MODELLING SOCIO-SPATIAL DYNAMICS FROM REAL-TIME DATA

Towards a context-aware framework for modelling behaviour change in urban space from mobile data

by Sharon Richardson

May 2020

A thesis submitted for the degree of Doctor of Philosophy at the Centre for Advanced Spatial Analysis, The Bartlett Faculty of the Built Environment, University College London.

PAGE INTENTIONALLY LEFT BLANK

Declaration of Authorship

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

Signed: 

Dated: 31st May 2020

PAGE INTENTIONALLY LEFT BLANK

Abstract

This thesis introduces a framework for modelling the social dynamic of an urban landscape from multiple and disparate real-time datasets. It seeks to bridge the gap between artificial simulations of human behaviour and periodic real-world observations. The approach is data-intensive, adopting open-source programmatic and visual analytics. The result is a framework that can rapidly produce contextual insights from samples of real-world human activity – behavioural data traces. The framework can be adopted standalone or integrated with other models to produce a more comprehensive understanding of people-place experiences and how context affects behaviour.

The research is interdisciplinary. It applies emerging techniques in cognitive and spatial data sciences to extract and analyse latent information from behavioural data traces located in space and time. Three sources are evaluated: mobile device connectivity to a public Wi-Fi network, readings emitted by an installed mobile app, and volunteered status updates. The outcome is a framework that can sample data about real-world activities at street-level and reveal contextual variations in people-place experiences, from cultural and seasonal conditions that create the 'social heartbeat' of a landscape to the arrhythmic impact of abnormal events. By continuously or frequently sampling reality, the framework can become self-calibrating, adapting to developments in land-use potential and cultural influences over time. It also enables 'opportunistic' geographic information science: the study of unexpected real-world phenomena as and when they occur.

The novel contribution of this thesis is to demonstrate the need to improve understanding of and theories about human-environment interactions by incorporating context-specific learning into urban models of behaviour. The framework presents an alternative to abstract generalisations by revealing the variability of human behaviour in public open spaces, where conditions are uncertain and changeable. It offers the potential to create a closer representation of reality and anticipate or recommend behaviour change in response to conditions as they emerge.

PAGE INTENTIONALLY LEFT BLANK

Impact Statement

The contribution of this thesis is a framework for modelling population behaviours and behaviour change in urban space from multiple and disparate sources of real-time data. It produces contextual insights that demonstrate the quantifiable variability in human behaviour across space, time and situation, and offers numerous benefits both inside and outside academia.

Within academia, the thesis lays the foundations for advancing the field of contextual analysis and visualisation of social phenomena. It offers a new lens for studying human behaviour and opens up several routes for researching behavioural aspects of urban landscapes, both in developing new data-driven theories of social and spatial behaviour, and in developing new techniques for the analysis and visualisation of social phenomena. Findings from the research have already been accepted and presented at several international academic conferences and published in a leading peer-reviewed journal: IEEE Pervasive Computing. The research has also been featured in UCL's The Bartlett Annual Review and was the department's winning entry for 2016 in the annual Three Minute Thesis competition open to doctoral researchers worldwide.

Within industry, this thesis demonstrates the potential for building real-time decision systems that are context-aware and able to adapt to uncertain and changeable social atmospheres. It has also provided an evidence-based approach to studying social phenomena in urban public space at new spatial and temporal scales. Research findings have generated substantial interest, particularly within government agencies responsible for the public realm. There is a demand for knowledge about people-place experiences and how they vary during different situations such as large-scale events and unexpected incidents. Visual outputs have featured in presentations delivered by the London Legacy Development Corporation, Intel Research Labs, and the Chief Digital Officer of London. The author has also been invited to present the research at various industry conferences including the annual European IoT Forum in Brussels, to an audience of senior leaders across commerce and government. It has generated investment interest to continue development.

The number of mobile and embedded devices connected to the Internet is forecast to grow to 75 billion by 2025, a fivefold increase in the decade since this research was first envisaged. It will generate social and environmental data at unprecedented scales and create demand for new techniques to convert such data into contextual insights about urban social landscapes. The greatest impact from this research in both theory and practice is in advancing a scientific method that embraces context and uncertainty when forecasting human behaviour in real-world situations where decisions carry consequences for the lives and livelihoods of people.

PAGE INTENTIONALLY LEFT BLANK

Acknowledgements

This thesis is the culmination of four years of doctoral research that would not have been possible without the support and encouragement of my family, friends, supervisors, fellow doctoral candidates, and the many people with whom I have had the opportunity to work with and consult for advice across University College London, the wider academic community and industry partners. I have been incredibly fortunate to call on the help of so many. Any flaws in the work are my own.

First and foremost, I wholeheartedly thank my supervisors, Professor Andrew Hudson-Smith and Dr Hannah Fry, for their continued insights and guidance throughout the process, from cultivating the seed of an idea to the production of this thesis and the many ups and downs that occurred in between. I am also immensely grateful to the department and administrative team led by Sonja Curtis for providing a such a supportive working environment and the opportunity to present the findings from my research at several industry and academic conferences.

I owe a debt of gratitude to the EPSRC for their grant to help fund this research, to the London Legacy Development Corporation and OpenSignal for providing access to data sources that made the research possible, and the Intel Collaborative Research Institute for supporting a programme of study at the Queen Elizabeth Olympic Park. I also owe thanks to the open-source community for providing the tools and programming guidance that have been crucial to the research.

I dedicate this thesis to my parents, whose love and support has been overwhelming. For them, this has not been four years but more like twenty-five years from when I first became fascinated, bordering on obsessed, with how access to information via new technology can change behaviour. This has been a long journey. I am indebted to my mother for patiently listening to my worries and frustrations and reading through multiple drafts. Any foibles that remain are the result of my never-ending tweaks. A final mention goes to Sammy, who was so often by my side at those times when completing this thesis felt insurmountable and helped to keep everything in perspective.

PAGE INTENTIONALLY LEFT BLANK

Table of Contents

List of Figures

List of Tables

PAGE INTENTIONALLY LEFT BLANK

List of Abbreviations

AP Access Point. Wireless Access Point (or router) for transmitting data between devices and the Internet via a wireless communications network

API Application Programming Interface. Programmatic method to submit data to, or receive data from, a web site

CCTV Closed-circuit television

CE Connected Environment, a physical location with networked sensors and actuators

GDPR General Data Protection Regulation

GIS Geographic Information Science/System

GPS Global Positioning System, a satellite-based navigation system that provides geolocation information to a GPS receiver with unobstructed line of sight to four or more satellites orbiting the Earth

IoT Internet of Things

LLDC London Legacy Development Corporation

MAUP Modifiable Area Unit Problem: The outcomes from an areal analysis can be influenced by the boundaries drawn to aggregate data spatially

MCUP Modifiable Context Unit Problem: The outcomes from a context-specific analysis can be influenced by the categorisation and combination of circumstances to form contexts

MIUP Modifiable Interaction Unit Problem. The outcomes from using a population measure can be influenced by what interactions are included to calculate the population present

MLUP Modifiable Language Unit Problem: The outcomes from language-based analysis can be influenced by what words are excluded or combined to represent a single meaning

MTUP Modifiable Time Unit Problem: The outcomes from time-series analysis can be influenced by the time range and interval chosen, such as hourly across a 24-hour period

OECD The Organisation for Economic Co-operation and Development

OED Oxford English Dictionary. The version used this thesis was the Concise Oxford English Dictionary, Eleventh Edition, 2004. (Print edition)

ONS Office for National Statistics. Public agency in the United Kingdom.

ORNL Oak Ridge National Laboratory (producers of LandScan™)

OS OpenSignal (not Ordnance Survey unless specified)

QEOP Queen Elizabeth Olympic Park

UCL University College London

URL Uniform Resource Locator (internet web site address beginning with http:// or https://)

WHO World Health Organisation

Wi-Fi Wireless network

PAGE INTENTIONALLY LEFT BLANK

1 Introduction

1.1 Introduction and Motivation

This thesis introduces a contextual framework for learning and modelling the social dynamic of urban public space. It uses real-time data emitted from mobile devices as a digital source of real-world observations about people-place interactions.

The broad research question is: 'To what extent can mobile data, by providing a continuous sample of real-world observations, be used to sense and learn the social dynamic of a landscape?' In the context of this research, social dynamic refers to the size, distribution, actions and experiences of people present, and how it varies for different contexts, including both recurring activities such as peak versus off-peak hours during the working week, and infrequent or unexpected situations such as large-scale events and emergency incidents. Landscape refers to an area that can be comfortably traversed on foot, such as a local neighbourhood. Historically, data about human dynamics at street-level has been difficult to acquire and models have mostly relied on theoretical approaches such as applying Newtonian principles to simulate behaviour (Ball, 2015) supported by infrequent field studies to capture real-world observations. Mobile data offers the potential to take an evidence-based approach at scale and develop new models that incorporate the variability of behaviour in response to uncertain and changeable circumstances.

Under pressure from growing and denser populations, and increasing climate uncertainty, many cities have begun to explore the use of data and digital technology to create environments able to adapt in real-time to changeable conditions. Indeed, the Mayor of London, England, launched the Smart London Plan in 2013 (Smart London Board, 2013) outlining the need to use such methods to create 'smart sustainable districts' that can support citizens with a good quality of life whilst consuming fewer resources, generating less waste or pollution, and being resilient to unpredictable phenomena such as extreme weather events. Central to creating such environments is an emerging field called context-aware computing, defined as: "*a general class of mobile systems that can sense their physical environment, and adapt their behaviour accordingly*" (Mohan & Singh, 2013). For such systems to be effective in urban environments requires the ability to sense and respond to the social conditions of the landscape. That is the motivation for this research.

Within the Smart London plan, it was announced that the Queen Elizabeth Olympic Park (QEOP) in Stratford, East London, would be a testbed for data-driven innovations, launching in 2016. QEOP is one of the largest urban green spaces in Europe and is within a neighbourhood currently undergoing substantial redevelopment as part of the London 2012 Olympics legacy (LLDC, 2016). Thanks to a collaboration with the London Legacy Development Corporation (LLDC), the QEOP is the primary landscape for the core research in this thesis. Multiple real-time data sources acquired during 2016 and 2017 are explored for the potential to reveal context-specific local population behaviours. The resulting framework is then applied to three other landscapes within London that each experienced a major incident during 2017. Two London bridges suffered terror attacks whilst

Oxford Circus, one of the busiest shopping destinations in Europe (BNP Paribas Real Estate, 2017), experienced a situation initially believed to be a terror attack but that was later found to be a false alarm. Using the framework, a sample of digitised interactions is used to profile the social dynamic of each landscape and anticipate social conditions at the time of the incident.

There are multiple potential applications for this research. First, learning the social dynamic of a landscape would result in an improved indicator of the real-time population within a landscape. The effectiveness of an intervention or response to a situation may depend on knowing the size of the population present and their motivations for being present. Second, public places such as parks and green spaces are known to benefit the health and wellbeing of people and help mitigate pollution and climate change in cities. However, evidence of their use and value has been lacking (Orr, Paskins, & Chaytor, 2014). The framework has the potential to provide a low-cost and scalable source of insights to inform public policy. Third, the analysis of real-world behaviours from digitised interactions could be brought full-circle and assist those present within the landscape. We intuitively sense the vibe of our surroundings, but our senses have a limited horizon. The framework could extend that horizon and provide insights about current conditions direct to individuals present, via their mobile devices or through adaptive interfaces such as digital signage embedded within the landscape, to inform and assist spatial choices.

A novel benefit of using real-time data is the ability to continuously learn about a landscape. This can include both cultural shifts that take place over longer periods and sudden acute shocks. For example, the arrival of mobile phones has been shown to have altered walking patterns: people viewing their phone tend to walk more slowly whilst people wearing headphones walk faster and people are more likely to pause before entering subways (Vanderbilt, 2012). Whereas the typical working week in the UK during the 20th century would involve being present at a single place of work from Monday to Friday, flexible working has recently become more prevalent and may have a noticeable impact on population dynamics in terms of weekday variation. The most popular days for remote working appear to be Monday and Friday. By continuously or frequently sampling reality through digitised spatial and social interactions, it becomes possible to revise and recalibrate models to accommodate changing population behaviours. It also becomes possible to undertake ‘opportunistic’ analysis (Miller, 2017) of unpredictable real-world phenomena that create acute shocks, as demonstrated by three major incidents across London that form part of this research.

1.2 Background

This research began as an intellectual curiosity: can data being emitted as the by-product of using mobile devices during human activities within physical environments provide a new scalable source of data to reveal how context can affect people-place experiences.

Psychologist Kurt Lewin proposed that all behaviour (B) is a function of a person (P) and their environment (E) (Lewin, 1936), expressed as equation [1].

$$B = f(P, E) \quad [1]$$

Whilst this expression does not provide a mathematical formula to quantify different behaviours, Lewin used it to emphasise that behaviours arise from a complex interaction between human and environment, producing different behaviours at different times from the same person, and different behaviours at the same time from different people. The same argument was outlined in the book 'The Sciences of the Artificial' (Simon, 1996), by economist and cognitive psychologist Herbert Simon. Simon posited that natural phenomena have a necessity, the inevitable, about them whereas the artificial – man-made phenomena – have a contingency about them. Much of human behaviour can be considered artificial because it occurs within a designed environment, both physically and socially. As such, to make a forecast about human behaviour requires knowledge about initial conditions. The rapid rise and adoption of mobile devices have created the opportunity to acquire and analyse data about conditions and behaviours at new spatial and temporal scales.

1.2.1 Against generalisation

Whilst it has been generalised that people are happier when in more scenic locations (Seresinhe, Moat, & Preis, 2015), such a generalisation does not accommodate the dynamics of the location and people present. A person may not be happy if they urgently need to be somewhere else, or current conditions reduce the attractiveness of the location. An urban park in the UK is more likely to be frequented during the day than the night and more attractive during summer than winter. The park may generally be the preferred route as part of somebody's commute from home to work but there may be circumstances when preferences change, such as an unexpected and threatening social atmosphere, or an abnormal event obstructing the pathways or creating noise and crowds that repels the regular commuter away from the landscape.

When modelling real-world phenomena, all models are necessarily simplified representations of reality. The challenge is whether or not they are an over-simplification when used in real-world decisions, as articulated by statistician George Box:

> *"Remember that all models are wrong; the practical question is how wrong do they have to be to not be useful." (Box & Draper, 1987).*

Figure 1 shows the varying levels of confidence depending on approach when making estimates for systems that are sensitive to conditions. The ability to produce the closest possible representation of reality depends on what data is available about real-world conditions.

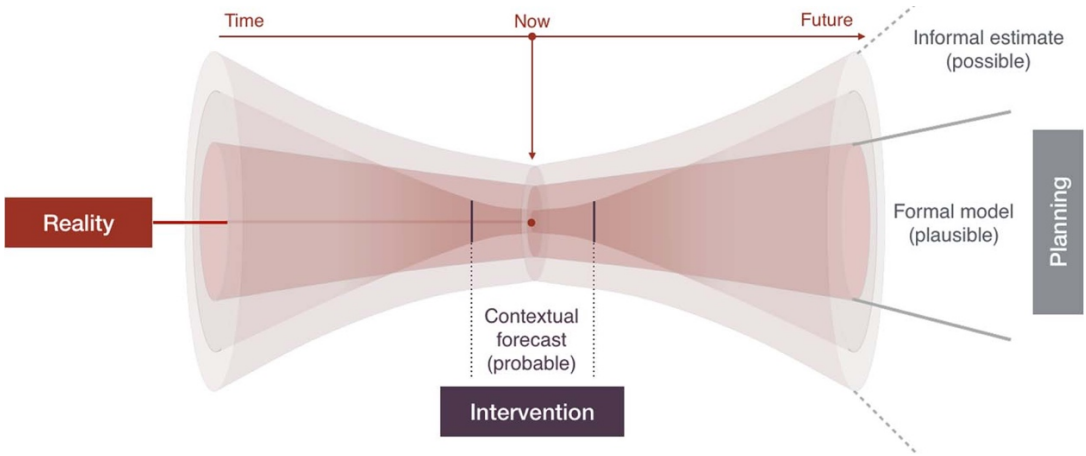


Figure 1. Funnel of prediction – from plausible to probable behaviour

Three levels of estimation accuracy: informal (possible, no matter how likely); formal (plausible, most likely given unknown conditions); contextual (probable within a limited time window given current conditions).

Assuming that the impossible is eliminated, an informal estimate can be made for any possible outcome, no matter how unlikely. It has the widest margin for error. Formal models produce generalised predictions and indicate the most plausible outcome, given limited information. Studies have shown that formal models consistently outperform informal human estimates (Dietvorst, Simmons, & Massey, 2015). However, generalised models exclude context and thus can still have a wide margin for error when applied to a specific situation. A contextual forecast produces an estimate based on data about current conditions, estimating the most probable near-term outcome. It is the closest representation of reality but useful only within a limited time window, beyond which there is too much variability in potential future states produced from the current conditions.

To give an example, climate modelling is a generalised model that can be used to estimate the weather conditions most likely to be experienced at a given location at a given time in the future. A weather forecast will produce a more accurate estimate than a climate model for what the weather is going to be like over the next few days because it is based on real-world observations about current conditions. However, it cannot accurately forecast beyond a few days because the weather is a chaotic system and small changes in initial conditions can profoundly affect outcomes. Beyond a few days, the forecast will be no more accurate than the simpler generalised model. It is argued that social systems are also chaotic, with an added complication: whereas a forecast does not change the weather, it can change behaviour (Simon, 1996). Thus, any forecast about behaviour needs to be continually revised and updated to be effective.

If models about social phenomena are to be used for near real-time interventions, this research posits that ignoring context and relying on abstract generalisations and theoretical assumptions will result in models that are too wrong to be useful. Furthermore, it is arguably inexcusable to ignore context when such information is accessible and can contribute to a more informed decision. By 2019, at least 76 out of 176 countries had begun to use artificial intelligence for monitoring urban environments with concerns raised about the reliance on generalised algorithms and theories developed using historical data that may contain representation bias (Bradford Franklin, 2019). The purpose of this research is to demonstrate that context now can and should be incorporated when

modelling human preferences and behaviours within real-world environments. A generalised model or average that does not consider context may not be representative of any real-world situation. Mobile data creates the possibility of producing forecasts about social phenomena in real-time.

1.2.2 Contextual mapping

Modern interactive urban maps are increasingly incorporating dynamic conditions such as traffic flows and temporal variation in the popularity of local venues. Figure 2 contains an example of an online map showing a detailed satellite image of the landscape overlaid with live traffic data for the roads and markers for four food and drink venues.

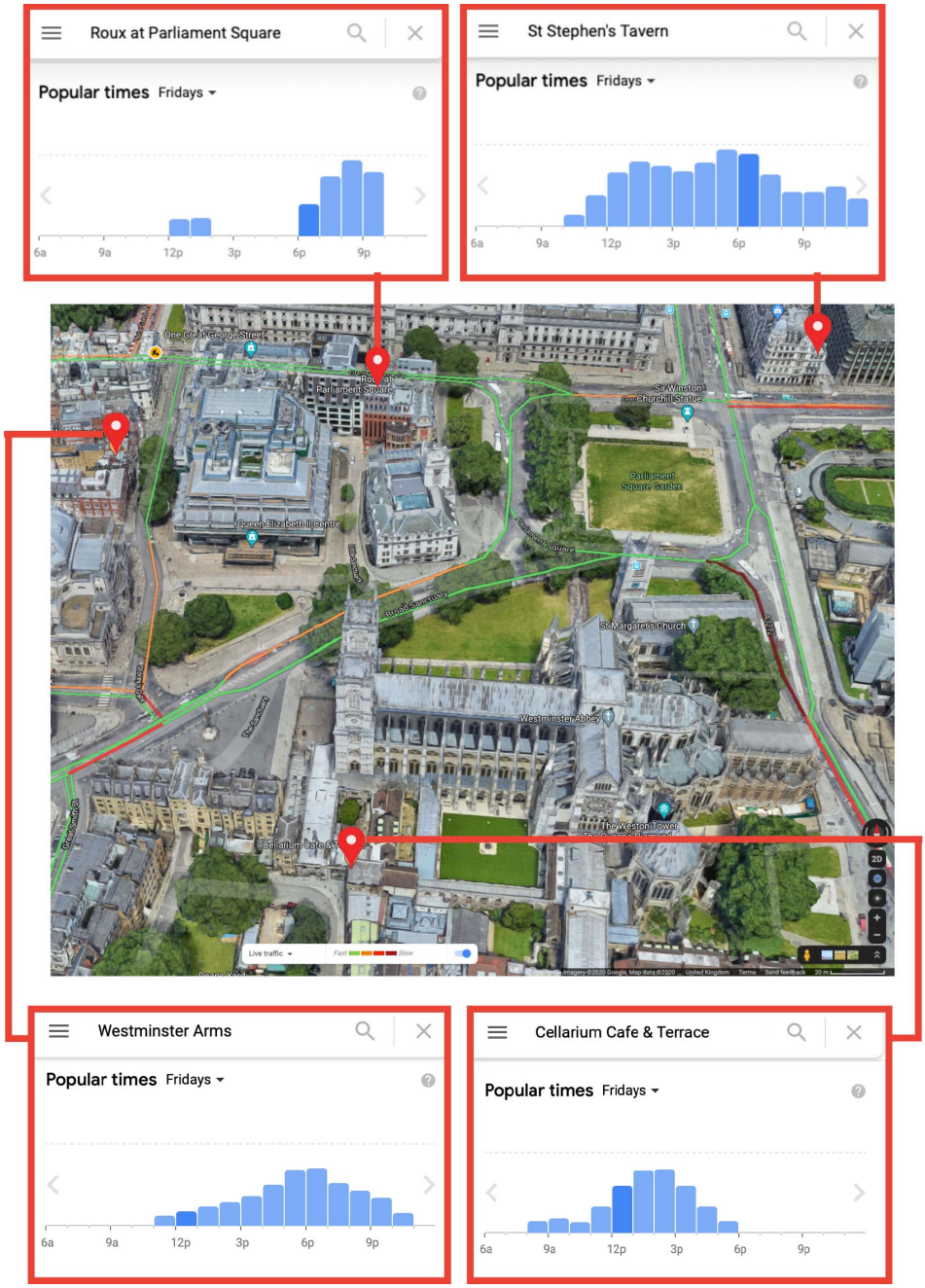


Figure 2. Example of a web-based interactive map

Map showing a 3D satellite view of Westminster Abbey in central London with live traffic data. Above and below are profiles of venues highlighted with markers. Map data as of 16 February 2020 at 18:58, © 2019 Google.

Clicking on individual food and drink venues within the map reveals a profile for the venue showing popular visiting times. What the map lacks are people and context. There is no indication of the current social atmosphere, nor the real-time street population, or how they may vary during different times of the day, on different days of the week, seasonally throughout the year or due to changeable social and environmental conditions. Furthermore, the satellite image appears to have been taken on a dry day during the summer based on the foliage visible. The live traffic data is from February and during an extreme weather event – Storm Dennis – with London experiencing prolonged heavy rain (BBC News, 2020). The live traffic is for a Sunday evening. Traffic on Monday morning would be likely to be much heavier due to the morning weekday commute. How might people's experience of the location differ under such changeable conditions? The landscape not only experiences extreme weather such as Storm Dennis. In 2017 it was the location of a major terror attack. Whilst the landscape is scenic, containing historic landmarks and green spaces, people's happiness whilst physically present will depend on the current conditions of the landscape and reason for visiting.

A fundamental step towards sensing the social atmosphere of a landscape is to have a measure of the real-time population: its size and spatial distribution. Traditionally, the size of a population for an area of interest has been estimated using the residential statistics from the most recent census. However, it has been acknowledged that people are unlikely to be at home during the working weekday. Recently, ambient population measures have been developed that disaggregate and redistribute the residential census according to demographics and spatial attributes to produce an estimate that reflects where people might be during a typical weekday. However, they apply a top-down approach, redistributing census data according to physical attributes of the landscape and lack contextual variation, producing a generalised ambient average in place of a residential count.

Historically there has been no method to capture such activity other than to deploy observers at street-level. The recent and rapid adoption of smartphones has created the opportunity to study human behaviour at new spatial and temporal scales. Smartphones are simply mobile devices containing internet-connectivity, built-in sensors including GPS, and the ability to install software applications, referred to as mobile apps. They expand the functionality of a mobile device from being a simple two-way communications device to becoming a personal digital assistant, able to both capture and retrieve information in real-time. Ownership of smartphones by UK adults increased from just 27% in 2011 (Ofcom, 2011) to 76% by 2017 (Ofcom, 2017), the period covered by the research in this thesis. They are now so pervasive that they are no longer described as a separate class of phone within countries such as the UK with high adoption routes. When using the term 'mobile device' in this thesis, it is assumed that the device has smartphone capabilities.

The demand for improved knowledge about the real-time population was highlighted at a Crime, Policing and Citizenship (CPC) workshop held at University College London in September 2016. The workshop presented advances in using artificial intelligence in predictive policing, to identify potential locations for criminal activity (Cheng, et al., 2016). However, outdated residential statistics are still being used to quantify the population potentially at-risk or affected.

> *“A key data challenge is knowing the street population.” - Chief Superintendent Dave Stringer, Metropolitan Police Force, speaking at the CPC workshop held at University College London, September 2016.*

This quote frames the opportunity for this research: to determine if new sources of data generated at street-level by mobile devices in real-time can provide knowledge about the street population to aid near real-time interventions and produce more representative models of behaviour.

1.2.3 Research opportunity

The very concept of a built environment means that human behaviour is not completely random. The question is whether or not it follows a predictable distribution. Behaviour maps produced from human observation at street-level have revealed some spatial and temporal regularity, such as peak and off-peak times when people travel between home and a place of work and observing that people tend to dwell at the edges of plazas rather than in the middle. However, they have also revealed substantial variations due to variable conditions and motivations. The historical challenge has been one of scale. Such studies are time and resource-intensive, conducted infrequently at few locations. Data emitted from mobile devices is a potential source of real-world observations that can scale spatially and temporally. The data may lack the comprehensive detail of field studies, but also introduces new capabilities, such as being able to detect the number of times the same device (assumed to be being carried by the same person) enters the landscape to create a measure of spatial familiarity. However, perhaps the greatest opportunity is in being able to continually learn over time, incorporating a wider range of contexts and recalibrating findings to reflect physical, cultural and social changes that affect human behaviour within the same landscape.

1.3 Research Outline

The research question is:

To what extent can mobile data be used to model the social dynamic of a landscape, to help better understand how context can change human behaviour in urban environments?

The hypothesis is that mobile data, whilst assumed to represent only a sample of real-world activity, is sufficient to detect variations in population behaviours within a landscape by continuously producing readings across space and time. Thus, it can provide a more representative estimate of human dynamics for a specific scenario than a generalised model that is absent of context. Furthermore, such data can be used to detect changes in conditions and behaviours as they emerge, creating the potential to recommend near real-time interventions to assist spatial choices.

The research comprises of four objectives:

1) Utilise accessible mobile data sources to determine if sampling of local digitised interactions is sufficient to improve understanding of how context affects the social dynamic of a landscape.

2) Develop a context-aware framework to model the social dynamic of a landscape: variations in the size, spatial distribution, actions and experiences of the population present.

3) Leverage the unique properties of digitally generated real-world observations to design a framework that can be continually learning and self-calibrating.

4) Apply the framework to produce a near real-time forecast of behaviour and behaviour change in response to learned contexts and sensed conditions as they emerge.

These aims are grounded in the belief that a landscape will attract or repel different populations and behaviours due to changeable environmental and social conditions.

1.3.1 Scope and definitions

The research presented in this thesis explores how population behaviours within an urban public space vary for different contexts using the data emitted from mobile devices. It is inter-disciplinary and early-stage research, applying theories and developments from cognitive, urban and data sciences. Definitions and methods are not always established or consistent across the fields. The following scope, assumptions and definitions have been adopted throughout this thesis.

1.3.1.1 Space

The spatial scope for each study is an area that can be comfortably traversed on foot, such as a small neighbourhood or large urban public space. This is referred to as a ‘landscape’ throughout this thesis. The largest landscape studied has a bounding box of 3 kilometres by 2.4 kilometres. The landscape may be sub-divided into zones to identify active spaces within it. A zone or active space is considered to be where a cluster of human activity occurs, such as travel routes through the landscape and dwell spots or attractions within it.

A landscape may be defined by a naturally-occurring boundary, such as a park. Otherwise, it is defined manually using a bounding box to enclose the area. Human activities are only captured and analysed whilst present within the landscape. The term location may be used interchangeably for landscapes, active spaces and landmarks when referring to a specific and identifiable place.

1.3.1.2 Time

The objective of this research is to detect variations in behaviour occurring within the same space at different times throughout a day and on different days, and to determine whether the variations can be associated with specific and recurring contexts. The temporal scope for each study will be to aggregate real-time data at daily and hourly intervals. Sub-hourly is only feasible if using a source that generates a large volume of minute-by-minute behaviours. If readings are sparse throughout the day, periods such as comparing off-peak and peak hours may be preferable.

Time is used as two measures throughout this research. It can refer to a point in time when an event occurred. It can also refer to a recurring time interval for comparing different intervals, such as comparing the population present within a landscape on weekdays versus weekends, or during peak hours compared with off-peak hours. When referring to the social dynamic of a landscape, it means how population behaviours vary over time. When referring to the social atmosphere of a landscape, it refers to population behaviours at a specific point in time.

1.3.1.3 Context

The term ‘context’ is used to refer to the circumstances that form a setting for measuring population behaviours. The circumstances include a physical location, time interval, and environmental and social conditions. When studying the social dynamic of a landscape, the constant is the physical location. The expectation is that different behaviours will occur at different time intervals and due to different environmental and social conditions. For example, within an urban park, the population present during the day would be anticipated to be different than during the night. During the day, it is likely to vary depending on the weather and time of year, and whether or not outdoor attractions are taking place. Without context, a static population measure would suffice.

As will be introduced in chapter three, this research proposes that there is a hierarchy of context when measuring real-world population behaviours: spatial affordances (space), temporal rhythms (cyclic time) and situated actions (acyclic time). The spatial context is a static measure for a given landscape independent of time of day or day of the year. Cyclic time produces dynamics, showing changes over time for routine behaviours, such as comparing hours of the day, days of the week, and months of the year. Situated actions refer to variable conditions that do not follow a predictable cycle and can disrupt routine behaviours, such as special events disrupting access to a location.

1.3.1.4 Population

The phrase ‘population behaviours’ is deliberate. Whilst the data being analysed is generated by individual mobile devices, this research is not studying any individual people. All sources are aggregated in space and time for analysis. The research is focused on the social atmosphere of a

landscape that is generated by the combined actions of people present. For example, the research seeks to answer questions such as: By how much does the size of the population present vary on different dates and at different times? What percentage of the people present are dwelling within the landscape versus moving through it? How many of the people present regularly frequent the landscape versus are first-time visitors, possibly tourists who lack familiarity with their surroundings? How do these statistics vary for different contexts? The focus is on attributes of a population rather than the attributes of any individual.

A landscape may be occupied by multiple populations at the same time and at different times, where each population shares an attribute that is mutually exclusive from the other populations. For example, attendees of an event being held at a venue, commuters travelling between locations, tourists visiting local landmarks, locals taking a break from routine activities and so on. Mobile data allows identifying such population attributes without compromising the privacy of any individual, provided the data is processed ethically, ensuring device identifiers are fully and irreversibly anonymised. Ethical considerations are discussed later in this chapter.

1.3.1.5 Behaviour

Behaviour as a term can be applied to any action that is the response to a situation or stimulus (OED, 2004), such as raising an eyebrow or steering a car. This thesis is interested in people-place interactions that form the social dynamic of a landscape. The behaviours being analysed are categorised as presence and experience.

Presence refers to being able to detect that a device is present within the landscape because it is emitting data with spatial coordinates and a timestamp. Assuming multiple readings are emitted whilst present, presence can produce a count of devices within the landscape and describe the locomotive action of each device whilst present: moving, milling or dwelling. Moving is the act of traversing the landscape between locations. Milling is when stationary for an undetermined time, such as pausing to consult a map or waiting to meet someone. It can also include movement being delayed by an unexpected disruption, such as traffic temporarily blocking a route. Dwelling refers to the intention to remain within the landscape for an activity. It can be for ad-hoc activities, making use of features provided for rest and recreation such as playgrounds, seating areas and paths or trails intended for walking or fitness activities rather than as travel routes. Dwelling can also occur for planned activities such as events at venues located within the landscape. This research is focused on exploring differences between moving and dwelling population behaviours, and for analysing dwelling for differences between ad-hoc uses of a landscape and planned activities. Milling is the in-between, where there is uncertainty about whether or not a person is intentionally dwelling or is temporarily stationary during movement between locations.

Experience refers to information about conditions and/or subjective feelings expressed within mobile data sources. This research will consider if mobile data sources are a viable means to learn about behavioural aspects such as mood and how they vary due to different circumstances.

1.3.1.6 Data

Five terms are used to describe the data sources used for this research: reality data, behavioural data, mobile data, data traces, and social media.

'Reality data' refers to all real-world observations generated from within the environment where the situation and/or behaviour occurs. Such data contains a minimum of coordinates and a timestamp or interval for when reading was created. 'Behavioural data' refers to a subset of reality data: observations that represent human actions and experiences whilst present within a landscape.

'Mobile data' refers to readings emitted by mobile devices. They can include data that is transmitted automatically whilst the device is in use, communicating over a wireless network, and data that is posted manually by the device user, such as choosing to post a text message to an online service. 'Data traces' refers to data emitted from mobile devices as the by-product of some other action as opposed to being produced intentionally to study presence and/or experiences. All behavioural sources used for this research are data traces.

'Social media' refers to subjective experiences posted by people from their mobile devices and shared publicly with online social networks. The social media of interest to this research are those that produce location-based readings in real-time. Media can include text, images, audio and video. It can also include markers of presence, such as 'checking in' to a venue.

1.3.2 Privacy, ethics and limitations

This research is specifically studying social phenomena, not individual actions. However, to achieve the aims of the study requires individual data points that can be attributed to a device. No personally-identifiable information is retained or used unless mandated by the source of the data. All device identifiers are anonymised within each data set. A key principle of this research is to store only the minimum data required for analysis and for the minimum time. Individual readings are analysed only to be tagged with attributes that associate the reading with a population. Device identifiers can then be removed. Data are aggregated to a spatial and temporal scale for analysis. The research method adheres to the same ethical guidelines adopted by field studies deploying human observers to produce behaviour maps (Gifford (Ed.), 2016).

To anonymise device-identifiable readings, a pseudo-id is generated within the dataset and the original device identifier is deleted. The pseudo-id cannot be reverse-engineered to recreate the original device identifier or to associate the device with any other data set. The same device present in more than one data set will receive a different anonymised identifier within each data set. Data sets contain, at most, three months of data from a single source. An exception to anonymisation is made only when mandated by a third-party. One third party data source – Twitter, an online global social network – requires that any individual messages displayed must be reproduced in their entirety and original form, including the screen name of the person who posted the tweet. Only publicly-shared tweets are included in this research. Thus, any message displayed within this study was knowingly shared within a public forum and can be viewed online.

On 25 May 2018, a new European-Union General Data Protection Regulation (GDPR) came into effect (GDRP.EU, n.d.). It introduced new requirements for the collection, processing, storage and sharing of data containing personally identifiable information. All data used in this study was generated and acquired before this date. Whilst some data sources contain unique devices identifiers during the preparation of data for analysis, none contain personal information other than the third-party source mentioned. All analyses are aggregated for population behaviours and do not involve the identification of any individual or the manipulation of any aspect of the environment.

Whilst the lack of any personal information within the datasets ensures that the research adheres to legal, ethical and privacy expectations, it does introduce a limitation to this study. There is no demographic information about the owners or carriers of the mobile devices generating the data for this research. It is not possible to validate whether or not the devices are representative of the wider population present within the landscape or whether demographics affect population behaviours and contexts occurring within the landscape. Ownership of mobile devices is prevalent amongst UK adults but not children and is higher in working adults than non-working adults (Ofcom, 2017). As such, it is possible that the data generated may be skewed towards working adult behaviours in public space and is less effective if studying environments and contexts for family activities and scenarios involving other demographics such as children, the elderly and those unable to make use of public space due to accessibility issues created by its physical conditions.

1.3.3 Thesis structure

The remainder of this thesis is organised as seven chapters, described below. Chapters two and three cover relevant literature and methods, chapters four to six contain the core research, developing a contextual framework that is then applied in chapter seven. Chapter eight reviews the outcomes from the research and concludes the thesis.

Chapter two provides an overview of the literature that has informed the research. It includes contributions spanning psychology, geography and data science that have informed the approach and methodology applied to the studies presented in this thesis.

Chapter three introduces the definitions, data and methods that are applied throughout the research. Whilst some of the techniques were developed during later research stages, all are presented together in this chapter for ease of reference and to avoid duplication. The chapter includes the contextual framework developed during the research.

Chapter four introduces the landscape being studied in this and the next two chapters: the Queen Elizabeth Olympic Park (QEOP) in Stratford, East London, United Kingdom. It then presents two studies utilising a range of data generated within the park. The primary data source is mobile device readings recorded by the park's public Wi-Fi network. The chapter concludes with the first version of the contextual framework, identifying social and environmental attributes that have the most influence on population behaviours within the landscape based on digitised interactions.

Chapter five uses a different data set to evaluate the same landscape. This time, readings are acquired from a mobile app carried by participants during everyday activities. Its reach extends beyond the border of the park enabling relationships with neighbouring landmarks to also be studied, and for readings to be compared with administrative data sources by aggregating at the same areal scale. This introduces the potential to convert variations in mobile device activity into an active population measure. The chapter also explores the potential to reveal the spatial distribution of presence across the area covered by an aggregated measure.

Chapter six shifts the focus from studying the spatial and temporal distribution of people present to whether or not their digitised interactions can provide contextual information about situations and their effect on people-place experiences. A social media source is used to evaluate the potential to extract latent information from the content of text messages shared publicly whilst present within a landscape to learn about people-place experiences and whether it is possible to capture the mood of the landscape during different sets of circumstances.

Chapter seven builds on the findings of the previous chapters and tests its potential for opportunistic data science, the study of unexpected real-world phenomena as and when they occur. Three incidents triggered a major police response within Central London in 2017. The study examines the ability to use the contextual framework to analyse each landscape and learn its social dynamic during ambient conditions to estimate the active population at risk when the incident occurred and to study if there is any lasting effect of such an incident.

Chapter eight concludes the thesis. It provides a summary and discussion of the findings and potential applications, explores some of the challenges experienced during the research, proposes recommendations for future research directions and presents closing thoughts.

A final note regarding formatting and terminology throughout this thesis. It is the preference of the author to use the term ‘their’ in place of ‘his/her’ or a default gender when referring to a human subject and the subject’s gender is unknown or ambiguous. The term ‘data’ is the plural of ‘datum’, indicating many data points. Some publications use it in its strictest definition – ‘data are’ – whilst others are more relaxed and also use ‘data is’. This thesis takes the latter approach. It will use ‘data are’ when referring to data as multiple data points, and ‘data is’ when referring to the set of data points as a single entity. Single quotation marks are used for the first use of a domain-specific term or to draw attention to a phrase. Double quotation marks and italics are used to reproduce a quote or definition from a third-party. For spatial visualisations, unless otherwise specified, maps are oriented so that north is up. Colour scales in charts have been chosen to be colour-blind friendly. Cross-references to figures and tables within the text are capitalised. Unless otherwise specified, all web links were accessed and accessible on or before 31 May 2020.

PAGE INTENTIONALLY LEFT BLANK

2 Relevant Literature

The research undertaken for this thesis utilises data emitted from mobile devices to learn how the social atmosphere of a landscape can vary to different circumstances. For this research, social atmosphere refers to human presence, actions and experiences whilst landscape refers to an area the size of a small neighbourhood or large urban open space that can attract people for multiple different activities depending on circumstances, from simply traversing the landscape as part of a journey to dwelling within it whilst visiting an attraction. This chapter summarises the core literature that informed the research. It is organised into three parts. The first covers the key concepts from the study of human-environment interactions, drawing on contributions across psychology and geography that have established the language and approaches to model and map people-place interactions. The second part outlines the recent and growing trend for using mobile data to study real-world behaviour. The third part discusses advances in computational social science and the opportunity to move towards data-driven theories of behaviour.

2.1 Modelling and Mapping Social Landscapes

This research is about learning how much the social atmosphere can vary within the same physical space by sampling the data traces being emitted from mobile devices in real-time. The term social atmosphere is used to emphasise that the focus is on population behaviours, not any individual present within the landscape. Attributes of interest are the presence, actions and experiences of people and how they vary for different contexts, producing a social dynamic for the landscape. The approach and terminology are based on theories about human-environment interactions and how they have been applied to model and map behaviour.

2.1.1 Human-environment interactions

Systematic approaches to studying how humans perceive and interact with their environment began from early in the 20th century, spanning multiple disciplines including psychology, urban studies such as planning and architecture, and geography. The shared theme is the study and representation of how humans perceive and respond to their environment.

2.1.1.1 Spatial affordances

In his book ‘Remembering’, published in 1932, psychologist Frederick Bartlett believed that one of the most important human abilities is to perceive ways to escape from immediate circumstances, an action that requires imagery, language and recall (Canter, 1977). In 1948, this ability was termed ‘cognitive mapping’ by psychologist Edward C. Tolman (Devlin, 2012) and led to a focus on spatial cognition and wayfinding: the ability to navigate a landscape from origin to destination.

However, creating a spatial representation for navigation says little about how conditions may influence behaviour. In the 1960s, the concept of ‘affordance’ was introduced by psychologist James J. Gibson, first in a 1966 paper and later described in his 1979 book ‘The Ecological

Approach to Visual Perception' (Gibson, 1979). Gibson described affordances as being the actionable and adaptable properties of an environment:

> *"The affordances of the environment are what it offers the animal, what it provides or furnishes, either for good or ill." Gibson (1979, p127)*

The term affordance gained widespread adoption after its use by Donald Norman in his book 'The Design of Everyday Things' (Norman, 1999). Norman described affordance in the context of user-machine interactions and extended the definition to include a user's goals, beliefs and experiences that influence actions. Norman later clarified that his interpretation should be defined as *perceived affordance*, in that a designer's role is to focus on what actions a user perceives are possible. In extending the definition to 'perceived affordances', Norman emphasises that many environments offer more than one choice and that preferences can be malleable to circumstances, whether the motivations of the user or conditions of the environment. Norman posited that there is only one way to know what a person will do when offered arbitrary choices: data. Norman argued that even a small sample about how people actually behave is better than assuming how they might behave.

The relationship between human and environment is interdependent, as described by Lewin's equation [1] introduced in chapter one. People can react differently to the same conditions in part due to different motivations for being present. Architect Jan Gehl proposed that three types of activity occur in public space (Gehl, 1987):

- Necessary activities
- Optional activities
- Social activities

In Gehl's taxonomy, necessary activities are those that are more or less compulsory, such as going to work or school. Their necessity means such activities are assumed to be mostly insensitive to the affordances of, or changes to, the environment. Optional activities are those undertaken if desired, and the time and place appropriate, such as going for a walk, sitting and reading for relaxation or deciding to go shopping for non-essential items. These activities are only likely to occur in environments that attract or facilitate such behaviours, and when conditions are favourable. Social activities are those that depend on the presence of others, also referred to as 'resultant' activities because they mostly develop in connection with (due to the occurrence of) the other two types. They include greetings and conversations, meetings, play, and communal activities. They generate dwell-time at locations and create a disruptive unpredictability even in routine activities.

Whilst a location may have a single physical representation for wayfinding, the types and volume of interactions will vary at different times due both to human motivations for interactions and changeable environmental and social conditions. To model and map the variety of behaviour that can occur in space requires incorporating time.

2.1.1.2 Time Geography

Around the same time as Gibson was publishing his work on visual perception (Gibson, 1979), geographer Torsten Hägerstrand is credited with introducing 'time geography' and the need to study space-time trajectories to better understand spatial choices. In his 1969 paper 'What about people in regional science?' (Hägerstrand, 1989), Hägerstrand introduced the concept of a space-time prism (Figure 3), within which a person will have a range of possible behaviours defined by their capabilities coupled with constraints – limitations that "define where, when, and for how long" the individual has to participate in an activity. Hägerstrand's definition focuses on individuals and spatial preferences for how they spend their 'time budget'. However, the concepts can be applied to study the dynamic affordances of the landscape and how they affect population behaviours. For example, a public park near an office building is likely to have a peak usage that coincides with work breaks at nearby offices. Its attractiveness is also likely to vary by daylight and weather. As such, the landscape also has a space-time prism where capabilities are coupled with constraints.

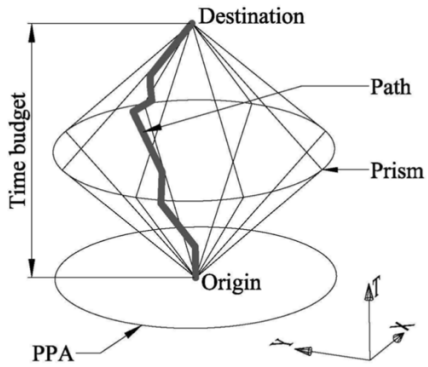


Figure 3. Hägerstrand's space-time prism, reproduced from Neutens et al

Figure 1. Time geographical concepts: space-time path, space-time prism, and potential path area (PPA). P.337 (Neutens, Witlox, & De Maeyer, 2007).

Whilst it introduced a time dynamic to spatial analysis, Hägerstrand's early work was criticised by humanists for lacking a human dynamic: meaning and experiences that influence choice (Sui, 2012). Linking to Gehl's taxonomy of necessary and optional activities, the time budget for optional activities is likely to be influenced by environmental and social conditions. For example, an office worker may prefer to spend their lunch break in a local park if weather conditions are pleasant. However, if weather conditions are unpleasant, they may choose to go to the office cafeteria instead. Whilst the space-time prism introduces a time dynamic and reveals that there are opportune times for certain behaviours occur, it does not provide the means to explain variations in behaviour for the same time interval on different dates.

Hägerstrand's later work linked perspectives of time and space with human agency, visualised as three inter-related dimensions that represent the landscape, or 'diorama' (Figure 4). The space dimension defines the area being studied (Choros) and its physical features, or affordances, (Topos) that enable or constrain action. The time dimension enables the occurrence and duration of actions to be measured (Chronos) and defines the time intervals that are 'opportune' for action to occur (Kairos). Hence activities can occur at any time within an open space, but there may be

opportune times when certain actions are more likely to occur. The Agency dimension considers the variability in human actions. These include individual personality traits, social influences (cultural, economic and political factors), the use of technology to adapt or augment capabilities, and human-environment relationships influencing capabilities and constraints.

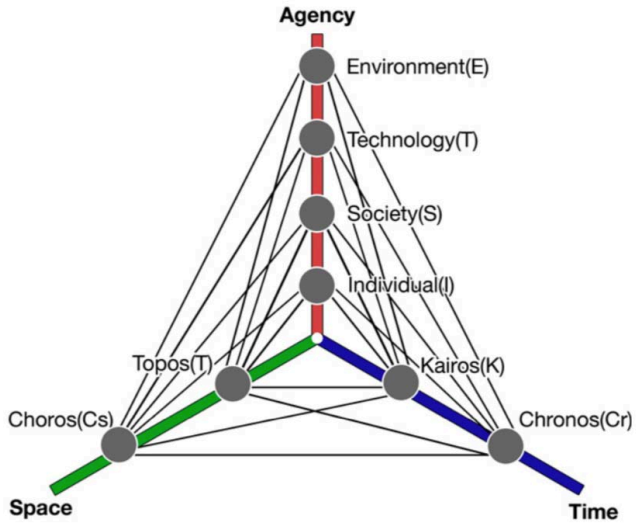


Figure 4. Conceptualising space, time and agency, reproduced from Sui

Figure 1, page 10: Alternative conceptualisations of space, time and agency for time geography (Sui, 2012).

