal limitations. The data were collected at only ten locations across the study area, thus represent a limited geographic area. As the study enumerated footfall counts at three sites at which footfall cameras were located, these three sites were pre-determined and could not be altered. However, these sites were in close proximity to one another and, consequently, covered a small geographic area, further limiting the

representativeness of the sample. The small sample size is a significant limitation of the data as footfall is a heavily localised phenomenon and is often dependent on land use. Data were collected at each of the ten sites for six hours; thus, they are representative of a limited period of a single day and cannot account for features such as seasonality. Due to the high costs, both temporal and fiscal, of collecting manual footfall counts, it was not possible to conduct counts over a longer period or at a greater number of locations. However, this presents an opportunity for further research, and a more extensive investigation of the accuracy of footfall camera counts would be a novel contribution to the literature.

Human error is a significant limitation of manual data and can impact data accuracy. Two forms of human error may occur in the data collection process: skill-based errors and mistakes. Skill-based errors commonly occur when a task is repetitive, does not require a significant thought process, and if there are distractions in the environment. Due to the nature of capturing high volumes of manual counts in a busy, outdoor environment, these errors are highly likely to occur and cannot be prevented. Data collectors are susceptible to slips of action, such as counting an individual who does not qualify as a pedestrian (for example, someone riding a scooter) and lapses in attention, for example, not counting a pedestrian due to a distraction. Mistakes occur when rules are misapplied or when there are no rules which apply to a given situation. To avoid mistakes, the data collectors undertook a training session to ensure they had a clear understanding of the data collection process and guidance was provided on how to take appropriate action in unforeseen circumstances. While appropriate measures were employed to ensure mistakes were minimised, skill-based errors likely occurred during the collection of the manual footfall counts.

Despite the limitations of the manually collected footfall counts utilised in this research, alternative datasets that can be used to validate footfall camera counts are not currently available. Consequently, these manually collected footfall counts remain

valuable for validating the accuracy of both the footfall camera counts and the model estimates.

6.2.4 The modifiable areal unit problem

Substantial variations in the spatial distributions of crime rates when alternative (i.e., non-residential) measures of the population at risk are employed were demonstrated in Chapter 5. However, it is important to recognise the effects of the modifiable areal unit problem (MAUP) on the results of the research. The MAUP is a source of error that occurs when geographical boundaries are imposed on data. The spatial patterns of aggregated data are affected by the boundaries as the variance structure of the data is altered (Charlton, 2009). Therefore, when different geographical boundaries are employed, the spatial patterns of the data will also differ.

While there are no ways in which the effects can be prevented, it is important to acknowledge the presence of the MAUP and how its effects can be minimised (Wong, 2009). One approach which can be employed to explore the effects of the MAUP is to produce visualisation and conduct spatial analyses at multiple geographical levels. Within Chapter 5, all visualisations and analyses were conducted at the LSOA level as it is the commonly utilised level of aggregation within crime studies. Additionally, smaller units of aggregation could not be employed as a study by Tompson et al. (2015) demonstrated that below the LSOA level, police recorded crime data for England contains spatial error, while employing larger units of aggregation would limit the utility of the results. The alternative solution is to use scale-independent analysis; however, as crime rates are specific to a geographical area, this solution is not appropriate. While these approaches are not suitable for use within this study, it is important to note that the results are likely to be affected by the MAUP.

6.2.5 Generalisability

The work presented in this thesis may be limited by its generalisability. While the estimates of the size of the ambient population produced in Chapter 4 were validated in two locations, the work has not been tested in other urban areas within the UK or internationally. The indicators of the size of the ambient population may vary in different geographical locations due to factors such as variations in human activity patterns and the structures of urban centres. Consequently, there is the need for future research to explore variations in these indicators.

6.2.6 Data equity

The work in this thesis relies heavily on the use of footfall counter data. The presence of footfall counters, as part of wider sensor networks, in urban centres has increased significantly in recent years. These sensor networks and the data they collect aim to collect high-resolution spatio-temporal data which can be used to inform policy-making across a range of areas. However, it is important to note that the spatial distribution of these networks is often uneven and may reinforce growing social injustices within cities. Consequently, those locations which do not have sensor data available will be further impacted due to their inability to produce robust estimates of the size of the ambient population which would benefit public safety.

6.3 Recommendations for future work

There are several avenues for potential research identified by the work in this thesis. The most beneficial avenue for future work may be the further development of the model produced in Chapter 4. To improve this model, through an increased predictive capacity, a better understanding of the indicators of the size of the ambient population is required. Despite the extensive work undertaken to identify indicators that have a statistically significant relationship with footfall counts, the identification of additional indicators would allow estimates of the size of the population to be produced with higher levels of accuracy. Assessing the relationship between the number of footfall counts and relevant indicators of the size of the ambient population post-pandemic would be a valuable contribution to the literature. The

model estimates produced would, consequently, be a more valuable asset within crime studies and other areas of research and policymaking.

This thesis explored the impact of different measures of the population at risk on the spatial distribution of crime rates. However, the study could be expanded, by examining other crime types. This thesis only explores theft from the person and violence and sexual offences, as these crimes target individuals and, therefore, the locations of offences will vary as populations move throughout space. Consequently, residential measures of the population at risk are inappropriate, as they do not accurately reflect the number of individuals at risk from these particular crime types. However, there remains an opportunity to investigate the impact of the ambient population on the spatial distribution of crimes that do not target individuals, such as daytime residential burglary, vehicle crime, or shoplifting. The exploration of these crime types would be a valid avenue for future research, as routine activity theory states that for a crime to occur, three things must converge in space and time; a target, an offender and the absence of a capable guardian (Cohen and Felson, 1979). As the size of the ambient population increases, there is likely to be an increase in the number of capable guardians present who may deter an offender from committing a crime. For example, an offender may be less likely to commit a daytime residential burglary if there are high numbers of people in the area to witness the crime. Therefore, the size of the ambient population may also be an appropriate measure of the population at risk for crime types that do not target individuals.

Another opportunity for future research is the production of fine-grained temporal estimates of the size of the population at hourly intervals, which would be particularly valuable within crime studies and hazard management. Due to the highly temporal nature of crime events and hazards, hourly estimates of the size of the ambient population would allow for more fine-grained analysis and ultimately allowing for a more in-depth understanding of these phenomena.

Additionally, there remains an opportunity to investigate the utility of estimates of the size of the ambient population within research areas outside of crime studies. Relevant areas of study include exploring the exposure of the ambient population to air pollution and investigating the use of estimates of the size of the ambient population within hazard management.