Incorporating human agency introduces the potential to model behaviour change.

2.1.1.3 Behaviour Change

At its simplest, a behaviour is simply an action in response to a stimulus. Behaviour change is the study of what stimuli lead to different behaviours. For this research, the stimuli are contexts that produce different population behaviours within the same physical landscape at different times, focusing on presence: the size and distribution of the population present, and experiences – conditions and feelings expressed whilst present.

Behaviour change theories are dominated by psychological approaches that coalesce around the criminal investigation technique for identifying suspects as having the 'means, motive and opportunity' to act (Innes, 2007). One example is the COM-B system (Michie, van Stralen, & West, 2011). As visualised in Figure 5, there needs to be an opportunity (O) for the behaviour (B) to occur. There also needs to be the motivation (M) and capability (C) to act, both of which can also be applied to create the opportunity. Motivation represents a person's desire to behave in a certain way. It assumes that humans are goal-seeking to achieve wants or needs. Capability represents the means to act. This can include requiring certain skills, resources or information. Being motivated is not enough if a person lacks the capability and vice versa. Having both the motivation and capability to act will also not result in a behaviour change unless there is an opportunity for a behaviour to occur. Opportunity is considered to be all the factors that lie outside the individual that make the behaviour possible or prompt it. All three elements must be present for a change in behaviour to occur.

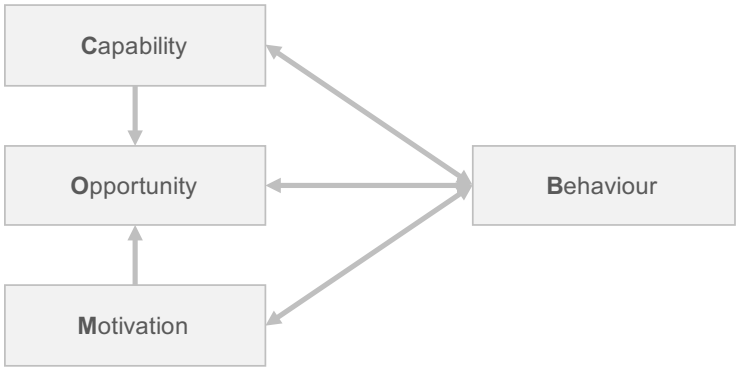


Figure 5. COM-B system for understanding behaviour, reproduced from Michie et al

Figure 1, page 4: COM-B system – a framework for understanding behaviour (Michie, van Stralen, & West, 2011).

Linking the COM-B model to Lewin's equation for behaviour [1] introduced in chapter one, opportunity is oriented towards the environment. Capability and motivation are oriented towards the person. Despite research indicating that environment has the largest influence on behaviour (Wortley, 2012), the model emphasises the personality traits of the individual and the individual as the protagonist creating a behaviour change.

The approach taken by the research presented in this thesis shifts the focus from the individual to the landscape. Combining Hägerstrand's time geography with the COM-B system, the landscape can be studied in terms of three contexts. Capability represents the spatial affordances of the landscape, the boundary of the area and its physical features. Opportunity represents the temporal dynamic – the opportune moment for different behaviours to occur. Motivation represents agency – the individual, social, technological and environmental conditions that can determine and alter the attractiveness of the landscape, its motivator for behaviour to occur.

Viewing the landscape as an actor in behaviour analysis shifts the focus from individual spatial cognition to population behaviours arising from socio-spatial dynamics, the focus of this thesis.

2.1.2 Behaviour modelling

Whilst Hägerstrand's work could be argued as the most comprehensive theory for modelling spatial behaviour, it has been said that the concepts failed to gain traction in urban modelling due to a lack of data and/or computational tools to represent complex relationships (Neutens, Witlox, & De Maeyer, 2007). Instead, urban models have emphasised the physical attributes of locations as being the dominant drivers of spatial behaviour, irrespective of time.

Space syntax is one such method, conceived by architects Bill Hillier and Julienne Hanson based on the belief that all social processes are realised in space (Hillier & Hanson, 1984). Focusing specifically on the built-environment, research has found that the spatial configuration of a behaviour setting, a location with defined boundary, can explain a substantial proportion of the variation in human movements along different paths within the behaviour setting (Penn, 2003), as visualised in Figure 6. Without going into specific details, a neighbourhood of streets can be converted into a graph, or network, and the popularity of different possible routes can be quantified

based on properties of the network, such as the number of intersections between streets. Whilst such an approach has proven effective at modelling human movements at a general level, it does not provide the means to quantify different movement behaviours such as milling or dwelling with the landscape, or variations over time due to different dynamic conditions that can affect the attractiveness or accessibility of routes. The focus is on wayfinding and making trips with purpose from origin to destination to aid architecture, urban design and planning decisions.

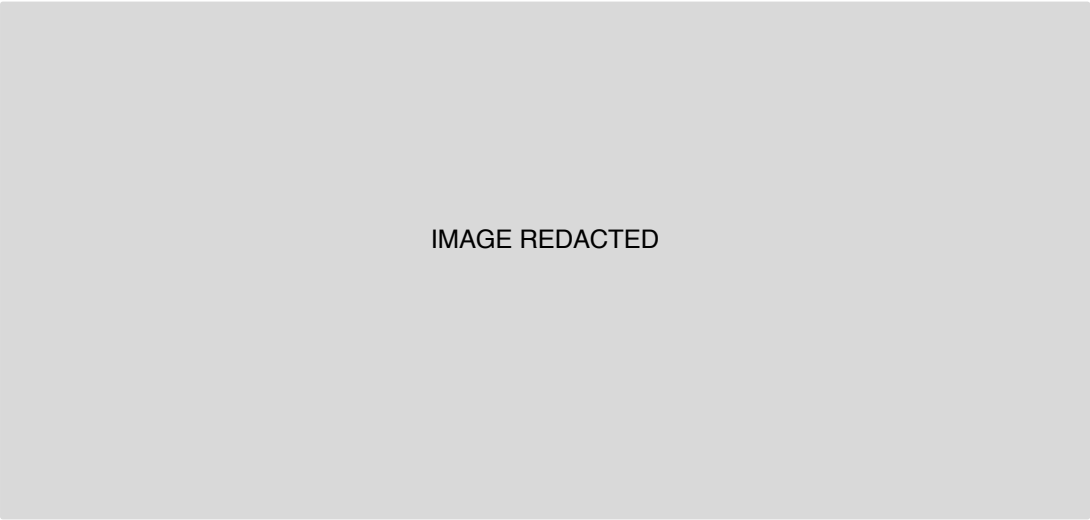


Figure 6. Modelling street-level pedestrian routes, reproduced from Space Syntax

The image on the left is the space syntax model based on the physical configuration of space (street network). The image on the right is a pedestrian model based on survey data from 2011. Location: Woolwich, East London. Source: https://spacesyntax.com/project/woolwich-squares/. Copyright © Space Syntax Ltd.

Agent-based modelling (ABM) is an interactive approach that reveals emergent behaviours over time. ABM is a computer simulation that generates an artificial population of individual agents each following a set of rules (Cheng, et al., 2016). A popular pedestrian modelling software application, LEGION, uses ABM to simulate pedestrian flows (Still, 2000). The rules are based on the psychological theory that humans seek to minimise dissatisfaction. It decides their next step using a mathematically weighted combination of the tolerance for, and wish to avoid, inconvenience, frustration and discomfort (Vanderbilt, 2012). Figure 7 shows a screenshot from the LEGION software used to model event attendees leaving a stadium (AECbytes, 2018).

Agent-based models are perhaps the most advanced method to simulate human-environment interactions. Whereas spatial interaction models and spatial representation models produce generalised averages, ABM simulations reveal how behaviour emerges over time as a result of individual agents interacting with one another and the environment. However, the models still focus on goal-seeking behaviours that themselves do not change, as articulated in a 2012 interview with LEGION's head of business development, Dan Plottner (Vanderbilt, 2012):

> *'Modelling pedestrians works best for discrete flows in concentrated spaces when masses of people are trying to get somewhere with purpose. We can't do people loitering in Times Square. We don't understand why their behaviour is what it is" Dan Plottner, LEGION (Vanderbilt, 2012).*

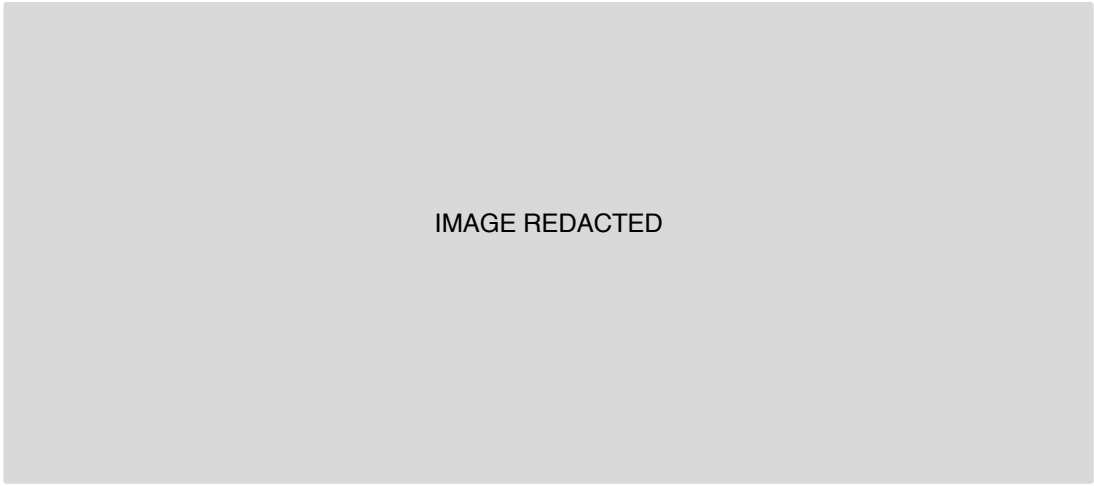


Figure 7. 3D simulation of egress from a stadium in LEGION, reproduced from AECbytes

A screenshot from LEGION pedestrian modelling (AECbytes, 2018).

Common to all theory-driven models is a lack of data about how many people are likely to be present and whether or not the people present are likely to share the same motivations. Models rely on an artificial population that is disaggregated from administrative statistics such as the census or based on assumptions to test thresholds such as the maximum capacity of stairways or the optimal time to cross a street junction. The individuals are typically identical, all sharing the same attributes, a single goal and observing the same set of rules. Taking just one limitation, how many people are able-bodied or have accessibility needs? In the UK, one in five people in England and Wales reported some form of disability in the 2011 census (ONS, 2015). How many people leave a stadium immediately after the event concludes versus leaving early or waiting for crowds to clear? If given an arbitrary choice between multiple exits, how many will leave by the same door they arrived through regardless of whether it is the most convenient or suitable? How many will be influenced by the choices of others nearby, following them rather than thinking independently?

An alternative to theoretical simulations is to produce data-driven models of behaviour. To achieve the latter has historically required human observers at street-level collecting real-world data about people-place interactions. This is referred to as place-centred behaviour mapping.

2.1.3 Behaviour mapping

The phrase behaviour map was first proposed in 1970 by W.H. Ittelson et al, to describe a visual summary of observed frequencies of real-world activities (Canter, 1977). It emerged within the field of environmental psychology during the 1960s as a technique to study real-world behaviours by subjects within a landscape, as an alternative to surveys, self-reported behaviours and direct observations of participants (Gifford (Ed.), 2016). A behaviour map is defined as consisting of five elements: 1) a base map identifying the physical features of interest; 2) behavioural categories with definitions and codes; 3) a schedule of observation; 4) a systematic procedure of observation; and, 5) a system of coding and counting (Gifford (Ed.), 2016). Developed during the same era as Hägerstrand's time geography (Hägerstrand, 1989), it incorporates the three dimensions of spatial affordances, temporal measures and human agency when studying spatial interactions.

The objective was to produce more robust studies of real-world behaviours, recognising that traditional surveys and focus groups could contain cognitive biases, whether deliberate or unintentional. For example, asking people to maintain a record of their movements resulted in only 13% of participants accurately reproducing their routes (Hill, 1984). A study of visitors to a museum (Betchel, 1967) noted: “People who know they are being studied, observed, or evaluated generally act differently.” Gifford concurred: “…participants are often motivated to describe their behaviour in the best possible light or to tell the researcher what they think he or she wants to here.” (Gifford, 2016). Behaviour mapping using unobserved tracking methods seeks to avoid these effects. The methods have been used to test and revise theoretical assumptions. For example, a study of store layouts found that the best-selling location is not at the entrance, as was commonly assumed. Observations showed that people do not begin shopping or looking for items until they are some distance inside the store (Gifford, 2016 citing Underhill, 2000).

Behaviour mapping can take two forms: person-centred or place-centred. A person-centred behaviour map follows the trajectories of individuals. A place-centred behaviour map creates a defined boundary of space, the landscape, and observes actions of people whilst present within the landscape. This research is focused on place-centred approaches.

From the 1970s onward, place-centred behaviour mapping steadily increased its use of technology. An in-depth study of downtown Manhattan in the 1970s by Boris Pushkarev with Jeffrey M. Zupan combined street-level observations and administrative records with helicopter aerial photography (Pushkarev & Zupan, 1975). The study revealed both the regularity of routine behaviours and the variability within routines in response to different conditions. One-third of home-oriented travel occurred between 4pm and 8pm whilst travel was lightest between 1am and 5am, accounting for less than 3% of all trips. Even hourly aggregates masked local peaks due to work shifts starting and ending on the hour. The report found that many trips in the central business district had intermediary stops, such as picking up a newspaper or window-shopping, and that such behaviour was substantially underestimated in existing pedestrian models. Temperature influenced optional activities, with 13 degrees Celsius being the threshold above which a significant amount of pleasure walking occurred. This result was supported by a separate study in Copenhagen by Jan Gehl (Vanderbilt, 2012) that also observed walking speed increased by 35% on colder days compared with temperate summer days. Precipitation was more disruptive, affecting pedestrian flow by up to 60%, in most part due to people making shorter journeys and diverting to covered transport. The effect was most pronounced in shopping areas and least pronounced during the morning peak commuting period, reinforcing the expectation that optional activities will be more sensitive to changes in environmental conditions than necessary activities such as commuting to work.

In 1980, William ‘Holly’ Whyte published a study of 14 plazas and two small parks in New York, (Whyte, 1980). The study combined the use of time-lapse cameras with human recorders at street-level. Amongst the findings, the number one activity when sitting in a plaza was people-watching. People crossing the plaza rarely stopped to talk in the middle, preferring instead to walk first to the edges (Figure 8).

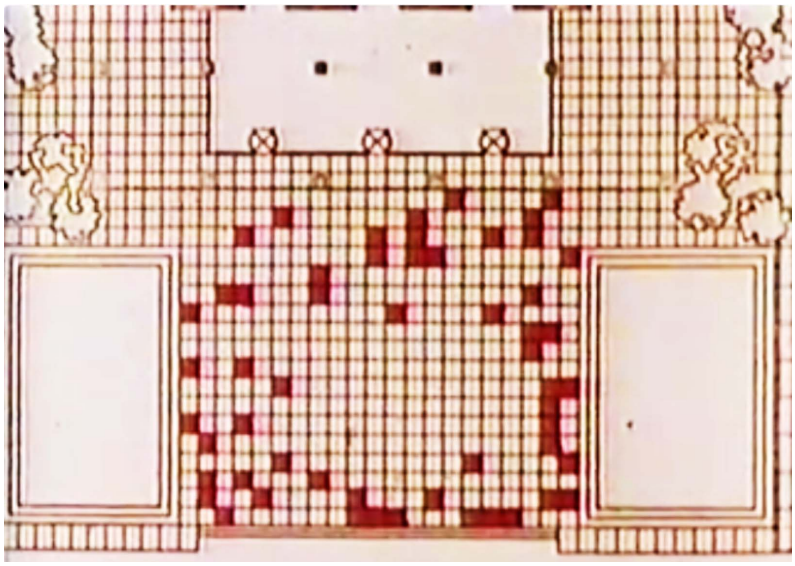


Figure 8. Form to observe and map behaviour, reproduced from Whyte

Screenshot from the video 'The Social Life of Small Urban Parks and Plazas' demonstrating the method used to record the areas of a landscape where people stood and dwelled, showing a preference for the edges of an open space rather than its centre (Whyte, 1980).

Such detailed studies of unrestricted open spaces have been limited historically due to the time and resource-intensive methods required. However, the findings have proven beneficial to urban planners. Both Pushkarev & Zupan's and Whyte's reports were incorporated into New York zoning codes for pedestrian spaces in city centres, mandating an increase in the use of building frontages for retail and food outlets, improving street-level access to facilities and increasing the use of seating and greenery in urban plazas. A separate study published in 1971 used time-lapse photography to establish six levels of service (LOS) to support pedestrian needs for walkways and stairways in the built environment (Fruin, 1971). It became the standard for urban design and pedestrian modelling tools (Erkan & Hastemoglu, 2016).

The challenge with such studies is that they are limited to observations collected at few locations during a specific period of a single year. As well as only being able to collect data about contexts that happen to occur during the period of the study, the findings can become outdated if cultural and social norms change. This has already been demonstrated in observations collected since the widespread adoption of mobile phones, revealing changes in pedestrian walking behaviours (Vanderbilt, 2012). It is also a concern when making assumptions about the population present. The 'Nuclear family' of the mid-20th century, (The Editors of Encyclopaedia Britannica, 2015), was a generalisation based on the most common North American or European family unit of the time, assumed to consist of a working father and stay-at-home mother raising a family of 2 or 3 children. Models based on such an assumption may be ill-suited for 21st-century behaviours.

2.2 Digitally Sensing the Landscape

As is explored in this thesis, the recent and rapid adoption of mobile devices generating data digitally about human-environment interactions, combined with new computational methods to analyse such data for contextual insights, have created the potential to overcome the limitations of infrequent field studies, enabling the continuous and ongoing sampling of real-world behaviours. The use of mobile data generated during everyday interactions has been termed 'social sensing' (Liu, et al., 2015) as part of a new field of computational social science (Lazer, et al., 2009).

2.2.1 Social sensing

Social sensing refers to the sensing of human activities located in space and time from mobile data sources. Early studies have demonstrated the potential to challenge theoretical assumptions. For example, location data taken from mobile phones showed that movements were more regular spatially and temporally than predicted by computer simulations that used the popular Lévy flight and random walk models (González, Hidalgo, & Barabási, 2008). In 2012, a study of multiple cities of varying sizes across the United States of America using device-identifiable mobile data challenged the effectiveness of distance-based gravity models for the study of human dynamics across a city, arguing that opportunities within different places matter more than distance once people are within city limits due to availability of diverse public transport options and proposed an alternative rank-based model (Noulas, Scellato, Lambiotte, Pontil, & Mascolo, 2012).

Population measures are also being challenged by access to social sensing. A 2016 study demonstrated that mobile data could be used to produce alternative population measures, quantifying the population at risk of exposure to air pollution in New York by comparing an 'Active Population Exposure' measure generated based on a sample of mobile device activity with a 'Home Population Exposure' measure generated using residential statistics (Nyhan, et al., 2016). Social media has also been used to challenge the reliance on residential data for modelling population size. A 2013 study (Birkin & Malleson, 2013), used social media messages posted on Twitter from locations across the city of Leeds in the United Kingdom and revealed five social archetypes – domestic living, education, work, recreation and shopping – to consider when modelling human dynamics. A 2016 study of geo-located tweets across Greater London identified seven clusters of activity: London Residents, Commuting Professionals, Student Lifestyle, The Daily Grind, Spectators, Visitors, and Workplace and tourist activity (Longley & Adnan, 2016). Each cluster exhibited different levels of activity across five temporal categories: morning peak hours (Monday to Friday, 7am to 9:30am), weekday (Monday to Friday, 9:31am to 3:59pm), weeknights (Monday to Friday, 7:01pm to midnight, and midnight to 06:59am) and weekends.

The growing use of mobile devices to capture data about people-place interactions has led to the creation of the terms 'living lab' and digital 'social observatory'.

2.2.1.1 Living labs and social observatories

The term 'living lab' is credited to William J. Mitchell, Kent Larson and Alex 'Sandy' Pentland of the Massachusetts Institute of Technology (MIT) who first began applying the concept in the early 2000s (Pentland, 2014). Whilst there is no universal definition, in its simplest form, a living lab is a location that enables the analysis of situated human behaviour in real-world settings. A variation of the term, sometimes used interchangeably due to the lack of formal definitions, is the 'social observatory'. Social observatories have been described by geographer Harvey Miller as creating the potential for 'opportunistic data science' (Miller, 2017). By continually monitoring the real-world, they provide the opportunity to study unexpected incidents and behaviours.

For this research, the terms 'living labs' and 'social observatories' are defined separately. A living lab undertakes periodic studies within a defined behaviour setting to uncover patterns of behaviour. A behaviour setting is an area of the real-world that has a natural boundary to contain action, such as a building or sports field (Barker, 1968). Such studies may include experiments, such as deliberately altering the environment and observing responses, because those being monitored have provided informed consent and are knowingly participating in a living lab, they are participants rather than subjects. The data being generated is a primary data source, intentionally collected for the study. A social observatory undertakes ongoing monitoring of any landscape where behavioural data traces can be captured. It may not involve direct informed consent and the people involved are subjects, not active participants. This is comparable with traditional behaviour mapping techniques, as outlined earlier in this chapter. However, to adhere to the same ethical guidelines adopted by human observers producing traditional behaviour maps (Gifford (Ed.), 2016), it should not involve conducting experiments or manipulating the environment to affect behaviour or identifying individuals. The data is collected as passive observations of real-world conditions and behaviours, to create an indirect, or secondary, data source for analysing people-place interactions.

Whilst the terms living lab and social observatory refer to different approaches to social sensing, they both rely on the use of mobile and embedded devices that produce behavioural data traces. The data can be analysed both as signals for presence and movement patterns, and as semantic content that can reveal conditions about the environment and peoples' experiences of situations occurring within it.

2.2.2 Signals of presence

Studies utilising mobile communications as signals of presence and movement appear to have begun in 2004, pre-dating the mass adoption of smartphones. Instead, many of the earliest studies were run within living labs where participants could be given pre-configured devices. They took place within behaviour settings such as university campuses (Eagle & Pentland, 2006), and holiday resorts and conference centres (Jayarajah, Balan, Radhakrishnan, Misra, & Lee, 2016). A 2013 study explored multiple sensor-based methods to estimate the active population for a large metropolitan university campus in Queensland, Australia (Charles-Edwards & Bell, 2013). Techniques included manual counts at seven entry points by a team of 25 human observers over

a single day, infrared cordon counts over a two-week period and a travel survey completed by staff and students. The research found a strong correlation between infrared counts and the travel survey, and that both were three times the estimate of the working population based on administrative statistics taken from the 2006 census.

Whilst producing new sources of information about human mobility, a challenge with such studies is that they typically involve participants of a specific demographic, students for example, and/or movements and activities constrained by the setting.

As adoption of smartphones expanded, more studies have been published using mobile data, either provided by the telecommunications provider (Telco) or produced by apps installed on the device, enabling the study of urban public space. Telco data is usually provided in aggregated form using a measure known as 'Erlang'. A single Erlang represents one person-hour of mobile phone activity recorded at a cell tower. It can be a single phone call lasting an hour or multiple devices making multiple calls that total one hour and can include SMS text messaging. One of the earliest studies utilising Erlangs to analyse population dynamics was of Milan based on sixteen days in 2004 (Ratti, Pulselli, & Williams, 2006). The research produced a diurnal rhythm: activity during daylight hours (Figure 9), one that has been reproduced independently in other studies. Whilst the findings may appear common sense, such as day-to-night and peak-to-off-peak variations, previously there was little evidence to quantify the changes. However, there is a noticeable and unexplained wide variation in readings on different dates during daytime hours, with readings varying by over 100 per cent.

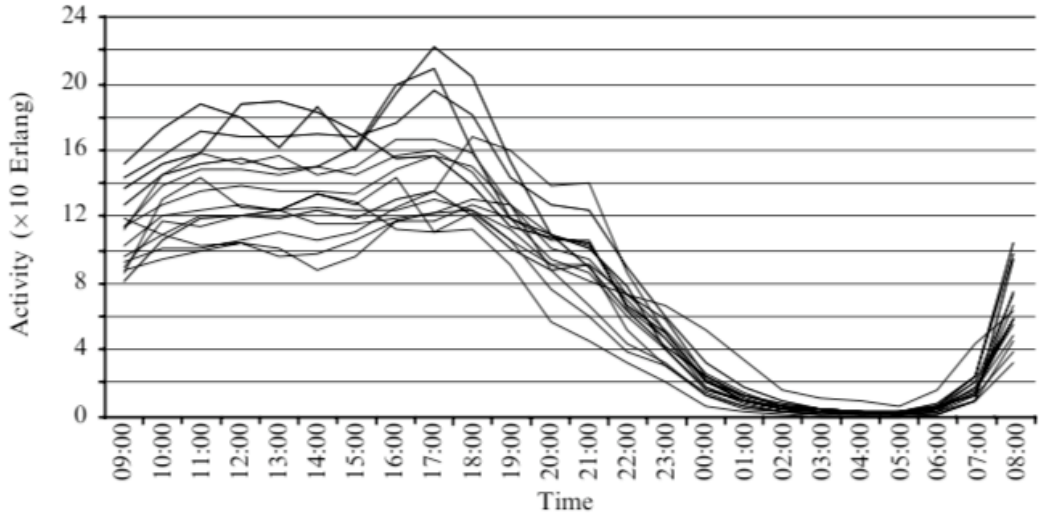


Figure 9. A single day of cell activity in Milan, reproduced from Ratti et al

Image is a direct reproduction of figure 4: 19 April 2004, Milan metropolitan area. Cell activity, absolute values in Erlang on 16 dates, (Ratti, Pulselli, & Williams, 2006).

A 2006 study involved just under 300 participants in Tallinn, Estonia, who permitted the TelCo to release their detailed call data records (CDRs) over eight days in 2006 (Ahas, Aasa, Silm, & Tiru, 2010). Spatial tracking was provided at 15-minute intervals, essentially turning the city into a living lab for eight days for the 300 participants. Despite the short period, the study revealed diurnal

rhythms of the city centre and surrounding districts. It also revealed gender differences and cultural effects, for example showing that female jobs were significantly closer to home than male jobs in the suburbs. A 2008 study utilised CDRs over six months for 100,000 users within the US (González, Hidalgo, & Barabási, 2008). The study challenged the popular Levy random walk algorithms used in urban modelling by showing high spatial and temporal regularity to movements. A later study analysed 15 months of mobility data for 1.5 million people and found that 95% of the people could be uniquely identified from four spatio-temporal data points (de Montjoye, Hidalgo, Verleysen, & Blondel, 2013).

The majority of Telco studies produce uniform grid-based surface models for spatial analysis, with 500m2 and 250m2 appearing to be the most popular cell size based on studies reviewed here. The use of local wireless communications networks enables spatial accuracy to be reduced to street-level. A 2006 living lab study utilised sensors deployed at just under 100 locations across the city of Bath in the United Kingdom, recording devices detected within range of the sensor in real-time across three months (O'Neill, et al., 2006) and evaluating movements against the Space Syntax method to incorporate a time dynamic. A study of the Manhattan area of New York analysed 20 million Wi-Fi readings across 53 fixed access points from March to October 2015 (Kontokosta & Johnson, 2017). The readings contained device IDs enabling visitors to be categorised as daily, weekly, occasional (visited at least once before) or first-time visitor. The period revealed weekday hourly trends, seasonal variations and effect of public holidays (Figure 10). As with other studies, the time intervals had substantial variation in counts during daytime hours, in this case, both within and between the categories of a visit.

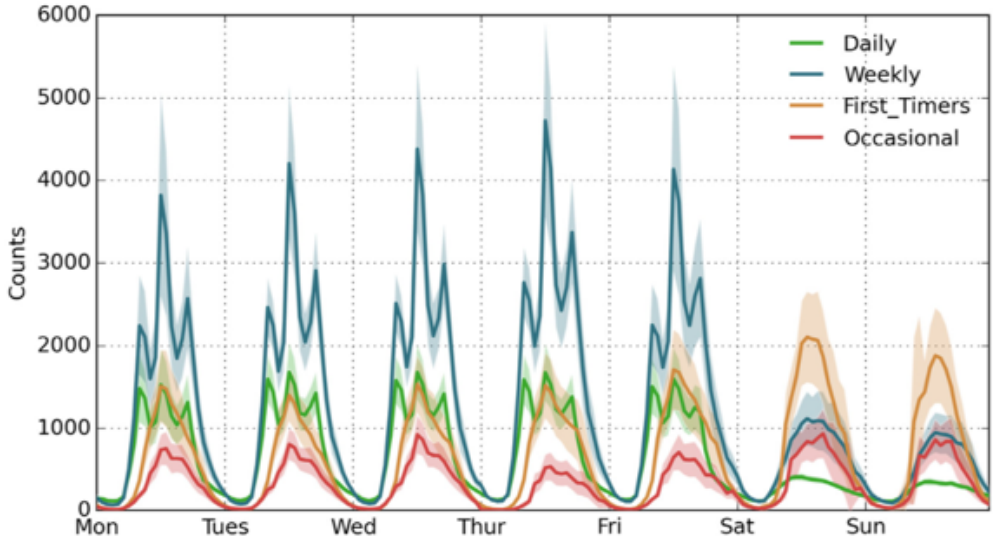


Figure 10. Categorising Wi-Fi activity, reproduced from Kontokosta & Johnson

Image is a direct reproduction of figure 3: Weekly Wi-Fi activity, by user group, mean counts by hour, 2015 data, Water Street network. Source: (Kontokosta & Johnson, 2017).

2.2.3 Semantics of experience

The second aspect of mobile data is to go beyond analysing the readings as individual data points in space and time and inspect the content for descriptions that may reveal environmental conditions and human experiences of place. Studies utilising mobile data for semantic insights appear to have begun in 2011. This followed the arrival of social media networks and mobile apps from 2009 that encouraged people to share experiences of physical places with online communities via their mobile devices. Analysing the content can reveal contextual information about the conditions of the environment that may explain changes in behaviour at different times.

The two earliest social media networks generating location-based data were Foursquare, launched in 2009, and Twitter, which introduced a geotagging feature in 2009. Foursquare is a globally-available mobile app that people used to 'check-in' to venues. Twitter is a global online social network that enables people to post short messages with attachments such as images and videos. Other sources that have provided publicly-available data used in urban analysis include Flickr and Instagram, two online photo-sharing sites. The rise and popularity of social media expanded the use of mobile devices from providing signals of presence to sharing situated experiences.

Semantic content generated by people posting updates using their mobile devices can be broadly divided into three categories: activity markers, visual imagery and text-based language. Activity markers such as Foursquare venue check-in data can enhance presence analysis because venues are categorised based on types of activities they are assumed to attract, such as 'park', 'shop', 'hotel' etc. It enables the analysis of different populations present based on the locations they visit within the landscape. A 2012 study used Foursquare check-ins during 2011 to analyse if such a source can reveal the urban character of an area and confirm mental maps made by residents (Cranshaw, Schwartz, Hong, & Sadeh, 2012). A 2017 study analysed plazas within cities in Spain to evaluate their successfulness as liveable, social and attractive places, (Martí, Serrano-Estrada, & Nolasco-Cirugeda, 2017). A potential limitation is whether or not such data represents the full range of potential activities for a landscape. Are there venues that people would prefer *not* to check-in to? Who uses Foursquare? Is its use biased demographically? These are unknowns that must be considered when conducting research using such sources.

Studies involving location images are problematic. The same physical location can have a very different social and psychological effect as well as visual appearance, depending on the time of the day and time of the year it was taken and what activities were occurring – the combination of space, time and agency, to use Hägerstrand's time geography described earlier. Furthermore, there may be curation effects. People sharing photos may seek to take the best possible picture or most dramatic visual for effect. Nonetheless, several studies have produced interesting spatial results from analysing large volumes of images taken across cities and even countries, measuring generalised outcomes such as ambience (Graham & Gosling, 2011), smells (Quercia, Schifanella, & Aiello, 2014), greenness (Li, et al., 2015), and 'scenic-ness' (Seresinhe, Moat, & Preis, 2015). That said, none have been able to incorporate temporal distributions throughout the day or explore variations in experiences within the same location due to contextual differences. Access to images

in real-time is also proving difficult as social networks adjust their terms of service. For example, Instagram, the most popular photo-sharing mobile app when this research study was first proposed in 2015, closed access to its public API in 2016.

The ability to instantly share life experiences has maintained the popularity of social media networks such as Twitter, where people can post short text-based updates. The content can be mined computationally using natural language processing to create linguistic representations of people-place experiences. The ease of creating and posting brief messages shared with spatial coordinates and a timestamp automatically generated enables both spatial and temporal analysis of the content. A 2013 study utilised Twitter to detect real-world events that could potentially impact city services such as traffic, transport and public safety (Zhou, De, & Moessner, 2016). A real-time streaming analytics model was proposed in 2013 to study and recommend contextual topics relating to venues, with the model tested on the London 2012 Olympics (Balduini, et al., 2013) and Milano Design Week 2013 (Balduini, Bozzon, Valle, Huang, & Houben, 2014).

Studies describing the mood of a place, such as positive and negative experiences, began to appear in 2013. In these studies, the mood is either directly expressed or inferred. A 2016 study (Paldino, et al., 2016) analysed sets of photos shared publicly on two photo-sharing web sites popular at the time – Flickr and Picasa – that contained location data. They used descriptive tags attached to the photos to infer the attractiveness of the place, linking attractiveness to subjective well-being. A 2015 study combined Foursquare venue categories and Flickr images within the proximity of the venue to infer the safety and walkability of street segments (Quercia, Aiello, Schifanella, & Davies, 2015).

Analysing the direct expression of affect is dominated by the use of sentiment dictionaries that score words based on emotion value (Liu B. , 2015). An early study analysed geo-referenced Twitter status updates to measure state and city-level happiness across an entire country, the United States of America (Frank, Mitchell, Dodds, & Danforth, 2013). A 2015 study analysed media posted on social networks including Twitter and Facebook to produce CrowdPulse, a ‘real-time semantic analysis of social streams’ (Musto, Semeraro, Lops, & de Gemmis, 2015) giving two examples of its real-world deployment: to identify at-risk areas in Italy (Figure 11), and to monitor the recovery of social capital at L’Aquila, the site of a major earthquake in 2009.

Figure 11. Italian Hate Map from social media posts, reproduced from Musto et al

Image contains figures 11, 13 and 15 from pages 140-142 (Musto, Semeraro, Lops, & de Gemmis, 2015). From left to right: anti-Semitism, violence against women, and racism as expressed in geotagged tweets.

A growing interest in subjective well-being within cities has generated interest in studies that can reveal human emotions (Leyden, Goldberg, & Michelbach, 2011). However, administrative and self-reported data in response to survey questions currently dominate (Glaeser, Gottlieb, & Ziv, 2016). Such approaches have historically been challenged as failing to represent reality on the ground, perhaps most famously by the work of Jane Jacobs (Jacobs, 1961) who challenged claims made about the experiences of those living in slums of major American cities.

The studies presented here raise two concerns. First, whether or not subjective feelings can or should be generalised for a physical location absent of context. As has been expressed earlier in this thesis, variable social and environmental conditions can generate different behaviours. Second, when relying on algorithms that have been developed using historical and/or generalised data, such as sentiment dictionaries, whether or not they contain a bias towards a particular culture or style of linguistic expression. Other research has indicated there is bias when using generalised language dictionaries to infer sentiment of individuals (Sap, Card, Gabriel, Choi, & Smith, 2019).

The examples presented here are a selection of the studies that have been published from 2006 to 2017 using data generated from mobile devices to reveal real-time populations and people-place experiences. Whilst there are undoubtedly limitations, they have demonstrated that samples of human-environment interactions created by mobile devices can reveal diurnal rhythms and subjective experiences. Whilst some have proposed concepts such as producing a real-time census of the city (Kontokosta & Johnson, 2017), there is little evidence for such systems being actively produced at scale. Instead, many of the findings are generalised, even when including time variations. This thesis builds on these approaches but focuses on what mobile data can provide that theories lack: contextual variations that reveal how much behaviour can change within the same space, at the same time interval (hour of the day, day of the week, month of the year) due to different circumstances affecting the landscape and those present.

2.3 Urban Informatics

The literature presented here is just a sample of theoretical and evidence-based approaches to modelling and mapping population behaviours. A review of spatial cognitive research identified that the volume of theoretical research far exceeds practical uses (Kitchen & Freundschuh, 2000). It could be argued that the volume of data-driven research is heading towards a similar position, with numerous individual studies using increasingly advanced quantitative methods but few practical outcomes. The potential of mobile data is in informing real-world decisions by incorporating local contextual knowledge not available when using static counts or generalised assumptions.

As mentioned in chapter one, context-aware computing is an emerging field of mobile systems that can sense their physical environment and adapt their behaviour accordingly (Mohan & Singh, 2013). It is a part of a larger field referred to as urban informatics, a phrase is believed to have been first used in 2003 by Howard Rheingold, in an article titled 'Cities, Swarms, Cell Phones: The Birth of Urban Informatics' (Foth, 2009). In 2011, the Urban Informatics Research Lab at the Queensland University of Technology proposed the following working definition:

> *Urban informatics is the study, design, and practice of urban experiences across different urban contexts that are created by new opportunities of real-time, ubiquitous technology and the augmentation that mediates the physical and digital layers of people networks and urban infrastructures (Foth, Choi, & Satchell, 2011).*

Broadly speaking, it is the analysis and visualisation of urban life from real-world observations located in space and time, captured digitally and continuously, creating the potential for connected environments and context-aware computing.

2.3.1 Context-aware computing

The concept of context-aware computing was proposed in 1991 by computer scientist Mark Weiser. Weisner (Weiser, 1991) imagined what the digital computer might become in the 21st century, coining the phrase 'ubiquitous computing' to imagine future human-computer interactions:

> *"The most profound technologies are those that disappear. They weave themselves into the fabric of everyday life until they are indistinguishable from it." (Weiser, 1991)*

A follow-up article with John Seely Brown introduced the phrase 'calm technology' (Weisner & Seely Brown, 1995), defined as technology that can "*move easily from the periphery of our attention, to the centre, and back*", learning and supporting our wants and needs in the physical environment. The need for calm technology was inspired by the same affordances (Gibson, 1979) that influenced the development of spatial representation models (Hillier & Hanson, 1984). It was seen as creating the potential to "*enhance our peripheral reach,*" (Weisner & Seely Brown, 1995).

Weisner's and Seely Brown's ideas were conceptual, imagining the next stage of technology design and development. A decade later, Mahadev Satyanarayanan (Satyanarayanan, 2001) reviewed the progress made. Satyanarayanan outlined some of the requirements and challenges to achieve

Weiser's vision. He believed that ubiquitous computing needed to be proactive, self-tuning, and minimally intrusive. The challenges to be met include:

- *User intent*: To be proactive, a pervasive computing system needs to be able to track user intent, expertise, and familiarity with the environment to determine which actions will help versus hinder. For example, to respond to a travel disruption may require more information before a decision can be made, or a fast response with some tolerance for error, or it may require cancelling the action. The correct approach will depend on what the user needs or wants to accomplish. Furthermore, expertise and familiarity develop over time. Navigating a familiar place is a different experience to navigating a new location. The system needs to be self-tuning to avoid annoying the user with unnecessary interventions that were previously helpful but are no longer relevant.
- *Context-awareness*: To be minimally intrusive, a pervasive computing system needs to be context-aware. In Satya's words: '*cognizant of its user's state and surroundings and must modify its behaviour based on this information'*. This could include physical location, environmental and social conditions, physiological state, emotional state, personal history, plans, routine activities and preferences. To be context-aware requires knowing the minimum information needed and how to balance between immediate and historical information given no two contexts will ever consist of identical conditions.
- *Transparency, Privacy and Trust*: To function, a pervasive computing system will need to store and analyse substantial amounts of information about a person's behaviours, movements, intentions and preferences. To continue using such a system will require a user to trust it, demanding both transparency and privacy. Algorithmic decisions should be open to inspection and personal data must be protected from abuse or misuse.

The challenges focus on ubiquitous computing operating at an individual level. This thesis adopts a different and less-intrusive approach that is arguably an intermediary step. The focus is on achieving context-awareness but not user intent. Furthermore, context-awareness extends as far as the physical location, environment and social conditions, and potentially some aspect of the combined emotional state of the population present, but not to the extent of a detailed analysis or knowledge of any individual present.

To achieve some level of context awareness requires adopting new approaches to data analytics, increasingly referred to as 'pervasive data science'. A 2017 article defined pervasive data science as characterised by, "*a focus on the collection (sensing), analysis (inference) and use of data (actuation) in pursuit of the vision of ubiquitous computing,*" (Davies & Clinch, 2017).

2.3.2 Big data analytics

Data analysis as a field was envisaged in the 1960s by mathematician John Tukey, differentiating from statistics in its emphasis on procedures relating to the acquisition and preparation of data to be modelled (Tukey, 1962). It arose from the arrival of the digital computer in the 1950s, enabling faster and more complex calculations than had previously been possible. In a 2007 talk (Gray,

2009), computer scientist Jim Gray described data exploration as the fourth paradigm of science, following empirical, theoretical and computational approaches (Figure 12).

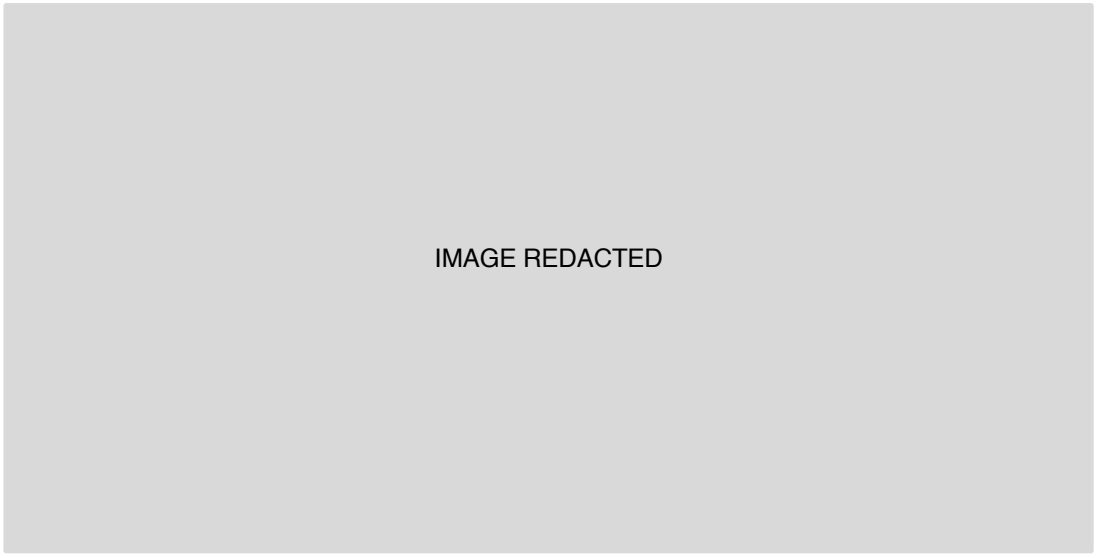


Figure 12. The fourth paradigm of science, reproduced from Hey, Tansley & Tolle

The slide is taken from a transcript of a talk by Jim Gray to the NRC-CSTB in Mountain View, CA, on January 11, 2007. Figure 1, Jim Gray on eScience: A Transformed Scientific Method (Gray, 2009).

Gray posited that we are moving from conducting field studies, developing theories, and running computer simulations reliant on synthetic data, to using computers to capture, analyse and interpret observations about the real-world continually, referred to as 'big data' and requiring a new form of data-intensive science, shortened to data science.

The phrase 'big data' was first mentioned in a 2001 article describing new types of data that cannot be analysed using traditional data analysis or statistical methods due its volume, velocity and variety (Laney, 2001). The 'V's (Table 1) were later expanded to include veracity, virtue and value in recognition of the increasing analysis of behavioural data (Williams, Burnap, & Sloan, 2016). Further attributes have been proposed including visualisation, validity and visibility (Li, et al., 2016). Preliminary analysis of big data can benefit from visualisation to reveal patterns and relationships that can inform hypotheses and identify features of interest to partition the data.

Table 1. The 'V's that differentiate 'big data' from traditional data

'V'	*Definition*
Volume	The amount of data being captured and made available for analysis
Velocity	The speed of data generation, propagation and change
Variety	The range of data formats to analyse, including text, images and audio/video
Veracity	The quality, authenticity and accuracy of the data readings
Virtue	The ethical considerations when using the data
Value	The applicability of the data in decision-making
Visualisation	The ability to identify patterns and form hypotheses to guide analysis

First three introduced by Laney (2001), second three mentioned by Williams, Burnap and Sloan (2016), seventh mentioned by Li et al (2016).

Volume refers to the amount of data being generated and available for analysis. It can refer to the number of records within a dataset, and/or the number of characteristics, or attributes, for each record in a dataset. It would be easy to assume that all 'big' data must be big in terms of the size of the dataset. However, when analysing a specific location at a specific point in time, the number of records may be small, but the data can still be big in other ways. I refer to this as 'little big data' and it is the focus of this thesis.

Velocity refers to how quickly data are being generated but also can change and decay due to being replaced by new data. Variety refers to the different formats of data being created – numbers, text, images, sounds, video, sensor readings. Veracity introduces the potential for uncertainty – the quality, authenticity and accuracy of different data sources. Virtue raises the concern about privacy and ethics if the data refers to human activity and is used to analyse human behaviour, as is the case for this research. An industry report by Deloitte (Guszcza, 2015) highlighted that the data generated by mobile devices is often behavioural:

> *"Much of what we call 'big data' is in fact behavioural data… the 'digital breadcrumbs' that we leave behind as we go about our daily activities in an increasingly digitally mediated world." (Guszcza, 2015).*

The final V is value. Big data is only of use if the results of an analysis can be applied in decision-making. To analyse big data has required the development of machine learning algorithms able to analyse data in ways and speeds not possible using traditional statistical methods. The phrase 'machine learning' was defined by computer scientist Arthur Samuel in 1959 (Russell & Norvig, 2014) as part of the emerging field of artificial intelligence. Machine learning produces algorithms by learning from data without requiring any theory or assumptions about the distribution of the data. However, concerns are being raised about such algorithms when they are used to aid real-world decisions that carry consequences (Bradford Franklin, 2019). When analysing massive amounts of data for patterns, given a sufficiently large data set, any pattern becomes possible:

> *"Over-reliance on data analytics without considering its context risks creating a data-driven technocracy where any unexpected behaviour is suspicious." (Duarte & Ratti, 2016)*

The opportunity that location-based sources of real-time data provide is the ability to tune algorithms to the environment in which they are to be used, to create the context needed to understand whether or not a behaviour being observed is random or deliberate and normal or abnormal. However, such data sources are only samples of the real world, not a complete set of observations as are produced in behaviour maps produced using human observers. A concern that is considered throughout this thesis is whether sufficient statistical rigour can be achieved, or whether an alternative approach is needed when presenting outcomes. One approach is to adopt 'middle range' theories.

2.3.3 Middle-range theories

An in-depth look at the potential of and challenges for data-driven geography (Miller & Goodchild, 2015) cited social scientist Duncan Watts, who made the argument that such volumes of data will not produce general laws but rather will produce 'a more modest type of theory that would include general propositions'. The approach was linked to a call made by sociologist Robert Merton in the 1960s for middle-range theories:

> *"Middle-range theories are empirically grounded: they are based in observations and serve to derive hypotheses that can be investigated. However, they are not endpoints: rather, they are temporary stepping-stones to general conceptual schemes that can encompass multiple middle-range theories." (Merton, 1967)*

A decade earlier, philosopher Nelson Goodman articulated similar ideas in the 1954 first edition of his book 'Fact, Fiction and Forecast' (Goodman, 1990). Goodman anticipated that the arrival of the digital computer would result in the ability to produce complex models involving large volumes of data far beyond the capability of a human to process using traditional statistical methods, and that such models would require a new kind of knowledge, an 'over-hypothesis' to define what the space of reasonable hypotheses might be for a complex, unknown, or uncertain situation.

By the beginning of the 21st century, a similar call was expressed by biological anthropologist Melvin J. Konner:

> *"We crave simple clear comprehensive explanations, the elegance of a Euclidean proof. …A good textbook of human behavioural biology, which we will not have for another 50 years, will not look like Euclid's geometry but more like a textbook of physiology or geology, each solution grounded in a separate body of facts and approached with a group of different theories, all the solutions connected in a great, complex web." (Konner, 2003)*

This is the approach adopted by this thesis: to produce a framework that functions as an adaptive toolkit, creating contextual weights that can adjust a generalised average of behaviour to measure changes that occur given a set of circumstances. The outcomes are not expected to be precise. Rather, they will be indicators of what impact context can have on population behaviours, narrowing a set of over-hypotheses to the most probable outcome for a learned context.

The benefit of utilising real-time data is that such a model can continue to recalibrate on an ongoing basis, learning new contexts as and when they occur, and adjusting recurring contexts as cultural and social norms change. Furthermore, it does not require historical data, nor does it require long-term storage of data. The model is generated within an environment for use within it. The generalisation is in applying the same computational technique to any landscape but with each landscape populating the model with data generated from within the landscape.

2.3.4 Research approach

To summarise the literature that has informed this thesis, there have been substantial developments in the 20th century towards our understanding of spatial cognition: how humans perceive, recall, imagine and react to their environment. The need for such knowledge has become urgent in the 21st century due to the growth in urban populations and demands for cities to create resilient, healthy and sustainable environments. To meet those demands, there is a need for tools to aid decision making in near real-time response to uncertain and changeable conditions. The challenge with simulated models has been a lack of data or theories about real-world conditions and the variability of spatial choices. Behaviour maps generated by human observers have provided such details but have occurred infrequently due to resource constraints and their findings can become outdated as land-use and cultural norms change. Developments in computer science and the miniaturisation of technology have resulted in the generation of real-world observations at scale, and computational tools to collect, process, analyse and visualise such data at new spatial and temporal scales. Early research using mobile data and social media has demonstrated that it can reveal socio-spatial dynamics. However, the outcomes are often presented in the same way as traditional planning models, as generalisations. The focus of this research is to instead explore contextual variation and continuous availability as the key benefits that mobile data traces can bring to the study of people-place interactions. If context can be measured and incorporated into urban models, with continuous or frequent recalibration to ensure it is tuned to its environment, it can enable a closer representation of reality to assist real-world decisions, the contextual forecast in the prediction funnel introduced in chapter one (Figure 1).

The next chapter presents the design of the proposed contextual framework and how it is applied to the case studies presented in the chapters that follow.

3 Profiling the Landscape

This chapter presents the approach taken to learn the socio-spatial dynamic of a landscape using real-time data. The objective is to develop a context-aware framework for studying the variability of human behaviour in urban spaces. The phrase 'profiling the landscape' is deliberate to emphasise that the focus of this research is on the social dynamic of the landscape, not the actions of individuals. All analyses are performed on aggregated data to study population behaviours across space and time. This research mimics the approach taken by behaviour mapping using human-observers, as outlined in chapter two. It draws on contributions from cognitive, data, and urban sciences. Some familiarity with terminology across these fields is assumed. The chapter is organised into three parts. The first part introduces the contextual framework. The second part details the definitions for each layer of the contextual framework. The third part describes the steps taken to acquire and process real-time data used to develop the framework. The technique adopted throughout all studies was to conduct visual analytics using programmatic methods. A substantial amount of code was developed and refined during the research. Samples are included in Appendix B and referenced in this and the following chapters.

3.1 A Contextual Framework

The contribution of this thesis is the development of a contextual framework to model socio-spatial dynamics through the sampling of real-world observations generated by mobile devices. The framework was developed iteratively through a series of case studies presented in chapters four, five and six. Concepts are organised here for reference across the chapters.

3.1.1 Context hierarchy

Applying the concepts introduced in chapter two, this thesis proposes that a hierarchy of three location-based contexts influence people-place interactions: space, time and situation (Figure 13).

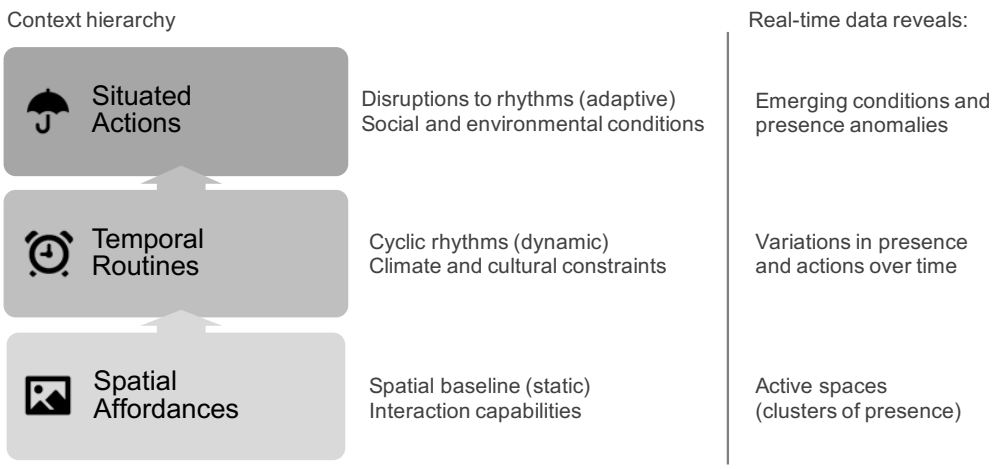


Figure 13. Location-based hierarchy of contexts affecting people-place interactions

Three layers of circumstances within the landscape that can affect behaviour. Each builds on the preceding layer and can benefit from an analysis of real-time data generated within the landscape.

The base layer of the hierarchy is spatial affordance: how the physical features of a landscape create capabilities for action. For example. a green space affords different activities compared with a shopping mall, and a home is used for different purposes than a factory. A local neighbourhood may consist of multiple different buildings and open spaces. Spatial attributes are assumed to change infrequently. A physical landscape may remain relatively unchanged for decades. However, when development does occur, the change can be transformational, such as converting a derelict brownfield site into a residential area with parks and playgrounds. Real-time data can be used to identify active spaces within the landscape from clusters of readings generated by mobile devices. The measure produced at this layer would be a static value such as the average number of people present daily or hourly. It is neutral of time of the day or day of the year.

The second level introduces cyclic time: routine behaviours that have a regular recurrence associated with a time interval, such as time of the day, day of the week and season of the year. It creates the 'social heartbeat' of the landscape, converting a static spatial baseline into a dynamic distribution over time. For example, urban environments would be expected to follow a diurnal pattern, with movements matching the circadian rhythms of human physiology: we are typically more active during the day and rest at night. The types of activities that occur are constrained by climate and culture. Climate defines the expected environmental conditions that produce different behaviours, such as comparing summer and winter activities. Culture influences when and where activities occur, such as defining what is a working day and when school holidays are scheduled. Real-time data can be used to measure the temporal distribution of presence in a landscape. Furthermore, it can potentially reveal different behaviours, such as dwelling within versus moving across the landscape. The measure at this layer is dynamic but generalised, representing the ambient or 'normal' behaviours that occur assuming no unexpected conditions or circumstances.

Whereas the properties of the physical location may change infrequently, routine behaviours evolve to reflect changes in climate and cultural norms. For example, the standard working pattern in the UK for much of the second half of the 20th century was a Monday to Friday presence in a single location. By the second decade of the 21st century, working patterns have changed. In 2014, people meeting certain employment criteria were given the legal right to request flexible working including part-time, job-sharing and working from home. An increasing number of people are working on zero-hour contracts for multiple different companies. Seasons are being altered by climate change. Summers in Australia from 2014 to 2018 were 50% longer than in the mid-20th century (BBC News, 2020). If extreme weather events increase, they may lead to further changes in cultural expectations, such as increased remote working due to travel disruptions. Continuously or frequently sampling data on an ongoing basis creates the potential to produce self-calibrating algorithms able to update and accommodate evolving routine behaviours.

The third level incorporates all acyclic conditions local to the landscape that temporarily disrupt routines. Such changes can occur due to planned events, unexpected disruptions and sensory conditions that affect motivations for being present, attracting people to, or repelling people from, the landscape. This can include unseasonable weather conditions or changes to air quality or noise

levels. These are all situated actions, specific to a location in both space and time. Real-time data can be used to measure the amount of disruption caused by such conditions. It also offers the potential to sense unexpected conditions as and when they emerge, enabling 'opportunistic data science', as envisaged by geographer Harvey Miller (Miller, 2017).

Modelling real-world situations requires embracing uncertainty. People react differently to the same conditions due to malleable preferences. The same people can react differently to the same conditions occurring on different dates and different people can react differently to the same conditions on the same date. The likelihood of changing behaviour will be influenced by whether or not the activity is considered to be necessary or optional, as described by architect Jan Gehl (Gehl, 1987). Poor weather may deter visits to an outdoor space unless the visit is for a special event. As shown in chapter two, precipitation has been found to have a bigger impact on activity levels in shopping areas than in business districts (Pushkarev & Zupan, 1975). This introduces a fourth aspect of contextual behaviour, oriented towards the person: the sensitivity of human reactions to different conditions.

These four aspects form the framework for using real-time data to study socio-spatial dynamics and learn how different sets of circumstances can affect people-place interactions.

3.1.2 Context formula: P-STAR

A formula is proposed [2] to describe measuring variations in population behaviours for different contexts using the context hierarchy presented in Figure 13. It is an adaptation of Lewin's behaviour equation [1] introduced in chapter one.

$$P = f(S, T, A, R) \qquad [2]$$

where:

$P = population\ behaviour$
$S = spatial\ baseline\ (static\ value)$
$T = temporal\ routine\ (ambient\ dynamic)$
$A = situated\ action\ (adaption)$
$R = reaction\ sensitivity\ (uncertainty\ range)$

P represents the population behaviour being measured for the landscape, whether a signal of presence or the semantics of experience. *S* represents all attributes for producing a static baseline for the landscape. *T* represents all cyclic attributes that can change behaviour, converting the static spatial baseline into a rhythmic distribution over time, the social heartbeat of the landscape under normal conditions. This can include season, day of the week, hour of the day, whether or not the day is part of a working week or a holiday etc. *A* represents all situated actions that produce temporary anomalies and disrupt space-time patterns in the landscape. This can include events occurring at venues and different weather conditions that affect human-environment interactions. *R* indicates the variation that can occur for the same set of circumstances due to the malleable preferences of people. Rather than produce an absolute or precise measure, it is anticipated that outputs will need to incorporate some measure of uncertainty to identify whether or not variable population behaviours are within normal ranges or an indication of an abnormal situation.

The formula is not presented as a mathematical equation. Population behaviours will emerge from a complex interaction between attributes and the nature of the relationships is uncertain. The objective of this research is to investigate those relationships and determine the feasibility of modelling population behaviours from a sample of interactions captured digitally via mobile devices. Computational statistics and machine learning will be applied to various sources of real-time data throughout chapters four, five, six and seven.

3.1.3 Conceptual model

Recalling the prediction funnel from chapter one (see Figure 1) the impact of this research is in creating contextual forecasts tuned to the conditions of the environment. Such an approach would enable informed decisions and interventions grounded in reality. Figure 14 visualises the concept.

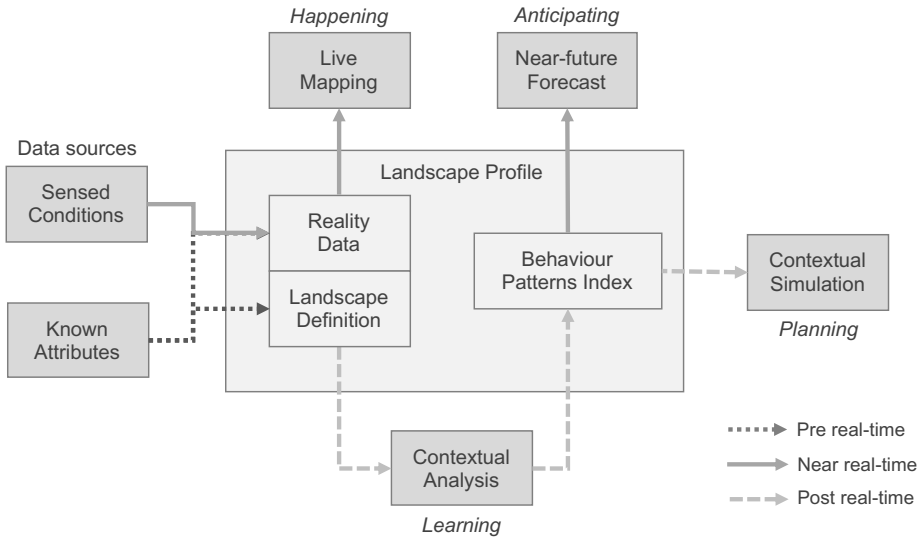


Figure 14. Concept for producing and applying contextual insights from reality data

The proposed contextual framework consists of three elements: a definition of the landscape, reality data that continually or frequently update, and an index of behaviour patterns learned from the contextual analysis.

In its simplest form, the framework consists of three elements: the landscape definition that determines the baseline attributes for the P-STAR formula, reality data sensed in real-time and an index of behaviour patterns learned from the contextual analysis. Reality data can be used on its own to map live behaviours. A contextual analysis learns from the reality data to measure how behaviours within the landscape change for different contexts and produces behaviour patterns. Behaviour patterns can be used to simulate a learned context to aid planning. Combining reality data with behaviour patterns can enable a near-future forecast given current sensed conditions and potentially make recommendations based on previous experiences of similar contexts.

To implement the conceptual model requires defining the landscape, acquiring and processing reality data, and performing a programmatic analysis to measure context-specific behaviours. The remaining three sections of the chapter detail these three elements. The methods were developed and refined iteratively through the studies presented in chapters four to six. Final versions of techniques are presented here.

3.2 Defining the Landscape

Adopting the terminology of behaviour mapping introduced in chapter two (Gifford (Ed.), 2016), to profile a landscape requires defining its boundary, defining the time intervals for comparing behaviours; and developing categories to describe the different contexts that can occur, organised as three levels: space, time and situated action.

3.2.1 Spatial context

The spatial context of the model defines the boundary containing the landscape and active spaces within it that may experience different population behaviours for different contexts. To study population behaviours requires performing some level of aggregation to determine who or what is counted. An outer bounding box defines the boundaries of the landscape being profiled. The landscape can then be studied as a whole or sub-divided into zones using either a thematic or node-based approach (Figure 15).

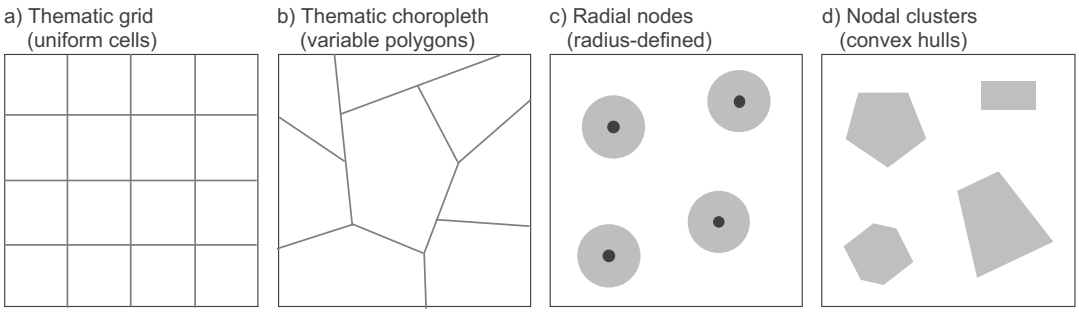


Figure 15. Examples of thematic and nodal-based areal aggregation

Thematic maps divide space into contiguous segments. Nodal maps focus on non-contiguous locations of interest.

Thematic aggregation divides the landscape into contiguous segments, either as a grid of uniform cells or as a set of polygons that can individually vary in shape and size. All data points that fall within the boundary of the landscape will be allocated to a segment. A limitation of this approach is that the choice of cell size and shape can affect aggregated outcomes. For example, mis-representing a small but dense cluster if it is positioned across two cells. This is referred to as the modifiable areal unit problem (MAUP). The MAUP can result in what is referred to as an ecological fallacy, where aggregate averages for an area do not reflect the actual values for any individual within the area being averaged (Openshaw, 1984). This is discussed further in chapter five.

Nodal aggregation identifies locations of interest – nodes – within the landscape and counts only data points that fall within the range of each node. For this research, the range is defined either as a radius around the centre of each node or drawn as a convex hull to form a polygon around clusters of data points. A challenge is how to set the range. Altering the radius around nodes is a variation on the MAUP and drawing convex hulls around clusters of data points assumes the data sample is representative of the population. To perform spatial aggregation of data requires making assumptions that may affect outcomes. These assumptions are communicated throughout the case studies presented in the following chapters.

Table 2 outlines the spatial data scales and assumptions adopted in this thesis. Initially, an arbitrary bounding box and cell-size were used to create a grid for thematic analyses of the QEOP landscape described in chapter four. Later studies use a grid aligned with a global system for population measures, LandScan, introduced in chapter five.

Table 2. Spatial scales for measuring population behaviours

Category	*Notes*
Thematic	
Grid	Bounding box to define landscape outer limits.
LandScan Cell	Standard LandScan cell size across London. LandScan cells are 30 arc seconds tall (South-North) and wide (West-East) globally. For London, this equates to 1km tall by approximately 0.62km wide.
Pixel Cell	1/16th the size of a LandScan cell, 250 metres tall by approximately 155 metres wide.
Polygon	
Output area	Administrative output area used for formal population measures such as the census and socio-economic indicators.
Manually drawn	Manually drawn polygon to enclose a geographic feature of interest.
Computed	Data-generated outline such as using a Voronoi model or connecting outer points as a convex hull to enclose a cluster of data points.
Points	
Raw data point	Coordinates written in latitude and longitude provided up to 7 decimal places, sometimes with a location accuracy estimate.
Rounded point	Coordinates written in latitude and longitude rounded to 3 decimal places. This creates a location accuracy of approx. +/- 55 metres around the data point (points will be plotted approx. 111 metres apart).

Manually drawn polygons are created as shapefiles using QGIS.

The spatial context can also define land-use potential as an attribute to describe the actions expected due to physical features and attractions within the landscape. For example, behaviours are anticipated to be different if the landscape or a zone contains a park than if it contains a shopping mall. It may be possible to learn social dynamics associated with land-use potential by categorising the landscape or zones within it and comparing landscapes with similar land-use potential. Such a potential is represented in the conceptual model but is beyond the scope of the studies presented here. This thesis focuses on contextual variation within a landscape.

3.2.2 Temporal context

Time is incorporated in two ways: as intervals to measure duration, and as an attribute that affects behaviour – the opportune time for actions to occur due to climate and cultural constraints.

3.2.2.1 Time intervals

A unit of time is an interval used to measure the occurrence and/or duration of events. Individual points in time are recorded in years, months, days, hours, minutes and seconds. Smaller and larger scales are possible but, for everyday human interactions, the duration of an activity or an individual point in time on a given date is measured in hours and minutes.

Dates and times recorded digitally are captured in Coordinated Universal Time (UTC) or using a local time zone. UTC is set at the same time zone as Greenwich Mean Time (GMT) and neither change for daylight saving hours (DST). The UK time zone is Greenwich Mean Time (GMT) during the winter and British Summer Time (BST, GMT+1) during the summer for daylight saving. The clocks move forward one hour on the last Sunday in March, and back one hour on the last Sunday in October. If a time zone is not specified for a data source, it is assumed to be UTC. For analyses, data timestamps will be converted to match the local time zone.

Table 3 provides a list of intervals and assumptions used in this research. The smallest temporal aggregation will be hourly. Sub-hourly intervals were considered for evaluating peaks times, such as on the hour, when most working shifts start and finish. However, some of the data sets available lacked temporal granularity below the hour and others were too small in volume for comparative analysis at that scale. When readings are received pre-scaled to hourly, and it is not known what method was used, a timestamp coded as 14:00 could be for a reading from 13:31 to 14:59 depending on if the reading is rounded to the nearest hour or simply cut-off to the hour during which it was generated. When hourly counts are small, groups of hours as periods may be preferable to minimise errors by containing expected peak hours within the same period.

Table 3. Temporal intervals for measuring context and behaviour

Category	*Format*	*Notes*
Date	YYYY-MM-DD	Date identifier, Y = year, M = month, D = day
Day of week	Mon to Sun	Day of the week, Monday (1) to Sunday (7)
Wk/We	wkday (Mon-Fri) or wkend (Sat and Sun)	Day average across weekdays (Monday to Friday) or weekends (Saturday and Sunday)
Hour	HH (00 to 23)	24-hour day. Readings will be aggregated to the hour of the time stamp
Day Period	Peak, Off-peak day, Off-peak eve, night	Peak is 07:00 to 09:59 and 16:00 to 18:59 Off-peak day is 10:00 to 15:59; Off-peak eve is 19:00 to 23:59; Night is 00:00 to 06:59
Detailed Period	As above but with off-peak more detailed	As above but Off-peak day is separated into Off-peak am (10:00 to 11:59), lunch (12:00 to 13:59) and off-peak pm (14:00 to 15:59)
Minute	00 to 59	HH:MM. Each device in a dataset will have a maximum of one reading per HH:MM per day.

3.2.2.2 Opportune time

Whilst time units are universal, activities in urban outdoor spaces can vary by country due to variations in climate and culture (Carr, Francis, Rivlin, & Stone, 1992). All exhibit a cyclic pattern: they recur at known, or expected, times and repeat on an annual basis within the location of interest, producing routine behaviours or habits.

Climate

Climate is a naturally occurring phenomenon that produces different seasons annually and determines the weather conditions anticipated during each season. Winter is expected to be cooler and wetter, summer is expected to be warmer and drier. Expectations vary regionally due to the topography of the planet. The UK is located in the northern hemisphere. January is the middle of winter and July is the middle of summer. The UK does not typically experience extended periods of hot or cold temperatures or extreme events such as hurricanes or pro-longed droughts. Instead, the climate is generally cooler and wetter from October to April and warmer and drier from May to September but there are substantial variation and overlap. Appendix A contains a summary.

Culture

Cultural factors produce human-generated expected and recurring contexts, such as religious and public holidays, the typical working week (e.g. Monday to Friday), working day (e.g. 9am to 5pm), and non-work patterns (e.g. shopping on Saturday, rest on Sunday). They are regional rather than neighbourhood scale and different demographic and socioeconomic groups may observe different routines and religious or cultural festivities. For example, some work may be organised on a regular schedule lasting approximately from 9am to 5pm whilst other work may be organised on an 8-hour shift rotation to provide 24-hour coverage. Age groups also have different routines and priorities, typically grouped as children/students, adults of working age, and adults over 65 years old, the majority of whom are assumed to be in retirement. This age may need to be extended over time as the retirement age when people become eligible for a state pension is changed. 65 is the category used within the 2011 UK census but that age will have increased at least 67 when the next census is conducted in 2021. It is further evidence for the need for algorithms that can recalibrate over time. Some cultural factors are linked with climate. For example, tourism tends to peak in summer months, which is June to August in the UK (see Appendix A).

Cultural attributes include differentiating between holidays and work/study periods (‘term-time’), and different types of holiday. A public (or bank) holiday is a single weekday considered a holiday from work and education, although retail and leisure facilities usually remain open. The majority of public holidays within the UK are scheduled to occur on a Monday. Two public holidays, Easter and Christmas Day, differ from all others. All non-essential facilities are closed except for some food/drink establishment. School holidays consist of week-long half-terms at the mid-point of each school term, two weeks over Christmas and Easter, and six weeks over the summer (July to August). The summer period is the traditional time when working families are likely to take a vacation, and when tourism within the UK is at its highest for outdoor attractions.

3.2.3 Situation context

Situated actions are the third level of the context hierarchy and include all temporary conditions of a landscape that are acyclic. They do not follow a predictable cycle and can temporarily disrupt both the spatial attributes of the landscape and routine behaviours that occur within it. Attributes are organised into two categories: sensory conditions and social conditions.

3.2.3.1 Sensory conditions

Sensory conditions are elements that create or alter the attractiveness of a location to people. This is described by phenomenologists using terms such as ‘ambience’, ‘atmosphere’, or ‘vibe’ (Bille, Bjerregaard, & Sørensen, 2015). They are changeable aspects of the environment and include air quality, visual appearance, light, noise and olfactory experiences. Outdoor environments are also exposed to variable weather conditions such as changes in temperature, precipitation, wind, and general outlook (sunny, cloudy, overcast etc.).

Weather conditions are known to affect pedestrian behaviours and can vary substantially within a season and from day to day. As described in chapter two, studies in New York and Copenhagen indicated that pleasure walking increases noticeably when the temperature reaches 13 degrees Celsius, and that persistent precipitation results in shorter and fewer pedestrian journeys, although optional activities such as shopping are more affected than necessary activities such as commuting to work. The likelihood of experiencing different weather conditions is associated with climate. For example, in the UK, it is unlikely to snow during August or to experience a heatwave in January.

3.2.3.2 Social conditions

Social conditions are human-generated activities that occur within the landscape. They can affect sensory conditions, temporarily change physical attributes, and disrupt routine behaviours. They can be divided into two categories: scheduled events and unplanned incidents. Whilst neither follow a recurring cycle, scheduled events enable predictions and interventions to be made in advance whilst unplanned incidents are unpredictable and can only be reacted to.

Scheduled activities are planned and can be divided into two sub-categories: organised events and planned disruptions. Organised events attract an audience to the location that is not part of the normal routine population and may temporarily create a different atmosphere such as attracting a large crowd and/or generating loud noise for a sustained period. Planned disruptions temporarily affect physical features, such as construction work blocking a route across the landscape. They may repel people from a location or displace people to other nearby alternatives. In both cases, the location in space and time is known in advance including its expected duration.

Unexpected incidents are not planned and are unpredictable. They include emergency incidents such as a major road traffic accident that can affect access to, or alter, routes within the landscape, extreme weather events that repel people from the landscape, and unexpected public gatherings. The location is not known in advance or is discovered with an insufficient warning to design or plan an intervention. Instead, decisions need to be made in response to conditions as they emerge, with limited time to coordinate a response.

To summarise, Table 4 lists contextual attributes that can affect behaviour spatially and temporarily. Each attribute may be further sub-divided into properties to describe the attribute. For example, different types of event, and different start times, may have different effects on population behaviours. Combined with Table 2 and Table 3, the three tables describe the attributes to define

a landscape and measure contextual variations in behaviour using the P-STAR formula. For completeness, the table includes additional attributes not directly investigated during this thesis due to lack of available data or due to scope but recommended for future work research directions: Potential (land-use that provides capabilities and constraints for behaviour), and Environmental (non-weather sensorial elements that can attract, repel or displace populations temporarily).

Table 4. Contextual landscape attributes for measuring behaviour change.

Level	*Attribute*	*Description/Properties*
Space (Static – general capabilities)		
	Land-use*	Land-use potential for different behaviours to occur (e.g. 'park', 'shops', 'office', 'school', 'transport hub', 'stadium')
Time (Cyclic dynamic – routine behaviour change)		
	Season	Affects the likelihood of certain weather conditions, tourism levels etc.
	Day status	Differentiating between working week/academic term and holidays
	Holiday type	Type of holiday: school, public, religious
Situation (Acyclic dynamic – adaptive behaviour change)		
	Weather	Temperature; Precipitation; Wind; Outlook (sunny, cloudy, overcast)
	Environmental*	Air quality index (UV, pollution); light; vision, noise; sound, smell
	Activity status	Is there an activity occurring (Y/N)
	Activity	Category (e.g. music, football etc.); Start time; Duration; Location etc.
Reaction (Variation in readings)		
	Uncertainty range	Unexplained variations in readings for a given set of circumstances

* not used in this research but potential for future research directions.

3.3 Data Mining Reality

Physical mining is the process of discovering and extracting valuable minerals from the earth. Data mining applies this analogy to the discovery and extraction of hidden patterns, referred to as latent information, from data (Miller, 2010). Data mining of reality is simply a subset, focusing specifically on real-world observations located in physical space and time.

The research is centred on three types of data: mobile device readings from the logs of a public wireless (Wi-Fi) network, readings emitted by an installed mobile app, and interactions generated for two popular social media networks, Twitter and Foursquare. The Wi-Fi data was generated within a single landscape, the Queen Elizabeth Olympic Park (QEOP) in Stratford, East London, which is introduced in chapter four and the focus of the studies in chapters four, five and six. The other sources generate readings on a global scale. They are applied to the QEOP as well as the final study that introduces and explores three other landscapes within London. Additional reality sources were acquired for specific studies, either as comparative measures or to provide additional context. Sources are summarised in Table 5 for reference and detailed in the relevant case studies.

Access to data sources was proprietary, academic, public or 'public-at-time'. Proprietary sources were provided for use in this research only and cannot be shared in their original form. They can only be viewed as aggregate analyses and visualisations. Academic sources are available for general academic research. Again, the data cannot be shared publicly, only the results of any analysis. Public sources are publicly available both for this and any other applications. Public-at-time refers to sources that were retrieved using public APIs for data that is made available for a limited amount of time after it is generated and must not be retained beyond the conclusion of the research. For reproducibility, Twitter allows sharing of status IDs. One source – Weather Underground – was acquired in 2018 and free access to the live data was withdrawn although archived data can still be viewed and downloaded.

Whilst a single landscape is used for the core research of this thesis, the methods to acquire and prepare reality data are generalised and can be applied to any landscape or accessible reality data source. The QEOP is a demonstrator to develop the contextual framework and test the feasibility of profiling a landscape using only samples of digitised real-world observations. The data sources used were those accessible for research at the time of the studies. Most of the data sources were acquired post-real-time. However, the same methods can be applied to live data feeds.

Table 5. Summary of reality data sources used in chapters four to seven

Source	Collection period	Time interval	Spatial range	Data coords	Provider	Access	Studies	Summary
Wi-Fi 1	Mar 2016; May to Aug 2016	Real-time	QEOP	Access point	LLDC partner WiFiSpark	Proprietary	Ch.4	Data provided as a file containing raw system logs from which individual device connections to access points can be extracted
Wi-Fi 2	Jan to Dec 2017	Day total	QEOP	Access point	LLDC partner WiFISpark	Proprietary	Ch.4	Data file retrieved from an online dashboard as a day count, the total number of devices that connected to the Wi-Fi network during the day.
Twitter 1	Mar 2016 to Jun 2017	Real-time	QEOP, Westminster Bridge	Device GPS or inferred	API	Public	Ch.4 & 6	Query to Twitter search API for tweets matching query criteria (geo-tag range or keyword matches) posted during the previous 6 to 9 days, performed weekly during 2016 for QEOP and in response to incidents at other London landscapes during 2017
Foursquare	Mar 2016 to Aug 2016	Hour avg	QEOP	Venue	API	Public at time	Ch.4 & 6	Query to Foursquare search API for 'here now' count at venues located within 2.5km radius around centroid in QEOP. Queried every 15 minutes. Counts aggregated to produce an average count 'here now' per hour
Mobile app 1	June 2017	Partial real-time	Greater London	Device GPS rounded	OpenSignal	Academic	Ch. 5 & 7	Data provided as a file containing readings with device ID anonymised, timestamp obfuscated to hourly, spatial coordinates reduced to 3 decimal places, and no more than one reading per 15 minutes within the hour
Mobile app 2	May 2017	Real-time	QEOP	Device GPS	OpenSignal	Proprietary	Ch.5	Data provided as a file containing raw readings: full timestamp and full coordinates with location accuracy estimate, and all readings generated
Mobile app 3	Mar 2017	Real-time	Westminster Bridge	Device GPS	OpenSignal	Proprietary	Ch.7	Data provided as a file containing raw readings: full timestamp and full coordinates with location accuracy estimate, and all readings generated
LandScan	2015, 2017	Static	Greater London	Grid cells, 30 arc seconds	Oak Ridge National Laboratory	Academic	Ch.5 & 7	Data downloaded from online service. Provided as a bitmap image. Cell-based population counts extracted from image using QGIS
Crime records	Mar to May 2017	Partial real-time	Greater London	Nearest public street	Metropolitan Police Service	Public	Ch.5	Data downloaded from London Data Store. Readings have timestamp obfuscated to month. Location jittered to centre of the nearest public street
Fire incidents	Jan to Jun 2017	Real-time	Greater London	Incident location	London Fire Brigade	Public	Ch.5	Data downloaded from the London Data Store. Individual fire incident records including timestamp and coordinates for location of the incident
TfL journeys	Sep - Nov 2017	Quarter-hour avg	Central London	Tube station	Transport for London	Public	Ch. 7	Data downloaded from Transport for London. Entry/exit counts at gates to selected tube stations on the London Underground. Counts provided per quarter hour, averaged across collection period excluding school holidays or industrial action
Webcam	Mar 2016; May to Aug 2016	Hour total	QEOP	Webcam location	LLDC partner Movement Strategies	Proprietary	Ch.4	Data provided as a file containing hourly counts per camera, incoming and outgoing, per day during collection period. No access provided to camera images or computer vision algorithm used for counting
Weather	Jan to Dec 2017	Day highlights	QEOP	Sensor location	Weather Undeground	Public at time	Ch. 4	Temperature high, low and average; Precipitation, Wind

3.3.1 Acquisition and preparation

Datasets are either provided as text files by a third-party, as is often the case with proprietary data sources or retrieved programmatically from an online service by submitting a query using an application programming interface (API). This is a standard technique for acquiring online data and is not documented here. The parameters used to construct queries are detailed in the relevant studies and code samples are included in Appendix B. Figure 16 describes the preliminary steps to prepare acquired data for use in contextual analysis.

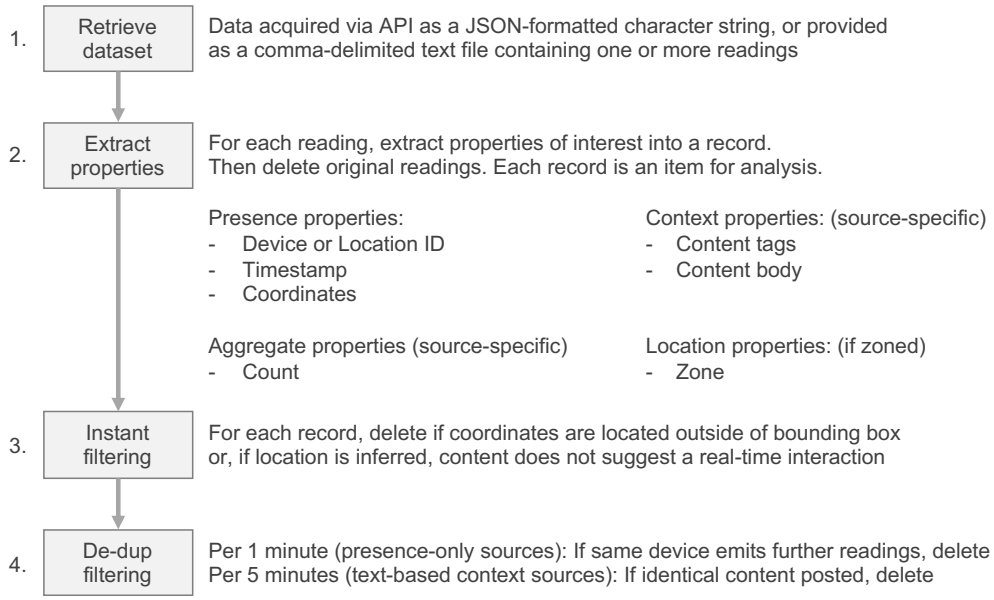


Figure 16. Steps to prepare reality data for use in contextual analysis

All data sources require a minimum of three properties: device or location identifier, timestamp, and coordinates.

For each dataset retrieved, readings are stored as delimiter-separated strings of text. The delimiter enables the separation of individual properties contained within the string for each record. Only properties of interest need to be retained. The rest of the reading can be deleted.

For presence measures, the minimum properties required are a unique identifier for the device or location the reading relates to, a timestamp capturing the date and time the reading was taken, and coordinates for the location where the reading was taken. If the data source produces aggregate measures, such as an hourly count of readings, then the count is stored as a separate property. If the data source contains contextual information, such as a sensor reading or text message, the information is stored as content tags and the content body, with content tags specific to each data source. The timestamp can be further sub-divided into time intervals of interest to assist analyses: date, day, hour and minutes. If the landscape has been segmented into zones, then each reading can be assigned to the zone which its coordinates fall within.

This process produces a two-dimensional array of data. Each row is a record, each column contains a property being retained. Once properties have been extracted, original readings are no longer required and can be deleted. The remaining data is filtered to remove any irrelevant records. When querying for geotagged data, the standard technique is to define a centroid and radius distance

around it. This produces a circle and all data falling within the circle will be returned. The landscape bounding box is a rectangle, hence some readings will fall outside the landscape, assuming the circle is drawn large enough to contain the rectangle. In chapter six, an additional technique is explored: inferring location from the content of text messages.

The first three steps can be performed in real-time if live data sources are available. The fourth step requires at least a one-minute delay to remove duplicate records, either identical readings retrieved twice or additional readings that are not needed or indicate an artificial bot is posting content rather than a human being. For text-based sources, if identical content is posted in quick succession, within the space of 5 minutes, it is assumed to be being automatically generated using an artificial bot, a computer program that mimics human behaviour. Only the first instance is retained, duplicates are deleted.

The prepared data can then be and comparisons can be made between known contexts. The data can be analysed for the landscape as a whole and, if the landscape has been sub-divided into zones, also as a spatial distribution across the landscape. To measure actions and assign spatial familiarity requires further preparation to assign visit attributes.

3.3.2 Assigning visit attributes

For data sources generating multiple readings per device, the readings can be assigned visit behaviours. A unique value of mobile data, when compared with human observers, is the ability to analyse action behaviours. When reliant on human observers, such an ability is only feasible for a very small area, where the human observer can maintain line-of-sight visibility of each person present. Once mobile data has been collected over a period, whether hour, day or month, the data can be analysed to identify trips and stages whilst present in the landscape, creating visit attributes (Table 6). The attributes are appended to the data as additional properties for analysis.

Table 6. Visit attributes assigned to behavioural data

Attribute	*Values*	*Period*	*Notes*
Visit Type	Habitual or Explorer	7 to 21 days	Explorer is the default. Updates to Habitual if present on 3+ days within the period
Trip ID	Auto-numbered	1 to 24 hours	Identifies different trips by the same device each day. A new trip is assumed if 3+ hour gap in readings
Stage ID	Auto-numbered	1 to 24 hours	Identifies different stages of a trip: movements between zones of the landscape during the trip.

Visit type is the simplest attribute to assign. It requires at least 7 days of consecutive data, preferably 21 days. A device is considered a regular visitor (‘habitual’) if present on at least three days during the period. If present on only one or two days, the device is assumed to be an infrequent visitor (‘explorer’). The terms ‘habitual’ and ‘explorer’ are adopted from the terminology of environmental psychologist Robert B. Bechtel (Betchel, 1967) who identified that people who have frequented a place several times exhibit different spatial cognition to those who are visiting for the first time. The choice of three as the differentiator is arbitrary. It is assumed that, after three

visits, a person would be familiar with the layout of the landscape. No more than 21 days is required, and familiarity could be tracked on a rolling period of 3 weeks. This would require device identifiers to be stored for a maximum of 21 days.

A device may visit a landscape more than once on the same day. For example, commuting from home to work in the morning, and then commuting from work to home in the afternoon. Each visit would be a different trip. A count of devices present daily may not reflect presence at any time interval because it does not consider devices making multiple trips or variations in durations whilst present. To analyse different trips and behaviours during trips, each reading is assigned a trip ID and stage ID programmatically (Figure 17).

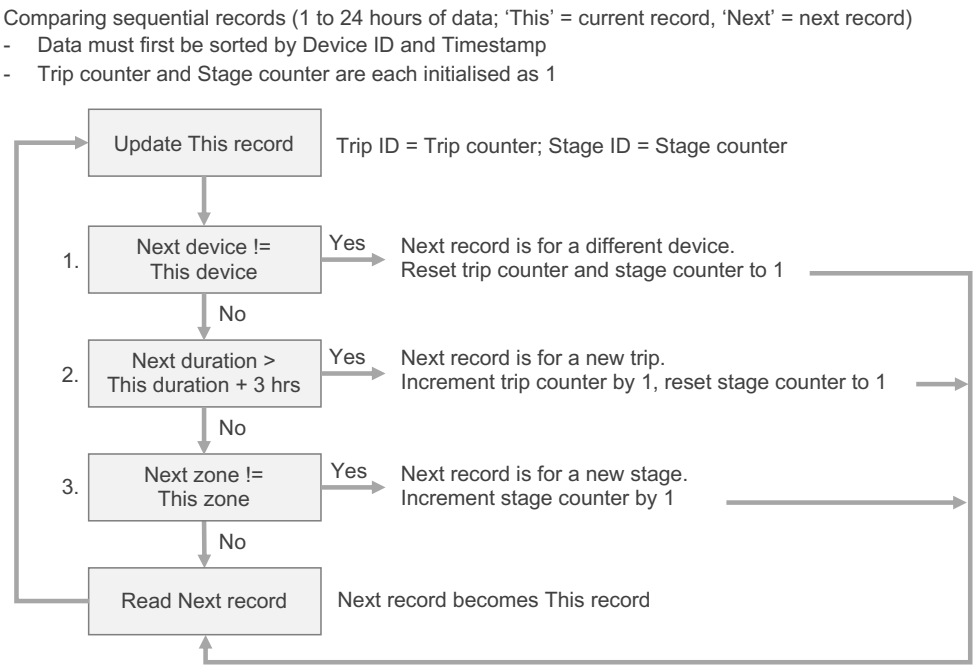


Figure 17. Steps to calculate trips and stages for device visits to a landscape

At intervals of at least one hour up to daily, a dataset is first sorted by device ID and timestamp. Records are then compared sequentially to determine if the next record is a new trip and/or a new stage. 'This' refers to the properties of the current record. 'Next' refers to the properties of the next record.

The data must first be sorted by device ID and timestamp. Two counters, a Trip counter and a Stage counter, are initialised each with a value of 1. The data is then read in sequence. The current 'This' record is assigned a Trip ID matching the current value of the Trip counter and a Stage ID matching the current value of the Stage counter. Three checks are then performed to determine if either counter should be incremented for the next record. If the next record is for a different device ID, then both counters are reset to 1. There are no further trips or stages for the current device. Else, if there is a gap of at least 3 hours between consecutive readings for the same device, then the device is assumed to have left the landscape between readings and the next visit is a new trip. The trip counter is incremented by 1 and the stage counter is reset to 1 because the next record will be the first stage of a new trip for this device. Else, if the next reading is in a different zone (if the landscape has been zoned), the trip counter remains the same, but the stage counter is

incremented by 1 because it is a new stage during the same trip. Else, the next reading will be for the same stage of the same trip for the same device.

Three hours is a long gap between readings. However, for at least one data source used in this research, the timestamp was provided as hourly for individual readings. Testing for two+ hours would mean readings barely one hour apart would become separate trips (for example, 11:59 and 13:01 would be stored as 11:00 and 13:00). Ad-hoc activities such as lunch breaks could last just over one hour and should be considered a single trip.

3.3.3 Programmatic analysis

All steps, from acquiring data to preparing for and conducting contextual analysis and visualisation, are performed programmatically using scripts developed as part of this thesis and using open software tools and languages. Open-source software is freely distributed and maintained non-commercially. All scripts are written in the programming languages Python or R. Animated visualisations are written in Java to run in the graphics software Processing. Web-based interactive visualisations are produced using a combination of HTML, JavaScript and interactive graphics libraries D3.js. Creation and editing of map shapefiles have been completed using QGIS. Samples of the scripts are included in Appendix B.

The use of programmatic analysis through scripts enables the framework to scale in terms of the volume of data, in types of data, in incorporating different algorithms as appropriate to the data type and analysis being performed, and in the integration of outputs with third-party tools and models. The scripts can be easily inspected and updated to modify any of the steps, and to incorporate new and different data sources and algorithms as and when they become available and of interest. The intention is that the framework developed in this thesis could be implemented as a real-time model integrating live data feeds. This is discussed further in chapter eight. The focus of the studies presented in the following chapters is in conducting a data-intensive analysis of real-time data sources for their value in modelling socio-spatial dynamics.

This chapter has described the contextual framework that was developed iteratively through a series of case studies presented over the next three chapters. Each chapter documents a stage in the development of the framework. The first case study explores a single behaviour setting – a large urban park. The primary data source is the park's embedded Wi-Fi network, behavioural data traces captured by access points deployed throughout the park. Further data sources are acquired for comparison. The second study expands from the behaviour setting to a bounding box encompassing it to enable real-time data to be aligned with administrative data sources that are aggregated to defined areas. It explores the potential to produce an active population estimate and spatial distribution by blending static administrative data with real-time observations. The third study focuses on extracting additional contextual information and subjective experiences from text messages posted whilst present in a landscape, exploring the potential to learn contextual insights without requiring advance knowledge about the landscape or events planned within it.

4 Case 1: A Connected Landscape

This chapter begins with an introduction to the landscape used for research spanning this and the next two chapters – the Queen Elizabeth Olympic Park (QEOP) in Stratford, East London in the United Kingdom. It then presents two studies using reality data located within the park to measure how visits to the park vary over space and time. The first study examines variations in the number and distribution of mobile devices connecting to the park's Wi-Fi network at daily and hourly scales. Additional data sets are also evaluated to consider if there is consistency across disparate real-time sources and whether or not combining them can provide additional context and reduce uncertainty. The second study focuses on the volume of device connections daily to the Wi-Fi network across a full year, to evaluate a wider range of contexts including seasonal differences. The chapter concludes with a summary of the findings and its contribution to the proposed P-STAR contextual framework.

4.1 Landscape Introduction

The landscape used for the case studies presented in this and the next two chapters contain the Queen Elizabeth Olympic Park (QEOP) located in Stratford, East London in the United Kingdom (Figure 18). As part of the Smart London plan launched in 2013 (Smart London Board, 2013), it was announced that the QEOP and surrounding areas would be a testbed for data-driven innovations, launching in 2016. Thanks to a collaboration between UCL and the London Legacy Development Corporation (LLDC), data from the park was made available for academic research during 2016 and 2017 and forms the basis of the research presented in this chapter.