6.4 Outlook and concluding remarks

Small area estimates of the size of the ambient population, particularly in urban areas, are essential to a wide range of applications, within both research and policymaking. Such estimates can allow a better understanding of urban phenomena to be developed and enable more effective decision making. The demand for easily reproducible, accurate estimates will increase as levels of urbanisation continue to rise globally, and climate change poses an increasing risk to urbanised areas. These estimates of the size of the ambient population can then be used to mitigate impacts of urbanisation and climate change through their use in informing emergency planning, improving hazard management, and monitoring exposure to noise and air pollution.

The COVID-19 pandemic and the subsequent restrictions to limit social interactions have profoundly impacted human activity patterns. Shifts towards remote working within many sectors, travel restrictions, and limitations on leisure activities to ensure social distancing have all resulted in significant changes in the size and locations of the ambient population. To explore and quantify the extent of the impact of the COVID-19 pandemic on human activity patterns, estimates of the size of the ambient population would be beneficial. The COVID-19 pandemic reinforces both the value of estimates of the size of the ambient population and the importance of the frequent and systematic collection of population data.

This thesis has explored the development of small area estimates of the size of the ambient population in an urban area. Through the use of geographically weighted regression and open-source data, an approach to produce estimates of the size of the ambient population has been developed. Estimates of the size of the ambient population produced using this approach were then successfully employed to explore the impact of different measures of the population at risk on the spatial distribution of crime rates. The approach developed and presented within this thesis enables the production estimates of the size of the ambient population which have the potential to be effectively utilised within research and policymaking.

Reference list

Charlton, M. 2009. Quantitative Data. *International Encyclopedia of Human Geography*., pp.19–26.

Cohen, L.E. and Felson, M. 1979. Social change and crime rate trends: A routine activity approach. *American sociological review*., pp.588–608.

Tompson, L., Johnson, S., Ashby, M., Perkins, C. and Edwards, P. 2015. UK open source crime data: Accuracy and possibilities for research. *Cartography and Geographic Information Science*. **42**(2).

Whipp, A. 2021. Manually Collected Footfall Counts in Leeds. *Consumer Data Research Centre*. [Online]. [Accessed 25 November 2021]. Available from: https://data.cdrc.ac.uk/dataset/manually-collected-footfall-counts-leeds.

Wong, D.W. 2009. Modifiable Areal Unit Problem *In*: *International Encyclopedia of Human Geography*.

Appendix

Appendix A: Chapter 5 supplementary tables and figures

Appendix A Table 1 Descriptive statistics for the data used to calculate the rates of theft from the person.

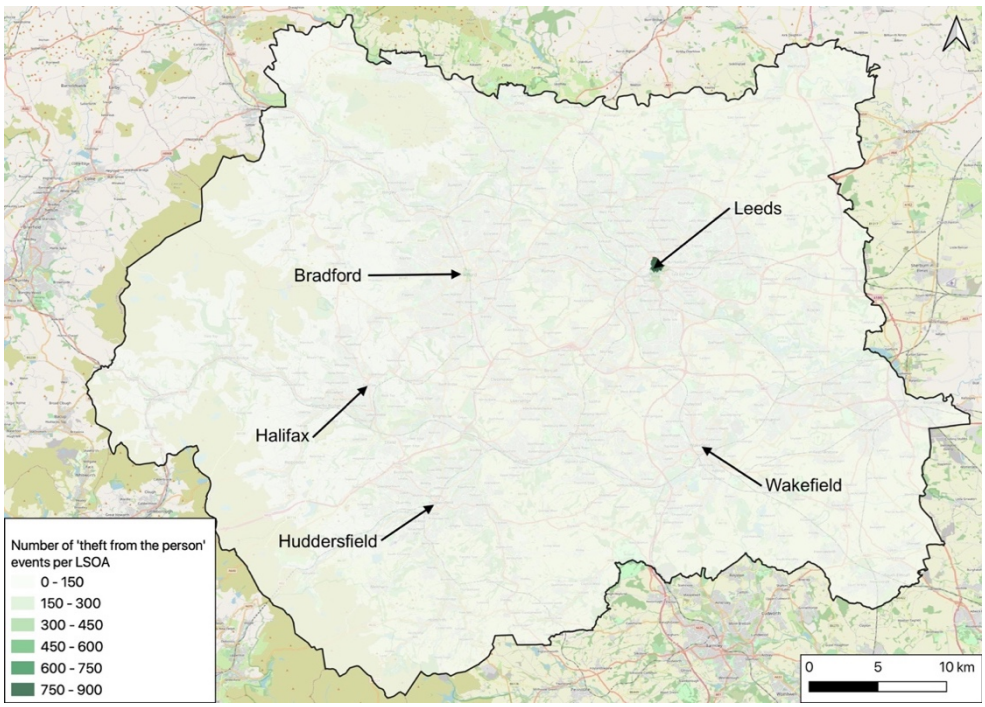
Data	Source, producer, and date	Sum	Mean	SD	Min. value	Max. value
Theft from the person data	Police, police.data.uk, 2019	3749.000	2.701	23.129	0.000	770.000
Violence and sexual offences data	Police, police.data.uk, 2019	1388.000	79.315	85.937	3.000	1781.000
Usual resident population	Census of England and Wales, Office for National Statistics, 2011.	2,226,058.000	1603.788	277.537	1011.000	4156.000
Workday population	Census of England and Wales, Office for National Statistics, 2011.	1,036,058.000	746.439	1602.690	58.000	32261.000

Footfall counts	Produced using the method outlined in Section 3.24, OpenStreetMap, 2021.	890432.897	1213.124	3190.626	0.000	57577.998
The ambient population	OpenStreetMap, 2021; Census of England and Wales, Office for National Statistics, 2011.	1,926,488.000	1387.959	3557.072	58.000	89839.000

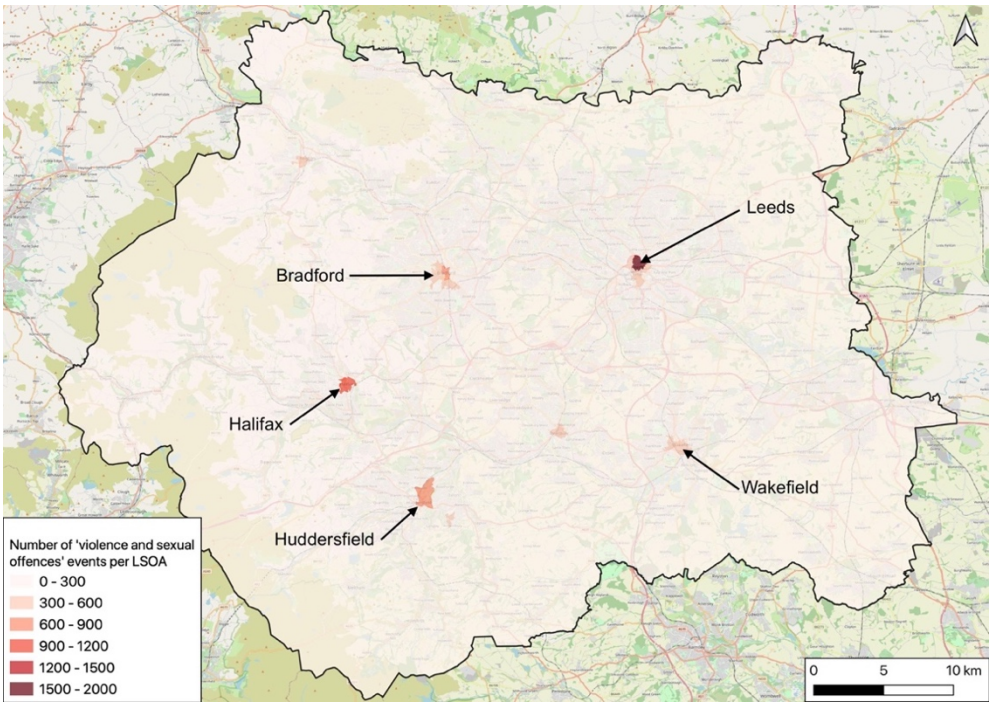
Appendix A Table 2 Descriptive statistics for the rates of theft from the person and violence and sexual offences, calculated using three different measures of the population at risk.

Variable name	Crime type	Measure of the population at risk	Mean	Standard deviation	Min. value	Max. value
Resident theft	Theft from the person	Usual resident population	1.594	10.877	0.000	274.314
Workday theft	Theft from the person	Workday population	3.119	5.952	0.000	74.084

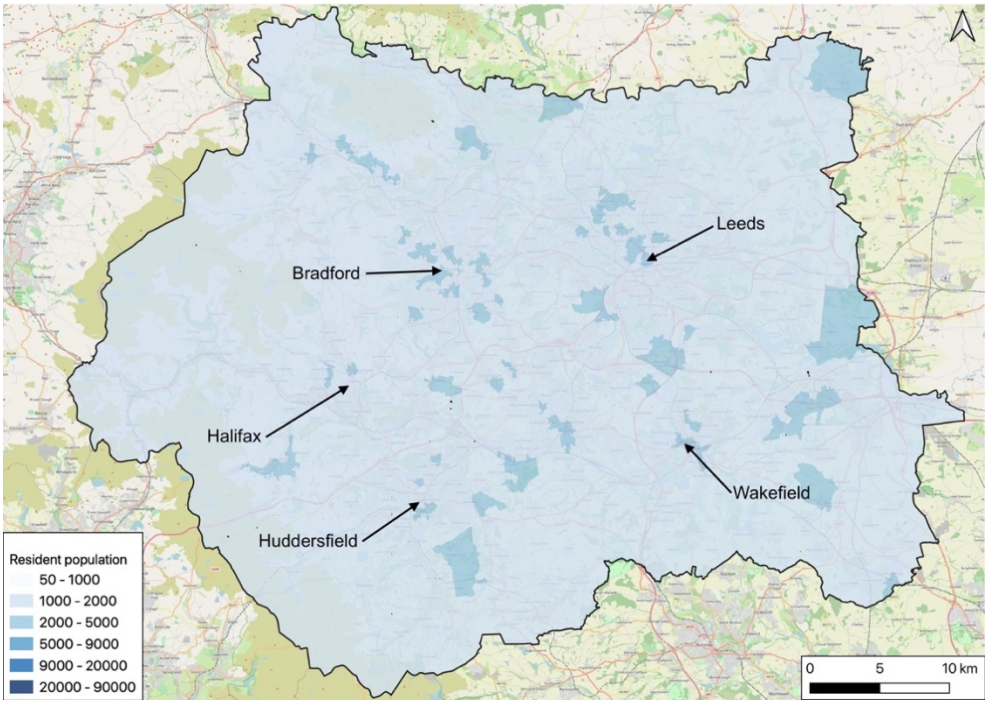
Ambient theft	Theft from the person	Ambient population	2.477	5.339	0.000	74.074
Resident violence	Violence and sexual offences	Usual resident population	49.118	48.785	1.856	764.499
Workday violence	Violence and sexual offences	Workday population	228.762	268.761	4.509	2916.667
Ambient violence	Violence and sexual offences	Ambient population	191.430	253.598	1.914	2916.667



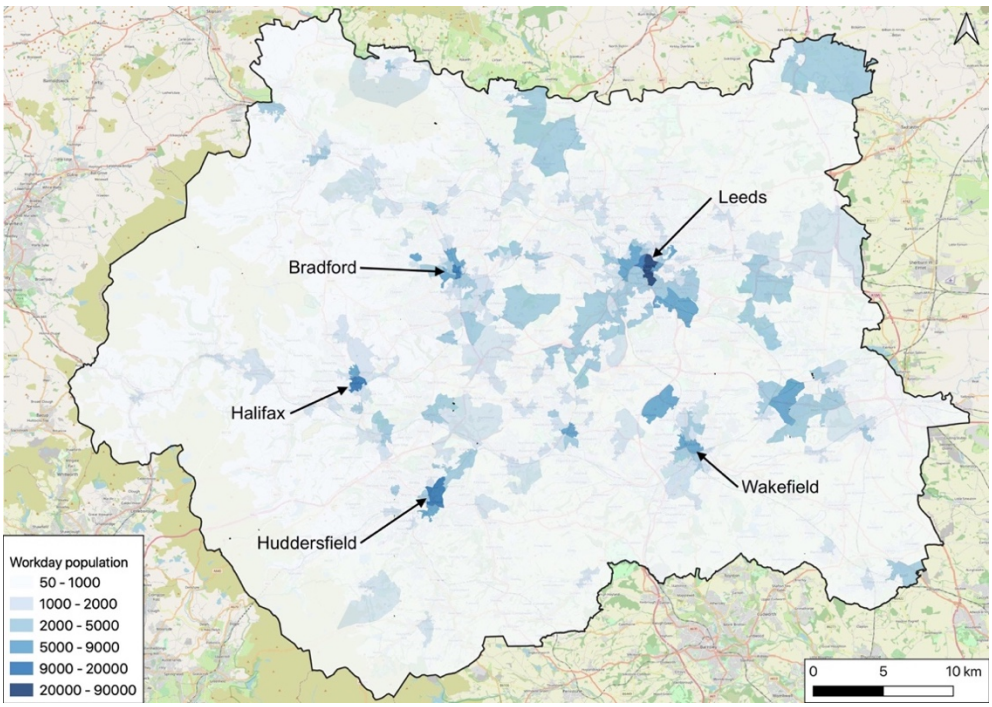
Appendix A Figure 1 The spatial distribution of the number of theft from the person and violence and 952 sexual offences events per LSOA (Basemap: © OpenStreetMap contributors, 2021 and 953 Ordnance Survey data © Crown copyright and database right 2010-19).