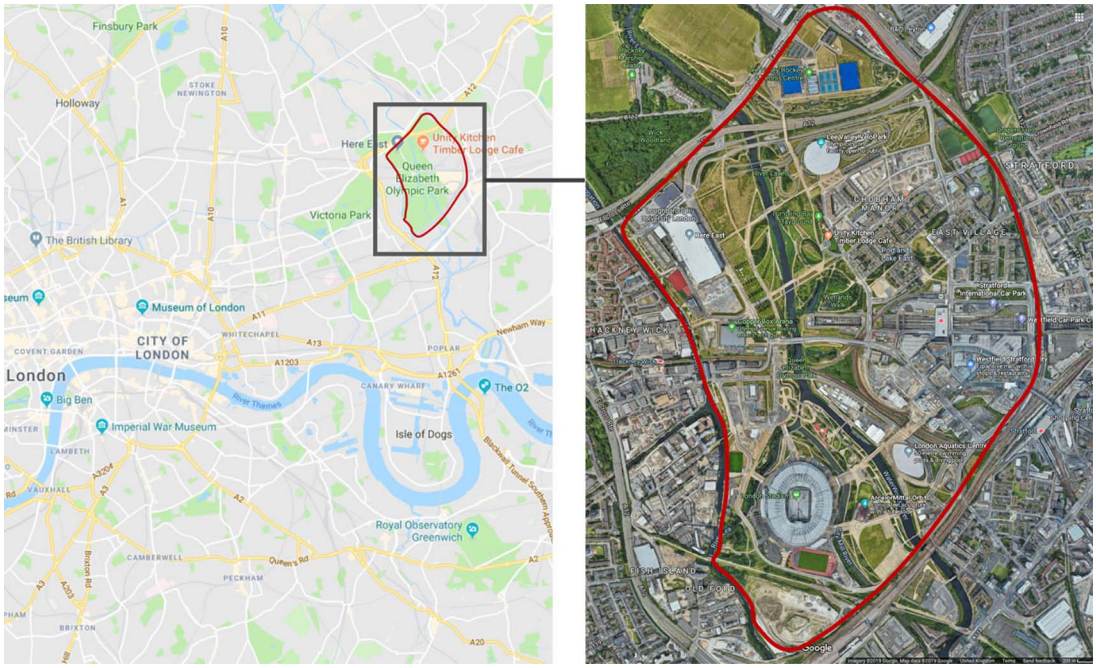
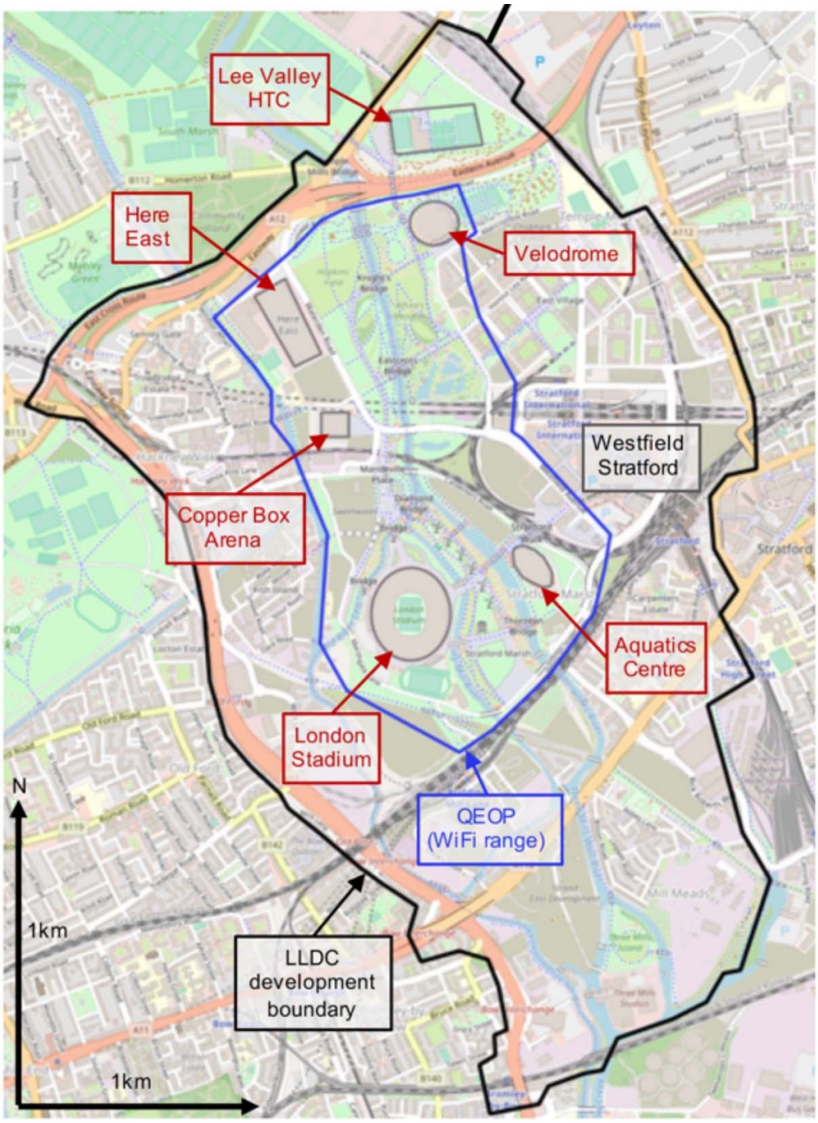


Figure 18. Map showing the location of the Queen Elizabeth Olympic Park

The image on the left shows the location of the QEOP in relation to the City of London (the centre of London). The image on the right shows satellite close-up of the bounding box containing the park. Map data © 2019 Google.

Host to the London 2012 Summer Olympics, the QEOP is the largest new urban park created within the UK for over a century (LLDC, 2012). The park is managed by the London Legacy Development Corporation (LLDC) as part of a post-Olympic legacy commitment (LLDC, 2016) to regenerate what was previously a contaminated industrial wasteland. The development area consists of 257 acres of green space with 6.5km of waterways and surrounded by five new neighbourhoods (LLDC, 2012) containing a mix of residential, industry, office, retail and educational buildings. The park provides a range of outdoor attractions including wetlands, parkland, paths and trails, playgrounds, cafes and seating areas, and it is the location of six built venues including a 66,000-seat stadium. There are national transport hubs and one of the largest retail centres within Europe – Westfield Stratford – located on the eastern border, with local transport hubs also to the west and south of the park. Figure 19 shows the boundary for the development area and the park itself.



Map background by OpenStreetMap, under ODbL, as of August 2019
Produced in QGIS 3.4.2. Map projection: EPSG:3857

Figure 19. LLDC boundary surrounding the QEOP

The outer boundary shows the complete development area that is the responsibility of the LLDC. The inner blue boundary outlines the area of the QEOP covered by the park's public Wi-Fi network.

The park opened to the public in 2014. The London Stadium (formerly called the 'Olympic Stadium') was reopened in June 2016 and became the home venue for a premier league football team – West Ham United (WHUFC) – from 4th August 2016. Table 7 provides names and details for the venues within the park. The park itself forms a natural behaviour setting within the landscape, bordered by roads and waterways. The Aquatics Centre pool is the only venue open to the public for general use daily, when not in use for events. Access to the London Stadium is controlled by gated bridges across waterways to the island where the stadium is located.

Table 7. QEOP venues and capacities during 2016/17

Venue	*Capacity*	*Description and Event Types*
London Stadium	Up to 66,000 fixed seats[1]. Up to 80,000 attendees for staged events[2]	Large stadium (open roof) Football; Rugby; Track and field athletics; Music concerts; Large-scale exhibition events
Copper Box Arena	Up to 7,500 retractable seats	Indoor arena Music concerts; Small-pitch sports (e.g. netball, basketball, and boxing); and exhibition events
LeeValley VeloPark and Velodrome	Up to 6,000 seated (indoor velodrome)	Indoor track cycling; Outdoor track cycling and BMX park
London Aquatics Centre	Up to 2,500 seated	Swimming pool and exhibition space. Swimming pool is open to the public on non-event days
Lee Valley HTC	3,000 seated, up to 15,000 attendees for large events	Hockey and tennis club. Excluded from research due to being outside the range of the park Wi-Fi

[1] Most events limited to 60,000 attendees max. WHUFC football matches regulated to 57,000 attendees max.
[2] Most staged events limited to 70,000 attendees max (based on 'sold out' attendance figures).

The park provides the opportunity to study multiple facets of human behaviour, from routine daily activities to the impact of large crowds attending events. An objective of the LLDC is to establish the zone as a smart sustainable district, utilising digital technology embedded within the park along with the digital twin concept to optimise building and infrastructure management. A more difficult challenge is to understand how the park itself is used by people. LLDC is seeking to make use of embedded technology to improve knowledge about uses of the park and facilities within it. and the effect of large-scale events on the park and surrounding area.

4.2 Revealing the Contextual Life of the Landscape

This study evaluates the potential for data generated within the QEOP to reveal how context affects visits to the park. The park is an urban public space designed for pedestrian outdoor activities. Aside from environmental benefits, such as mitigating pollution, its purpose is to attract people to spend time within its perimeter. Activities can include attending events, visiting attractions, undertaking fitness and leisure activities, meeting with people, working outdoors and relaxing. Some activities may be planned, such as a tourist trip or attending a ticketed event. Some activities may be part of a routine, such as visiting playgrounds after school or using paths and trails as part of a regular fitness routine. Other activities may be ad-hoc and dependent on circumstances, such as choosing to spend a lunch break in the park when the weather is pleasant.

Common sense might assume that more people visit the park on weekends than weekdays and during summer than winter and that numbers would increase substantially when large events take place. It might also be assumed that weather would impact visit numbers, although the relationship is likely to be more complex. Attendees who have paid for tickets may be less sensitive to unpleasant weather conditions than those making ad-hoc visits for rest and recreation. This study explores the potential for using data being generated by mobile devices whilst present within the park to move from common sense assumptions to measured observations and compare different uses of the park associated with different contexts.

The contexts evaluated in this study are:

- Is there an ambient signature, a ‘social heartbeat’, for ad-hoc (non-event) visits to the park?
    - Daily variations and weekday versus weekend comparisons
    - Hourly variations through the day, focusing on daylight hours
- Is there a seasonal or cultural adjustment to visitor levels?
    - The effect of term-time versus school, public and religious holidays
- What effects do large-scale events have on the ‘social heartbeat’ of the park?

A core consideration for this thesis is whether data traces emitted by mobile devices are sufficient to represent real-world population behaviours. Multiple sources are acquired, and contexts known to temporarily attract a larger population are used to evaluate the data.

4.2.1 Data and methods

4.2.1.1 Data sources

The primary data source for this study is from the free public wireless network operating within the park (‘Wi-Fi’), provided by LLDC. A second data source, also provided by LLDC, was a headcount generated from high-resolution cameras installed at entry points to the park (‘webcam’). Both of these sources were generated by sensors within the park. The data is proprietary and not publicly

available. Two sources of publicly-shared social media were also acquired for the periods covered by Wi-Fi data: Twitter and Foursquare. As introduced in chapter two (section 2.2.1), several studies have already been published using both sources to infer population dynamics and people-place experiences. They are considered here for revealing contextual variations. Code samples for the acquisition and preparation of these data sources are included in Appendix B.1.

Wi-Fi

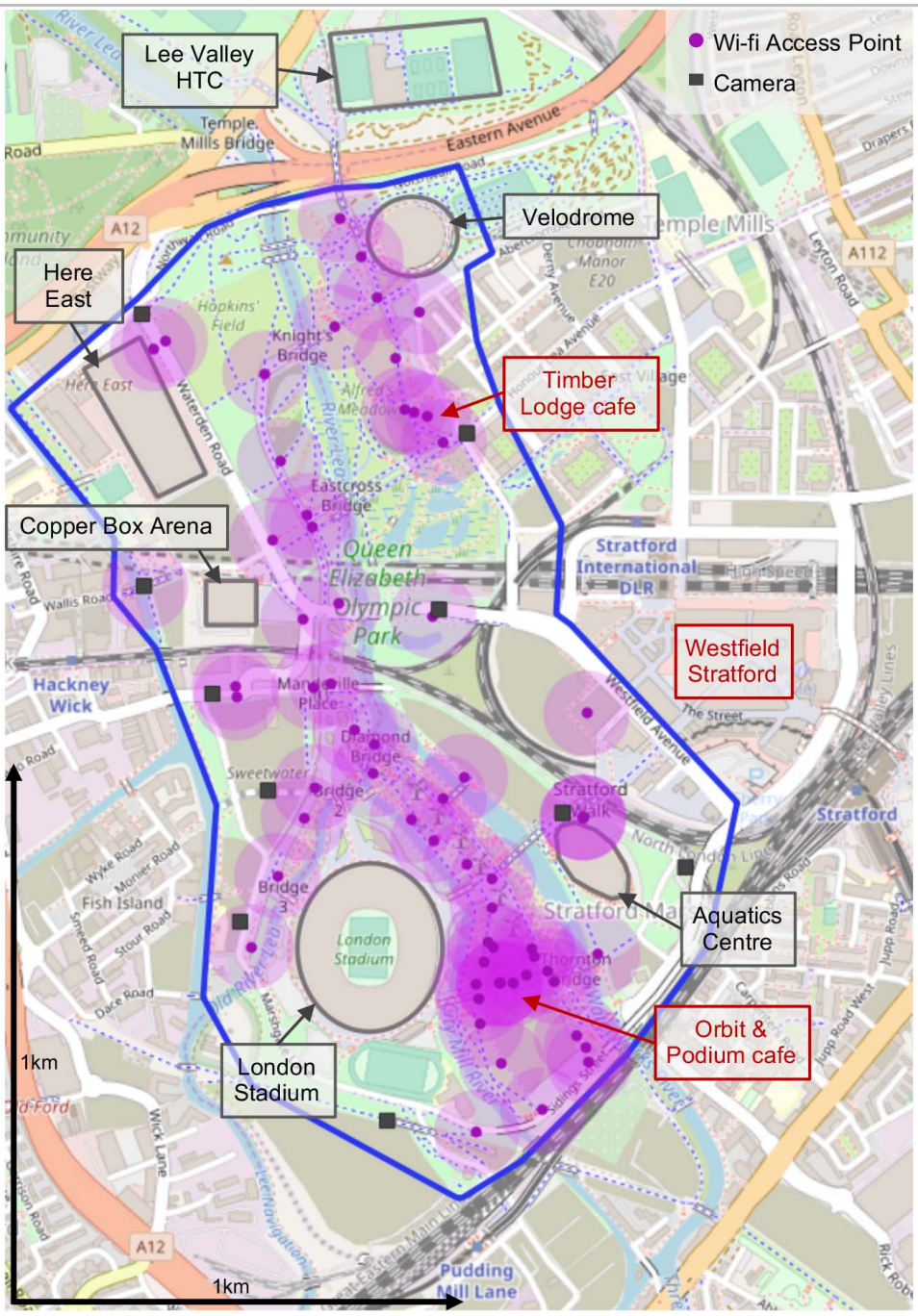
The park Wi-Fi network consists of 65 Wi-Fi access points (APs) located on paths across the park (Figure 20). They are not located within venues and they do not extend to the northern point of the park containing the Lee Valley Hockey and Tennis centre. According to the provider, WiFiSpark, APs are omnidirectional with an approximate signal strength radius of 67 to 80 metres (WiFiSpark, 2016). The actual range depends on the number of devices connected and can potentially extend beyond 100 metres or be less than 50 metres. In areas expected to experience high demand, APs are positioned approximately 50 metres apart. Some have been co-located, such as three stacked vertically near the entrance to the Aquatics Centre. To use the public Wi-Fi network first requires completion of a registration form. A device will then automatically connect to the nearest available access point to establish an internet connection. The connection is maintained whilst moving through the park by disconnecting from an AP when a closer AP is available to connect to.

For this study, two sets of raw system logs were provided as delimiter-separated text files, covering March 2016 and May to August 2016. The system logs contained every communication transmitted by every access point in the network, including device connections to use the Wi-Fi and administrative tasks such as synchronising time and receiving software updates. Table 8 contains a list of events generated by device connections to access points on the network.

Table 8. Wi-Fi event definitions for device activity recorded in the system log

Event Type	*Abbrev.*	*Notes*
Joins To	J	Device joins the network and starts a session to use the Wi-Fi network
Re-joins To	K	Device re-joins the network. For analysis, re-join is treated the same as a join, it is the start of a new session using the Wi-Fi network
Roams From/To	R	Device disconnects – 'Roams from' – from the last access point and connects – 'Roams To' – different access point
Disconnects From	D	Device disconnects from the last access point and does not connect to a new access point (device moves out of range of the Wi-Fi)
Leaves	L	On disconnecting, the device leaves the network. A session record is generated and statistics for the session calculated such as trip duration

The logs were parsed to extract device connection events. Every time a mobile device connects to or disconnects from an access point, an event is recorded that includes the MAC address of the device – a unique identifier – the access point, a timestamp, and event type. Details and code samples are provided in Appendix B.1.

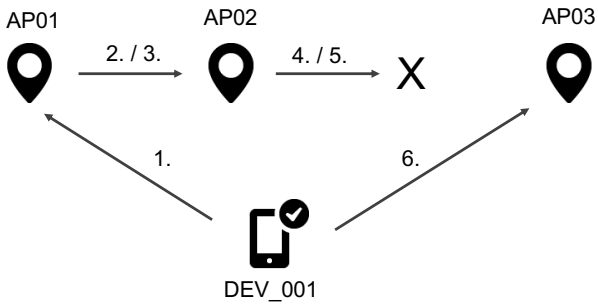


Map background by OpenStreetMap, under ODbL, as of August 2019
Produced in QGIS 3.4.2. Map projection: EPSG:3857

Figure 20. Locations of Wi-Fi access points and entry/exit cameras in the QEOP

Purple dots are Wi-Fi access points. The semi-transparent circle surrounding each indicates approximate signal range. Grey squares are locations of cameras for head counting (no longer installed, as of 2019).

Figure 21 provides a simplified workflow to demonstrate the events recorded in the Wi-Fi system log. The first time a device connects to an access point on the Wi-Fi network, a 'Join to' (J) event is generated to initiate a session. When the device disconnects from one AP and connects to another, a 'Roam from' and 'Roam to' (R) event is generated for each AP respectively. When the device disconnects from an AP and does not connect to a new access point, a 'Disconnect' (D) event is generated as well as a 'Leave' (L) that provides the session duration statistics.



	Timestamp	Device	Event	AP	Duration
1.	2016-06-01 09:00:00	DEV_001	JOINS TO	AP01	
2.	2016-06-01 09:04:32	DEV_001	ROAMS FROM	AP01	
3.	2016-06-01 09:04:32	DEV_001	ROAMS TO	AP02	
4.	2016-06-01 09:06:28	DEV_001	DISCONNECTS FROM	AP02	
5.	2016-06-01 09:06:28	DEV_001	LEAVES	AP02	420 seconds
6.	2016-06-01 10:12:04	DEV_001	JOINS TO	AP03	

Figure 21. Simplified example of Wi-Fi network log of device activity

Example of a set of readings extracted from the QEOP Wi-Fi network system log to demonstrate the different events used when a device connects to or disconnects from an access point.

Once the device events had been parsed from the system logs, the data can be prepared for analysis following the steps described in chapter three. All device IDs were anonymised and the original MAC addresses deleted from the data. Only properties relevant to analysis were extracted.

Inspection of the data revealed that the sequence of events was not always followed. A device could generate two J events in succession without an R or D event in between. The number of Disconnect events did not match the number of Join events. For estimating duration for incomplete sessions, if the gap between events was more than one hour, it was assumed that the device had left the park between events and the duration for the last event was set to 60 seconds (to allow for still being within the boundary of the park for a further minute after moving out of range of an access point). This enabled trips and stages to be calculated.

For analysis, the focus was on people visiting the park and making use of its facilities, even if only to traverse the park as part of a route between locations. Devices that connected for less than 5 seconds or remain connected for more than 12 hours were assumed to be devices that either briefly were in the range of an access point but outside the park boundary, or were devices permanently located within the park boundary and therefore not reflecting human interactions with the park. They were removed from the dataset as not relevant to the analysis.

Following processing, a total of 4,515,188 records remained, generated by 98,591 devices of which 80,149 visited on less than three days and 18,442 visited on three or more days.

Webcam

A concern with using Wi-Fi data as a measure of population behaviours within the park is that not everyone within the park will make use of the network. Less than 100,000 devices connected to the network for more than 5 seconds during the four months of data retrieved. It had been forecast

that the park would attract as many as 9 million visitors annually from 2016 (LLDC, 2016). If the forecast was accurate, then very few visitors made use of the Wi-Fi network.

LLDC, via their partner Movement Strategies, provided an additional data source as a potential validator for the Wi-Fi data: an hourly headcount. At ten entry points to the park, high-resolution bi-directional cameras were installed for crowd monitoring and management during the London 2012 Olympics. The cameras were still operational in 2016. The data provided was an hourly headcount incoming and outgoing at each camera for each day during the same period as for the Wi-Fi data. The headcount is performed automatically using a computer vision algorithm to detect and count the number of faces visible. No information was available regarding the cameras or algorithm and no access to images was permitted. In theory, the count should produce a complete population and be representative of variations in visitor numbers associated with different contexts. However, the cameras do not cover all access routes into and out of the park, and it is unknown what assumptions are applied to produce an hourly count. It is, however, the most comprehensive data source available to compare with the Wi-Fi data over the same period.

Foursquare

Foursquare is a location-based app and website. Participants can check-in to venues and indicate that they are 'here now' if they have the app installed on their smartphone. Following its launch in 2009, participants competed to become the designated virtual mayor of a location by having the highest frequency of check-ins. As of 2017, 50 million people were still participating in the service daily[1]. There is no information to indicate how long a device remains 'here now'. Devices can check-in to a venue but do not check-out. It is assumed that people are more likely to check-in to venues where they plan to dwell for some time. The online service includes a social network and prompts people to rate and provide reviews about venues they check-in to.

Foursquare has an open API for viewing active venues and the number 'here now' at each venue in real-time. Foursquare data were collected for this research by requesting a list of all active venues within a 2.5km radius of the centre of the QEOP and then filtering to venues within the park boundary. Readings were captured every 15 minutes. The average number of check-ins throughout the hour was then calculated and recorded as an hourly count for each venue. This is a crude method but considered sufficient for the research. In terms of indicating the size of a population present, Foursquare was not anticipated to be a viable data source. However, it could prove useful for sensing different contexts that may affect presence. Every venue listed on the service has spatial coordinates, a unique identifier, and a category describing the venue. This metadata has the potential to indicate types of activities being undertaken and whether or not different types of venues are popular during different contexts.

[1] Source: Foursquare https://foursquare.com/about accessed October 2018

For the two periods being analysed, 18,944 ‘here now’ actions occurred across 8 venues within the boundary of the park, matching the same area covered by the Wi-Fi network.

Twitter

Twitter is a global social media site for posting short messages that can include text, images and other media as attachments (Figure 22).

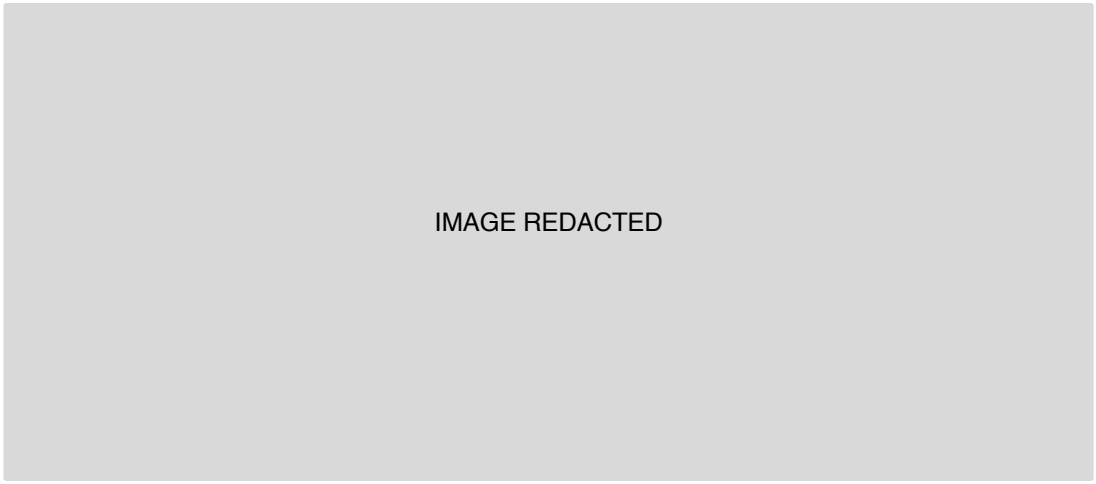


Figure 22. A single tweet highlighting its content and metadata

Highlighted elements are metadata: from left, a thumbnail image for the user profile, the user profile preferred name (in bold) and screen name (preceded by @), and the timestamp for the tweet. The rest of the image is the content of the tweet, consisting of text and links. When links connect to images, the image is displayed.

First launched in 2006, Twitter introduced geotagging in 2009, the ability to automatically attach coordinates based on a device’s GPS sensor if available. At the beginning of 2017, Twitter had 328 million monthly active users[2]. Originally, messages were limited to 140 characters in length. In September 2017, the limit was doubled to 280 characters. The data for this study were collected during March and May to August 2016, when the original limit was in place.

Twitter data were retrieved by submitting a query to Twitter’s Search API weekly to return tweets containing geotags posted within a 2.5km radius of the centre of the QEOP. The Twitter Search API will return tweets matching the query posted within the previous 7 to 9 days (Twitter is not specific about the cut-off point for returning samples of tweets). During the periods of data collection used for this research – 2016 to 2018 – the sample was believed to be 1% of all tweets posted. Researchers evaluated performing analyses on a sample versus the full Twitter repository (Twitter provides access to the full feed for a fee) and concluded that the 1% sample is sufficient to be representative, However, they acknowledge that geotagged tweets were less than 1% of data and did not evaluate specifically for geographically-located tweets (Burnap, et al., 2015).

The text string returned includes the full record for each tweet. Only properties of interest for analysis are extracted, detailed in Table 9, and the rest of the record is deleted. As with Foursquare,

[2] Source: Statista https://www.statista.com/topics/737/twitter/ accessed 10 September 2018

the dataset is filtered to exclude tweets outside the boundary of analysis. An additional preparation step is performed on Twitter data: each tweet also has its content extracted for analysis using natural language processing (NLP). NLP extracts terms and phrases from the text and enables linguistic analysis such as term frequency, term similarity and topic modelling to compare and summarise sets of terms. The process is described in detail in chapter six.

Table 9. Attributes of a tweet retained for analysis

Attribute	*Data Type*	*Notes*
Status ID	numeric	Unique identifier for the tweet (auto-generated)
Tweet	character string	Content of the tweet.
Date/Time	timestamp	The data and time the tweet was posted in UTC
Latitude	Numeric	Coordinate when tweet was posted (-998 if not available)
Longitude	Numeric	Coordinate when tweet was posted (-998 if not available)
Source	character string	Application used to post the tweet
User ID	Numeric	Unique identifier for account posting the tweet
Screen name	character string	Screenname that is prefixed with an @

For the two periods being analysed, 2,929 geotagged tweets were posted by 1,736 accounts.

4.2.1.2 Contexts for analysis

Park data sources were initially provided for March 2016 to perform a pilot analysis and determine the feasibility of using Wi-Fi data to analyse local socio-spatial dynamics. Following a presentation of the findings to LLDC in August 2016, an additional set of data for each source – Wi-Fi and webcam - was provided for the period May to August 2016 (up to 21 August 2016 for Wi-Fi). This opened up the opportunity to examine seasonal as well as event effects. In March, temperatures in the UK are typically cooler and wetter whilst, from June to August, the days are usually warmer and drier. March 2016 included the Easter public and religious holidays from Friday 25th to Monday 28th. The school summer holidays began during the week commencing 18th July. Approximate weather conditions were retrieved based on Met Office archives for London and the South East.

The following special scheduled events (national or international in scale) took place at QEOP venues during the two periods:

- World Track Cycling Championships (WTC) at Velodrome, 2nd to 6th March. Finals scheduled during the evening of Saturday 5th March
- National Sports Relief Charity event at multiple venues on Sunday 20th March
- AC/DC music concert at the London Stadium, Saturday 4th June
- International Athletics at the London Stadium, 22nd to 23rd July, with finals on the evening of Saturday 23rd
- WHUFC football matches at the London Stadium on Thursday 4th (unofficial first event), Sunday 7th (official first event), and Sunday 21st (first Premier League match)

Other events included basketball and netball matches at the Copper Box arena, and swimming competitions and exhibitions at the Aquatics Centre. These events are assumed to be too small to create a noticeable impact across the entire park. The Lee Valley Hockey and Tennis club (HTC) is outside the range of the Wi-Fi network and its activities are excluded from the analysis.

It had been hoped to incorporate data from two weather stations that had been installed within the park in 2016. However, they were off-line for much of the periods being analysed. Instead, to explore for potential weather effects, daily summaries for London and the South East hosted online by the Met Office were reviewed (Met Office, n.d.). The following dates appear to have been very wet, or hot, in the park according to written descriptions in the summaries and Met Office definitions:

- Very wet (>2mm rain per hour)- 9th March, 27th March, 11th May, 31st May, 7th June, 16th June, 19th June, 20th June, 1st August, 2nd August
- Hot or very hot (> 26 degrees Celsius)– 19th July, 24th August

Weather can be very localised. A summary spanning London and the South East covers a large area, including the coastline and high-density built districts inland. It can only be assumed that the weather on these dates occurred within the QEOP. It is also possible that the QEOP experienced such conditions on other dates. Furthermore, weather can have two separate effects on visitors to an outdoor landscape: the conditions forecast before visiting, and the weather experienced whilst present that may or may not match the forecast. Studies incorporating weather effects presented here contain uncertainty and, at best, are an indicator for future research directions.

4.2.1.3 Analysis steps

To analyse, first, the data sources are aggregated and compared as daily counts representing the total number of devices visiting the landscape. Second, the data sources are aggregated and averaged hourly for the two time-based levels of context: recurring cycles with no expected disruptions – the ambient context, and disruptions to the ambient context: abnormal weather and scheduled events. Third, the data sources are aggregated spatially by segmenting the landscape into zones and explore contextual variations in the spatial distribution of presence.

4.2.2 Results

Before performing quantitative analysis, the prepared data was explored visually as a space-time animation using the open-source software package Processing. Appendix C contains a series of screenshots from the animation for demonstration. The visual analysis confirmed that the data exhibits a diurnal rhythm similar to those produced by literature presented in chapter two (section 2.2.2) and that the data is sensitive to events, showing both spatial and temporal variations in readings. The visual analysis was then explored as day and hour variations.

4.2.2.1 Day variation

The daily counts for each of the two data capture periods, March 2016 and May to August 2016, are presented in Figure 23 and Figure 24 respectively to provide an instant visual of the sensitivity

within each data source to known contexts: scheduled events and holiday periods, and to identify any abnormally low readings that may indicate data collection issues.

The counts are plotted as a combined line chart with values scaled using z-score standardisation to compare daily variations between sources, and as bar charts plotting individual counts per date. Key events and holiday dates are identified using vertical lines (straight and dotted respectively). Weekends are indicated by a pale grey background. In both plots, social media includes readings beyond the Wi-Fi boundary.

Figure 23. QEOP daily data volumes, March 2016

The line chart is a plot of all sources, each scaled using z-score standardization to enable direct comparisons. The bar charts are plots based on actual counts for each source, for visual inspection only. Note: line chart includes an additional data source – Twitter keywords - not analysed in this chapter but referred to in chapter six.

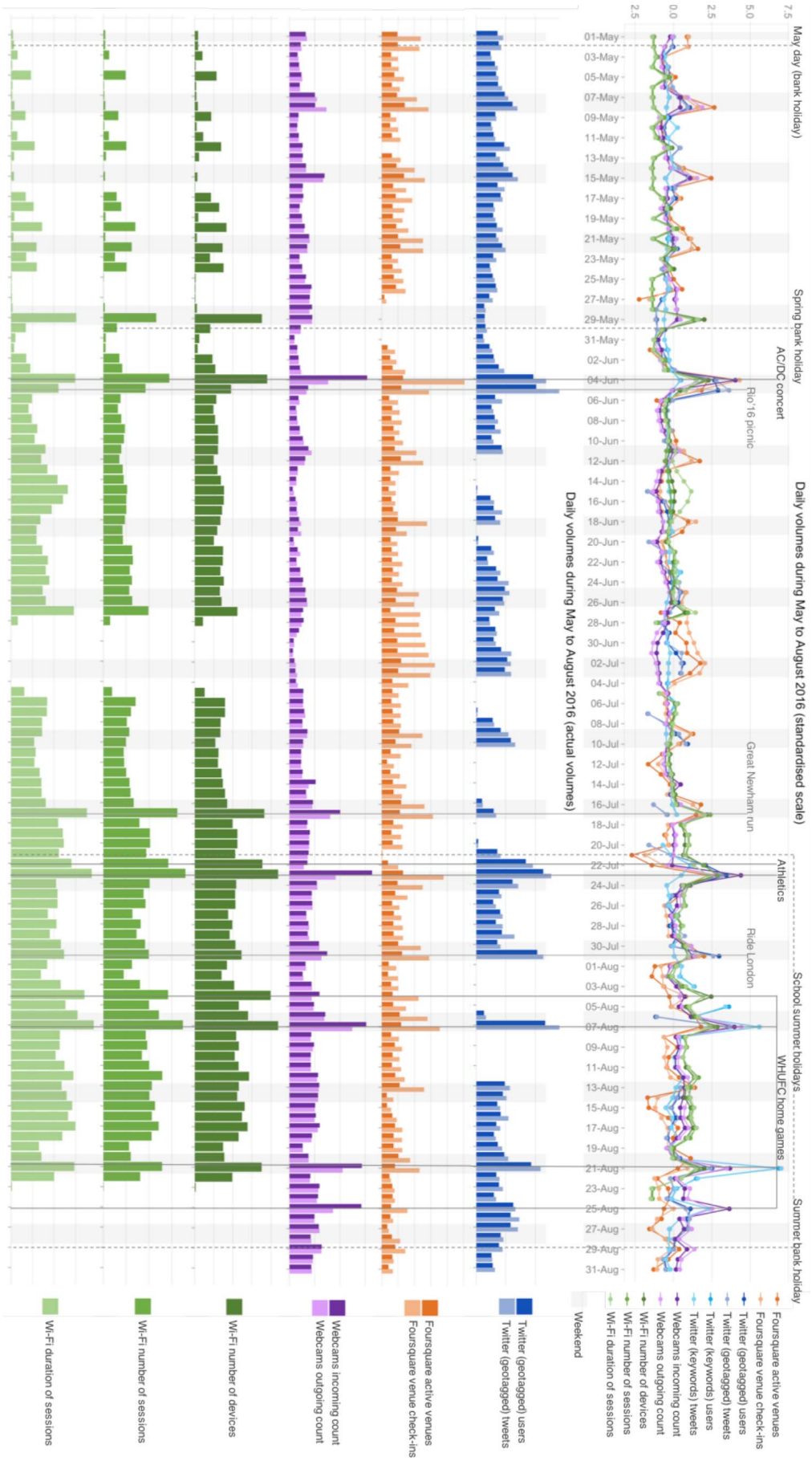


Figure 24. QEOP daily data volumes, May to August 2016

The plots in Figure 23 and Figure 24 provide an immediate visual to explore if each data source is sensitive at least to large-scale events. The visual scan shows all data sources had peaks on large-scale event days, suggesting at least some contextual sensitivity. Figure 25 shows the mean day counts summarised monthly for non-event days, showing the variation in counts across the months (Figure 25a) and comparing weekdays with weekends (Figure 25b). It had been expected that there would be a clear difference between weekday and weekend counts. A common-sense assumption is that the park will be busier at weekends when longer leisure activities can take place, particularly during the summer months of June to August. However, only one source, Foursquare, exhibited a substantial increase. If weekend activities are family-oriented, Wi-Fi data may be under-representing families, given young children are less likely to be carrying mobile devices. It could indicate that use of the Wi-Fi is biased towards people who work in the area and such bias should be considered when interpreting results. However, the same effect is evidenced for headcounts captured by the installed cameras.

An additional attribute available when studying Wi-Fi device activity is that devices can be categorised as 'habitual' or 'explorer' to identify devices that frequent the location and are familiar with it versus those who are not. The percentage split between habitual visitors (visit at least three times) and explorers (visit once or twice) does vary through the months and could be interpreted as demonstrating a tourism effect within the park. Explorers make up less than one-third of visitors on non-event days during March and June but over 40 per cent of visitors during the August summer school holiday (Figure 25c).

Figure 25. Comparing median day counts across data sources, QEOP, 2016

Red line indicates mean across all months in figure a) and uniform distribution for figures b) and c).

The counts for webcams are suspiciously low during June suggesting a partial data outage may have occurred similar to the one that occurred partway in March. However, the Twitter readings also dip in June, whilst Foursquare has the least variation across the four months.

The counts reveal a key challenge with using indirectly captured data. Each source has gaps or unexplainable abnormal readings. For example, in the March 2016 period, the webcam headcounts look suspiciously low from the 9th to 15th. An inspection of the data revealed zero counts for a number of the webcams during that period. For the longer period covering May through August 2016, each source has data quality issues at different times. The Wi-Fi data appears to have sporadic gaps throughout May into early June, and prolonged gaps at the end of June and August. The webcam counts drop during mid-June and early July. The geotagged Tweets have several gaps that were not expected as there was only one known data collection issue during that period. Foursquare also shows dates with zero readings when activity is recorded in other sources.

A benefit of capturing disparate data sources is in the potential for one source to 'gap-fill' another. If one source has a count of zero but other sources register readings, it is more likely that the zero count is a data collection error rather than the absence of people. To explore for correlations in daily counts, each data source is plotted against the other on a scatter plot (Figure 26). Even without quantifying the regression, correlation is visibly weak for all combinations except for comparing Wi-Fi devices with Webcams. There is a visible separation between weekdays (blue squares) and weekends (orange circles) for all sources. The outliers to the far right of all plots are the major event dates. The noise in the Wi-Fi and webcams indicates Wi-Fi data outages in May.

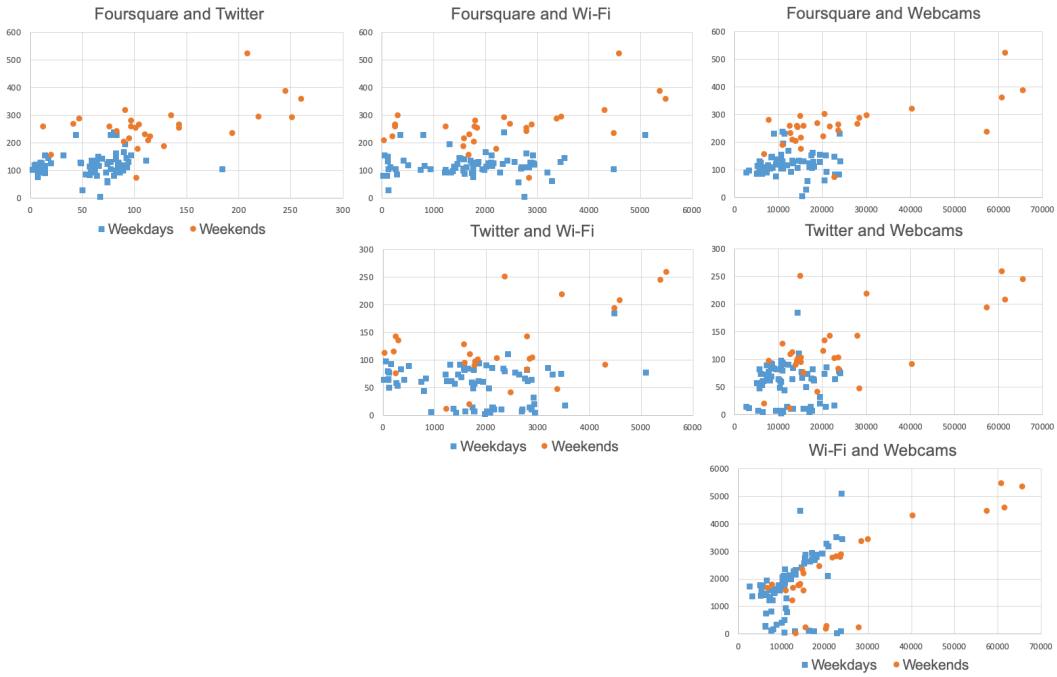


Figure 26. Scatter plot of daily volumes comparing data sources

First source listed in chart title is plotted on the Y-axis, second source is plotted on the X-axis. Data points are colour-coded to identify weekdays (blue squares) and weekends (orange circles). Readings from May to Aug 2016.

Social media sources registered low readings on non-event days, with each exhibiting a different rhythm. Foursquare showed increased readings at weekends. This could indicate that people use Foursquare more for leisure than work activities and should be a consideration when using Foursquare data to infer the urban character or vitality of a landscape. Both sources are sensitive to events, including smaller events. Twitter also exhibits high activity on the day after the music concert, potentially indicating a lagging effect if tourists visit the area for a special event and extend their stay before or after when events occur at weekends.

Whilst the results from social media are mixed, one potential benefit with text messages is that they contain information that may reveal conditions within the park that may affect population behaviours. Figure 27 contains word clouds containing the top 50 most frequent terms, excluding common or shared words such as 'and', 'the', and 'london' for three periods exhibiting higher than normal readings. Terms are sized by relative frequency.

Figure 27. Top 50 words in Tweets per date range, QEOP 2016

Frequency of terms, excluding terms common to all dates, for geotagged tweets on dates of interest. Terms sized as relative frequencies within each date.

Twitter messages on 5th June (a non-event day) are dominated by terms referring to the AC/DC music concert that took place the night before. It could indicate a lagging tourist effect for large-scale events that attract non-locals to the area. People who travel to attend an event may extend their visit overnight and also visit the park the following day. For the week when both social media sources exhibited high readings but with no Wi-Fi data available, it appears that there was a 'make the future' event taking place within the park and an 'ecomarathon' possibly sponsored by the global energy company Shell. On Sunday 31st July, 'ridelondon' is the dominant term. Consulting local news sources, a cycling event was taking place in the park. These findings indicate that access to a live social media or real-time news source may be beneficial for providing additional contextual information, particularly for local events that do not attract national press attention. This possibility is explored in detail as a separate case study in chapter six.

The social media sources appear to be too low in volumes on non-event days to contribute to a meaningful analysis, with several gaps in the Twitter daily counts. Whilst other studies have used social media such as Twitter for everyday population dynamics, it could be that it is better suited to built-up areas or larger spatial scales rather than open spaces and street-level comparisons.

4.2.2.2 Hour variation

The Wi-Fi data is the only source that provides granularity in both space and time, generating readings in real-time at 65 locations across the park. It is the sole focus for the remainder of this study, exploring differences in the hourly and spatial distribution of presence in the QEOP.

Figure 28 contains a series of plots showing the hourly device connections to the Wi-Fi network for a range of contexts. Figure 28a shows the averages per day of the week for non-event days. It has a visible variation in the distributions for weekdays versus weekends. Weekdays exhibit three peaks around the morning commute, lunchtime and afternoon commute periods, with counts rising through the day. Weekends have a single peak spanning the lunch period and early afternoon, with a later peak on Sunday compared with Saturday and both having a higher peak than on weekdays. On weekdays, there is also a noticeable variation between Monday and Tuesday compared with Wednesday, Thursday and Friday from 10am until 5pm.

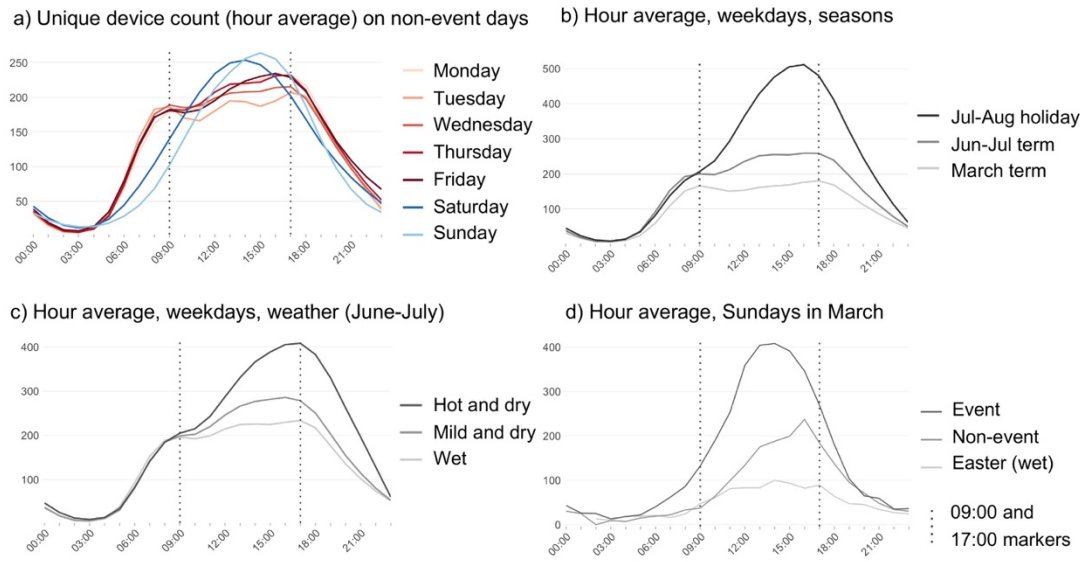


Figure 28. Comparing hourly averages for different contexts

When comparing weekdays across months (Figure 28b), all periods exhibit similar behaviours from 00:00 through to 08:00. From 09:00 to 21:00, there is a difference between the term-time hourly average for March and June to mid-July with March showing activity during off-peak daytime hours. However, there is a noticeable increase during the afternoon for the school holiday compared with term-time. Examining for weather effects during the summer term (Figure 28c), hot weather shows large increases afternoon visits, whilst the morning commute period is unaffected. Whilst this may sound like common sense, it is the first time there has been a data source for the park to quantify it. Comparing dates for a single day in the March data set (Figure 28d), an event has a substantial impact on park visits compared with a non-event day. Easter Sunday is the quietest day but also experienced wet weather. It is not possible to conclude that the reduced readings are due to Easter affecting shop opening times and also affecting visits to the park as a result, or whether the weather deterred visits, or if it is seasonal. More data is needed to verify the findings.

The last temporal analysis explores event effects and how they may differ depending on the start and end of the event. Figure 29 shows hourly device activity for events at the London Stadium: the music concert in June and two of the football matches in August. For each image, the nearest non-event equivalent day is plotted alongside the event day for comparison.

For each of the evening events (Figure 29a and Figure 29b) activity in the park during the daytime is unaffected until the afternoon. For the music concert, there is a visible and gradual build-up from 14:00 whilst the football match has a much shorter and more intense build-up from 17:00. The same intensity is also visible for the two football matches held in the afternoon (Figure 29c and Figure 29d), with activity also returning to normal almost immediately after the football match concludes. The plots show the potential to forecast how and when population dynamics change on event versus non-event days with more nuance than just considering differences in daily counts.