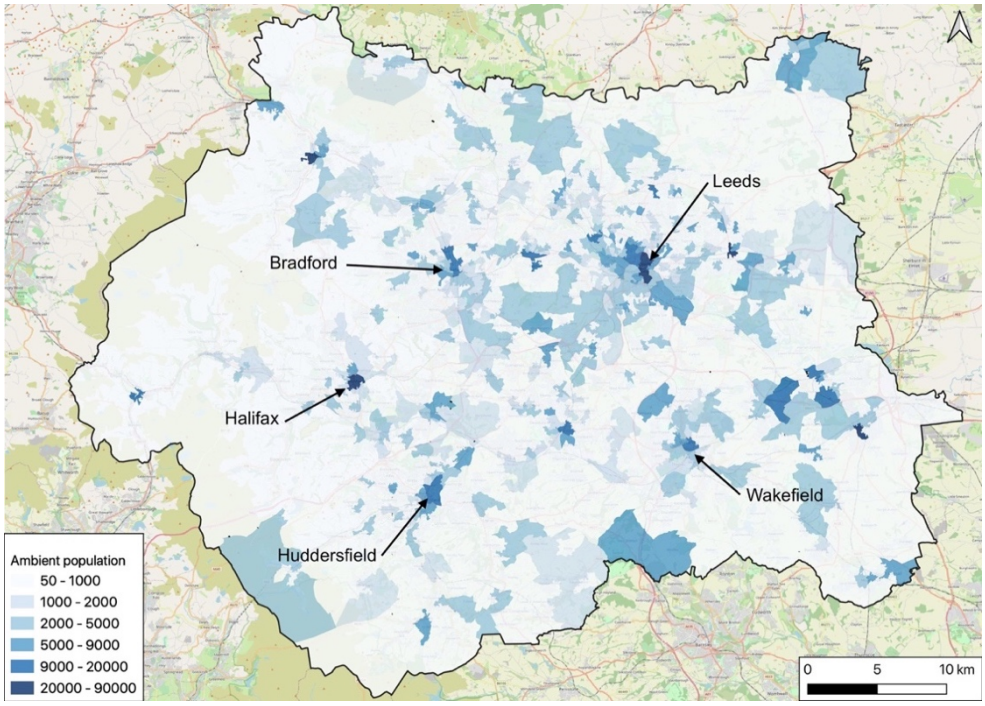
Appendix A Figure 2 The spatial distribution of the number of theft from the person and violence and sexual offences events per LSOA (Basemap: © OpenStreetMap contributors, 2021 and Ordnance Survey data © Crown copyright and database right 2010-19).



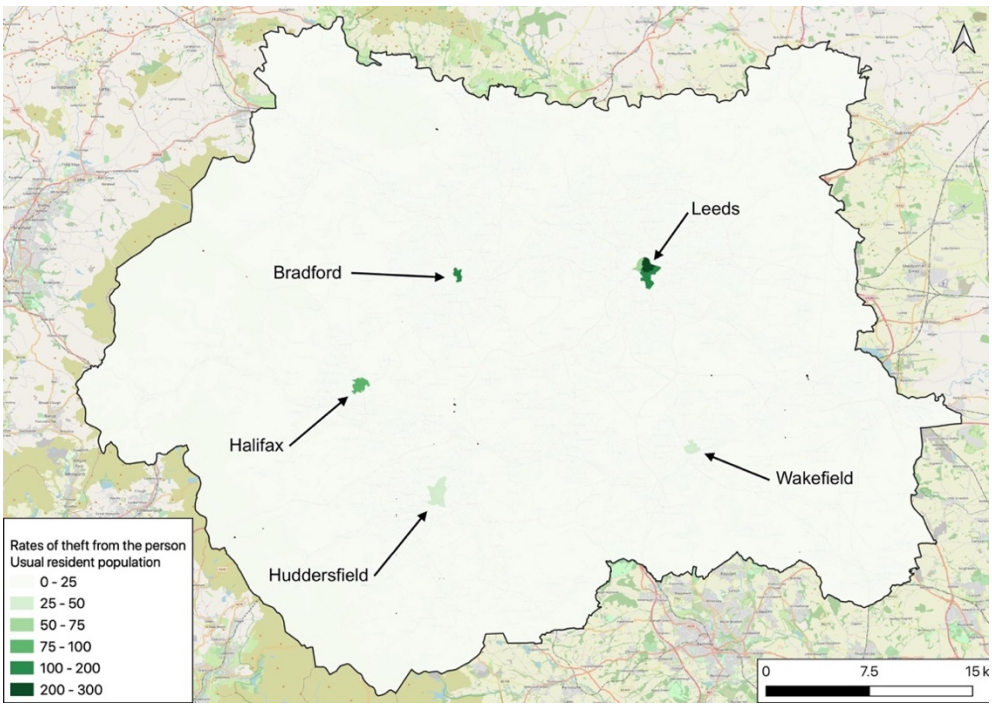
Appendix A Figure 3 The usual resident population per LSOA (Basemap: © OpenStreetMap contributors, 2021 and Ordnance Survey data © Crown copyright and database right 2010-19).



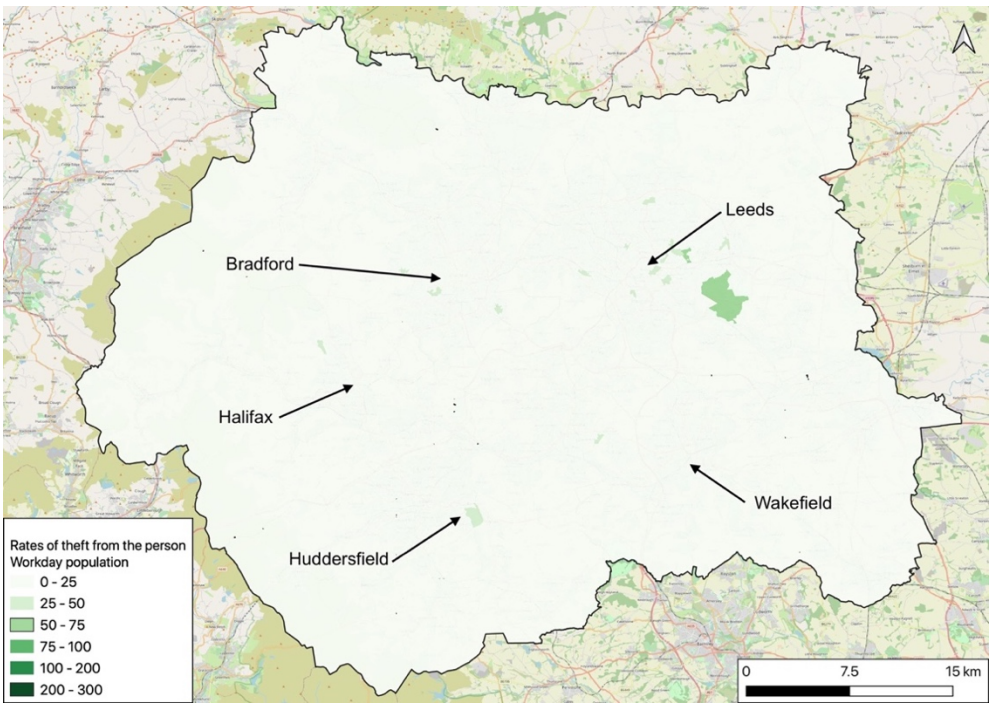
Appendix A Figure 4 The workday population per LSOA (Basemap: © OpenStreetMap contributors, 2021 and Ordnance Survey data © Crown copyright and database right 2010-19).



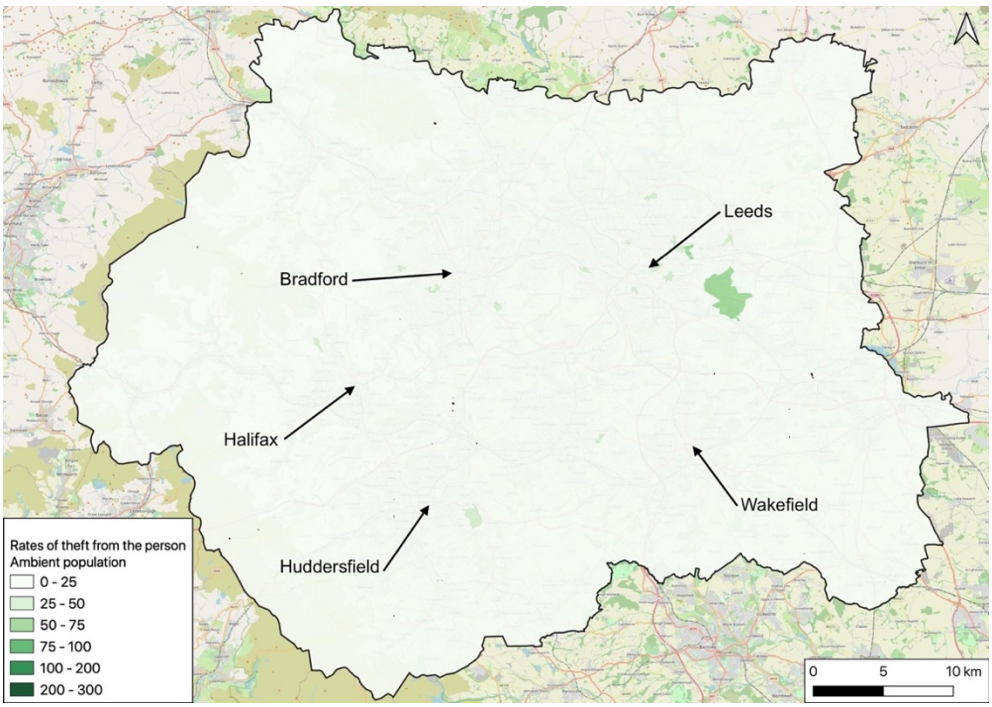
Appendix A Figure 5 The ambient population per LSOA (Basemap: © OpenStreetMap contributors, 2021 and Ordnance Survey data © Crown copyright and database right 2010-19).



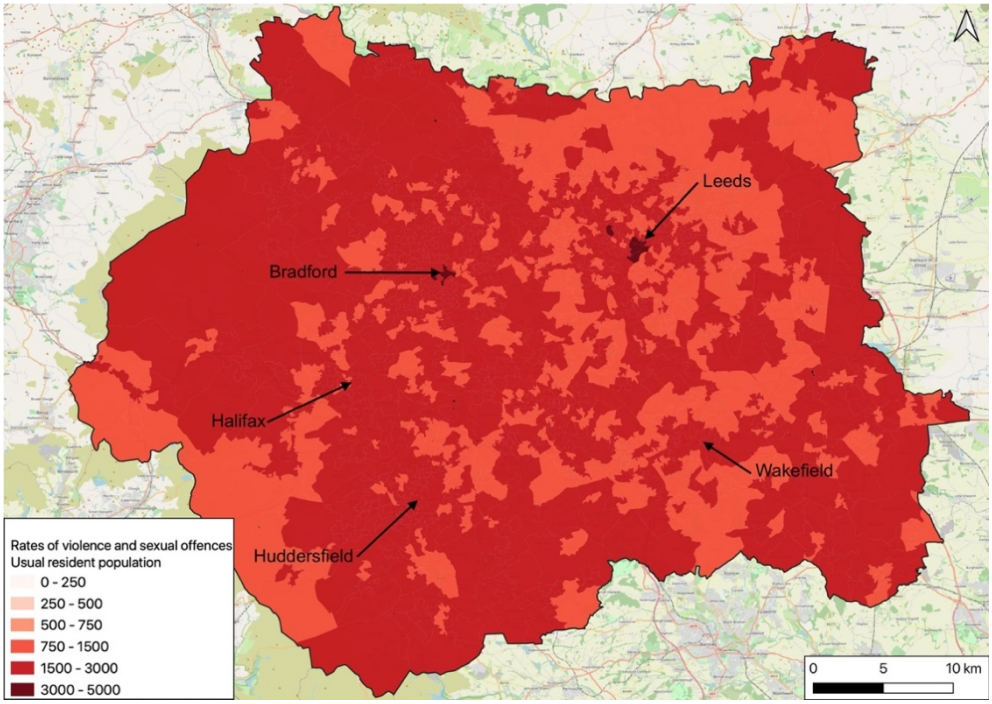
Appendix A Figure 6 Rates of theft from the person per 1000 people per LSOA, calculated using estimates of the resident population (Basemap: © OpenStreetMap contributors, 2021 and Ordnance Survey data © Crown copyright and database right 2010-19).