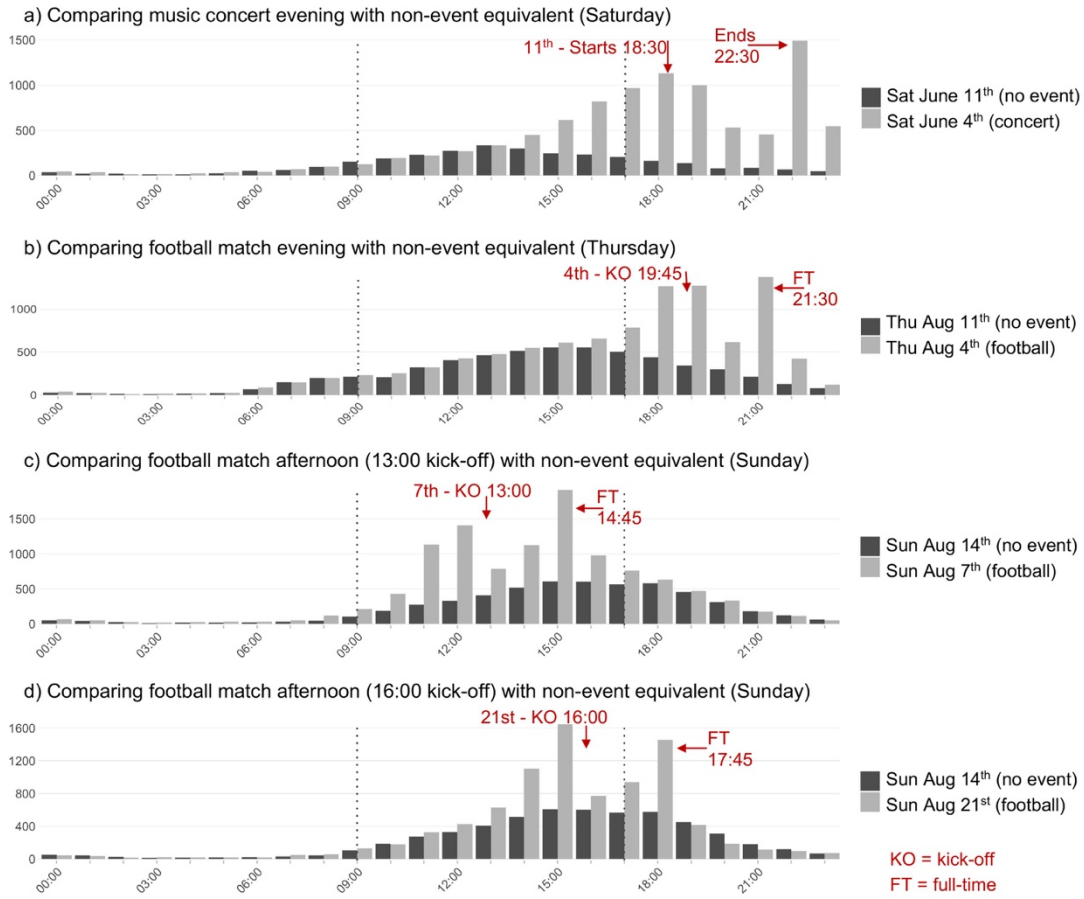


Figure 29. Hourly readings on stadium event days at the London Stadium, 2016

4.2.2.3 Spatial distribution

The previous section showed that large-scale events, expected to create a substantial increase in the size of the population within the park for the duration of the event, have an impact that is concentrated around the duration of the event on the day it. This analysis considers whether or not these effects are also spatially concentrated around the venue. The range of the Wi-Fi network is

segmented into nine zones representing different areas of the park and forming a choropleth map for comparing aggregated counts across the zones (Figure 30).

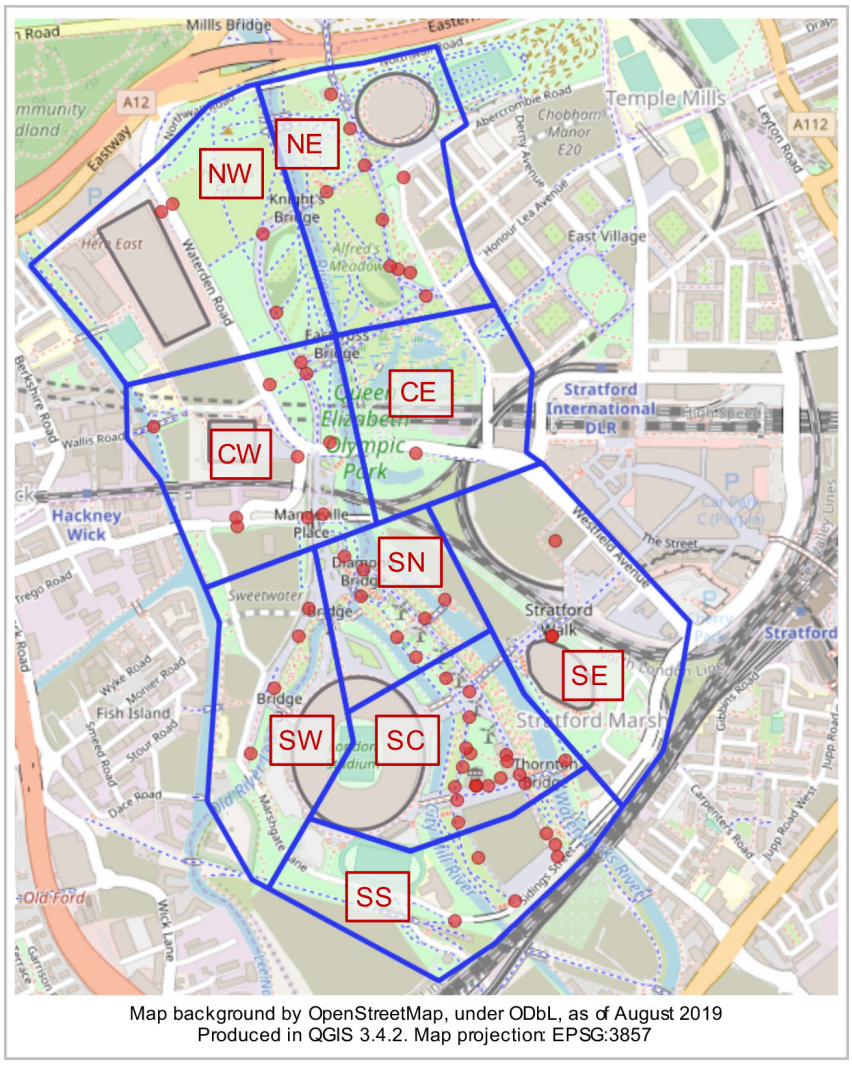


Figure 30. Segmenting the park into Wi-Fi zones for spatial analysis

Wi-Fi range is divided into three areas: North (Nx), Centre (Cx), and South Park (Sx). Each area is sub-divided into group access points within the area of the park. Hence SW is the west sub-area of the south area of the park.

The nine zones are organised and labelled as three broad areas of the park: North park (N) sub-divided into East (NE) and West (NW). Centre park (C) sub-divided into East (CE) and West (CW); and, South park (S) sub-divided into North (SN), East (SE), South (SS), West (SW) and Centre (SC). The boundaries are roughly drawn based on the locations of the access points. For each analysis, the count is of unique devices per area. Note that this means the sum of the areas may exceed the sum when counting park-wide due to devices visiting more than one area of the park.

Daily comparison

Looking at the spatial distribution on non-event days (Figure 31a) for the north, centre and south areas of the park, all show increases in readings from March to August. However, only the South park shows a substantial increase during July and August, the peak tourist period in the UK. It suggests that tourism activities are concentrated in the south. On large event days (Figure 31b),

the majority of the increase is concentrated in the South as would be expected, since it contains the event venue. However, the other areas of the park do also show minor increases. Examining just the south park, on non-event days (Figure 31c), the majority of the summer activity is in the Centre and East zones and, to a less degree, the North zone. On large event days (Figure 31d), there is variation between the events for the three most active zones. The East zone is much more active on the football dates than for the music concert. Both the Centre and the East zones have similar levels for both football dates. The North zone has a different variation, showing a bigger increase in activity on Sunday 7th August than on Friday 4th August. This could be due to higher attendance on Sunday (it was the first official match at the stadium), or it could be due to crowd control measures directing people to certain routes.

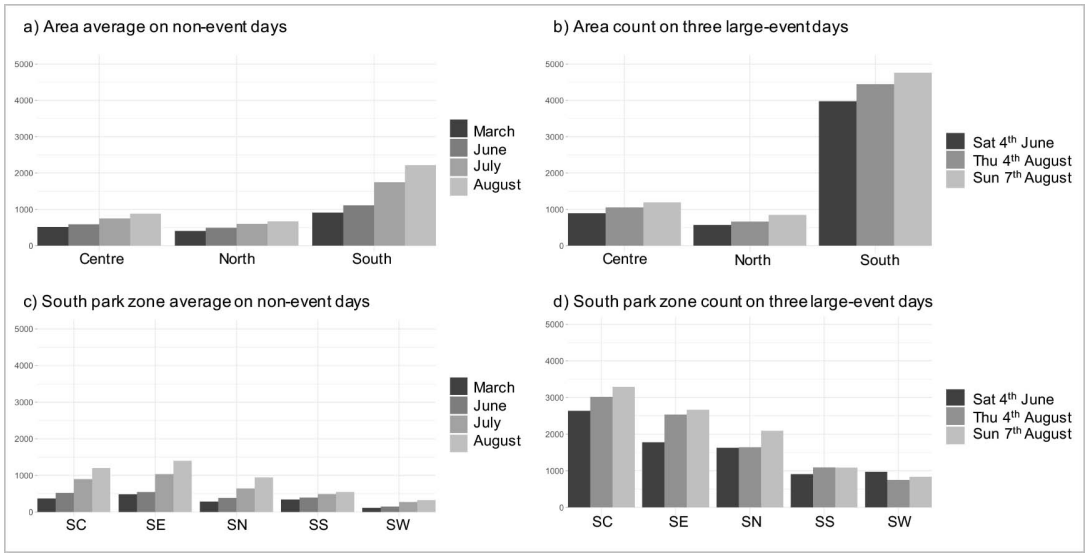


Figure 31. Comparing the spatial distribution of Wi-Fi device activity

All plots have the same y-axis scale of 0 to 5000. Average counts are median values. Events are individual event day counts. Counts are of unique devices present in zone or area.

When exploring for variations between similar events, Twitter can again be a useful source for additional context. Several tweets regarding the first football match indicated that people were being routed away from Westfield to a route running along the south edge of the park (Figure 32).

IMAGE REDACTED

Figure 32. Tweets expressing conditions on football match day at QEOP, 4th August

Hourly comparison

To focus on comparisons between areas at the hourly level that are not influenced by daylight and seasonal effects, the date range is reduced to cover from 2nd June to 21st August. Three averages are calculated: weekdays, weekends and comparing large events with the nearest non-event day equivalent, plotted hourly (Figure 33).

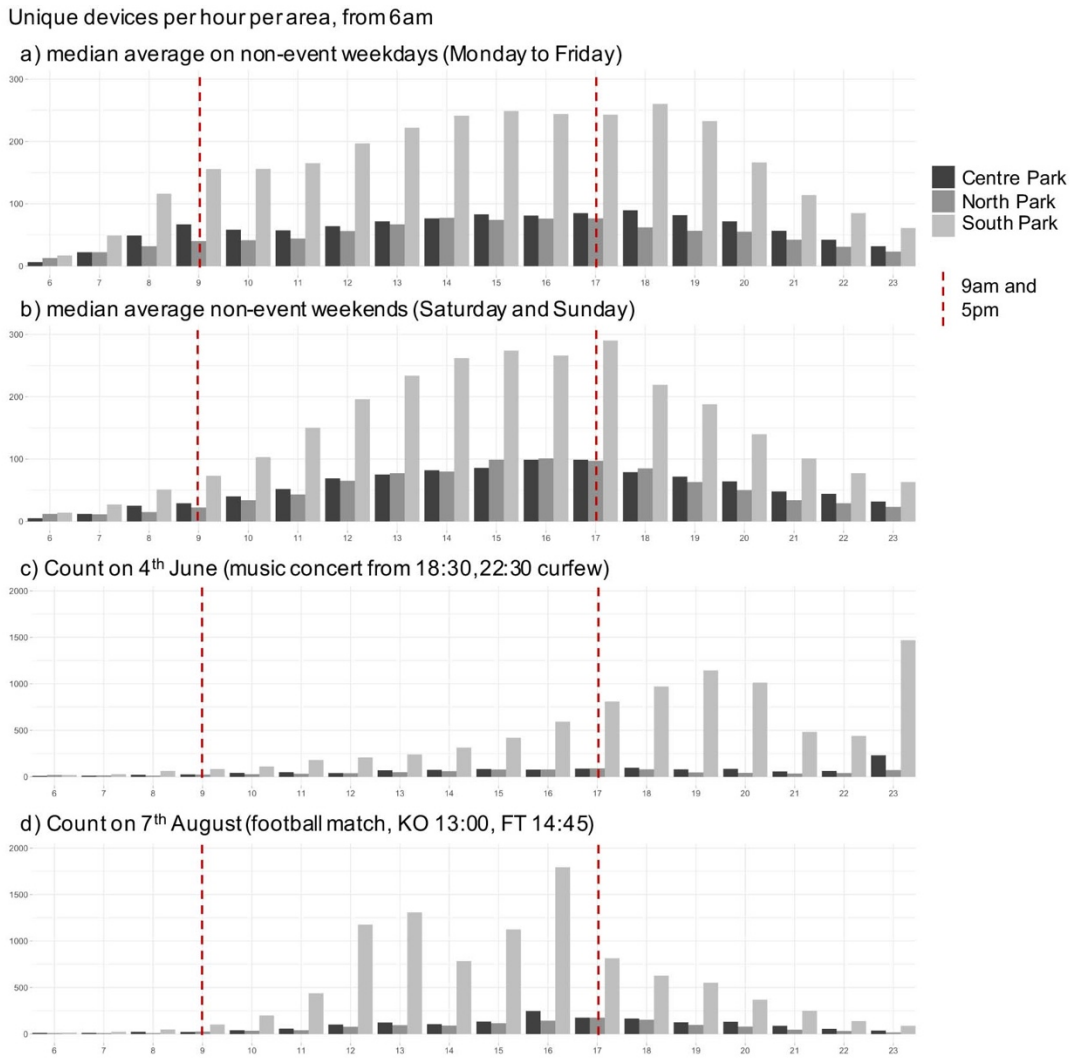


Figure 33. Devices connected to Wi-Fi per zone in the QEOP

Vertical dotted lines indicate 9am and 5pm. Plot starts from 5am. images a) and b) have a maximum of 300 on the y-axis, images c) and d) have a maximum of 2,000.

Figure 33 shows that each zone of the park has a different temporal profile through the day on weekdays. The South park dominates the readings. After a morning peak at 9am, readings increase from midday and continue to increase through the afternoon and into the early evening. The centre park has a morning peak at 9am but then declines before increasing again from midday onwards and peaking again at 6pm. This could be due to picking up connectivity by people travelling through the park. The centre area contains a busy road that separates the north and south park. The North park steadily increases through the day and peaks at 4pm before declining. It suggests that the north area of the park is not affected by commuting activity but perhaps is more closely associated with post-school routines given the presence of a large playground and café.

On event days, the South Park dominates again but event-related increases are visibly concentrated, as indicated previously when studying the park as a whole. However, the centre of the park also experiences a noticeable increase at the end of each event. The data shows how people arrive across a large interval before the event starts, with readings diverging from non-event days up to four hours before the start time. However, when the event concludes, the large majority leave at the same time.

The final analysis compares counts between each zone of the South park on event days (Figure 34). For the music concert, the same concentration of effect is seen at the end of the concert with all five areas recording their highest counts during 23:00 to 23:59. A similar effect is evidenced for the evening football match finishing one hour earlier. The two football matches held on Sunday afternoons both exhibit an intense build-up before the event but different behaviours after the event. The earlier afternoon event has an extended increase in readings through the afternoon whereas the later event has a much shorter post-match effect. This indicates that event effects in the park are sensitive to the specific finish time, at least to an hourly scale, and are not just a simple comparison of afternoon versus evening events.

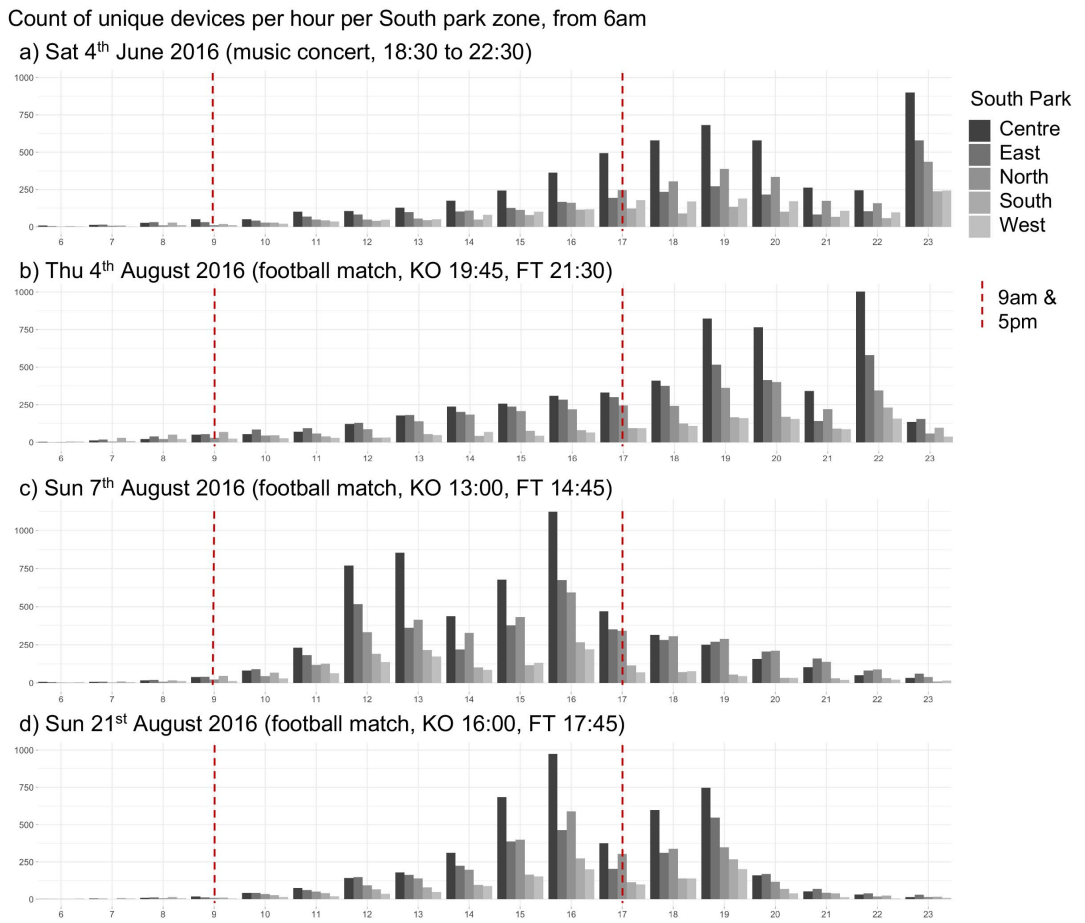


Figure 34. Devices connected to Wi-Fi per South Park zone in the QEOP

Vertical dotted lines indicate 9am and 5pm. Plot starts from 5am. All images have the same y-axis scale with a maximum of 1,000.

Comparing the different zones across the south park, people visiting the stadium can enter and exit the south area of the park via either the North, South, East or West zones. The plots indicate that the majority go east for all events although the second football match on Sunday 7th (Figure 34c) shows a higher reading in the north of the park. This is also repeated for the third match (Figure 34d). This could be an indication of the routes available to attendees or that home fans head back towards Leyton to the northeast of the park, which is within walking distance. It is assumed that the majority of people who are fans of the home football team live within the local area. The music concert is likely to be more attended by a mix of locals and people travelling from much further afield. National and regional transport hubs are all located in Stratford, close to, bordering or beneath the Westfield retail centre to the east of the stadium.

The Wi-Fi data shows promise for revealing how context affects visits to the QEOP, showing both daily and hourly variations. The data indicates there is a seasonal effect, both from the cooler climate of March to the warmer summer period and between term-time and school holidays. The data also indicates that different types of event held at the London Stadium will have a different impact on the park and surrounding area. The next study explores these findings further, with a study of readings across 12 months.

4.3 Learning Variations in Visits to the Landscape

The second study explores a simplified set of Wi-Fi data over a longer period to examine seasonal effects and a wider range of contexts within the QEOP. The research was completed in 2018 using Wi-Fi data generated throughout 2017. An edited version of this study has been published in the journal IEEE Pervasive Computing (Richardson, 2019).

4.3.1 Data and methods

To explore if seasonal effects can be quantified, data was retrieved for an entire year – 2017. Raw Wi-Fi data logs were not available. However, a daily count of unique devices connecting to the network was provided. Furthermore, an installed weather station emitted readings throughout the year with daily summaries provided under open access by the Weather Underground network (www.wunderground.com). Temperature readings were extracted and compared with Met Office daily summaries. When dates had gaps, the temperature was approximated between readings either side of the gap. Precipitation proved unreliable. Recordings only began partway through the year with several gaps. A diary of events held at the London Stadium was retrieved from the park web site (www.queenelizabetholympicpark.co.uk). All processing and modelling of the data were performed in Python using the Scikit-Learn package. Code samples are included in Appendix B.4.

This study considers only one measure – the total number of people visiting the park during the day, based on the number of unique devices detected. It cannot indicate the spatial distribution across the park or the temporal distribution of presence throughout the day. The objective is to create a predictive model of visitor numbers based on discoverable contexts. Machine learning is used to identify and measure which attributes have the most influence. Whilst the volume in terms of the length of the dataset is very small – just 365 records representing each day of the year, the dataset can be comprised of many different attributes retrieved from disparate sources.

Machine learning is commonly associated with analysing very large data sets in terms of the number of records. However, it is also beneficial for analysing data sets that are small in volume but contain a large number of features and are thus difficult or time-consuming to model using traditional statistical methods. I call this ‘little big data’. A machine learning model can rapidly iterate through multiple combinations of features to develop a predictive model. Two approaches were tested: building a single model for the full year and building a split model to produce separate predictions for scheduled event days versus non-event days, given the substantially larger number of visitors that would be expected on stadium event days.

Machine learning requires all data values to be in numerical format. This can be an issue, and a potential source of error, for attributes that are categorical and/or expressed as text and attributes that are numeric as an identifier rather than a measure, such as the use of numbers to represent calendar months. For example, it is unlikely that there will be a linear relationship between visitor numbers and the month of the year since that would result in the lowest readings occurring during January (month 1) and the highest readings in December (month 12) with a continual increase

between the two. Thus, following an initial analysis of the data, presented in the next section, a ‘season’ category was created to mimic the changes across the year. The highest months were given the highest score of 6 (July and August), followed by 4 (June and September), 2 (April and October), and 0 (January to March and November to December). To incorporate weather, temperature values were initially categorised as hot, pleasant and cool. However, early tests identified that individual temperature readings produced better results. To categorise events, an ‘event_rank’ was introduced: 0 for no event, 1 for West Ham football matches, 2 for athletics and park-wide events, and 3 for music concerts. This was based on the variation in categorised event distributions visible in Figure 35b. However, these are assumptions that could influence outcomes and must be considered when interpreting the results of the model

To build the regression model, the data was divided in two, with 85% used for training and 15% set aside for testing. The split was performed using stratified random sampling. This enables specifying an attribute whose values need to be evenly distributed between the training and test sets. For the single model, ‘stadium_event’ was used, to ensure that the proportion of event versus non-event dates is consistent across the training and test sets. For the split model, ‘day_of_week’ was used for non-event days, and ‘event_rank’ was used for the event days. Evaluations were completed during training using cross-validation with ten folds – randomly splitting the training data into ten subsets and then training across nine and testing using the tenth, iterating so that the tenth fold is a different subset for each test and then producing an average score across the ten evaluations. This helps indicate how well the model will generalise to potentially refine feature selection before evaluating the final model using the set. The choice of ten was arbitrary and considered sufficient for this study given the small number of records in the dataset. The higher the number, the larger the performance overhead to produce the model.

4.3.2 Results

4.3.2.1 Contextual variations in daily visitor counts

Figure 35a shows the daily count of unique devices connecting to the park Wi-Fi network (in lighter grey) and the monthly mean count (dark grey lines). The highest count was recorded on Saturday, June 17th, when a music concert was held at the stadium. The lowest count was recorded on Monday 25th December. The day count ranged from 45 to 2070. The day average (mean) across the year was 480 whilst the month average ranged from 182 to 664. There was a visible increase from January to August and a decrease from September to December. However, there is also a wide variation within each month. The majority, but not all, peaks occur on known event days, including football matches held during the Autumn, Winter and Spring, music concerts held in June, and two two-week long athletics championships held in July and August.

Figure 35b is a plot of mutually exclusive categories. Non-event days are divided into four seasons and event days are separated by type of event. There were three international athletics events during July and August: The World championships and Para World Championships, each lasting two weeks, and a single-day London 2012 Anniversary games. Four music concerts were held

during June. West Ham football matches were held from January to May and September to December. The stadium was also used for the start of the annual 'Great Newham' fun run detected in the first study within this chapter. There is a near 300% difference in the mean count for 'no event' days compared with the mean count for 'events' days during 2017. However, the range of possible values has substantial overlap between contexts. Furthermore, some of the contexts have very large ranges, specifically sporting events, and the Spring and Summer seasons.

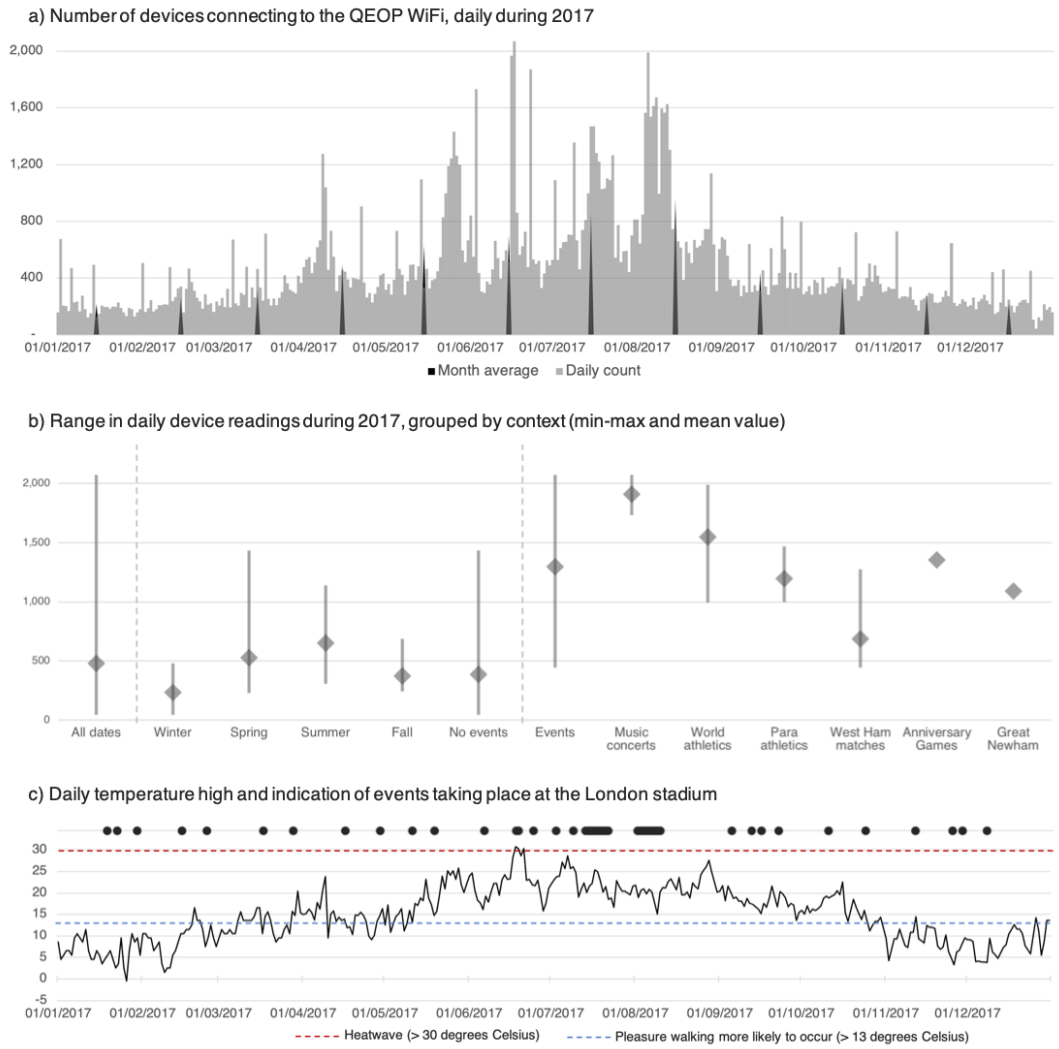


Figure 35. Wi-Fi unique device connections daily at the QEOP during 2017

a) shows daily count of unique device connections to park Wi-Fi network; b) shows mean and min-max range of daily readings grouped by contexts; c) shows daily high temperature and indicates whether or not an event was taking place at the London Stadium (black dots). Red horizontal line indicates heatwave (>30 degrees Celsius).

As detailed earlier in the chapter, audience size at the stadium can vary by event type. The largest attendance during 2017 was for Guns N' Roses with a combined attendance over two nights of 139,267. Athletics events can potentially use the full 66,000 fixed seats. However, most sporting events are capped at 60,000 and research suggests that music concerts are limited to 70,000 tickets. West Ham football matches are currently regulated to a maximum of 57,000. For Premier League football matches, 52,000 seats are held by season ticket holders. It is not known in advance how many season ticket holders will attend each event.

Figure 35c contains a plot of the daily high temperature recorded within the park and an indicator for whether or not an event was held at the London Stadium. The upper horizontal dotted line indicates 30 degrees Celsius, the daytime temperature for heatwave conditions. The lower horizontal dotted line indicates 13 degrees Celsius, the temperature above which pleasure walking is expected to increase, based on previous studies (Pushkarev & Zupan, 1975).

4.3.2.2 Predicting daily visitor numbers

Building the prediction model

Four regression algorithms were tested: linear, decision tree, random forest and support vector machine (SVM). They were chosen for being suitable to use with small data sets containing a mix of discrete and continuous attributes (Géron, 2017). The decision tree was anticipated to be the best because it is non-parametric, suits categorical data and non-linear relationships, and makes no prior assumptions about the distribution. Trees, however, are susceptible to overfitting. Random forests can overcome this issue to some extent by training across multiple decision trees. SVM constructs a hyperplane that maximises the distance between classes. It is the most versatile but better suited to continuous data. Linear assumes the distribution is normal, that there is a linear relationship, and that the features are independent of one another which may not be the case with this dataset. The results are presented in Table 10a (single model) and Table 10b (split model) and includes a tuned version of the Random Forest algorithm. Tuning was performed using a grid search to automatically find and select the best performing combination of features. It is a technique within machine learning to rapidly iterate through a model evaluating all possible combinations of hyperparameter values. It can be configured to evaluate the maximum number of features and estimators and will indicate what combination produces the best result. For measuring the error in predictions, root mean squared error (rmse) is used. Lower values indicate better performance.

For the single model (Table 10a), the decision tree performed best for all models in training, as expected, but performance degrades substantially during cross-validation, indicating the model is over-fitting. The Random Forest has larger variance on the training data but less degradation during cross-validation and with the smallest standard deviation. The Linear Regression model, whilst performing poorly during training, is consistent in cross-validation and final testing. The SVM model performs the worst and is discounted from further analysis. The tuned random forest produces the best test results. However, it is close to the simpler Linear model. This was unexpected given there are known dependencies between attributes such as temperature and month and a linear model assumes all dependent variables are independent of one another.

The story is different when the data is split into two contexts: non-event days, and event days (Table 10b). For the model based on non-event days, performance is similar to the single model in training when comparing algorithms, and results are improved across all algorithms. Again, the simpler Linear model performs as well as the tuned Random Forest model. The model based on event days only does not concur. The tuned random forest performs with a variance half that of the linear regression model. It is likely that certain event attributes have dependencies that limit the

performance of the Linear model. For example, holders of high-cost tickets may be much less sensitive to poor weather conditions, particularly special events that may attract visitors travelling long distances who have also booked overnight accommodation. Attendance by season ticket holders to football matches may have variable sensitivity to weather, depending on the importance of, or interest in. the match. For example, playing against a local rival or later in the season when consequences of match outcomes are becoming known, such as the likelihood of promotion or demotion within the English football leagues.

Table 10. Performance of machine learning regression algorithms on QEOP data

a) Single Model	Train RMSE	Cross-validation		Test RMSE
		Mean	St. Dev	
Linear Regression	138	137	35	136
Decision Tree	29	178	39	207
Random Forest	61	146	28	155
Support Vector Machine	347	339	83	427
Tuned Random Forest	56	134	29	132

b) Split Model	Non-event days				Event-days			
	Train RMSE	Cross-val		Test RMSE	Train RMSE	Cross-val		Test RMSE
		Mean	St.D			Mean	St.D	
Linear Regression	129	127	36	92	180	219	57	161
Decision Tree	30	171	41	130	33	205	81	198
Random Forest	61	132	33	95	83	240	74	127
Tuned Forest	57	127	31	93	103	233	105	80

RMSE is the root mean squared error (rmse) of the predictions. Mean and Standard Deviation (St.Dev) are for the range of rmse scores. Lower values indicate better performance.

The mean daily count across the year is 480. Using the single model, the prediction would be +/- 27 per cent. The mean day count for non-event days is 388 and for event days is 1,103. The average split model results in a confidence range of 24 per cent and 7 per cent respectively. The providers of the park Wi-Fi – WifiSpark Ltd – estimated that each unique device connecting during 2017 represented 33.3334 actual people in the park. There is no information provided as to why this number was chosen, and it increased in 2017 from 25 when the portal was first launched in August 2016. Assuming that the multiple is fair for the dataset, the confidence will be +/- 4,400 for a prediction built on estimates that ranges from 6,000 to 21,000 on non-event days.

Interpreting the model

The best performing model overall for predicting presence was the tuned random forest algorithm. However, it is a 'black box' algorithm making it difficult to inspect how influential the different features are. Whilst the decision tree performed poorly due to over-fitting, it can be easily inspected and visualized for human interpretation. This is an important consideration if outcomes are to be used in real-world decision-making. It can provide an indication of which parameters are influential.

Figure 36 shows the first four levels of the decision-tree for non-event days.

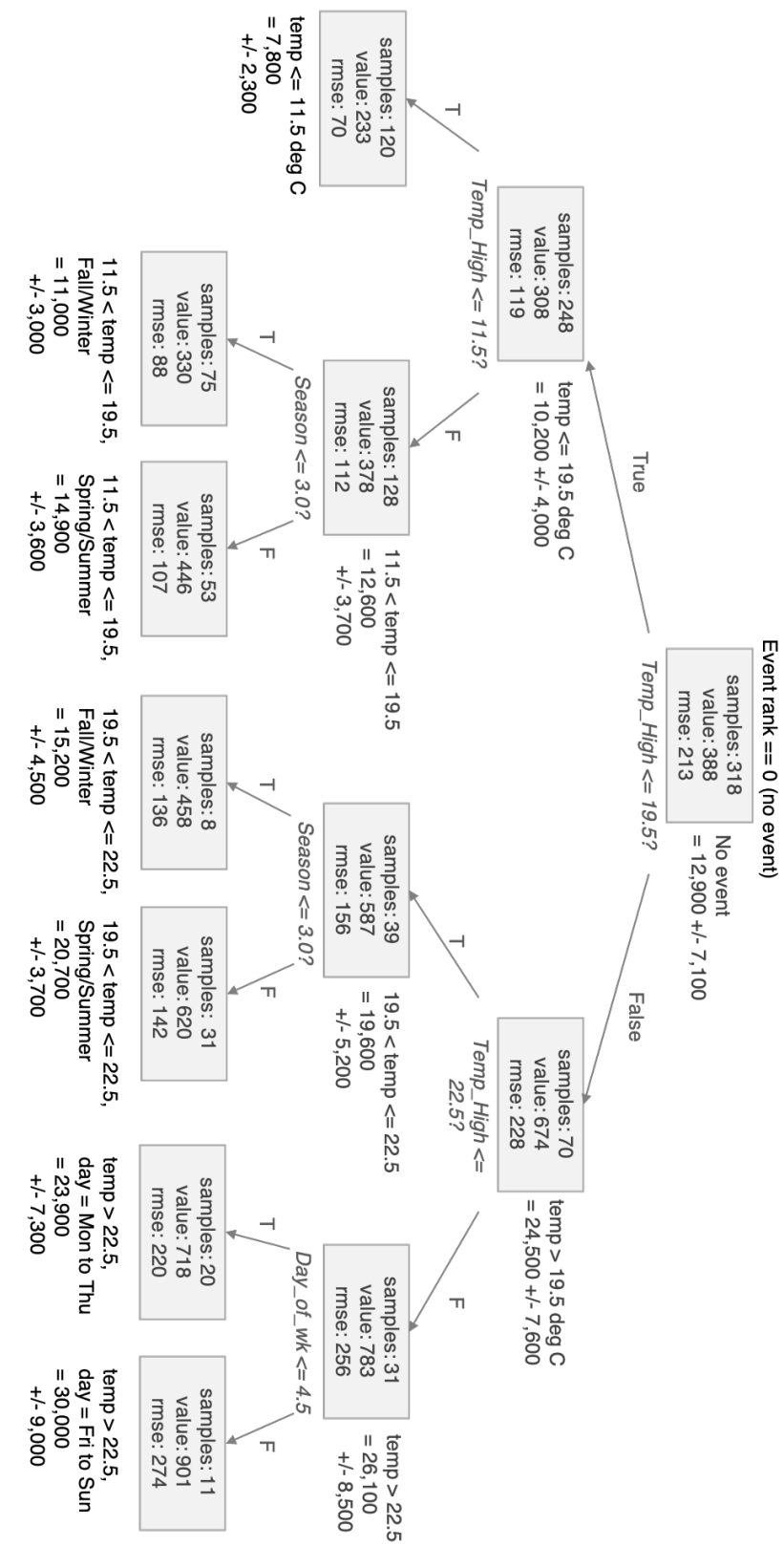


Figure 36. Decision tree regression model for non-event dates at the QEOP

Four levels of a split model, showing split for non-event days. samples = number of dates included. value = average (mean) number of devices. rmse = root mean squared error, indicating variance in the mean.

Figure 37 shows the first levels for football matches during 2017.

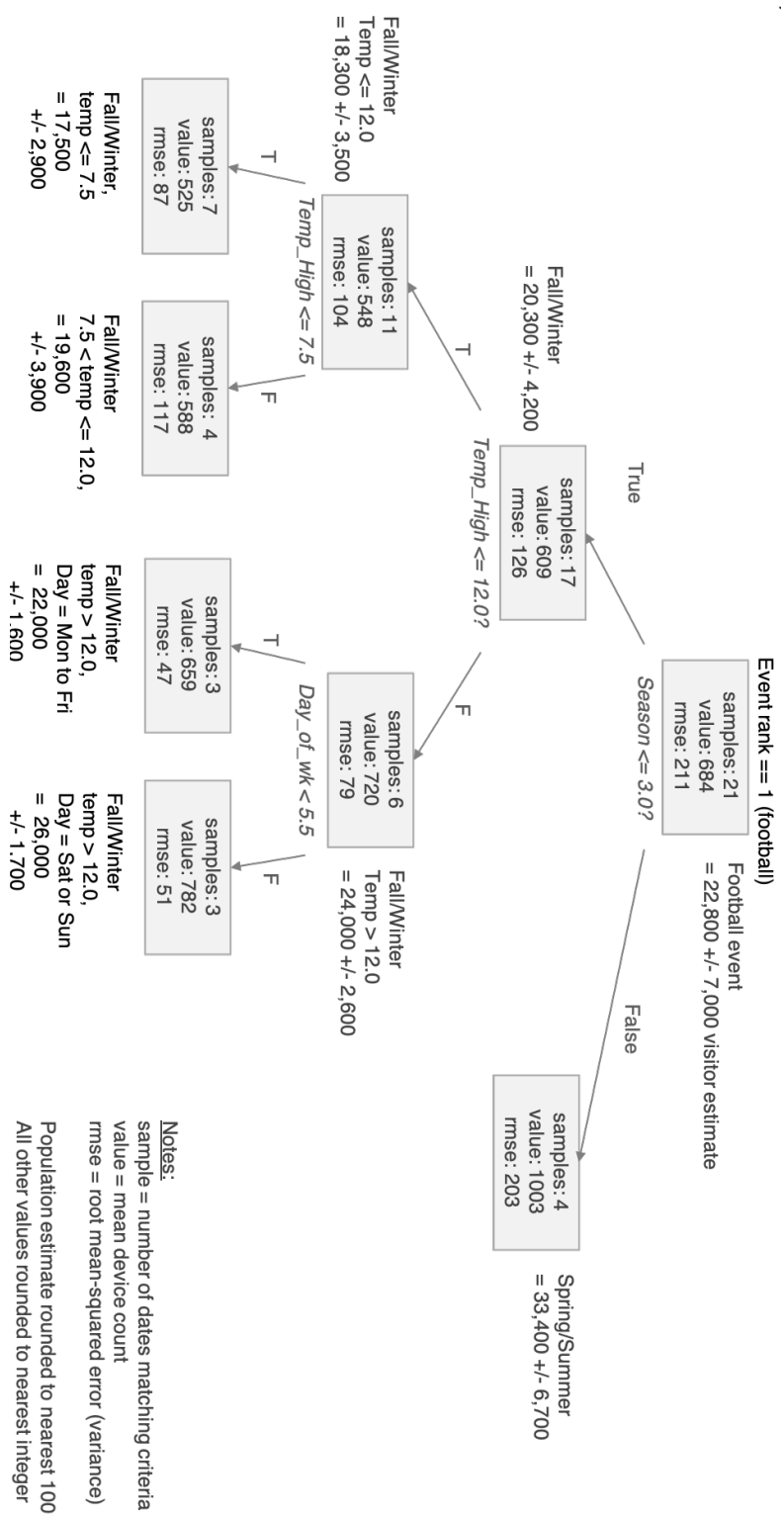


Figure 37. Decision tree regression model for event days within the QEOP

Four levels of a split model, showing split for football match days. samples = number of dates included. value = average (mean) number of devices. rmse = root mean squared error, indicating variance in the mean.

Both figures show the population estimate if using the WiFiSpark multiplier of 33.334 indicate the variation in presence within the park, assuming the multiplier is representative. For non-event

dates, the first two levels of division are temperature-based. By the fourth level, it is possible to see the combination of both season and temperature on visits. For the temperature ranges of 11.5 to 19.5, and 19.5 to 22.5 degrees Celsius, the population varies by 4-5,000 when comparing Fall/Winter with Spring/Summer seasons. For the hottest conditions, exceeding 22.5 degrees Celsius, visits on Fridays to Sundays are 6,000 higher than Mondays to Thursdays, an increase of approximately 25 per cent, but with high uncertainty.

For dates when large-scale events take place in the park, it could be questioned that a prediction model may not improve on a simple calculation based on advance ticket sales for the venue. However, football matches involving season tickets require the assumption that all season ticket holders will attend all matches. There may also be uncertainty for large-scale events with low-cost, discounted or free tickets distributed for less popular dates to fill the stadium. Furthermore, ticket sales do not indicate whether or not visits by non-attendees to other attractions within the park are affected by the event. Football matches typically include a substantial police presence and segregated routes to control pedestrian flows. Figure 37 shows the decision tree for dates when football matches took place in the stadium. Matches during the Spring predict visits to the park will be 64 per cent higher than Fall/Winter on average, although the samples are not balanced, with just 4 matches occurring Spring/Summer. During the Fall/Winter, the prediction varies by over 8,000. For temperatures above 12 degrees Celsius, there is a difference of 4,000 in the estimate for events on weekdays compared with the weekend. This could be an indication that season ticket holders are more likely to attend weekend matches than weekday matches.

One noticeable outcome for the football estimates is that the conversion to a population estimate appears to be low, ranging from average estimates of 17,500 to 33,400. The stadium has capacity for up to 57,000 people during football matches and official figures typically state attendance throughout the season as between 36,000 and 56,000[3]. It is possible that regular fans do not use the Wi-Fi network as much as visitors to other events, or that the Wi-Fi multiplier of 33.3334 is too low in general across all events. It is also possible that the official figures are over-stated as has been claimed in local news coverage[4]. The club's average across 12 games was quoted as 55,309. Research by the local authority, Newham Council, stated that average attendance was 42,779. This does indicate that the Wi-Fi multiplier is incorrect and should be revised, but perhaps not to the level of the club statistics.

The benefit of using a machine-learning algorithm to evaluate the influence of contextual attributes is that the algorithm improves as more data is provided. To improve the model, the next stage would be to incorporate further contextual data. For example, including ticket prices for events to explore for correlations between the cost of the ticket and the likelihood of attending.

---

[3] Source: West Ham website. https://westhamunitedblog.co.uk/2017/10/25/west-ham-attendance-2017-18-season-crowd-figures-stats-for-whufc-home-games-at-london-stadium-17-18/ accessed November 2018.
[4] Source: BBC Sport website: https://www.bbc.co.uk/sport/football/45158878 accessed November 2018

4.4 Research Outcomes I

4.4.1 Summary findings

The two studies produced several interesting findings and also challenges for the use of mobile data as a sample of real-world human activity.

Use of the Wi-Fi network requires devices to complete a registration process when first connecting. It is not known what percentage of people present in the park are carrying devices and choose to register to use the Wi-Fi network, or whether there is a demographic bias. Whilst 76% of UK adults owned a smartphone at the time of the studies (Ofcom, 2017), the percentage of children carrying smartphones is unknown but assumed to be lower, particularly so for younger children. Thus, on days when families are visiting the park, the number of devices may under-represent the actual number of people present. However, the Wi-Fi data correlated with headcounts provided by an LLDC partner that uses cameras monitoring entrances to the park, indicating that it is representative. This finding does rely on the assumption that the head counting method is robust. Unfortunately, we have no way of testing due to the proprietary nature of the data and its collection. The alternative is to collect counts using human observers. However, for the size of the park and potential numbers of people entering simultaneously at the 10 entrances covered by cameras, such a study would be difficult and resource-intensive to undertake.

The usefulness of social media sources in predicting daytime populations proved limited for the study of a large open space due to the sparse readings on non-event days. It suggests such sources may only be suited to denser built environments or generalisations across larger spatial scales. That said, the sources were beneficial on days when activities occurred within the park, both known and unknown. Whilst readings may be skewed towards high-profile landmarks and events, the resulting data can produce insights into changes in environmental and social conditions.

Taking into consideration the challenges of using samples of mobile data to represent real-world populations, the analyses produced substantial contextual variations, both for infrequent large-scale events and also for recurring ambient rhythms, and at daily and hourly scales. For large-scale single events held at the stadium, such as music concerts and football matches, it appears possible to forecast the impact on population behaviours throughout the day, provided the start and end times of the event are known. There is less certainty for all-day and multi-day events such as athletics championships. The school summer holidays and peak tourist season, from mid-July through to the end of August, also had an impact on park behaviours, concentrated in the afternoon. Readings were much higher, and the percentage of visitors likely to be ‘explorers’ who are visiting for the first or second time increased. This may affect how people navigate the park. Weather also had a demonstrable and unexpected effect. Unseasonably warm conditions in the winter produced a higher count than mild, but still pleasant, conditions in the summer. These variations create the potential to quantify the volume of interactions that are likely due to necessary activities versus optional activities that are more malleable to conditions, to use Gehl’s taxonomy (Gehl, 1987).

4.4.2 Contextual framework update

The findings from the Wi-Fi data across the two studies suggest measures that could be used to populate a contextual framework for modelling variation in visits to the Queen Elizabeth Olympic Park at two temporal scales: total visits daily, and the hourly distribution of visits throughout the day. The model was developed with 12 months of aggregate daily device counts and four months (March and June to August) of individual real-time readings aggregated hourly. It can be used to replace a static or annual average with a rolling time-specific estimate that reflects known circumstances for a range of contexts across seasons, activities and weather states.

Figure 38 shows the P-STAR concept. The spatial properties are thresholds to constrain results (for example, to apply an upper and lower limit to the number of visitors who could be present), and the spatial distribution across the park. The temporal distribution reveals variations daily and/or hourly with adjustments for the time of year and the day of the week, assuming a normal term-time day with ambient weather for the time of year. The temporal distribution can then be adapted for activities taking place and adjusted for weather conditions. Only temperature has been evaluated. The landscape is a park and it is assumed that the majority of non-event visits would be optional rather than necessary, suggesting high sensitivity to conditions. There was an indication in the second study that ticket sales may not be a reliable indicator of event attendance. In particular, season ticket holders for football matches appear to be sensitive to cooler temperatures during the winter. Multi-day events such as athletics championships had the widest range of variation in attendance. It is likely that ticket prices varied substantially across the days and that the likelihood of attending correlates with the price paid for a ticket. However, this information was not available for analysis.

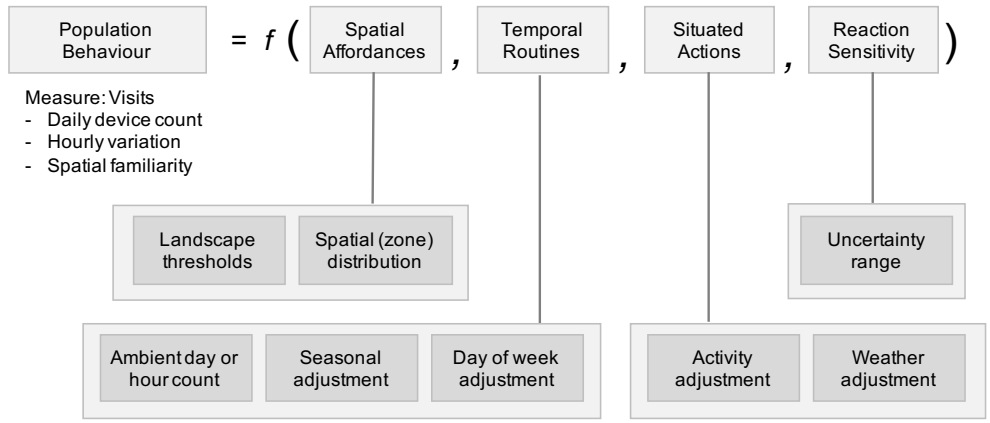


Figure 38. Using the P-STAR formula to forecast visits to the QEOP

At this stage, the P-STAR formula would produce an estimate of the number of devices connecting to the Wi-Fi network and the spatial familiarity (proportion of habitual visitors versus explorers). We do not know how representative device connectivity is of the actual population present. The next chapter builds on these findings and explores the potential to incorporate administrative data to convert device counts into an active population estimate.

PAGE INTENTIONALLY LEFT BLANK

5 Case 2: A Mobile Landscape

In chapter four, the most comprehensive real-time data source for learning population behaviours within the Queen Elizabeth Olympic Park (QEOP) came from the public Wi-Fi network installed in the park. The research had two limitations. First, the Wi-Fi range covers only the park itself making it difficult to explore what effect activities in the park have on the surrounding area. Second, it is not easy to convert the number of devices detected into a representative population measure because there is no administrative population source that corresponds to the area of the park. This chapter introduces a mobile data source that records the spatial coordinates of individual devices and is not constrained to a specific location. It has many benefits over Wi-Fi data. First, it enables the incorporation of administrative data by aggregating the mobile data to the same areal boundary as the administrative source. Second, it provides readings for where people are located rather than where access points are installed. Third, it enables the study of relationships between local hot spots of activity without being constrained by the range of embedded sensors. However, it is a much smaller sample of real-world activity compared with the park data sources.

The chapter comprises of three parts. The first introduces the mobile data set and compares it with administrative population measures for counting people in a landscape. The second part explores the potential to combine administrative and mobile data to produce an active population estimate that incorporates contextual awareness. The third part explores the potential to utilise the mobile dataset to go beyond a simple count of visits to detect active spaces within the landscape and analyse behaviours such as dwelling within and moving across the landscape. Such behaviours cannot be disaggregated from static administrative statistics.

5.1 Counting People

A core objective of this research is to evaluate the potential for a sample of mobile data traces to provide quantitative information about local population dynamics. Such knowledge has historically been resource-intensive, expensive and time-consuming to acquire. The arrival of smartphones has created the opportunity to develop population measures at new spatial and temporal scales.

5.1.1 Mobile device as a proxy count

For this research, a source of data was kindly provided by OpenSignal (www.opensignal.com) for academic research. OpenSignal's mission is: "*to combine real-world measurements with scientific analysis to provide independent insights on mobile connectivity globally*." OpenSignal has produced a mobile app to crowdsource the mapping of, demand for, and availability of public wireless connectivity worldwide. Measurements are collected from the devices of participants who download and install the OpenSignal app. The app runs continuously in the background, generating a reading whenever the device detects an access point for a public Wi-Fi network. The reading includes the device's coordinates, a timestamp and data about the Wi-Fi network detected, such as signal strength and capacity.

A working hypothesis of the app's senior management team (Webb, 2017) is that a small random sample of people – as few as 100 in a small town or city district – are enough to map the quality of public wireless networks. This thesis extends the hypothesis by exploring whether the data can identify fluctuations in demand and how it varies for different contexts. It is anticipated that the dataset will be better suited to public space than private environments due to only detecting publicly available Wi-Fi networks and this must be considered when interpreting results.

A benefit of this data source compared with data from the Wi-Fi network is that it contains the spatial coordinates of the device rather than the access point. A second benefit is that participants are not recruited to conduct experiments or constrained to a specific behaviour setting, as has been the case with living labs deploying mobile devices to participants. A third benefit is that it requires no intervention by the user, so is not likely to be skewed towards specific or self-curated activities as can be the case with social media.

OpenSignal provided a data set that has been prepared for use in academic research. It contains individual readings within the Greater London area during June 2017, with some pre-processing. Original device IDs have been replaced with anonymised values and no information is provided about the demographic characteristics of the participants. Coordinates have been spatially reduced to three decimal places, with no data regarding the location accuracy. This provided individual data points with an uncertainty radius of approximately 55 metres around each coordinate. The timestamp was reduced to date and hour and the dataset was randomly shuffled. Readings within each hour are not necessarily in time order within the hour. As with the Wi-Fi data set used in chapter four, a concern with this source is the potential for demographic bias. There is no demographic information about any of the participants. A previous study (Richardson, 2015) using an earlier dataset from OpenSignal compared the spatial distribution of readings with various demographic and socioeconomic measures across London. No bias was evident in the spatial distribution of readings. That does not mean that the dataset was not biased, only that the bias cannot be easily quantified using administrative data or other urban data sources.

Adoption of smartphones is now near pervasive amongst UK adults (Ofcom, 2017). However, the use of city apps may be skewed towards individuals active in the city for work and/or leisure. The dataset likely under-represents children and less active demographics such as the elderly, disabled and people not in work or education. Such under-representation must be considered when interpreting results. Other sources of real-time data, such as public transport smart card data suffer similar issues. For example, children under the age of 11 can travel free on London bus and rail services and so are not represented in travel statistics based on gate counts. A 2007 study of the travel patterns of London residents found that 75% of men have a driving license compared with 57% of women, women use the bus more than men and men use the underground more than women, women take 15% more trips and are 25% more likely to trip-chain (TfL, 2007). People with mobility issues may be deterred from public transport due to accessibility issues and may prefer private specialist transport providers. Digitised interactions afford new spatial and temporal scales of analysis of population behaviours, but they are only a sample of real-world interactions.

A second consideration is whether or not the mobile dataset concurs with other mobile sources, those explored in chapter four. Whilst the periods covered by each dataset are different, which will create uncertainty in the comparison, they are available for the same month albeit 12 months apart. Figure 39 compares OpenSignal readings during June 2017 with Wi-Fi readings and incoming headcounts during June 2016. The counts are the mean per hour scaled from 0 to 1 using min-max normalisation where the min is 0 and the max is the largest mean count across 24 hours.

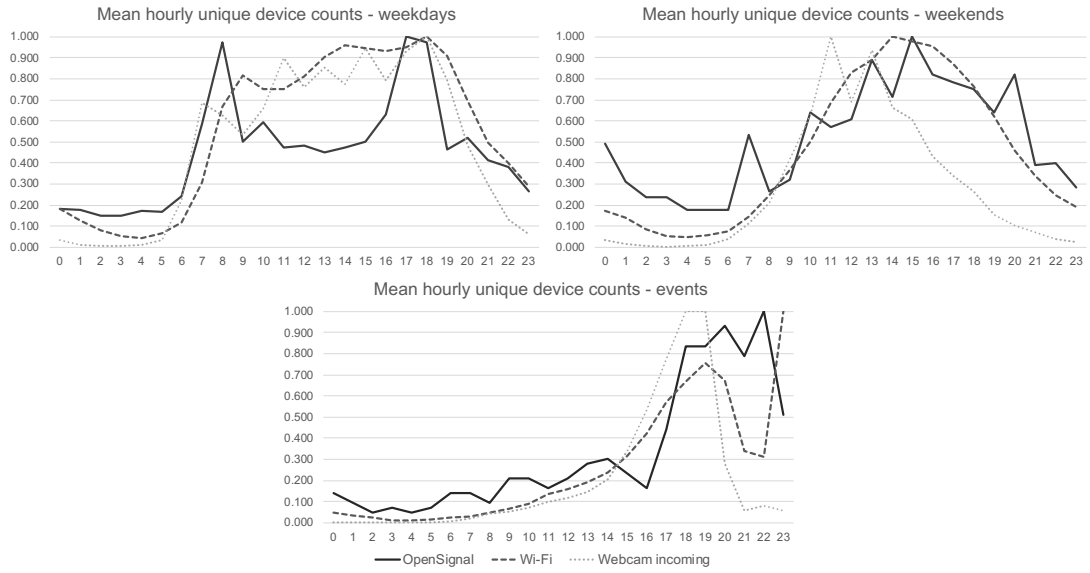


Figure 39. Comparing mobile app data with Wi-Fi and webcam readings in the QEOP

Hourly means have been scaled between 0 and 1 using min-max normalization, where the min is 0.

For non-event weekdays, the OpenSignal data source exhibits peaks during the morning and afternoon commuting periods but shows no lunchtime peak. The Wi-Fi has weaker peaks but does include a lunchtime peak. As can be seen in Table 11, there is an order of magnitude difference between actual counts for the different data sources and OpenSignal readings are very low within the park when aggregating hourly. It is possible that the volume is too small to reliably detect optional activities such as ad-hoc lunch breaks in the park but is sufficient for necessary activities such as travel between home and work. The OpenSignal data does broadly match the Wi-Fi data distribution at weekends and on the event day. Both OpenSignal and Wi-Fi show little weekday to weekend variation, suggesting a possible bias towards working adults visiting or traversing the park during the working week. This needs to be considered when interpreting results.

Table 11. Comparing mobile app data with Wi-Fi and webcam readings in the QEOP

	Non-event weekday		*Non-event weekend*		*Event Saturday*	
Mean hour count	*Min*	*Max*	*Min*	*Max*	*Min*	*Max*
OpenSignal	1.0	6.6	1.0	5.6	1.0	21.5
Wi-Fi	12.6	278.5	16.2	329.0	17.0	1,613.0
Webcam	2.8	560.0	5.2	1,673.0	19.0	12,550.0

OpenSignal data is from June 2017 (non-event: 5 to 28 June excluding 16, 17 and 23, events: 3 and 17 June). Other sources are from June 2016 (non-event: 6 to 26 June, event: 4 June).

5.1.2 The science of where people are

One objective of this chapter is to explore the potential to blend administrative data with mobile data traces to produce a near real-time population estimate. This section describes and compares the different population measures available.

5.1.2.1 Population measures

There are three approaches to measuring the size of the population present within a defined area: administrative data such as a census that provide a static formal statistic; ambient estimates that redistribute or disaggregate administrative data to account for behaviours that may not be captured in formal statistics; and, active counts generated from real-world observations that incorporate temporal variation by counting presence in the same space at different periods.

Administrative

Administrative data are official statistical records, such as a national census. They typically require mandatory participation and provide the nearest measure to a complete population aggregated at a range of areal scales. Administrative areas are drawn as a thematic map of contiguous polygons covering the entire country. Each polygon surrounds several people or houses within a defined range. This means administrative areas can vary substantially in size, spatially. Furthermore, output areas are revised and potentially amended before each census to reflect changes in the landscape. Table 12 contains the output area definitions used for the 2011 census in the UK.

Table 12. Administrative output areas within the UK for 2011 census

Administrative unit	*Definition*	*Number of areas*
Borough/Local authority (LA)	Largest administrative unit of division Population size ranges from 34,222 to 1,124,569	353
Medium super output area (MSOA)	From 5,000 people or 2,000 households Up to 15,000 people or 6,000 households	6,791
Lower super output area (LSOA)	From 1,000 people or 400 households Up to 3,000 people or 1,200 households	32,844
Output area (OA)	Smallest administrative unit of division From 100 people or 40 households Up to 625 people or 250 households Average population size of an OA in 2011 is 309	171,372
Workplace zone (WZ)	Created by splitting and merging OAs to produce a workplace geography that contains a consistent number of workers or businesses. Constrained to MSOA boundaries. Excludes Scotland	53,578

Source: (ONS, 2016)

The UK conducts a national census once every ten years with mandatory participation, recording the residential population at the household level including demographic, cultural, social, and economic information about each individual. The most recent census was completed in 2011 and included a workday population count for the first time, for England and Wales. It was a recognition that the use of residential counts to study social and environmental phenomena could produce

misleading statistics regarding the population affected or at-risk for incidents that occur in non-residential locations and at times of the day when people are less likely to be at home, such as daytime during the working week.

The workday population is a count of people who are employed in the area or not in employment but live there. The census included a question to indicate the place of work the week before the census was completed (27 March 2011) for all adults aged 16 to 74 in employment (nomis, 2014). It excludes people living in England and Wales but working in Scotland, Northern Ireland or outside the UK, people who work in England and Wales but reside elsewhere, and short-term residents. A new output area – workplace zone (WZ) – was also created to produce areas with a consistent number of workers or businesses contained within each area.

An administrative source may be considered a more robust workday population measure than a sample of mobile data due to its completeness. A limitation is that it provides only an indication of where people were working and studying at the time of the census and considers only one location each for residential and non-residential activities. Such an approach does not reflect the variety of activities and locations for those working in multiple part-time roles or whose work involves travel to multiple locations, or other daytime activities that can take place, both during the working week and at weekends such as travel, shopping, entertainment and leisure, socialising, community and care. It also does not consider tourism that may affect both day- and night-time counts.

Ambient

Over the past decade, researchers have begun developing methods for providing alternate measures to residential statistics, often referred to as 'ambient' population measures. An ambient population count is an estimate calculated by redistributing the residential population data from a census according to multiple local spatial attributes for a temporal category such as daytime. The term 'ambient' is used to exclude anomalies that would create substantial deviations from routine behaviours, such as attending special events or temporary disruptions within the landscape. Instead, an ambient measure is one that is temporally averaged and incorporates human mobility (Sutton, Elvidge, & Obremski, 2003).

Oak Ridge National Laboratory (ORNL) introduced a world-wide ambient population count in 2007, called 'LandScan™' (Bhaduri, Bright, Coleman, & Urban, 2007). Updated annually, LandScan is a dasymetric model. The count is generated using an algorithm that redistributes the residential population, taken from the most recent census, using ancillary datasets that include land-use attributes, satellite imagery, transportation data, socio-economic indicators and cultural factors that take into account regional differences in settlement practices. According to the documentation:

> *…cells are preferentially weighted for the possible occurrence of population during a day. Within each country, the population distribution model calculates a 'likelihood' coefficient for each cell and applies the coefficients to the census counts… The total population for that area is then allocated to each cell proportionally to the calculated population coefficient. The resultant population count is an ambient or average day/night population count. (ORNL, n.d.)*

The output is a grid-based count. Each grid cell has a height (y-axis) and width (x-axis) of 30 arc seconds (0.00833.. degrees latitude/longitude). 30 arc seconds represents a distance of 1km South-North by up to 1km wide West-East depending on latitude to accommodate the spherical shape of the planet. An ambient day is based on a regular weekday during term-time when academic institutions are in session.

In LandScan terms (ORNL, n.d.), differences between day and night populations are defined as:

Nighttime Population = Nighttime Residential Population + Nighttime Workers + Tourists + Business Travellers (+ Static Population)

Daytime Population = Workers + School Children + Tourists + Business Travellers + Residual Nighttime Residential Population (+ Static Population)

The creators of LandScan acknowledge there is subjectivity in dasymetric modelling and some assumptions about cultural settlement practices are impossible to validate. The LandScan data is copyrighted with license agreements available for commercial and academic use. Counts are provided under an academic license that represents the average present over 24 hours. Separate counts for day and night population estimates were not available when the research for this thesis was completed. The algorithm is proprietary and not available for inspection making it difficult to validate any assumptions or the accuracy of the individual data sources. There is also limited documentation provided. For example, it is not known what interval the ambient population represents, both for time of day and season of year.

Population 24/7 is a framework developed at Southampton University (Martin, Cockings, & Leung, 2009). It is similar to LandScan in adopting grid-based surface modelling and redistributing the residential population according to land-use and spatial attributes. However, the goal is to produce a near real-time ambient estimate. The framework utilises building information to identify schools, offices, shops and other non-residential structures. This enables the redistribution of school-age, working and retired populations that are likely to have different daily routines. The most recent development has reduced the spatial scale of the grid cells to 100-metre cells, analysing a 50km^2 area of the city of Bristol in the United Kingdom (Cockings & Martin, 2018) and is seeking to produce an ambient near real-time population count at 15-minute intervals, identifying variations by day of the week, time of day, and workdays versus holidays for different land-use profiles. Preliminary results for Bristol, focused on retail areas, mimic the diurnal patterns shown in studies presented in chapter two. Most activity occurs from 10:00 to 18:00 with variations across days of the week. The quietest period occurs at around 04:00. During the period this research was undertaken, the tool had been developed for two city regions – London and Bristol – and was not available for public use or other academic research.

A 2015 study (Greger, 2015) adopted a similar approach to Population 24/7 by labelling buildings in Tokyo into 6 categories likely to produce different activity patterns: home, business and office, education, retail and service, leisure and hotel, and public institutions. The study then combined

multiple static data sets to redistribution the population, including (residential) population census data, employment data and address point data. The study went further to incorporate temporal flows in building occupancy levels by acquiring data about peoples' trips from origin to destination across the Greater Tokyo Metropolitan Area. The survey was a paper questionnaire recording trips conducted a single date – Thursday, 1st October 2008 – to reflect a regular working day. It also collected demographic information to capture variation in trips by age, gender and occupation. The questionnaire recorded only the start and end locations for each trip so does not reflect trip-chaining activities such as visiting a shop or taking children to school en route from home to office. Furthermore, research using mobile data readings gathered over multiple days have revealed substantial variations in daytime readings aggregated hourly for weekdays and weekends (see section 2.2.2 in chapter 2). Such variation is not reflected for a count on a single day.

Both Population 24/7 and the study of Tokyo demonstrate that disaggregating administrative data by building category can produce ambient population measures that reflect variations throughout the day for recurring activities at a finer scale than is provided by LandScan or the workday population recorded in a census. However, a limitation of this approach is the need to gather and maintain detailed building information. Furthermore, there was no available dataset in this format at the time the research presented in this thesis was conducted. The only available sources of administrative or ambient population measures for locations across London in 2017 were the 2011 UK census containing residential and workday population statistics, and the LandScan ambient population estimates being revised annually.

Active

The alternative to administrative data and ambient population estimates that disaggregate administrative data is to count the number of people present within an area at different times. The traditional method is gate counting, also referred to as a cordon-count (Charles-Edwards & Bell, 2013), where people cross a physical threshold and are either counted automatically if a barrier is in operation, or manually by a person clicking a tally counter. Mobile data sources provide an alternative method that can cover areas where there is no physical barrier and/or when it would be resource-intensive to deploy human observers. Traditionally, active counts have been performed for specific events, or over a short period of time to capture detailed presence and movements within a behaviour setting. Urban digitisation and mobile data introduce the potential to extend active counting over longer periods and larger distances. It creates the potential to compare different contexts and recalibrate population measures on a frequent basis to accommodate urban and cultural changes that can affect presence and movement patterns.

As explained in chapter two (section 2.2), a number of studies have been published over the past decade to show presence and movements based on social media activity and mobile data provided by telecommunications providers (Telcos). For example, a 2014 study used a dataset consisting of over 1 billion mobile phone call records (Deville, et al., 2014) to demonstrate the potential to overcome the static limitations of census data, quantifying population density by calls made at cell towers to show variations during the day versus night and seasonal changes.

Social media-based studies have indicated the potential for producing population measures with temporal granularity when applied at city scales. For example, studies of geo-tagged tweets in Leeds (Birkin & Malleson, 2013) and London (Longley & Adnan, 2016) both demonstrated the potential to identify different activity types and their spatial distribution across a city. However, when reducing the focus to neighbourhood-level and urban outdoor space, the volume of tweets on a daily basis became too small to produce a measure of routine activity patterns.

Research utilising large volumes of real-time data has demonstrated that mobile data can reveal the diurnal rhythm of urban space, but also that there is substantial variation from day to day, such as the 2004 study of Telco activity in Milan (Ratti, Pulselli, & Williams, 2006) and 2017 study of wi-fi activity in Manhattan (Kontokosta & Johnson, 2017). This research is focused on studying those variations, seeking to quantify how much presence can vary within the same physical space at the same recurring interval due to changes in social and environmental conditions.

A limitation with using big data sources, when compared with cordon-count methods, is that they are only a sample of activity that occurs within the landscape. For example, data from a Telco provider will only represent customers of that Telco provider and not everybody posts to social media. Even with near-pervasive adoption of mobile devices amongst UK adults (Ofcom, 2017), we cannot assume that all adults themselves are mobile and must acknowledge the presence of others within the landscape who are not carrying mobile devices or participating in social media. Thus, a count of mobile data cannot, on its own, be converted into an active population estimate.

A hybrid approach is to combine an administrative or ambient source of data with an active source. The administrative or ambient source provides a baseline measure, and the real-time source can reveal contextual variations not captured by administrative or ambient sources. ENACT (ENhancing ACTivity and population mapping) was an exploratory two-year research project funded by the European Commission Joint Research Centre that adopted such an approach. It began in January 2016, coinciding with the start of the research presented in this thesis. The project's main objective was, "to produce a set of multi-temporal population grid maps that take into account the main seasonal and daily variations of population," (Batista e Silva, et al., 2016). Spatial aggregation was performed as a grid of uniform cells, at two scales: 100 metres (with the temporal resolution set to a single year, an ambient measure) and 1km (with a temporal resolution of seasonal and day/night to produce an active measure). The ENACT project blended administrative census data with big data sources to create a measure of tourist density at various locations across the European Union (Batista e Silva, et al., 2018). Data from two major online travel booking sites, Booking.com and TripAdvisor.com, were acquired to produce a measure of 'bed places' and 'nights spent' in bookable accommodation. These were combined with residential statistics to produce a measure of 'tourism intensity' and 'tourism seasonality' as variable population measures.

A hybrid approach is also proposed here, to blend a sample of mobile data with an administrative or ambient source as the spatial baseline, to help convert a mobile device count into an estimate of the number of people present within a landscape, and how it varies for different contexts.

5.1.2.2 *Comparing areal scales*

The population measures providing a spatial baseline and available for use in open research as of mid-2017 were the UK 2011 census – residential and workplace – and the LandScan 2015 ambient estimate. A 2016 UK-based study (Malleson & Andresen, 2016) evaluated a range of measures for improving quantifying rates of crimes against the person, a typically daytime phenomena. The objective was to improve quantifying rates of crimes against the person. The study concluded that the most reliable measure was the census workday population. However, crime rates are generalised and do not incorporate seasonality or temporal variation through the day. Furthermore, the structure of output areas may not be appropriate for aggregating data sources with spatial coordinates that contain uncertainty, as is the case with mobile devices using GPS to triangulate their position, and readings located at fixed access points or cell towers. Figure 40 demonstrates this challenge, showing the four different sizes of administrative output area for the UK census.

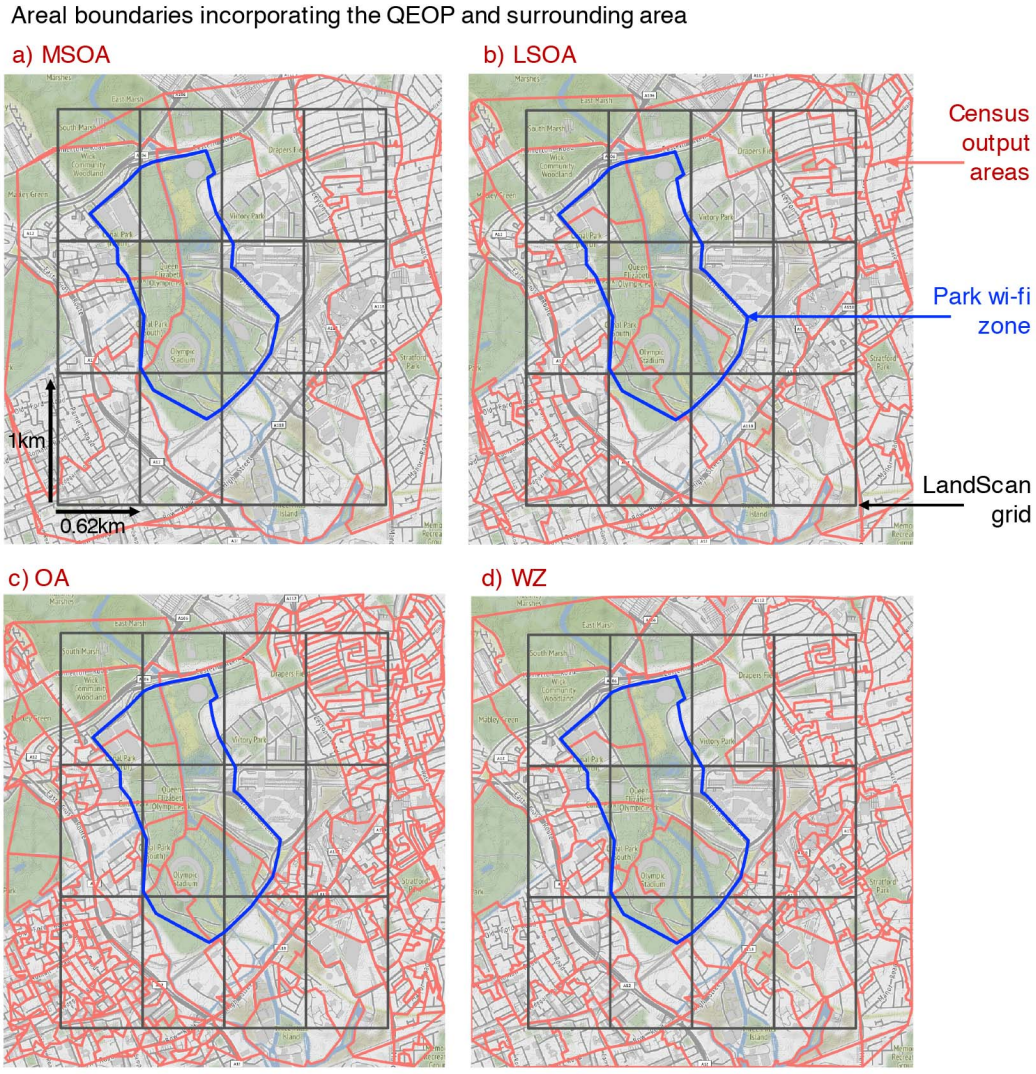
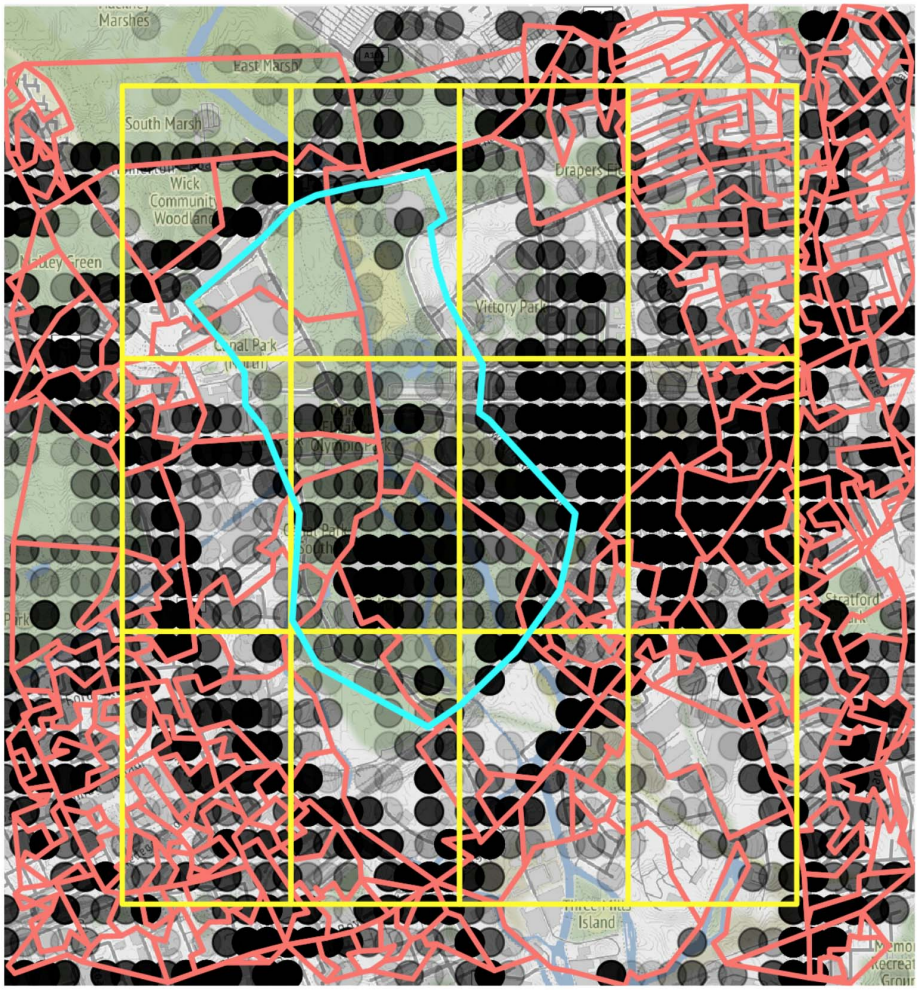


Figure 40. Areal boundaries incorporating the QEOP

Blue outline is the approximate area of the park covered by Wi-Fi; Black grid is LandScan uniform cells; Red polygons are shown for medium super output areas (MSOA), lower super output areas (LSOA), output areas (OA) and workplace zones (WZ) in images a), b), c) and d) respectively.

Across all four scales of census output area, drawing polygons to each contain a broadly consistent number of people based on their residential or work location can produce substantial variations in size and shape, but with different affects depending on the morphology of the landscape. For example, the outline of the QEOP park shows that the same boundaries apply at LSOA, OA and WZ scales. The same issue applies to the Westfield retail centre east of the park. Furthermore, there is a tendency to draw administrative boundaries along roadways to avoid dissecting buildings. This can affect mobile data sources captured at street-level, with reading falling on boundaries between output areas. Figure 41 demonstrates this challenge with a plot of three days of OpenSignal readings. Coordinates are provided in latitude/longitude format shortened to three decimal places, creating a spatial uncertainty of at least 55 metres around each data point. Each point plotted is sized to approximate this uncertainty. In denser locations, the actual coordinates could be located in one of up to four different output areas and there is a tendency for readings captured on roadways to fall on boundary lines.

Three days of OpenSignal readings across the QEOP map tile



Background: map tiles Stamen Design, under CC BY 3.0;
map data by OpenStreetMap, under ODbL

Figure 41. Plotting mobile data points for comparing with areal scales

Three days of OpenSignal readings during June 2017 with points sized to reflect the approximate uncertainty in the spatial coordinates due to being reduced to three decimal places latitude and longitude. Points plotted with an alpha of 0.2 (0 = transparent, 1 = opaque) to indicate volume of collocated readings. Outline of the park wi-fi zone, LandScan uniform grid and polygons for administrative output areas (OA) added for reference.

In recent years, there has been a move towards standardising the use of grids of uniform cells in preference over administrative output areas for producing ambient and active population measures (Batista e Silva, et al., 2016). Spatial resolutions range from 1km x 1km to 100m x 100m depending on the spatial and temporal granularity of the data being aggregated. The ENACT project proposed to disseminate population measures at a spatial resolution of 100m for generalised measures and a spatial resolution of 1km with a temporal resolution of monthly or seasonal and day- versus night-time counts (Batista e Silva, et al., 2016). This research has followed the latter scale, aligning with an available gridded baseline measure: LandScan. The LandScan cell size is 1km tall (South-North) by up to 1km wide (West-East) depending on location, to accommodate the shape of the planet. Across London, the width of cells is approximately 6.2km. As can be seen in Figure 41, at this scale, some cells are densely populated with mobile readings whilst others are much sparser. However, all cells contain readings, indicating the potential to capture temporal variations that may not be possible at smaller resolutions.

Uniform grids are also susceptible data accuracy issues. Readings with spatial uncertainty close to the edges of cells may be misclassified as being present in the neighbouring cell. However, the issue is consistent across all cells and, for mobile readings, is reduced when compared with output area that have borders along roadways. A benefit of uniform grids is that the areal boundary does not change, as can occur with administrative output areas. Furthermore, there is no need to produce population densities, given all cells are uniform in size.

To examine the use of the LandScan grid and ambient population estimate, data are plotted for OpenSignal and two other sources that capture human-environment interactions with spatial and temporal coordinates: the London Fire Brigade (LFB) and London Metropolitan Police Service (MPS). Fire records are openly available on the London Data Store (https://data.london.gov.uk accessed August 2017) and contain coordinates and timestamps. Crime records are provided by the MPS (https://www.met.police.uk/sd/stats-and-data/ accessed August 2017) with obfuscation. The timestamp is rounded to the nearest month, and the spatial coordinates for the location are 'jittered' to point to the centre of the nearest populated street. At the time of the analysis, fire records were available for January to June 2017, crime records were available from March to May 2017. The OpenSignal data was available for June 2017. Counts are scaled using min-max normalisation.

The ambient count for LandScan is highest in the south-west corner of the grid (cell A3). This cell contains a mix of residential and non-residential buildings. Of the three real-time sources, only the Fire incidents concur. Both the OpenSignal and Crime records are highest in the centre east cell (D2), which contains a mix of residential and non-residential buildings. Figure 43 shows the plots of real-time data as counts sub-divided into 'pixels' cells, 1/16th of a LandScan cell, 250 metres tall by approximately 155 metres wide. Counts are again scaled using min-max normalisation. Readings are so concentrated for crime statistics, a log transform (base 10) is also applied to enable visibility of readings in other pixels (Figure 43e, f, and g).

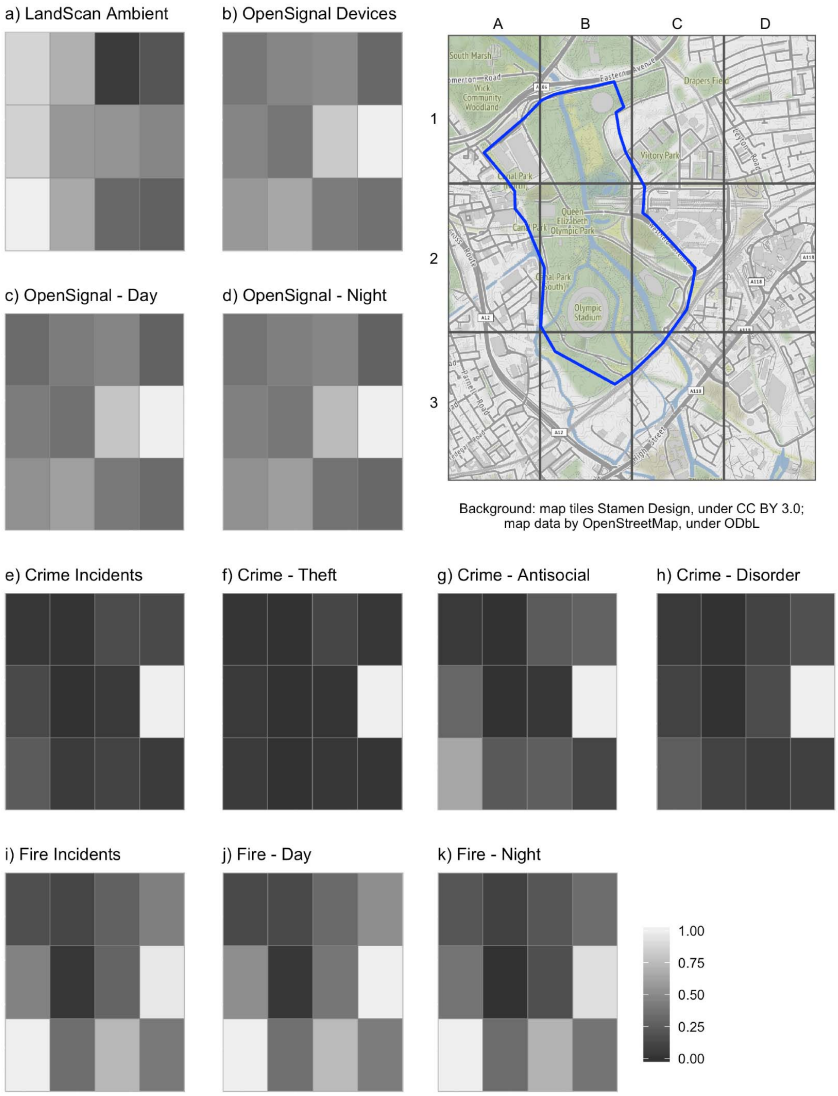


Figure 42. Compare population densities, LandScan grid of cells across QEOP

Values scaled from 0 to 1 using min-max normalisation, lighter grey indicates higher counts.

The reduced scale reveals potential bias in all types of data when aggregating. The highest OpenSignal counts are concentrated at the eastern edge of LandScan cell C2 and the western edge of LandScan cell D2. Crime readings are highly concentrated in a single pixel in LandScan cell D2. This is the location of Stratford Centre, an older retail centre that pre-dates Westfield. There is an apparent absence of crime at Westfield. This is possibly due to crime displacement resulting from the deployment of private security staff within the Westfield retail centre, which is a privately managed public space. Or it could simply be the jittering effect resulting in all crimes in the area being pinpointed to the nearest public street. A further aspect when studying at the finer grid-based scale is that the morphology of the landscape begins to visualise, most apparent in the OpenSignal mobile data set (Figure 43a, b, and c), but also visible as a rough outline in both log-transformed crime records and fire incidents. Interactions occur along the major roadways surrounding the eastern and western edges of the park, with minimal readings across the park itself.

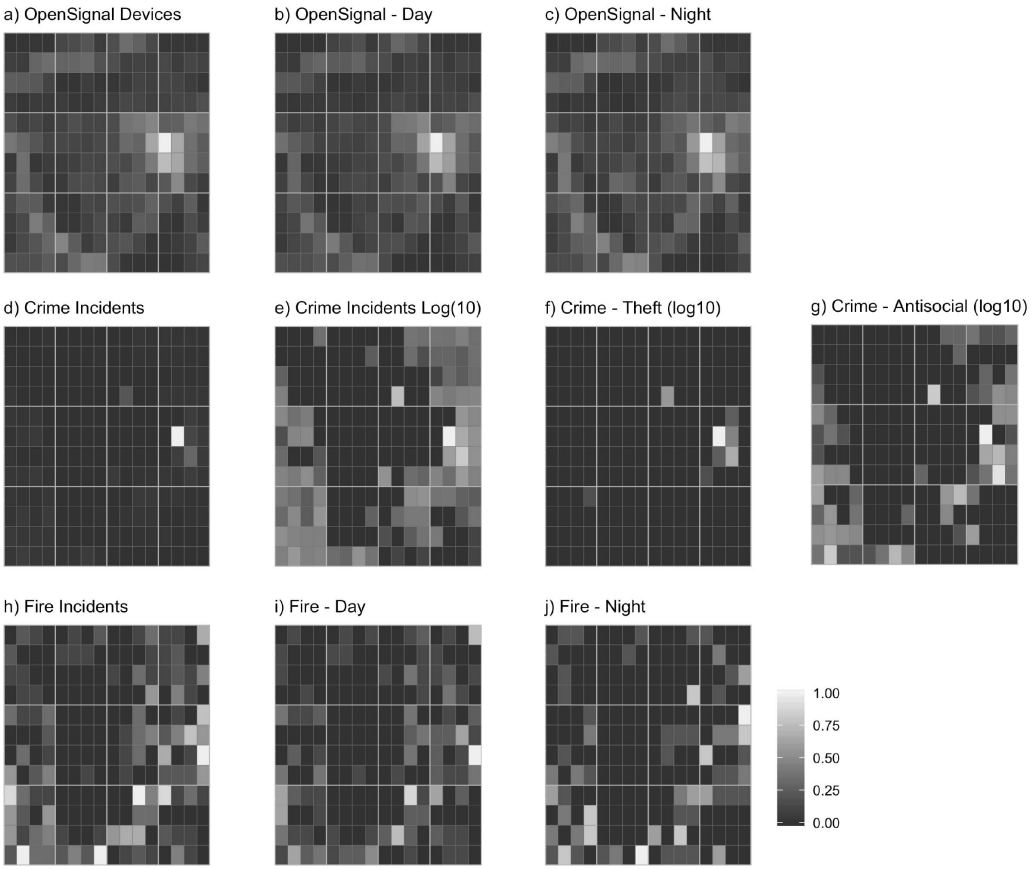


Figure 43. Compare population densities, Pixel grid of cells, across QEOP

Individual pixels are 1/16th LandScan cell size, 250 metres tall by approx. 150 metres wide. LandScan grid is outlined. Log transforms applied at Base 10 for crime data, where Log(10) is included in the label.

It is apparent that a cell resolution of 250 x 150 metres will be too small for producing multi-temporal population counts for all but a small number of cells within the grid. This concurs with the findings of the ENACT project (Batista e Silva, et al., 2016). However, it may be feasible to use the larger LandScan cell to produce a population measure, blending the temporal variation in the mobile data source with the spatial baseline provided by LandScan, and then use the smaller cell size to consider the distribution of the population across the LandScan cell.

The remainder of this chapter focuses on these two opportunities. First, to quantify the temporal variation in presence and produce an active population estimate, showing how much variation might be expected in any static population measure. Second, to identify the spatial distribution of presence across the area being counted, showing how concentrated or dispersed the population presently is within a target landscape.

5.2 Estimating the Active Population

The previous section of this chapter outlined the challenge to produce a representative population measure. This section examines whether or not a sample of movements captured individually using a mobile app can reveal the local population dynamics in a neighbourhood. The hypothesis is that a small and potentially biased sample of data can nonetheless be more representative of the target population as a whole than a static administrative or ambient count. However, a small sample of mobile data alone cannot provide a numeric count. This research proposes to use mobile data for its dynamic value, to produce a series of contextual weights for different recurring contexts. The weights can then be used to adjust a more robust but static baseline count and create an active population estimate. Whilst the mobile data source may contain some amount of bias regarding demographics, such as under-representing families with young children, detecting some variation in activity would be a potential improvement on using a static measure for the population targeted, at-risk or affected by a real-world dynamic urban phenomenon.

An edited version of the findings from this study was accepted and presented at the GISRUK 2018 conference (Richardson & Hudson-Smith, 2018).

5.2.1 Data and methods

5.2.1.1 Population sources

This study will blend a static baseline population measure with an active source that can measure temporal variations. The objective is to determine if it is feasible to use a sample of real-time data to produce a time-based weighting measure to reveal population dynamics through the day.

As explained in section 5.1, LandScan has been chosen as the static spatial baseline for this study due to being available as a grid-based aggregation of the ambient population and being updated on an annual basis to reflect land-use changes. Its role is to provide the mechanism for converting a device count from a sample of mobile data into a count of people. If an improved baseline measure becomes available for a target landscape, it could be substituted as a baseline without affecting the method to produce temporal weights using a mobile data set, provided the mobile data can be aggregated to the same areal scale as the baseline measure.

The active dataset for this study was provided by OpenSignal, introduced earlier in this chapter. The dataset contains all readings within the Greater London area during June 2017. Device IDs are anonymised and latitude/longitude coordinates where each reading was taken have been rounded to three decimal places, providing a spatial proximity radius of approximately 55 metres around the coordinates. The timestamp has been reduced to just the hour. It is unknown if the reading has been rounded to the nearest hour or whether just the minutes and seconds have been obfuscated from the timestamp. The latter is assumed. The dataset is prepared as described in chapter three. Each OpenSignal reading is tagged with the ID of the LandScan cell it falls within. A code sample is included in Appendix B.2.

During June 2017, 1,431 devices running the OpenSignal app recorded readings whilst located within the LandScan grid covering the QEOP. Of those devices, 754 visited the area just once and 667 visited on two or more days during the month (Figure 44). The mean number of devices present daily within the park was 151. If classifying the spatial familiarity of people based on their presence during June alone, more than half of the visitors were solo visits, likely attending for a special reason rather than as part of routine activities. Approximately one-third of the devices were present on more than three days and would be classified as 'habitual' visitors.

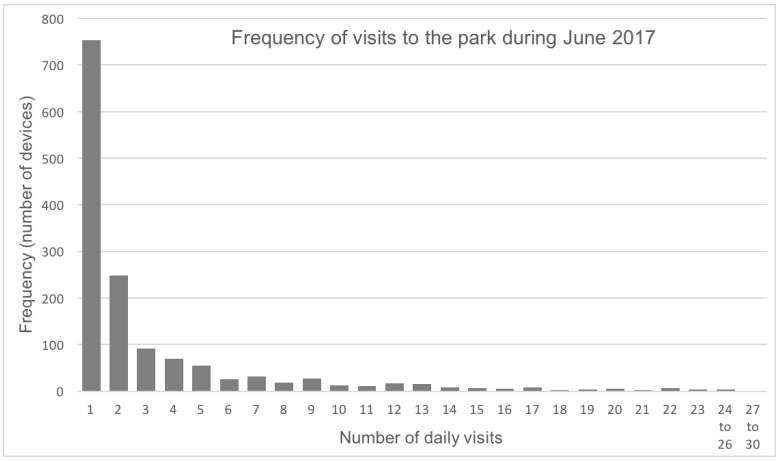


Figure 44. Frequency of daily visits to the QEOP LandScan grid in June 2017

Assumptions

Static population measures such as LandScan produce a count for an area independent of time. To produce a time weight for variations detected in mobile readings with timestamps, it is assumed that the LandScan count reflects the count of people present on average during an hour of the day, adopting the approach taken by Telcos when calculating Erlangs representing an hour of call activity. Whether or not the number represents a single set of people present for the entire hour, or a variable number of people present briefly during the hour that total up to this count is not specified in the LandScan documentation. Therefore, the assumption is that, over a typical hour, the count represents the total number of people who will have been present. Thus, the mobile data set can be aligned to the LandScan data by calculating an average hour count of devices present.

The mobile app is primarily used during the day, recording connections to wireless network access points and mobile communications cell towers. To provide an average hourly baseline count, 07:00 to 22:59 are included. This may produce conservative weights. There is little activity overnight, particularly from 2am to 5am. If including the full 24 hours, the hourly average for the day would be lower and the range between minimum and maximums would be larger.