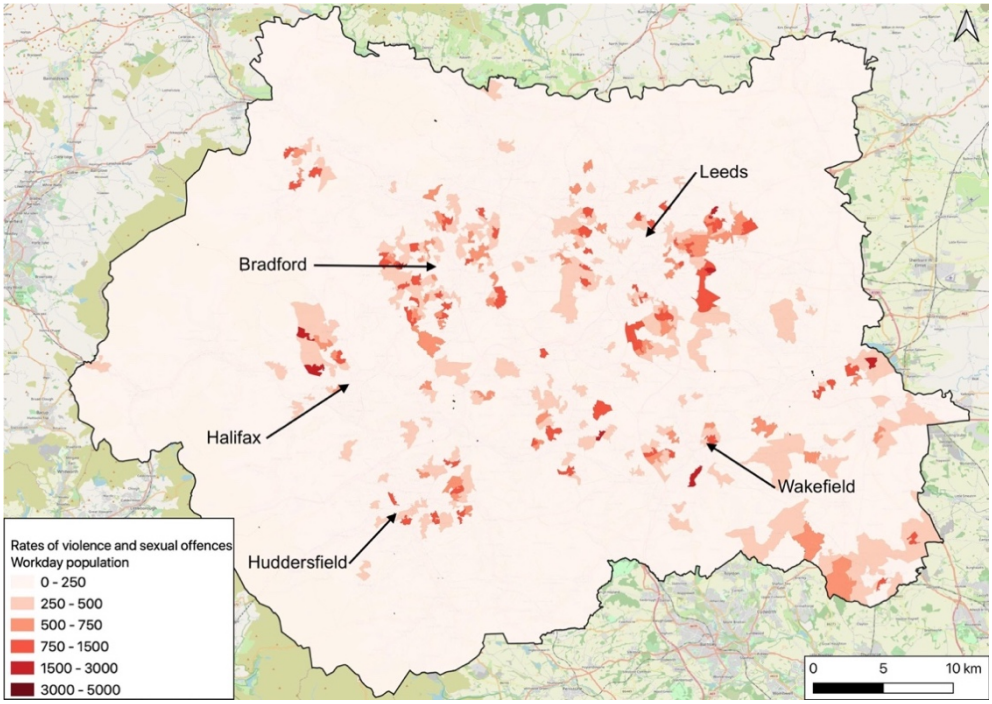
Appendix A Figure 7 Rates of theft from the person per 1000 people per LSOA, calculated using estimates of the workday population (Basemap: © OpenStreetMap contributors, 2021 and Ordnance Survey data © Crown copyright and database right 2010-19).



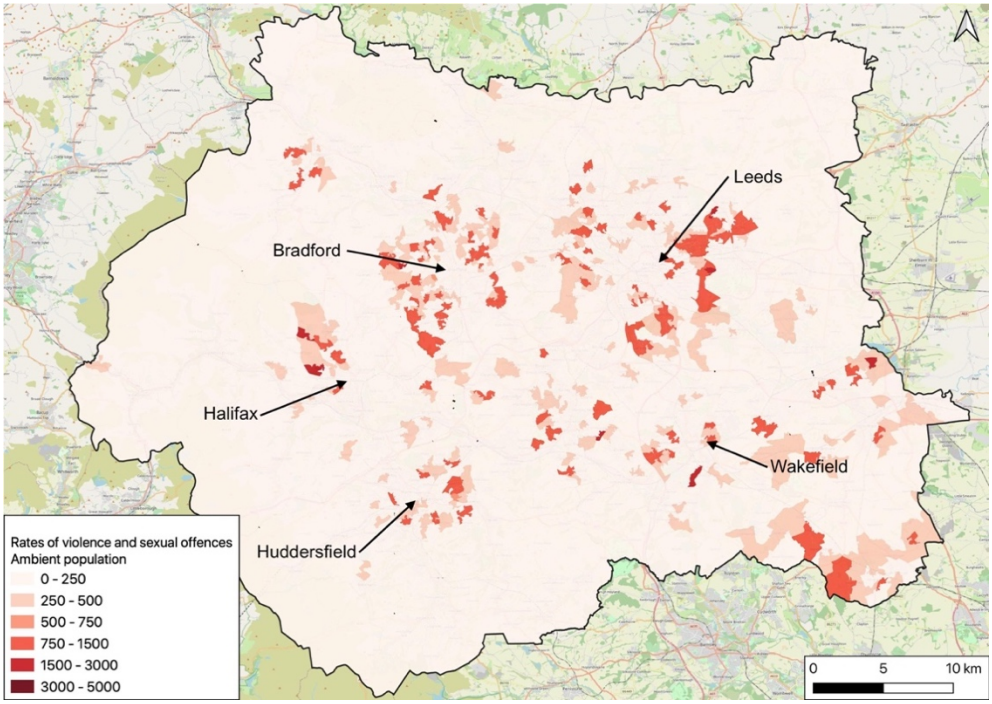
Appendix A Figure 8 Rates of theft from the person per 1000 people per LSOA, calculated using estimates of the ambient population (Basemap: © OpenStreetMap contributors, 2021 and Ordnance Survey data © Crown copyright and database right 2010-19).