Whilst a population count would typically be rounded to a whole number, decimal places will be kept for measuring variations in the samples. For low counts, the decimal place may reveal subtle differences under different conditions when averaging over multiple days. Rounding to the nearest integer would over-emphasise the effect. However, it is acknowledged that this may still result in small changes in presence being more influential than they should in calculations. To indicate this uncertainty, all results will be rounded to the nearest 500.

5.2.1.2 Scope

Spatial aggregation

The target area for this study is the Queen Elizabeth Olympic Park (QEOP) in Stratford, East London, and the immediate surrounding area containing mixed land uses. Details about the area have been described in chapter four. The QEOP and surrounding area provide a mix of residential and non-residential buildings and outdoor spaces, each of which will attract people at different periods, as well as inaccessible or transient areas such as the road/rail network, waterways and construction sites. It is assumed that people will not be uniformly or randomly distributed across the landscape but will be clustered at times and locations associated with land-use potential.

Aligning the boundary of the QEOP landscape with the LandScan grid referencing system produces a 3 x 4 grid of twelve cells spanning the park and surrounding area (Figure 45).

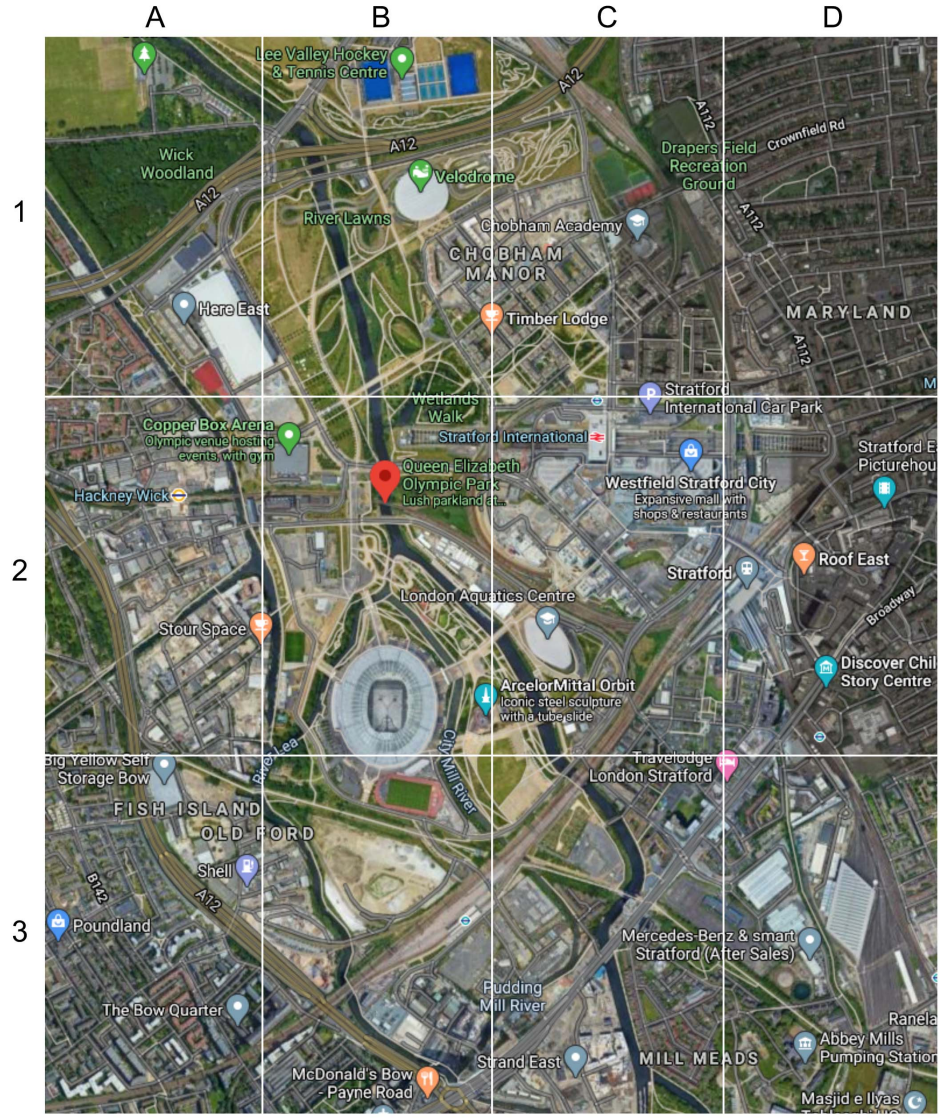


Figure 45. LandScan grid covering the QEOP with population estimates for 2015

Each cell is 1km tall by approximately 0.62 km wide. Imagery ©2020 Bluesky, Getmapping plc, Infoterra Ltd & Bluesky, Maxar Technologies, The GeoInformation Group. Map data © 2020 Google Maps.

Table 13 contains the LandScan ambient population estimates for 2015, the most recent estimates available when this study was completed. The percentages show the cell count as a proportion of the total count across the grid. The 2016 figures were reviewed upon release and contained only minor increases that would not alter the results significantly.

Table 13. LandScan ambient population estimates, 2015

A1: 9,967 (13.2%)	B1: 7,910 (10.5%)	C1: 593 (0.8%)	D1: 2,480 (3.3%)
A2: 9,844 (13.0%)	B2: 5,907 (7.8%)	C2: 6,818 (9.0%)	D2: 5,953 (7.9%)
A3: 11,485 (15.2%)	B3: 7,547 (10.0%)	C3: 3,588 (4.8%)	D3: 3,360 (4.5%)

LandScan cell counts and, in brackets, as a percentage of the total for the grid.

The size of the grid is 3.0km high (North to South) by approximately 2.5km wide (West to East). Each cell is 1km high by approximately 0.62km wide. The majority of the park green space is within cells B1, B2, and C2. The London Stadium and Copper Box Arena are both located in cell B2, although the edge of the stadium and southern part of the island it is located on also falls within cell B3. The total ambient population estimate for the grid is 75,452. The total ambient population estimate for the cells containing the majority of the park is 28,182. These cells also include a large retail centre, transport hub, road and rail infrastructure and surrounding residential and non-residential properties as well as construction sites and land targeted for redevelopment.

The cell counts highlight the potential weakness in the LandScan algorithm. The cell containing the Westfield retail centre (C2) is lower than the cell containing the VeloPark (B1). The retail centre is one of the largest in Europe and would be anticipated to have a high daily footfall. This suggests that, whilst the LandScan algorithm considers tourism and business travel within its redistribution of residential statistics, it is not reflecting other non-residential activities such as retail and leisure activities. This is a weakness with all workday or ambient population measures that concentrate on buildings used for business activities and do not incorporate ‘third space’ activities such as visiting parks, cafes, shops and leisure facilities to activities not related to either home or work.

In balance, a mobile data set may be skewed in the opposite direction, producing the majority of readings whilst mobile and connecting to public wireless and telecommunications networks, as opposed to being connected to private telecommunications networks whilst at home and work. It highlights the difficulties in producing a ‘street population’ measure in real-time.

A second concern is that one claim made by the LandScan providers is the use of satellite imagery to aid the annual updating of population counts to reflect changes in land-use (ORNL, n.d.). Cell C1 has been given a minimal count but, from a visit to the area in 2016 to ground-truth satellite imagery, the cell already contains populated high-density residential dwellings from 2015. Reviewing the 2016 LandScan counts indicated that this change had still yet to be detected. Whilst these issues raise concerns regarding the accuracy and usability of LandScan as an ambient population measure, there is no alternative available that could provide an improved spatial baseline. The nearest is the workday population from the 2011 census. However, it has similar issues regarding the representation of non-work activities and is substantially out of date given the

redevelopment of the area. It also would not register a residential population in cell C1 given the area was a brownfield site being redeveloped in preparation for the London 2012 Olympics. Cell C1 contained the athletes' accommodation, to be converted to residential accommodation as part of the post-Olympic legacy investment.

Contexts

The period being analysed is June 2017. June is an early summer month in the United Kingdom and can experience varied weather but the temperature is usually pleasant and above the level when urban outdoor activities are expected to increase – 13 degrees Celsius (Pushkarev & Zupan, 1975). Cultural assumptions are relevant to the UK and London. For example, parks are mostly frequented during daylight hours. In some areas of the world, parks have a more active night profile due to the outdoors climate being too hot to be outdoors during the day.

During June 2017, there were four music concerts hosted at the London Stadium in the park. Given the total ambient population estimate by LandScan for the park and the surrounding area in 2015 was 75,452, it is assumed that a far higher active population would be present on days when events take place in the stadium given the stadium's capacity of up to 70,000 for music concerts.

- Saturday 3 June 2017: Depeche Mode 'Global Spirit' tour
- Friday 16 and Saturday 17 June 2017: Guns N Roses 'Not in this lifetime' tour
- Friday 24 June 2017: Robbie Williams 'Heavy Entertainment Show' tour

Access to the London Stadium is controlled by bridges that restrict entry to the island the stadium is built on, and by gates to enter the stadium itself. For each concert, the bridges were opened at 3pm and the gates were opened at 5pm. The Depeche Mode and Robbie Williams concerts started at 7pm with the main acts due on stage at 8.15pm. The Guns-n-Roses concert started at 5:45pm with the main act due on stage at 7.45pm. For all the concerts, a curfew of 10:30pm was in place. Inspection of tweet content suggests it was adhered to (Figure 46)

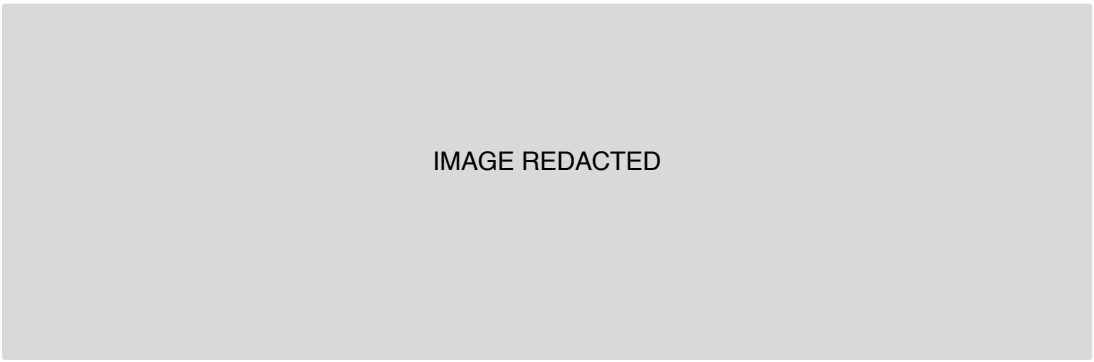


Figure 46. Tweets mentioning a concert curfew at the QEOP, 16 June 2017

The primary contexts of interest are a) dates and times when large-scale events take place in the London Stadium, and b) whether or not urban routines: peak versus off-peak periods are detectable and, if they are, how they vary between different zones within the landscape.

5.2.2 Results

Two studies are presented. First, a single LandScan cell is evaluated that is expected to show the most extreme contextual variation in population size – cell B2 containing the London Stadium. Second, the full grid of cells will be compared. The LandScan count is assumed to represent an ambient or 'normal' day, defined here as being a day when no special events take place that may attract more people to the area than would otherwise be present. An event day is considered to be 'abnormal' because it is anticipated to attract a substantial number of additional people to the area.

5.2.2.1 Quantifying LandScan cells

The first test is to quantify variations in OpenSignal presence within the LandScan cell B2 containing the London Stadium. Its use for large-scale events provides a benchmark for whether or not a small sample of mobile data is sensitive to contextual variations in presence. The cell currently contains no residential locations and a small number of workplaces but has event venues, open outdoor spaces, playgrounds, and seating areas. Multiple pathways and roadways connect the park with the surrounding area, with residential buildings, non-residential buildings and transport hubs on either side.

Daily readings

Figure 47 and Table 14 contain the daily mean average number of devices detected in cell B2 during June 2017, per day of the week. Counts are calculated for all days and separated for non-event ('normal', no event at the London Stadium) and event days ('abnormal'). The baseline is the mean daily count across weekdays.

Separating non-event and event days for comparison indicates the impact large-scale events have on park visits, substantially affecting the average count for days when events occur. The mean average for all days does not reflect either normal or abnormal conditions. On non-event days, Thursdays and Sundays are the only days of the week with an average device count higher than the baseline. This could be because the park is more popular for ad-hoc visits on these days. It also could potentially be a residual event effect, as was discovered with Wi-Fi readings in chapter four, with visits increasing on the days before and after the music concerts.

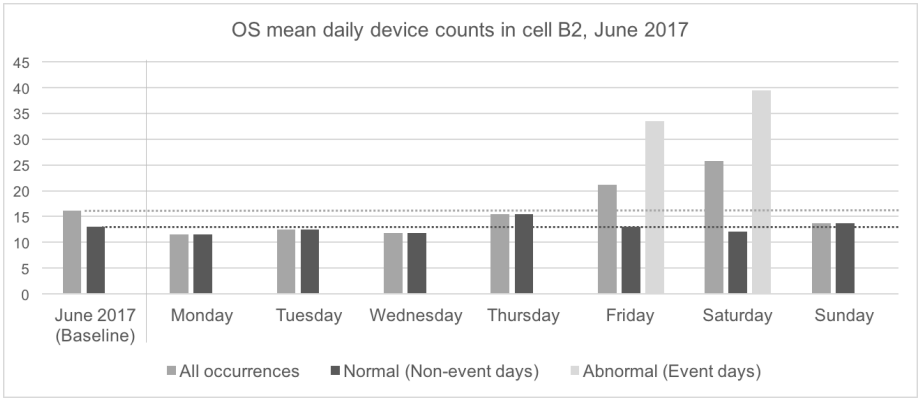


Figure 47. OpenSignal daily averages for LandScan cell B2, June 2017

Horizontal dotted lines for comparison of each day with the equivalent baseline.

Table 14. OpenSignal daily averages for LandScan cell B2, June 2017

June 2017	*All*	*Non-event*	*Event*	*All days (events)*
Day average (baseline)	15.97	12.84	36.40	30 (4)
Weekday	14.47	12.83	33.50	22 (2)
Weekend	19.75	12.87	39.50	8 (2)
Monday	11.50	11.50	---	4 (-)
Tuesday	12.50	12.50	---	4 (-)
Wednesday	11.75	11.75	---	4 (-)
Thursday	15.40	15.40	---	5 (-)
Friday	21.20	13.00	33.50	5 (2)
Saturday	25.75	12.00	39.50	4 (2)
Sunday	13.75	13.75	---	4 (-)

Baseline is the mean daily count across weekdays (Monday to Friday).

Hourly readings

To analyse how the active population changes through the day, four times at four-hour intervals are compared on non-event Fridays and Fridays when an event was held in the evening (Table 15). On the event day, bridges to the London Stadium were opened at 3pm and the event had a 10:30pm curfew. The counts are plotted in Figure 48 to visualise the variations.

Table 15. OpenSignal hour averages for LandScan cell B2, June 2017

Hour average	*All*	*Non-event*	*Event*
Baseline (07:00 to 22:59)	*1.84*	*1.18*	*6.13*
Friday hourly average	*3.00*	*1.10*	*5.84*
Friday 10:00 to 10:59	0.60	0.33	1.00
Friday 14:00 to 14:59	1.40	0.33	3.00
Friday 18:00 to 18:59	6.80	2.00	14.00
Friday 22:00 t0 22:59	5.20	0.67	12.00

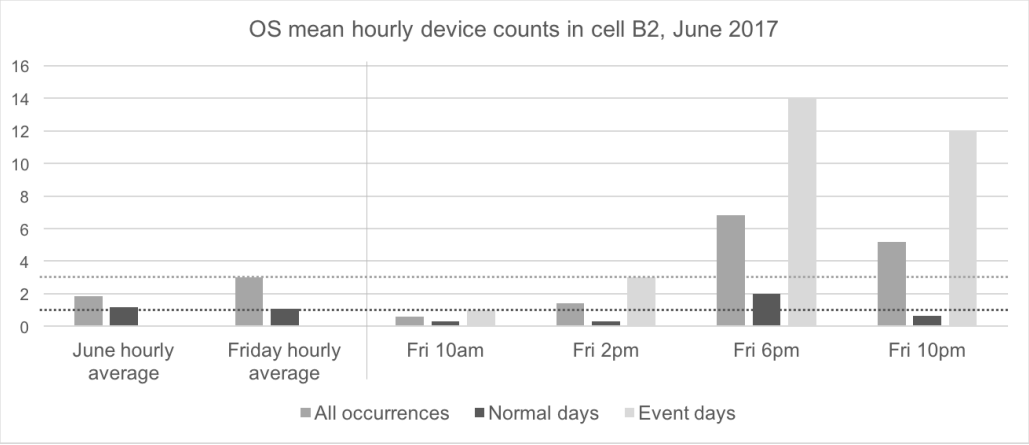


Figure 48. OpenSignal hour averages for LandScan cell B2, June 2017

Month hourly (June) and Friday hourly averages included for comparison with intervals through the day.

Whilst the device count is very low on normal days, there is variation between off-peak day, peak hours and off-peak evenings. As with the daily analysis, hourly averages are affected by large-scale events taking place in the park. The daily mean number of OS devices present in the park ranges from 11.50 to 15.40 on normal days in the park, and ranges from 33.50 to 39.50 on event days in the park, a difference of approximately 300%. The hourly mean number of OS devices shows sensitivity to temporal variation during the day. The baseline mean count for Friday is 1.10 on normal days and 5.84 on event days, an increase of over 500%. The variation is much larger on event days with a 700% difference at the start of the event compared with a non-event day.

5.2.2.2 Estimating active presence

To estimate the active population, the LandScan ambient estimate is multiplied by a weight representing the context being estimated, as expressed in equation [3].

$$P_i = A_i W_{ij} \quad [3]$$

where:

$P_i = active\ population\ estimate\ for\ LandScan\ cell\ i$
$A_i = ambient\ population\ estimate\ in\ LandScan\ cell\ i$
$W_{ij} = presence\ weight\ for\ context\ j\ in\ LandScan\ cell\ i$

To create weightings from the mobile data, an equivalent ambient average is calculated and set to 1. Each context is then divided by the ambient average. Table 16 contains the weightings, rounded to two decimal places, and the active estimate created by multiplying the LandScan 2015 ambient estimate (5,907) by the context weight, rounded to the nearest 500. The weightings are generated by dividing each context by the baseline. Actual values are shown in Table 14 and Table 15.

The fixed seating capacity for the London Stadium is 66,000. For a music concert, the floor area is also used, increasing the audience capacity up to 70,000. Combined with the ambient population count of 5,907 and event participants (non-audience), we would expect the active population during a music concert at the London Stadium to be in the region of 70,000 to 90,000. It is not known how much activity would vary at different times on ambient days such as peak versus off-peak hours.

Table 16. Context-weightings and population estimates for LandScan cell B2

	OpenSignal weight		*Active population estimate*	
Context	*Non-event day*	*Event day*	*Non-event day*	*Event day*
Day Baseline	**1.00**	2.83	5,900	16,700
Monday	0.89	---	5,500	---
Tuesday	0.97	---	6,000	---
Wednesday	0.91	---	5,500	---
Thursday	1.20	---	7,000	---
Friday	1.01	2.61	6,000	15,500
Saturday	0.94	3.08	5,500	18,000
Sunday	1.07	---	6,500	---
Hour Baseline	**1.00**	5.20	5,900	30,700

	OpenSignal weight		*Active population estimate*	
Context	*Non-event day*	*Event day*	*Non-event day*	*Event day*
Friday average	0.93	4.95	5,500	29,000
Friday 10am	0.28	0.85	1,500	5,000
Friday 2pm	0.28	2.54	1,500	15,000
Friday 6pm	1.69	11.86	10,000	70,000
Friday 10pm	0.57	10.12	3,500	60,000

Baseline is LandScan 2015 population estimate. All population estimates are rounded to nearest 500.
Weightings generated by dividing context by non-event day baseline.

Reviewing the variation in day estimates is ineffective for a comparison with the LandScan ambient population estimate. The LandScan measure estimates the number of people present on average, not the total number of people likely to visit over 24 hours, whilst counting the number of unique devices present provides a measure of the latter. An event may generate a substantially larger population. However, as discovered in chapter four, the increase is concentrated around the duration of the event, with a build-up occurring from three hours before.

The hourly estimates produce a more appropriate measure by showing how much the number of people present varies throughout the day, both for ambient conditions and due to event effects. The active population estimate based on OS device readings from 18:00 to 18:59 (6pm) is 70,000 on the event days, a more than ten-fold increase of the ambient measure. Applying the same method to ambient (non-event) days indicates that there is a substantial variation between peak and off-peak hours. Peak Friday commute hours are nearly double the ambient estimate whilst off-peak daytime hours are roughly half the ambient estimate.

The results are encouraging, and support findings from the analysis of the Wi-Fi data in chapter four. They suggest that mobile data can produce an active population estimate, approximated to hourly presence variations, and could be of particular value in open urban space where there is an absence of administrative sources and potential for substantial changes in population size and motivations due to different activities and interactions that can occur within the same location. Also, similar to the findings from chapter four, the results suggest that two scales of data analysis could be performed. Temporary access to real-time data over an interval incorporating multiple contexts may be sufficient to generate the hourly weights to approximate an active population distribution, with longer-term access to aggregated counts for seasonal adjustments.

5.2.2.3 Comparing LandScan cells

The second test is an exploration of the spatial distribution between LandScan cells. Cell B2 has a specific and rare profile due to the presence of a large stadium and absence of other buildings. If relying on residential statistics form the census, the population count for this area would be zero. Other cells within the LandScan grid have a more diverse profile. Some contain a large volume of residential and/or non-residential dwellings, as well as transport hubs.

For this test, two contexts are compared that are robust in variation within cell B2 – event versus non-event Fridays. Two periods are compared: mid-to-late afternoon incorporating peak commute time and event build-up; and, 7pm to 10pm representing off-peak evening on non-event days and the duration of the event on event days.

Figure 49 shows the average daily device counts per LandScan cell on Fridays in June 2017, comparing normal conditions (non-event days) with event days. The grid average is also provided for both contexts for comparison. All cells show a population increase on event days. Four cells are above average on both non-event and event days: A3, B3, C2 and D2. Cell B2 contains the event venue – the London stadium. C3 contains local hotels and one of the pedestrian routes from Stratford tube station to the London stadium. A3 and B3 contain south entry/exits to the stadium, local transport hubs and a mix of residential and non-residential buildings. A2 was anticipated to exhibit higher readings than detected due to the presence of Hackney Wick tube station and various entertainment venues surrounding it and providing the main western route into the park. However, the readings were below the grid average on both event and non-event days. Cell D3 registers the lowest population change. This is not a surprise. It is dominated by a large storage depot that is likely to be insensitive to contextual effects on human activities.

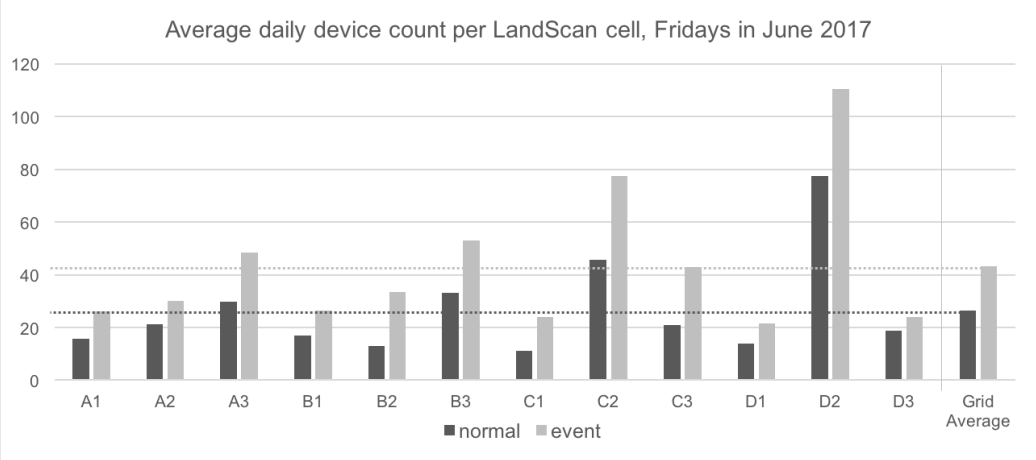


Figure 49. Average OS daily device count per LandScan cell, Fridays in June 2017

Horizontal dotted lines show mean counts across the grid for comparison with individual cells.

The challenge already discovered with daily averages is that they obscure temporal variations through the day – a change in population is unlikely to be distributed evenly throughout the day. Figure 50 shows the average hourly device counts on event days for two three-hour periods on event Fridays: 3pm to 6pm, when people will be arriving at the venue on event days; and 7pm to 10pm, when attendees will be present at the stadium.

The spatial distribution of device readings is affected by the concert timings. The hourly average from 3pm to 6pm shows that the majority of readings occur in cells C2 and D2 on both normal days and event days. This is expected due to the retail centre and major transport hubs being located in these cells. However, once the event is in progress, cell D2 is much closer to a normal reading whilst B2 has a very abnormal increase, as would be expected due to event attendees being

present at the stadium. B3 also registers a significant increase, possibly because the stadium slightly overlaps that cell. A2, A3, C2 and C3 also remain high possibly due to people leaving the concert early or non-attendees being routed around the park.

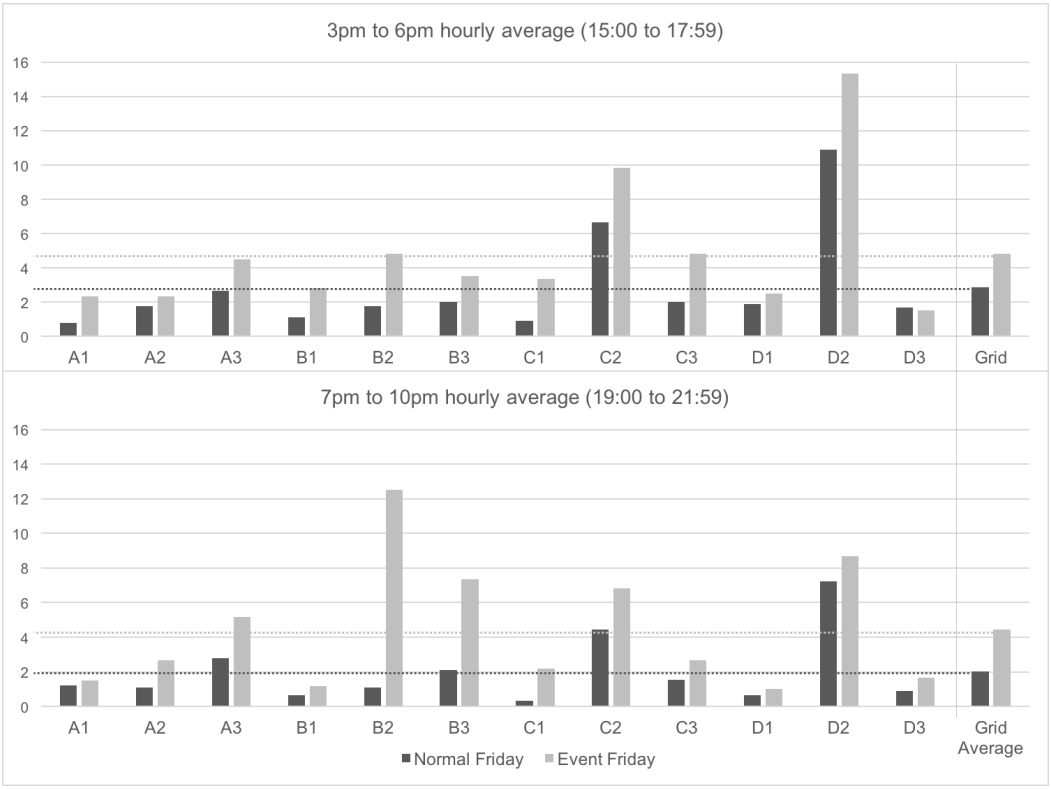


Figure 50. Average OS hourly device count per LandScan cell, Fridays in June 2017

In the previous chapter, it was discovered that certain events generate residual time effects: increases in population size on non-event days immediately before or after the event day. This study shows that it is possible to learn how other zones will be impacted by events both during and at different times, such as the transport hubs used to arrive at and leave from the park. This introduces a spatial residual attribute – that when there is an event scheduled at the park, the build-up and dispersal of presence will be leading and lagging at locations related to the event venue.

To conclude, the mobile data demonstrates temporal variation that indicates it is possible to produce an active population estimate. Based on the assumption that the LandScan count represents the number of people present on average during any hour, we can quantify how much variation occurs at different hours throughout the day. The next step is to consider what the count represents: is it a fixed number of people present for the entire hour, or the cumulative total of people present for brief intervals during the hour? This is the focus of the next study.

5.3 Learning Spatial Behaviours

Producing an active population estimate by combining mobile data traces with an administrative baseline creates the potential to improve knowledge about variations in the number of people present within a landscape for a given context. However, it offers no information about the spatial distribution within the boundary defined for the administrative measure. Are the readings clustered in a small area of the landscape or dispersed across it? Does the spatial distribution vary at different times of the day? Are people dwelling for long periods of time or present only briefly as they traverse the landscape? Data containing individual readings, each with a device ID, coordinates and timestamp can be used to learn spatial dynamics not represented in administrative data.

5.3.1 Data and methods

A data set containing individual device readings generated whilst present within a landscape creates the potential to identify different trips taking place within the landscape, and potentially to analyse the stages of a trip by segmenting the landscape into zones. This enables the study of aggregated population movement behaviours rather than individual trajectories.

5.3.1.1 Assigning trip attributes

Two trip-related classifications can be assigned to each reading within a mobile data set: trip ID and stage ID. The method for assigning trip attributes was described in chapter three and code samples are provided in appendix B.2. To summarise, if two readings for the same device in chronological order are more than three hours apart then the first reading is considered the end of one trip and the second reading is the beginning of a new trip. Once each reading has been assigned a trip ID, it is possible to count the number of trips within the landscape and calculate an approximate duration for each trip. There is uncertainty in that the device may have been present for some time before the first reading is generated and remain present for some time after the last reading is emitted for each trip. However, the time elapsed between the first and last reading sharing the same trip ID can provide some indication of whether the holder of the device dwells within the landscape whilst present or travels across it. It can also be used to exclude devices that cannot be analysed for spatial behaviours because their duration is too brief or consists of a single reading, or because the duration is too long suggesting the device is a static presence within the landscape. The second level of granularity for studying spatial behaviour comes from dividing the trip into stages. This requires deciding how to segment the landscape into different zones. Stages can then be calculated whenever a device moves from one zone to another.

There are three approaches to defining active spaces: grid-based zones, with each zone created as a uniform cell within the grid; geographically-defined zones, where a boundary is manually drawn around behaviour settings based on features expected to attract people; and, data-driven clustering, where a boundary is created by drawing a convex hull around the outer positions of people present in a cluster within the landscape.

A uniform grid is the simplest to implement. However, the size of cells may be influential, as demonstrated at the start of this chapter. Even a LandScan cell just 1km all by 0.62km wide can mask local clustering within it (see Figure 43). Ensuring the cell size is small enough to reveal local clustering may result in a large volume of zones, many of which may contain no or few readings. Clusters may be concentrated within a single cell or split across multiple cells, with difficulty identifying whether or not it is because the cluster falls on a boundary, splitting the count between cells, or whether it is spatially spread across the cells.

Geographically-defined zones involve drawing polygons to enclose zones. Availability of open-source GIS tools makes such an approach easy to implement through the creation of shapefiles. Boundaries enclosing zones may be drawn along geographical features, such as road and waterways, or arbitrarily decided. The zones may be contiguous, as used in thematic maps (choropleths), requiring all data points to be allocated to a zone even outliers at the very edge of a zone. Alternatively, geographically-defined zones may be nodal, with a perimeter drawn around a centroid that marks the centre of the zone and all data falling outside the nodal boundary being excluded as unclassified (see section 3.2.1 for a comparison of the two methods).

There are two immediate limitations to manually-defined zones. First, they require knowledge about the landscape to decide on a method to create spatial boundaries. For example, census output areas are drawn to produce a consistent a number of people within a range (ONS, 2016). A neighbourhood study may zone to focus on different attractions. Second, and related, manually drawing boundaries means that the outcomes from analysis can be substantially influenced simply by redrawing polygons to modify the shape and size of areal units. This is referred to as the modifiable areal unit problem (MAUP). It is a long-recognised issue within quantitative geography (Openshaw & Taylor, The modifiable areal unit problem, 1981) and considered endemic in areal census data (Openshaw, 1984).

Data-driven clustering is an alternative approach that uses reality data to determine nodal zones, defining boundaries for zones based on real-world observations. A clustering algorithm is used to identify clusters within the spatial distribution of readings. A convex hull can then be drawn around all readings within a cluster to create a polygon representing the nodal zone. There are various cluster detection algorithms. The DBSCAN (density-based spatial clustering of applications with noise) algorithm has been chosen for this research because it provides two benefits: it does not require the number of clusters to be specified in advance and it will ignore outliers. First introduced in 1996 (Ester, Kriegel, Sander, & Xu, 1996), the algorithm was designed specifically for clustering spatial data sources. It requires no assumptions about the distribution of data points. Instead, the assumption is that dense clusters will be separated from one another by sparser areas (Ester, 2014). The DBSCAN algorithm has two configurable parameters to determine the sensitivity of the algorithm to detecting clusters. Eps (‘eps’) accepts a numerical value to specify the maximum distance allowed between two points within a cluster. MinPts (‘minPts’) accepts a numerical value to specify the minimum number of points required to form a cluster. For this dataset, the eps value

defines the spatial distance between readings, recorded in decimal as latitude and longitude. An eps of 0.001 represents a spatial distance of approximately 110 metres.

This study explores the potential to perform data-driven clustering to identify active spaces as nodal zones within the landscape based on the readings emitted by mobile devices, and then use those zones to classify visit behaviour by calculating the duration of time a device spends continuously within a zone. Code samples are included in Appendix B.5

5.3.1.2 Spatial uncertainty

In the previous study in this chapter, an academic data set for June 2017 was provided by OpenSignal, with spatial coordinates rounded to three decimal places and timestamps reduced to hourly. This makes the calculation of trip durations and movement patterns difficult. To explore the potential for a sample of mobile readings to reveal spatial behaviours within the landscape, OpenSignal provided a raw data set containing readings with the original timestamp and spatial coordinates. The readings were limited to the QEOP landscape and provided for May 2017. As with the June dataset, device IDs were anonymised. Figure 51 demonstrates the difference in spatial accuracy. Each plot is a single day of data. In the first image, coordinates are plotted rounded to three decimal places. In the second, coordinates have been 'jittered' to 6 decimal places by randomly adding +/- 0.000499 to each reading. In the third, coordinates are provided for 6 decimal places and more clearly reveals spatial routes and cluster spots within the landscape.

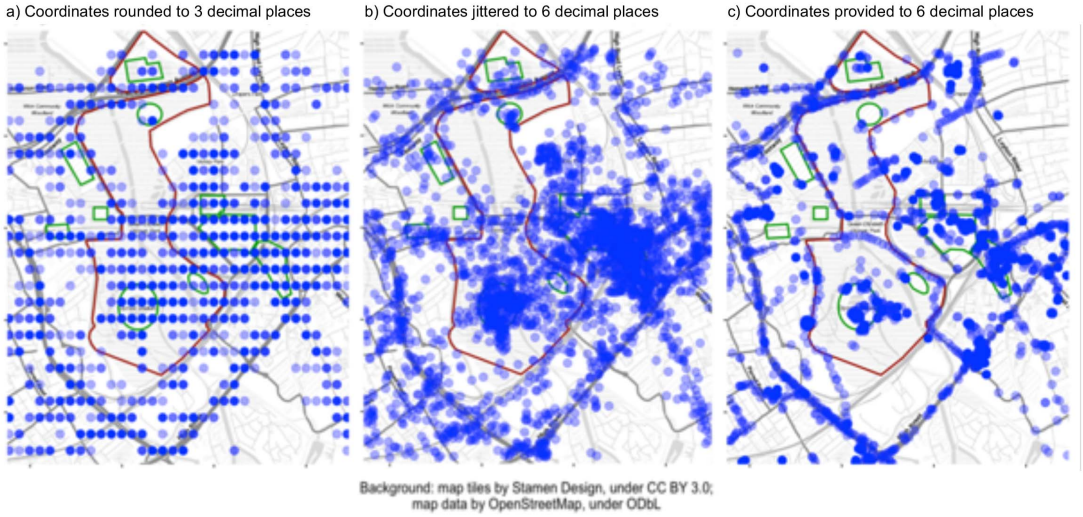


Figure 51. Comparing the spatial accuracy of data coordinates

From left to right, images a) and b) containing readings for 16 June 2017. Image c) contains readings for 14 May 2017. The park and key venues are outlined.

A second aspect of having a raw dataset is to consider how devices move across the landscape. For data generated by embedded sources, such Wi-Fi access points, it may be possible to provide a triangulated estimation of presence based on consecutive readings. However, plotting just three devices producing OpenSignal readings shows how varied human movement can be in open space (Figure 52). All three devices emitted more than 20 readings within the landscape on 14 May 2017. Figure 52a shows individual readings. The device represented by blue points appears to be visiting

the London Stadium. A West Ham football match took place that day with a 2.15pm kick-off. The device represented by orange points is recorded at various locations around the edges of the park but not at the stadium. It suggests a visit to the area but not attending the football match. The device represented by green points has left a trail of data along the west edge of the park, from north to south, suggesting they travelled along the edge of the park, possibly as an alternative to travelling along the major roadway nearby.

Figure 52b shows the individual readings connected by lines in time-sequence order. It confirms that the blue device did visit the stadium and shows that the device entered from one location (Stratford tube station) and then headed northeast after the football match. The orange device also entered and exited the landscape at different locations – Stratford high street at the south-eastern corner and Hackney Wick in the northwest. The green device appears to have had two visits across the landscape. The straightness of one line suggests that motorised transport may have been used in one direction versus a non-motorised means of travel along the pathway at the edge of the park for the opposite direction. The plots highlight that attempting to triangulate a device's position may be difficult in open spaces where movements are unconstrained.

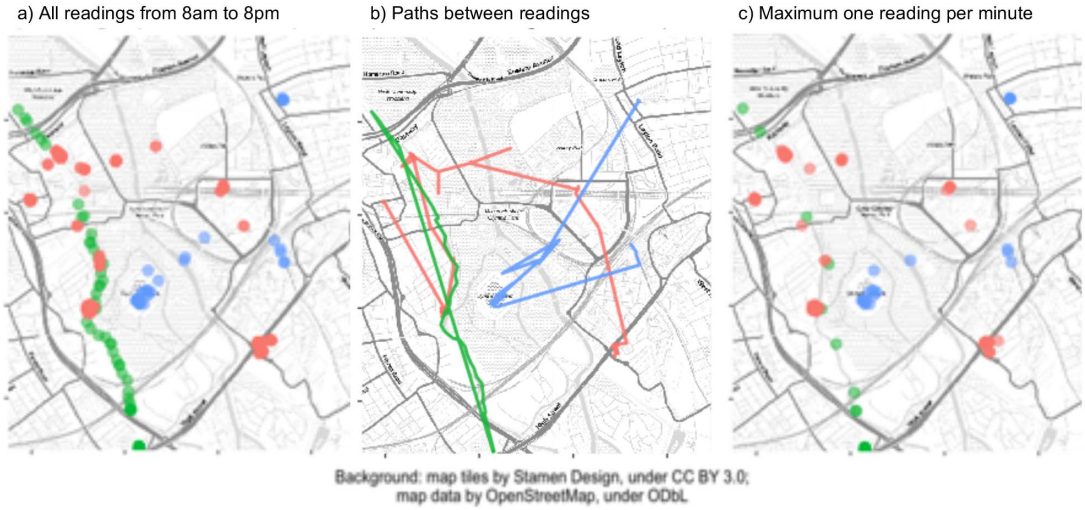


Figure 52. Plot of individual device routes across the park on 14 May 2017

Three OpenSignal devices emitting readings within the park landscape from 8am to 8pm on Sunday 14 May 2017. Westham football match kicked off at 14:15.

Figure 52a shows raw readings, some of which are emitted seconds apart. To perform spatial clustering based on the distance between points, the dataset is simplified to a maximum of one reading per minute per device (Figure 52c). Most of the locations for clusters of readings remain visible. The device represented by green dots appears to be moving quickly through the park with data points spread apart along the western edge of the park. In one minute, a person of average walking speed could travel 100 metres, the approximate length of a football pitch. The size of the park means that it would take several minutes to traverse on foot and 100 metres is smaller than the size of the smallest areal aggregation scale being used in this research.

5.3.2 Results

5.3.2.1 Analysing trip behaviours

The first study examines trips to the landscape based on trip duration. Durations are calculated from the time elapsed between the first and last reading in each trip. Figure 53 shows the frequency distribution of trip durations for the OpenSignal May 2017 dataset. The vast majority of trips last less than 5 minutes. To focus on people choosing to be present within the landscape as opposed to briefly travelling around it or being a semi-permanent fixture, only trips lasting from 5 minutes to 6 hours are included. This reduces the number of devices being analysed to 522. For this study, 6 hours was arbitrarily chosen as the cut-off point between a long visit or being considered to be permanently present within the landscape. An alternative approach would be to incorporate fuzzy logic to allow for overlap between categories.

The trips are categorised based on activities assumed to generate different durations (Table 17). 'Brief' is for trips lasting from 5 to 20 minutes, too short to be certain of dwelling within the landscape. It could be someone travelling through the landscape but delayed due to heavy traffic or traffic control measures. A 'Dwell' trip lasts from 20 minutes to 90 minutes. Whilst it could include some people travelling across the landscape, it is more likely that the trip involves some aspect of the landscape itself. It is still a short enough period that the visit is unlikely to be for a scheduled event. Dwell activities could include lunch breaks and after-school visits to playgrounds, or simply choosing to make use of the outdoor space for a rest from routine activities. The two 'Visit' categories are for durations that suggest a specific purpose for being present, such as attending an event or visiting tourist attractions. Visits are sub-divided into two categories: 'visit_half' for half-day visits, lasting from 90 minutes to 4 hours, and 'visit_long' for visits lasting 4 to 6 hours. Single events such as music concerts or football matches typically last from 2 to 4 hours. Events that include multiple activities, such as athletics championships or charity fun days can last all day.

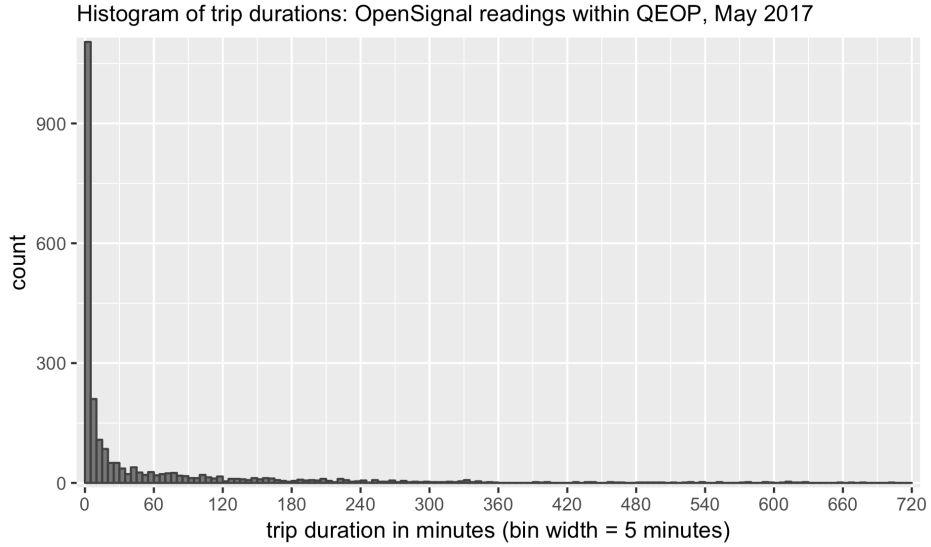


Figure 53. Histogram of trip durations in the QEOP, May 2017

The count is the number of trips. Trip duration capped at 720 minutes. 22 trips exceed 6 hours.

Table 17. Trip categories and statistics, OpenSignal data, May 2017, QEOP boundary

Category	*Duration*	*Trips*	*Devices*	*Trips/Devices*
Included				
Brief	>5 to 20 mins	403	202	1.995
Dwell	>20 to 90 mins	395	179	2.207
Visit_Half	>90 mins to 4 hrs	257	108	2.380
Visit_Long	>4 hrs to 6 hrs	75	33	2.273
Excluded				
Travelling	5 mins or less	1,104	380	2.902
All_day	More than 6 hrs	58	22	2.636

Brief assumes milling behaviour, such as waiting for a bus or train, or travel that is temporarily delayed. Dwell assumes ad-hoc trip (e.g. lunch). Visit assumes specific reason for presence (e.g. event attendance).

The duration periods, as with studying hourly counts in early studies presented here, introduces a second potential aggregation issue: a modifiable time (or temporal) unit problem (MTUP). Just as with the modifiable area unit problem (MAUP) mentioned earlier, it is possible that choosing different segmentation and aggregation boundaries for time intervals could produce different outcomes. Whilst recognition of the MAUP dates back to the 1930s (Gehike & Biehl, 1934), there is little mention of the MTUP prior to 2011, when it was posited that the MTUP could influence outcomes regarding the occurrence, frequency and duration of events (Cöltekin, et al., 2011). At the same time, a study of linear trends in seasonal vegetation time series also demonstrated the MTUP and the potential for models and outcomes to be altered by changing the definition of temporal units (de Jong & de Bruin, 2012). A 2015 study connects the MTUP with MAUP, identifying that aggregation occurring at both spatial and temporal scales requires consideration of both, referred to jointly as the modifiable spatiotemporal unit problem (Martin, Cockings, & Leung, 2015).

Whilst the research presented here is exploratory, consideration to both the MAUP and MTUP must be considered when interpreting outcomes. Figure 54 contains a plot of the readings for trips with durations assumed to represent intentional visits to dwell at features within the landscape – trips lasting from 20 minutes to 6 hours. Each individual reading is colour-coded for the category of trip the reading has been assigned to and only readings for trips categorised as ‘dwell’, ‘visit_half’, or ‘visit_long’ have been plotted. Such trips are assumed to involve a visit to some location within the park and are unlikely to be either travelling across the landscape without stopping or a fixed all-day presence. Note the park outline has been extended to include the immediate surrounding area including the Westfield retail centre and Stratford stations to the east of the park, where the majority of readings occur. The challenge with classifying movements based on the trip as a whole is that a visit to a specific location within the landscape will likely involve travel-like behaviours across the landscape until reaching the location to dwell at. The spatial plot of trips shows some clustering by category but also a mix of all categories along travel roads (roads and pathways) and at transport hubs. Classifying the individual stages of a trip would enable focus on locations within the landscape that attract different dwell durations. It first requires segmenting the landscape into different zones to identify active spaces.