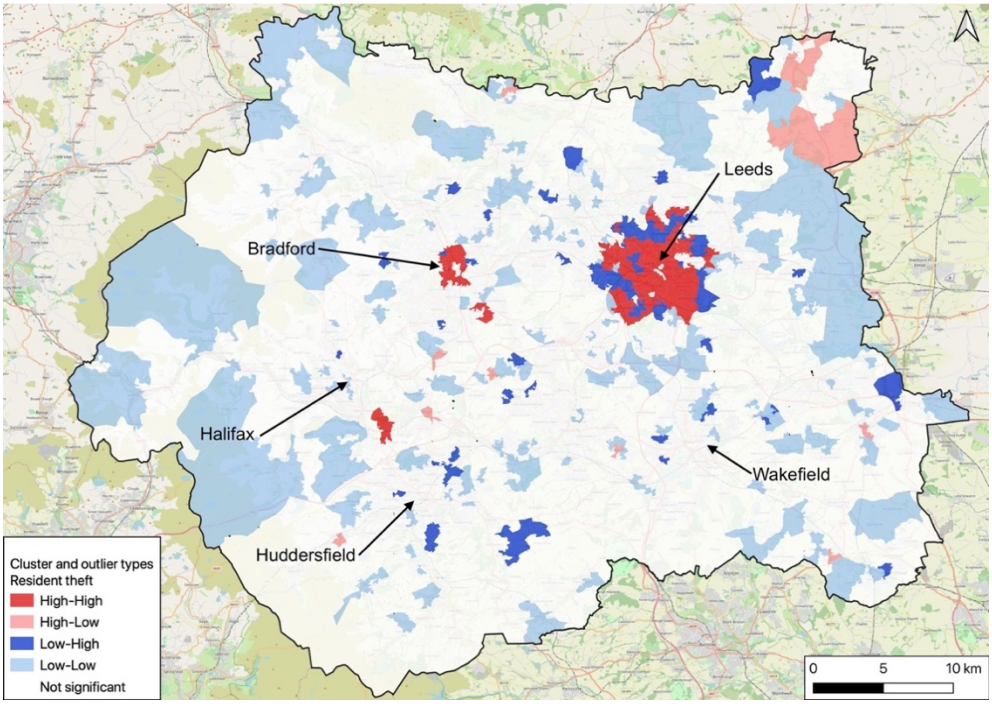
Appendix A Figure 9 Rates of violence and sexual offences per 1000 people per LSOA, calculated using estimates of the resident population (Basemap: © OpenStreetMap contributors, 2021 and Ordnance Survey data © Crown copyright and database right 2010-19).



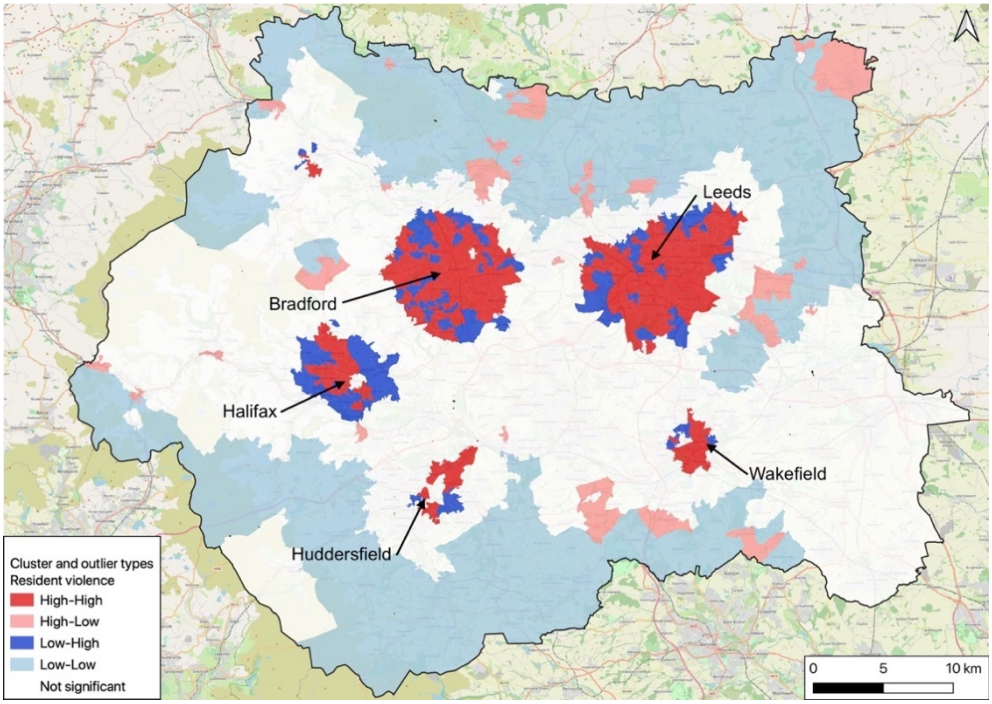
Appendix A Figure 10 Rates of violence and sexual offences per 1000 people per LSOA, calculated using estimates of the workday population (Basemap: © OpenStreetMap contributors, 2021 and Ordnance Survey data © Crown copyright and database right 2010-19).



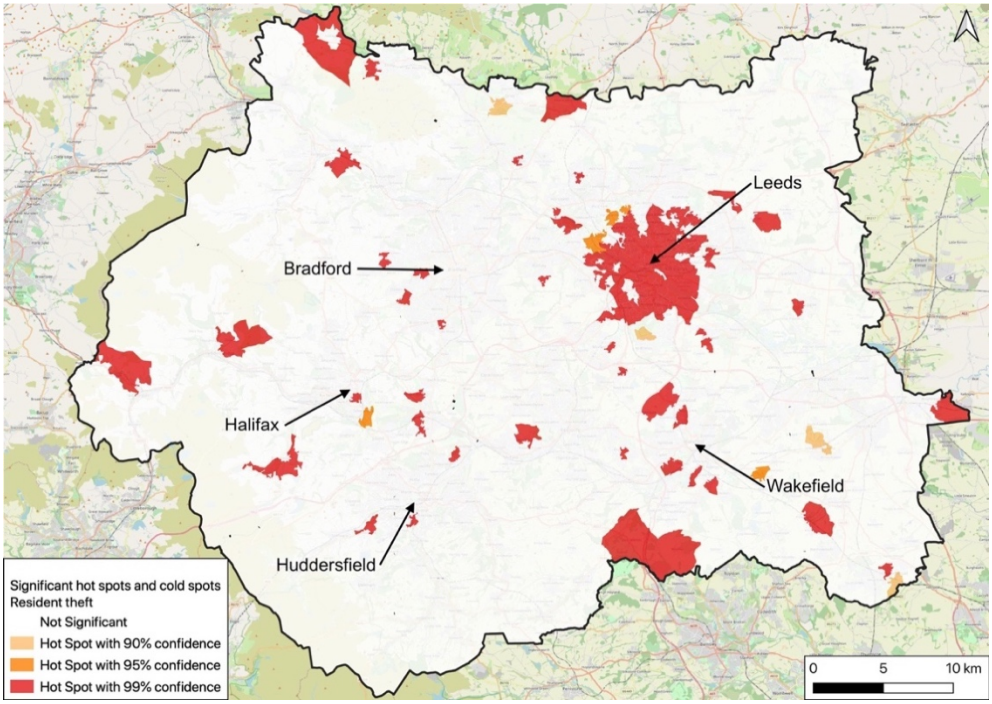
Appendix A Figure 11 Rates of violence and sexual offences per 1000 people per LSOA, calculated using estimates of the ambient population (Basemap: © OpenStreetMap contributors, 2021 and Ordnance Survey data © Crown copyright and database right 2010-19).



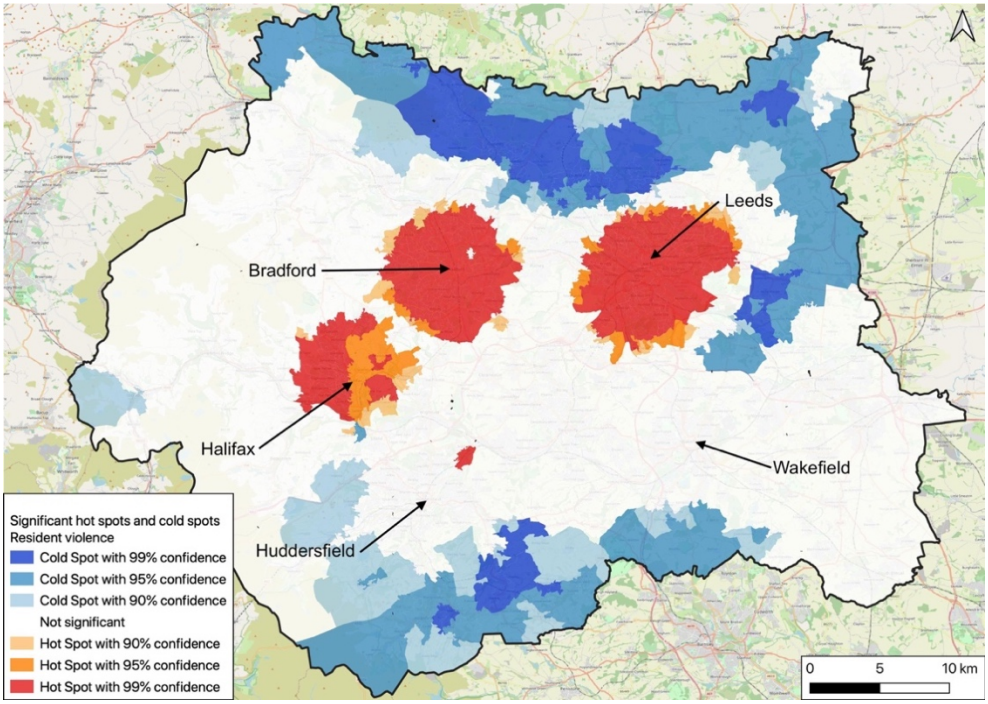
Appendix A Figure 12 The spatial distribution of clusters and outliers for rates of theft calculated using the resident population (Basemap: © OpenStreetMap contributors, 2021 and Ordnance Survey data © Crown copyright and database right 2010-19).



Appendix A Figure 13 The spatial distribution of clusters and outliers for rates of violence and sexual offences calculated using the resident population (Basemap: © OpenStreetMap contributors, 2021 and Ordnance Survey data © Crown copyright and database right 2010-19).



Appendix A Figure 14 The spatial distribution of hot spots and cold spots for rates of theft calculated using the resident population (Basemap: © OpenStreetMap contributors, 2021 and Ordnance Survey data © Crown copyright and database right 2010-19).



Appendix A Figure 15 The spatial distribution of hot spots and cold spots for rates of violence and sexual offenses calculated using the resident population (Basemap: © OpenStreetMap contributors, 2021 and Ordnance Survey data © Crown copyright and database right 2010-19).

Appendix B: Manual count metadata

This appendix includes an in-depth description of the manual footfall count dataset developed and used to validate work in Chapter 4 of the thesis. The data for this research were funded by the European Research Council under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 757455), and are openly available via the Consumer Data Research Centre, an Economic and Social Research Council Data Investment (grant ES/L011840/1;ES/L011891/1).

The manual footfall count data were utilised to:

1. Validate estimates of the ambient populations produced using a statistical model in three geographic areas within the Metropolitan Borough of Leeds, UK: Headingley, Wetherby and Leeds city centre.
2. Validate the accuracy of footfall camera data captured in Leeds city centre

Manual footfall counts were collected at ten sites between the 5th to the 9th of July 2021 between 10:00 and 16:00 each day. At the time of data collection, footfall cameras were installed at three of the ten sites: Briggate, Headrow and Commercial Street. Footfall counts were collected at two sites per day and three data collectors were located at each site. Data collectors were replaced during breaks to ensure continuity.

The dataset produced consists of 24 files providing manual footfall counts at ten locations. Each file contains the collection date and a timestamp, with each timestamp representing one count/one pedestrian. For each location, the files are numbered 1 to 3, representing the three data collectors. Due to data collection issues, timestamped data for the 5th of July 2021 at Bond Street and Vicar Lane are not available; however, the total counts recorded by each data collector at these locations are included in the dataset.

Counts were logged using the iOS application Counter+ which is available free of charge via the App Store. The application allows counts to be logged with an associated timestamp, enabling the data to be used in temporal analysis. The data can be exported from the application as a .txt file for further analysis. The application stores a maximum of 1000 records, thus the data must be exported prior to reaching over 1000 counts to avoid data loss.

The data collectors were instructed to count all pedestrians who were not cycling, skateboarding, scootering, or using any form of transport. Individuals using motorised mobility aids were counted. All children were counted including those in pushchairs or being carried. Data collectors were located in positions which did not impede the flow of pedestrians and allowed them to have an uninterrupted view of the count location. To ensure that all data collectors were enumerating pedestrians in the same geographical area, pedestrians were counted as they passed a pre-determined physical marker, for example a lamppost. At locations with a footfall camera installed, the footfall camera was the selected physical marker.

On Tuesday 6th July, there was heavy rain from 10:00 to 16:00 at both locations, North Street (Wetherby) and B6167 Otley Road (Headingley) which may impact the number of pedestrians. During the other data collection days, the weather was dry with no cloud between 10:00 and 16:00. On Friday 9th July at Commercial Street, a sales cart was in proximity to the footfall camera. This partially obscured the data collectors' view of pedestrian flows; it is unknown whether the sales cart would impact counts recorded by the footfall camera.

The data were collected by Annabel Elizabeth Whipp, Sedar Olmez, Ellie Marfleet, Deborah Olukan, Fredide Wallace, and Amandine Bodet Lefevre. Annabel Elizabeth Whipp was responsible for project management and data processing.