Park visits, OpenSignal data in QEOP, May 2017

Background: map tiles by Stamen Design, under CC BY 3.0;
map data by OpenStreetMap, under ODbL

Figure 54. Spatial plot active visits to the QEOP landscape, OpenSignal, May 2017

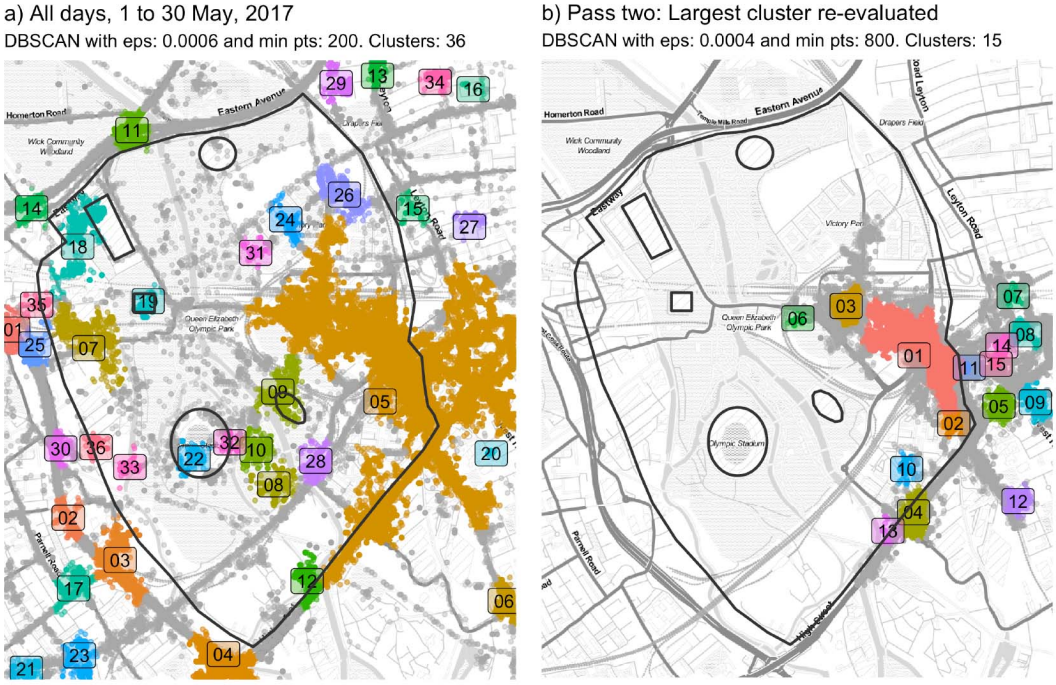
Data has been snipped to within the park boundary that includes immediate surrounding areas to include Westfield. Individual readings are colour-coded based on the duration of the trip that the reading is part of.

5.3.2.2 Detecting active spaces

As mentioned earlier, there are three approaches to defining active spaces: as uniform cells in a grid; as manually drawn polygons; or, as data-driven clusters. This study explores the potential for real-time data to enable the latter approach, using DBSCAN. Experimentation with the algorithm found that an eps distance of 0.0004 to 0.0008 generated clusters of varying sizes within the landscape depending on the minimum points required. This confirmed earlier research (Richardson, 2015) which established a measure of 0.0008 was the most suitable to study outdoor spaces using a March 2015 dataset provided by OpenSignal that spanned Greater London.

Figure 55a shows the results of clustering with a single set of parameters and introduces a limitation with using DBSCAN across a varied social landscape. It is unrealistic to use single eps and minPts

values for both open spaces with sparse activities and built areas that attract much higher footfall such as shopping centres and transport hubs. DBSCAN parameters that produce useful clusters within the park also result in a single mega cluster to the east of the park (cluster 05 in Figure 55a). Reducing the EPS to prevent the mega cluster results in no clusters forming at all within the park. Researchers have experimented with extending the DBSCAN algorithm to avoid the need for a single global set of parameters, such as first detecting variation in cluster densities (Ashour & Sunoallah, 2011). One option is to extract excessively large clusters and then re-run the algorithm on just data falling within that cluster, with an adjusted eps and minPts values (Figure 55b).



Background: map tiles by Stamen Design, under CC BY 3.0;
map data by OpenStreetMap, under ODbL

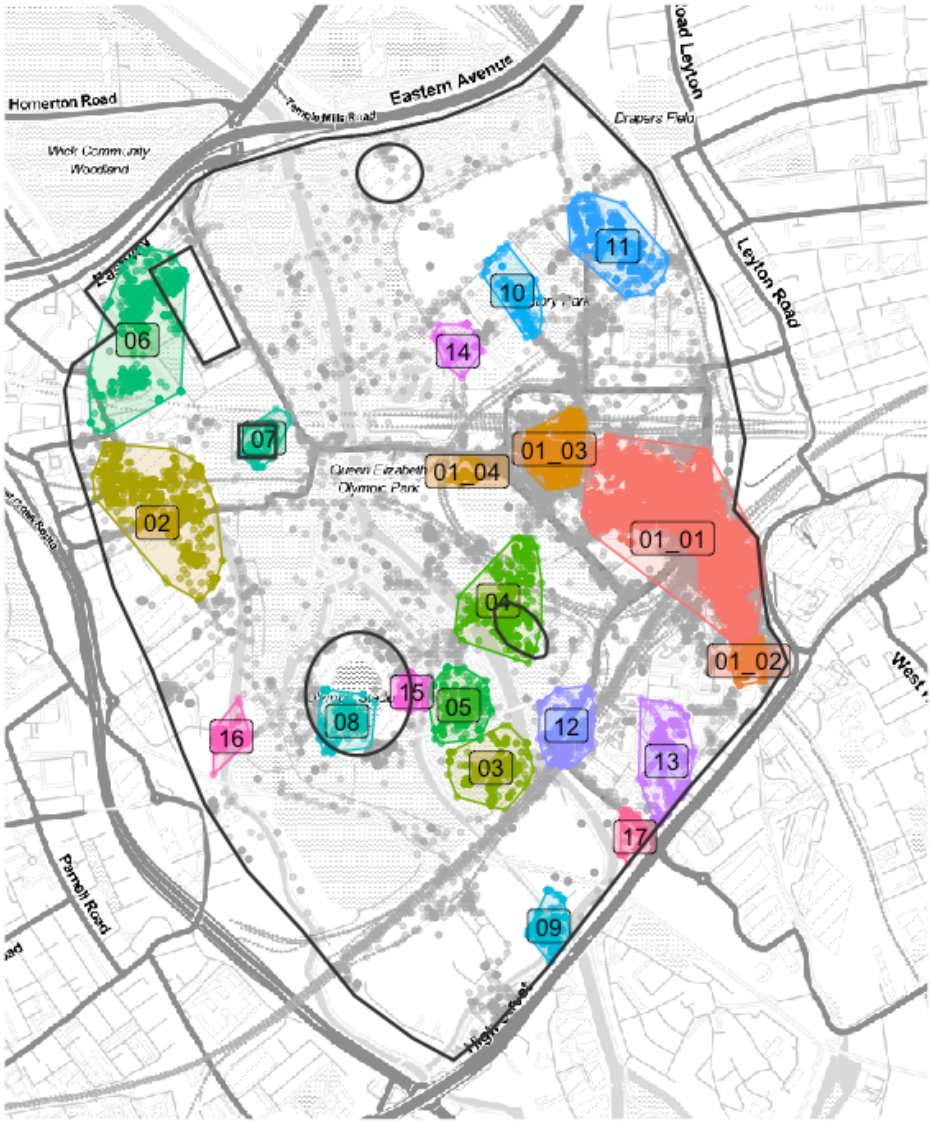
Figure 55. Detecting clusters of spatial activity using DBSCAN

Two-tiered DBSCAN clustering. Park and venue outlines displayed for orientation. Park outline includes immediate surrounding area. Un-clustered data points plotted in pale grey. Clustered data points colour-coded by cluster.

Running a second analysis on just the mega cluster divides it into multiple smaller clusters, one spanning Westfield and Stratford station (01 in Figure 55b), one located at Stratford International (03) and one at the bus station directly south of Stratford station (02). Other clusters are formed at the Stratford Retail Centre that is located opposite Stratford tube station, and south of Westfield, along Stratford High Street (04, 10 and 18).

Figure 56 presents the combined results from performing a two-tier DBSCAN cluster analysis. Data has been snipped to readings falling within a boundary that contains the park and immediate surrounding area to focus on clusters relating to park activities and neighbouring transport hubs or attractions. Each zone has been created automatically by drawing a convex hull connecting the outermost points within each cluster. The convex hull produces a polygon that defines the zone boundary. This is a crude approach to creating zones but one that does not require human assumptions about where to draw boundary lines.

After filtering to data within the boundary and tagging readings with trips and stages using the methods described in chapter three, the dataset was reduced to 124,981 records from 577 devices undertaking 2,319 trips comprising a total of 5,319 stages. Less than 10 per cent – 50 devices and 226 trips - do not visit any of the clusters. However, several zones are visited by only a few devices infrequently. Zones and device visits are detailed in Table 18.



Background: map tiles by Stamen Design, under CC BY 3.0;
map data by OpenStreetMap, under ODbL

Figure 56. Two-tier DBSCAN clustering of spatial points data

Two-digit clusters are from the first pass with parameters: eps = 0.0006 and min pts = 200. Clusters ending in _NN are from the second pass (performed on data points within cluster 01 from the first pass) with adjusted parameters: eps = 0.0004 and min pts = 800. Clustered readings colour-coded by cluster with a convex hull drawn to surround the cluster. Non-clustered readings plotted in grey.

Table 18. Data-driven zones created from data readings within QEOP, May 2017

Zone	*Label*	*Location notes*	*Devices*	*Visits*	*Ratio*
01_01	Westfield	Spans Westfield and Stratford tube station	423	1124	2.66
01_02	Stford_bus	Stratford bus station	73	153	2.10

Zone	*Label*	*Location notes*	*Devices*	*Visits*	*Ratio*
01_03	Stford_Intl	Stratford International railway station	120	201	1.68
01_04	Waterden	Edge of Stratford Waterfront and Waterden Rd	11	25	2.27
02	HkneyWick	Hackney Wick tube station and leisure venues	80	121	1.51
03	SouthLawn	South lawn in south park	47	79	1.68
04	Aquatics	The Aquatics centre in the east park	70	107	1.53
05	Podium	Podium, ArcelorMittal Orbit and café	26	31	1.19
06	HereEast	HereEast buildings and venue	23	60	2.61
07	CopperBox	Copper Box arena in the centre of the park	10	17	1.70
08	Stadium	London Stadium in south park	19	20	1.05
09	DanesYard	Danes Yard Tower (scenic point)	6	7	1.17
10	PelatonAv	Corner of Pelaton Avenue and Victory Parade	12	37	3.08
11	FortWalk	Top of Victory Parade, near Fortunes Walk	32	80	2.50
12	Montfichet	Bridge access at south end of park	64	100	1.56
13	StfordHotels	Hotel zone along Stratford High Street	18	56	3.11
14	OlyParkAv	Corner of Olympic Park Av and Victory Parade	10	14	1.40
15	StadiumEnt	Main entrance to the London Stadium	20	22	1.10
16	FishIsland	Includes location of WestHam supporters club	4	5	1.25
17	StfordHigh	Corner of Stratford High St and Warton Road	27	31	1.15

Devices = number of unique devices visiting cluster during May 2017 (a device may visit more than one zone); Visits = number of visits by devices during May 2017. Notes provides a rough location description for map reference. Ratio = number of visits/number of unique devices to calculate average number of visits per device.

Figure 57 shows the formation of clusters per date within the QEOP boundary outline. Each row shows the number of unique devices that are located within each zone per date. The two football dates are highlighted. Summary statistics across the month are also calculated per cluster. The total number of dates when a cluster forms is calculated by counting only dates when the cluster contains at least three unique devices. Three is chosen to identify locations that appear to attract multiple people on a given date. The darker the red shading, the more devices are present in the cluster. Grey cells indicate dates when no devices are present for a cluster.

Reviewing the clusters, there appear to be two classes of zone: stable and sporadic cluster formation (Figure 58). Stable zones are visited by at least 30 unique devices throughout the month, with clusters containing at least three devices forming on at least half of the dates. Sporadic zones are those that form infrequently, on fewer than half of the dates. They are visited by at least 30 devices across the month or have clusters containing at least five devices on at least one date. Five is an arbitrary number. It is the smallest value for the zone containing the stadium and would mean a single date contained at least one-sixth of all devices present during the month, suggesting an abnormal activity occurred on that date in that location. All other zones are unclassified due to containing fewer than 30 unique devices across the month and having a maximum count on any one date of fewer than five devices. They may exhibit either stable or sporadic behaviours but are underpowered in this data set. For example, zone 16 is located on Fish Island and does not have

a single reading of at least three unique devices. However, it has its highest count on a football match date. Analysis of Twitter activity on that date indicates football supporters visit pubs and restaurants located in the surrounding area (Figure 59).

	Westfield	Stford_Bus	Stford_Intl	Waterden	HkneyWick	SouthLawn	Aquatics	Podium	HereEast	CopperBox	Stadium	DanesYard	PelatonAv	FortWalk	Montfichet	SfordHotels	OlyParkAv	StadiumEnt	FishIsland	StfordHigh	
date/cluster	**01_01**	**01_02**	**01_03**	**01_04**	**02**	**03**	**04**	**05**	**06**	**07**	**08**	**09**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**wkday**
01/05/2017	21	1	5	1	3	0	1	0	0	1	0	0	0	2	0	1	0	0	0	0	Mon
02/05/2017	28	1	7	1	3	3	3	0	1	0	0	0	0	1	2	2	1	1	1	3	Tue
03/05/2017	40	6	5	0	3	2	1	0	2	0	0	0	0	3	3	2	0	0	0	1	Wed
04/05/2017	35	3	11	0	2	12	2	0	1	0	0	0	0	3	3	2	0	1	0	1	Thu
05/05/2017	40	4	5	0	3	3	9	2	2	0	10	0	1	4	1	3	0	4	0	0	Fri
06/05/2017	29	7	6	0	4	1	1	1	0	0	0	0	1	2	3	3	1	0	1	0	Sat
07/05/2017	21	3	7	1	3	0	5	1	0	1	1	1	1	2	3	3	1	0	0	3	Sun
08/05/2017	29	7	6	1	1	4	3	1	2	0	0	0	0	3	2	3	0	0	0	0	Mon
09/05/2017	29	6	6	1	3	2	3	0	3	0	0	0	1	2	4	2	0	0	0	1	Tue
10/05/2017	47	6	8	2	7	5	7	1	5	0	1	1	2	5	5	4	1	1	0	0	Wed
11/05/2017	52	8	8	2	5	2	3	1	4	2	0	0	1	3	3	3	1	1	0	1	Thu
12/05/2017	66	6	9	1	5	3	4	2	3	0	0	1	1	3	7	4	0	2	0	1	Fri
13/05/2017	32	4	10	1	1	0	2	2	2	0	0	0	2	5	0	1	2	0	0	1	Sat
14/05/2017	26	1	7	2	3	0	3	2	2	0	5	0	1	2	1	3	1	5	2	1	Sun
15/05/2017	43	8	8	2	3	0	4	0	1	1	0	1	1	3	3	2	0	0	0	0	Mon
16/05/2017	48	5	8	2	3	5	2	0	1	1	0	0	2	3	5	2	1	0	0	1	Tue
17/05/2017	42	8	7	1	5	1	1	1	3	0	1	0	1	4	1	2	0	1	0	0	Wed
18/05/2017	45	7	6	1	2	0	1	0	1	1	0	1	1	1	6	2	0	0	0	3	Thu
19/05/2017	51	6	8	1	3	5	3	2	1	0	0	0	1	2	5	2	0	1	0	0	Fri
20/05/2017	36	5	12	1	6	2	4	2	0	2	0	0	3	5	0	1	1	1	0	1	Sat
21/05/2017	25	3	4	1	2	1	1	0	2	1	0	0	2	2	2	1	1	0	0	1	Sun
22/05/2017	53	7	9	1	5	2	4	2	3	3	0	0	2	1	2	3	0	1	0	1	Mon
23/05/2017	41	5	8	1	3	5	7	2	2	1	0	2	1	2	8	1	0	2	0	3	Tue
24/05/2017	62	7	11	1	11	6	10	3	1	1	1	0	3	3	11	0	1	1	0	2	Wed
25/05/2017	53	9	5	0	8	2	7	2	4	0	0	0	1	2	7	0	0	0	0	2	Thu
26/05/2017	48	6	4	0	8	5	8	2	3	0	1	0	2	2	7	2	0	0	0	0	Fri
27/05/2017	28	8	7	0	4	4	4	1	3	0	0	0	1	2	5	0	0	0	0	1	Sat
28/05/2017	19	4	1	0	5	1	1	0	3	1	0	0	2	2	0	0	1	0	0	2	Sun
29/05/2017	12	0	1	0	5	0	1	0	4	1	0	0	2	4	0	1	1	0	1	1	Mon
30/05/2017	23	2	2	0	2	3	2	1	1	0	0	0	1	2	1	1	0	0	0	0	Tue
Min	12	0	1	0	1	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	
Mean	37.5	5.1	6.7	0.8	4.0	2.6	3.6	1.0	2.0	0.6	0.7	0.2	1.2	2.7	3.3	1.9	0.5	0.7	0.2	1.0	
Median	38.0	6.0	7.0	1.0	3.0	2.0	3.0	1.0	2.0	0.0	0.0	0.0	1.0	2.0	3.0	2.0	0.0	0.0	0.0	1.0	
Max	66	9	12	2	11	12	10	3	5	3	10	2	3	5	11	4	2	5	2	3	
St.Dev	13.4	2.5	2.7	0.7	2.2	2.6	2.6	0.9	1.3	0.8	2.0	0.5	0.8	1.1	2.8	1.1	0.6	1.2	0.5	1.0	
Num. clusters	30	25	27	0	24	13	18	1	11	1	2	0	2	14	17	9	0	2	0	4	
Classification	stable	stable	stable		stable	sporadic	stable		sporadic		sporadic			sporadic	stable			sporadic			
Wkday mean	41.3	5.4	6.7	0.9	4.2	3.2	3.9	1.0	2.2	0.5	0.6	0.3	1.1	2.6	3.9	2.0	0.3	0.7	0.1	1.0	
Wkend mean	27.0	4.4	6.8	0.8	3.5	1.1	2.6	1.1	1.5	0.6	0.8	0.1	1.6	2.8	1.8	1.5	1.0	0.8	0.4	1.3	

Note: Clusters only counted when contain at least 3 devices; Two stadium event dates are highlighted: 5th and 14th May

Figure 57. Formation of clusters per date, QEOP behaviour setting, May 2017

Count shows the number of unique devices present per cluster per date. Shading indicates a count from 0 to 30+. Bold outlines indicate large event dates (football matches at the London Stadium).

Stable	Sporadic	Unclassified
01_01 Westfield	03 SouthLawn	01_04 Waterden
01_02 Stford_bus	06 HereEast	05 Podium
01_03 Stford_intl	08 Stadium	07 CopperBox
02 HkneyWick	11 FortWalk	09 DanesYard
04 Aquatics	15 StadiumEnt	10 PelatonAv
12 Montfichet		13 SfordHotels
		14 OlyParkAv
		16 FishIsland
		17 StfordHigh

Figure 58. Classifying the formation of data-driven clusters, May 2017 readings

Unsurprisingly, the largest and most stable cluster is the Westfield cluster, indicating the high footfall that a large retail centre will attract. However, five other clusters meet the criteria to be considered stable whilst five clusters meet the criteria to be considered sporadic. The remainder are unclassified, primarily due to being underpowered.

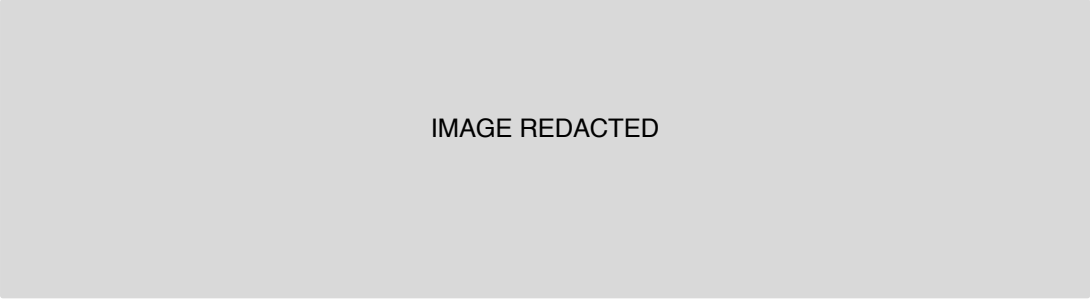


Figure 59. Tweets describing pre-/post-match activities in the QEOP, August 2016

Source: Twitter.com.

5.3.2.3 Plotting stage behaviours

Earlier in the chapter (see Figure 54), active trips were plotted spatially. With zones detected, it becomes possible to plot the readings for active trips based on stage behaviours. For example, whilst the trip overall may be classified 'visit_half', the trip will comprise of stages, some of which may be classified as 'travelling', some as 'dwell' etc. By plotting based on stage behaviours, it becomes possible to see the locations where dwelling activities take place and areas dominated by movement (Figure 60).

Plotting the durations of stages for each cluster showed a wide variety in times (Figure 61). Whilst the mean duration for Westfield (cluster 01_01) and Stratford Int'l (01_03) is low, there are a large number of outliers. It suggests that the cluster is dominated by people travelling through the transport hubs but shows the effect of a retail centre being co-located above a transport hub. Cluster 01_02 contains just a transport hub – Stratford Bus station – and exhibits few outliers.

The Copper Box arena cluster (07_00) is interesting in exhibiting a much larger interquartile range than any other cluster. The cluster was considered too weak to be classified. However, the box plot of durations suggests that the cluster does have a sensitivity to events taking place at the venue. Events at the venue were not considered as an attribute for detecting changes in population presence because changes were not visible in the preliminary analysis that took place using the Park Wi-Fi data logs for March 2016.

Identifying clusters of presence has revealed active spaces within the landscape. However, the results are weak. Only one cluster – Westfield – forms with sufficient size to have any confidence in the findings. The method has potential but requires a more comprehensive data set to produce results with statistical rigour. Its value here is in assigning stages to trips and enabling readings to be visualised as different population behaviours across the landscape, as exhibited in Figure 60. The green dots represent readings for presence within a zone lasting at least 90 minutes. The dark blue dots represent presence lasting 20 to 90 minutes. The pink dots highlight rapid movements. It enables the conversion of a points-based plot into a visualisation that indicates different behaviours occurring at different locations within the park.

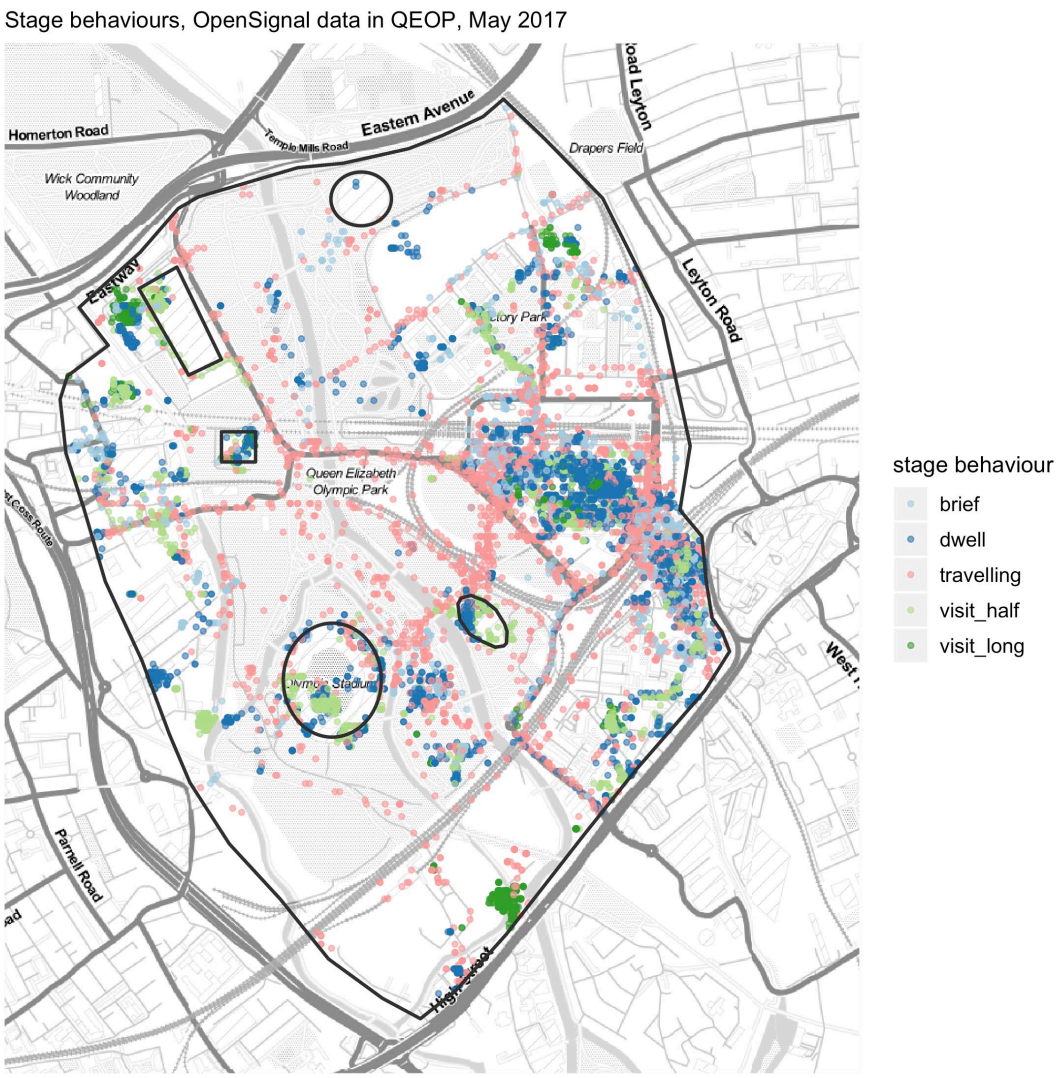


Figure 60. Spatial plot of stage durations during park visits, May 2017

Only trips lasting at least 20 minutes are plotted. Individual readings are colour-coded based on the duration of the stage they fall within. Same classifications are used for stage durations as for trip durations.

Figure 61. Dwell-times of stages at data-driven clusters, QEOP, May 2017

Selection of detected zones displayed including Westfield (01_01) and CopperBox (07_00).

5.4 Research Outcomes II

5.4.1 Summary findings

The two studies presented in this chapter provided insights that can be incorporated into the contextual framework developed in chapter four. First, incorporating administrative data as the spatial baseline and utilising mobile data as contextual weights enable the model to indicate the actual population estimated to be present. The contextual variations supported the findings from chapter four, including seeing residual effects of events on non-event weekend days after large events taking place on Friday or Saturday. Whilst such an estimate would lack precision, it is an improvement on administrative measures or ambient measures that redistribute administrative sources based on theoretical assumptions. Second, using a data source that is not bounded to the range of embedded sensors and contains coordinates for individual devices enables relationships to be explored between behaviour settings across a landscape.

Segmenting the landscape using a grid of uniform cells enabled the mobile data to be blended with an administrative data source. However, the spatial scale, even with cells at the relatively small size of 1km by 0.62km, masked local clustering. Sub-dividing the cells into pixels revealed local clustering but also generated a large volume of segments with minimal activity. Given the focus is on areas that attract people, an alternative approach was explored – discovering active spaces based on clustering within the dataset. The results were mixed. Clusters were detected and converted into zones using convex hulls. Treating each cluster as an active space made it possible to model different behaviours across the landscape. However, the sample size became too small when analysing relationships between zones at granular time scales. Further research is needed, preferably with a more comprehensive dataset. Further consideration for the modifiable unit problems, both areal and temporal, is also needed to progress this technique.

Whilst the use of data-driven clustering to identify active spaces was limited in its ability to analyse and compare durations of presence at different locations across the landscape, being able to assign each data point with a stage ID did produce an interesting visual outcome. Figure 60 reveals areas of the landscape where people are moving or dwelling, with dwelling categorised for the duration of presence within the area. It reveals many dwell spots surrounding the park. For example, on the west side of the London Stadium and a line of presence north of the Westfield retail centre. Visiting the areas confirmed the presence of food and drink establishments at both locations. Referring to the concept of over-hypotheses discussed in chapter two, a potential application of the technique could be in identifying areas of the landscape for conducting living lab research and/or the deployment of human observers to produce behaviour maps in more detail. This could include directing appropriate questions for qualitative research such as interviews and questionnaires to gain a more in-depth understanding of people-place experiences. This potential is discussed further in chapter eight under potential applications.

5.4.2 Contextual framework update

Given the findings, the contextual formula is revised (Figure 62). It now incorporates administrative data as a known attribute to provide a static spatial baseline. This enables a device count from mobile data to be converted into an active population estimate. The static baseline can also include identifying active spaces and their behaviour profiles. This can include spatial familiarity, explored in chapter four, and trip behaviours such as dwelling versus moving, explored in this chapter. It can also include relationships between active spaces. For example, when an event is scheduled at the London Stadium, it will affect zones containing transport hubs to the east of the stadium.

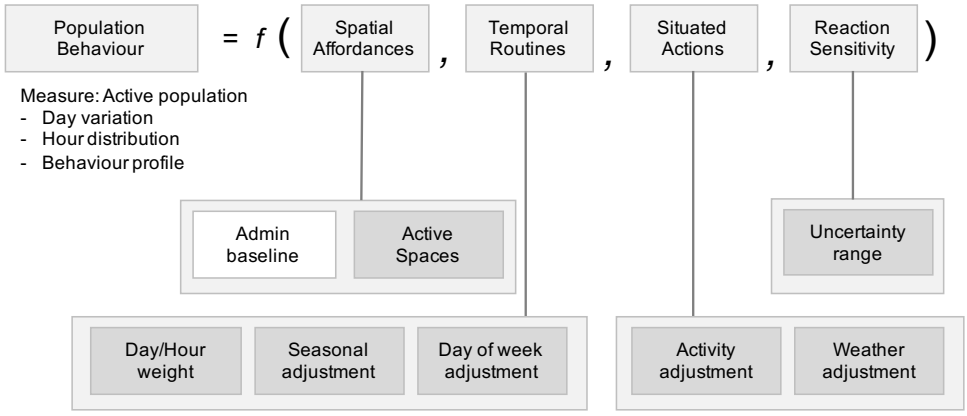


Figure 62. Using P-STAR to forecast active population behaviours

As with the version in chapter four, the temporal distribution and active conditions convert a static spatial baseline into a dynamic distribution to provide an active population estimate for a given time and set of circumstances, whilst an uncertainty range should ideally be included to indicate the sensitivity of the model. Interactions that are considered necessary will be less sensitive to unpleasant conditions than those considered optional.

An alternative to producing individual weights is to instead adopt a more simplistic multiplier to indicate the approximate impact a set of circumstances will have on the population behaviours expected for the landscape and active spaces within it. This would be akin to the Saffir-Simpson Hurricane Wind Scale (NOAA, n.d.) and the Enhanced Fujita Scale (NOAA, n.d.) used to indicate the amount of damage extreme weather events can create. A similar approach could be adopted to adjust an ambient population estimate, such as ranging from -5 to +5 where 0 is the ambient context: normal population behaviours are expected. On a declining scale from -1 to -5, the population weighting would be reduced in increments such as 0.9, 0.75, 0.5, 0.25 and 0.1 for circumstances that will repel people from the area. On an increasing scale from +1 to +5, the population weighting would be increased in increments such as 1.5, 2, 3, 5, 10 for circumstances that will attract people to the area. The scale -5 would represent a catastrophic incident that would result in no or little human activity. The scale +5 would represent an unprecedented mass gathering. More data and analyses are needed to produce such a scale and it is a recommendation for future research, discussed in chapter eight.

PAGE INTENTIONALLY LEFT BLANK

6 Case 3: A Social Landscape

Chapter five built on the findings from chapter four by focusing on mobile readings as a signal of presence within a landscape, to learn the active population and its spatial distribution. This chapter builds on the findings of chapter four by focusing on the semantics of mobile data. It considers whether or not the content of social media expressed in real-time can reveal information about subjective experiences whilst present. Just as it was hypothesised that the size and distribution of the population present within a landscape would vary for different contexts, so it is hypothesised that subjective experiences whilst present within a landscape will vary for different contexts

A single data source is used throughout this chapter: Twitter. The analysis of multiple location-based data sources within the Queen Elizabeth Olympic Park (QEOP) presented in chapter four found that the volume of geotagged tweets was too sparse to inform a presence estimate or make comparisons between recurring contexts such as different times of day or days of the week. However, Twitter did appear to be sensitive to abnormal situations occurring within the park, including smaller events that had not been anticipated to have a noticeable effect (see Figure 27 in chapter four). Thus, for this chapter, the time interval for comparisons between different sets of circumstances is date-based. The hypothesis is that a social media source such as Twitter will reveal a substantial variation in language used in the same physical location on dates when abnormal situations occur compared with normal conditions.

The chapter is organised into three parts. The first part introduces the foundations for analysing subjective experiences of a location and the basics for performing a text analysis that are applied throughout the studies presented here. The second part of the chapter presents the results from analysing situations and people-place experiences based on the words contained within location-based social media – tweets. It focuses on the differences in sets of terms used for different contexts. The third part explores a more advanced technique to consider whether there is a distinct vocabulary for the landscape and different circumstances. The chapter concludes with a summary of the research outcomes and its contribution to the P-STAR contextual framework.

6.1 Spatial and Social Cognition in Text

As described in chapter four, the London Legacy Development Corporation (LLDC), who is responsible for the redevelopment of the Queen Elizabeth Olympic Park (QEOP), are keen to learn how people are making use of and experiencing the park and its facilities. For subjective experiences, the common approach is to conduct surveys and questionnaires. An alternative is to analyse social media posts generated by mobile devices whilst present within the park. A review in chapter two of studies analysing location-based social media suggested that such sources can produce insights into human experiences of place. However, a concern raised in chapter four is that social media, when studied at the scale of a local neighbourhood, produces sparse readings daily for routine circumstances and a large volume of readings for infrequent situations such when music concerts are held at the London Stadium. Thus, an analysis that generalises for the landscape may not be representative of its everyday reality. This research mimics the approach taken in the previous chapter, positing that the value in samples of mobile data is in revealing differences in behaviour within the same landscape occurring at different times. Whereas the previous chapter focused on signals as an indicator of presence and actions, this chapter focuses on semantics as an indicator of conditions and experiences.

A single data source is used for this case study: Twitter. As described earlier in the thesis, Twitter is an online social network that enables people to post short text messages limited to 140 characters at the time of this study. Twitter provides an application programming interface (API) for retrieving what is believed to be a 1% sample of all tweets posted publicly for up to approximately one week after the date a tweet is posted. Continued access to all public tweets is available through the web site www.twitter.com. As learned in chapter four, the volume of tweets available for analysis posted daily from within the Queen Elizabeth Olympic Park (QEOP) – is very small except when large events take place. Figure 63 shows the sample retrieved daily during June 2016.

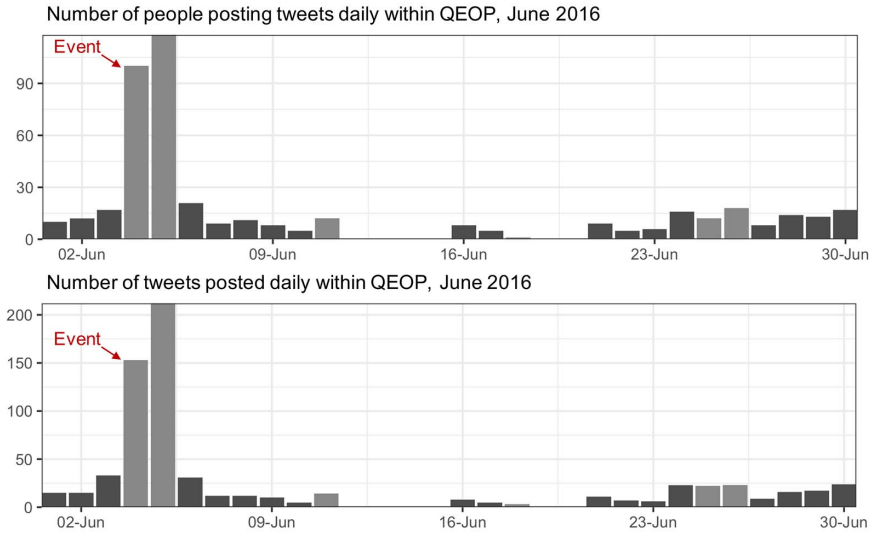


Figure 63. Count of tweets posted daily within QEOP, June 2016

Top image shows number of different Twitter accounts posting. Bottom image shows number of tweets posted. Weekends highlighted. Count is the sample of public tweets made available for download. Event: AC/DC concert.

The tweets used are those containing geotags positioning them within the Wi-Fi boundary of the park (see Figure 20 in chapter four). One large event took place during June 2016, an AC/DC concert at the London Stadium on Saturday 4th June. The question for this research is whether or not such a small sample is sufficient to reveal people-place experiences and if/how they vary on different dates. There is one of three possible outcomes: First, that language does not vary on different days and can be generalised for the landscape; Second, that language does vary but randomly or the sample is insufficient to make any inferences; Third, that language varies on different dates and the variations relate to different circumstances occurring within the same physical landscape, indicating that language is contextual rather than general for the landscape.

6.1.1 Hierarchical perception

The first consideration is whether or not a small sample of short text messages will be sufficient to infer experiences of the population present within a physical space on a given date. An argument in favour is that perception is hierarchical and, thus, people will focus on the most prominent elements of a landscape. This is based on a theory posited by cognitive scientist and economist Herbert Simon (Simon, 1996). When asked to draw a landscape, Simon argued that most people will start with a framing outline for the landscape, then include the most prominent features within, then add the next level of detail and so on. Whilst not a scientific study, an art project appears to demonstrate this finding by asking people to quickly draw a range of objects (Martino, Strobelt, Cornec, & Phibbs, 2017) and analysing the order in which different elements of the object are drawn. When asked to draw a face, most people start with a circle as an outline of the face, then draw the left eye, followed by the right eye, then the mouth, then the nose and so on (Figure 64).

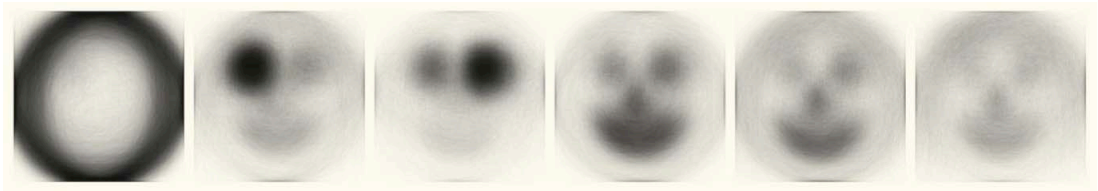


Figure 64. Time sequence for drawing a human face, reproduced from Martino, 2019

Time sequence on how 161,515 people draw a human face. From MIT presentation posted online (Martino, 2019). First viewed during the presentation 'AI and Data Visualisation' delivered at UCL, 28 November 2017.

If hierarchical perception applies also to language, it suggests the most prominent circumstances are likely to be the first to be expressed in location-based tweets, regardless of how many people post messages, suggesting the most common terms will be established with a small sample.

This theory is tested by examining the use of language in tweets posted within the QEOP. Figure 65 shows the frequency counts of terms used in tweets posted on four consecutive days, from Friday 3rd to Monday 6th June 2016. The AC/DC music concert took place on Saturday 4th June. The list excludes any terms listed as English 'stop words' using the Quanteda text analytics package in R (see Appendix A.2 for the full list of stop words) or any single characters such as 'I'. On all four dates, the distribution exhibits the curve of a power law when ordering terms used within the tweets posted on each date by their frequency, regardless of the number of tweets posted. The

only difference is that, when more tweets are posted, the curve becomes more pronounced and the tail becomes longer. Unsurprisingly, some terms are common to all dates, such as ‘london’ and ‘olympic’. The term ‘acdc’ only appears on the date of the AC/DC concert and the dates immediately following it. It is not mentioned on the day before. On Friday 3rd June, the terms ‘copper’ and ‘box’ are within the top three terms used. This indicates that an event or incident may have occurred involving the Copper Box arena on that day.

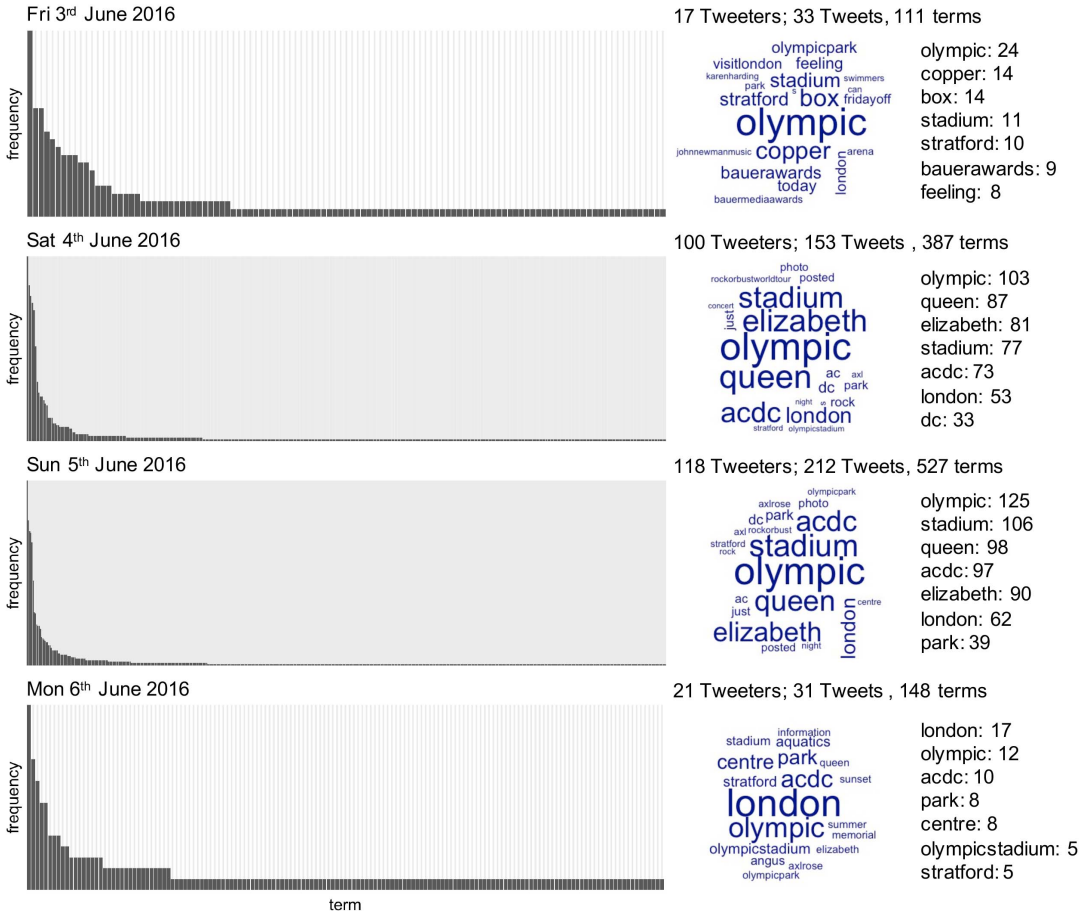


Figure 65. Term frequencies for four consecutive days in June 2016

Bar chart shows frequency of terms used in tweets on each date. Top frequencies listed on the right. Word clouds show terms sized relative to their frequency. Summary statistics across all four dates included underneath charts.

This preliminary review of term frequency suggests that it may be possible to extract information about people-place experiences from a small sample of location-based content and that the language will include contextual variations on different dates.

6.1.2 Inferring affect

The second consideration is to what extent human experiences during situations can be inferred from the language expressed when posting messages to an online social network. The LLDC, overseeing the redevelopment and management of the QEOP, expressed an interest in exploring whether or not location-based tweets could reveal how people feel about the park, its facilities and the different events taking place.

‘Affect’ refers to feelings, emotions and moods that influence behaviour. The terms are often used interchangeably. However, scientifically, emotions are considered to be intense and momentary feelings that have a salient cause whereas moods are considered to be milder feelings that endure longer and whose cause may be uncertain (Izard, 2010). In a 2005 book on happiness (Nettle, 2005), behavioural researcher Daniel explained how different emotions could generate different physical responses (Table 19). Being able to detect emotion in real-world observations could provide an indicator for anticipated behaviour or behaviour change when different circumstances occur within the same physical location.

Table 19. Physical responses to emotion felt, adapted from Nettle, 2005

Emotion	*Schema (situation type to evoke emotion)*	*Spatial Remedy*
Joy	Pleasurable experience	Carry on ‘as is’
Fear	Ongoing source of danger	Flee if possible, else hide
Anger	Violation of a norm or agreement by others	Fight, complain, disrupt
Sadness	Loss of valued support	Tread carefully until conditions improve
Disgust	Potential contamination	Move away and/or avoid
Anxiety	Anticipating or perceiving a threat	Be on alert

Adapted from Table 1, page 31 (Nettle, 2005). Note: the original table focused on the four negative emotions. Joy and anxiety have been added for completeness.

Furthermore, a recent and emerging goal for many countries and cities is to improve the health and happiness of citizens. In 2011, the United Nations (UN) adopted a General Assembly resolution to promote sustainable happiness (Oxford Poverty & Human Development Initiative, n.d.) and in 2013, the OECD published guidelines for measuring subjective well-being (OECD, 2013). Indexes have emerged to quantify happiness (PwC and Demos, 2016) and unhappiness (Glaeser, Gottlieb, & Ziv, 2016) of cities using self-reported data. As introduced in chapter two, studies have also emerged in the past decade that make use of social media as an alternative source to analyse spatial sentiment. For example, a 2013 study found associations between happiness in tweets and human mobility patterns. (Frank, Mitchell, Dodds, & Danforth, 2013).

There is, however, a need for caution in drawing inferences about human emotions expressions, whether visual through the face and body movements or verbally through language. There is extensive academic research on human emotions but, surprisingly, a lack of consensus about how to classify emotions (Izard, 2010). There are two broad approaches: discrete sets of categories or measures based on two or more dimensions. Table 19 is an example of discrete definitions, and are the most common in use currently, the basic universal emotions proposed by psychologist Paul Ekman (LeDoux, 1998). The dimensional approach most often uses dimensions of valence (positive to negative feelings) and arousal (calm to agitated state) to measure affect (Barrett, 2017).

Bearing these concerns in mind, the first study presented here expands on a preliminary finding from chapter three: that social media can provide salient information about social and environmental conditions that may affect people-place experiences.

6.2 Sensing People-Place Experiences from Text

This study examines whether or not a sample of location-based tweets can provide insights into people-place experiences and how they vary due to different circumstances occurring within the same physical space at different times.

It should be noted that the focus of this research is not on Twitter specifically but on the potential to extract information about people-place experiences from short location-based text messages. Twitter was an accessible source at the time of this research. Given the brevity of such messages, assuming the principles of hierarchical perception, the most influential circumstances and feelings expressed should rapidly emerge from even a small sample. If found to be the case, the method could be applied to any text-based service that encourages people to share updates whilst present within a physical location.

6.2.1 Data and methods

All analyses presented in this case study involve data collected from Twitter from March 2016 to June 2017 inclusive, when the message limit was 140 characters. Appendix B.1 contains code samples for retrieving and preparing a set of tweets for analysis

6.2.1.1 Text analytics

The studies presented here use text analytics to analyse the content of tweets. Computational analysis of a text is referred to as natural language processing (NLP, not to be confused with 'neural linguistic programming', an unrelated phrase). NLP applies principles from the field of linguistics to programmatically extract latent information from the text (Bird, Klein, & Loper, 2009). Methods include counting word frequencies, modelling topics from words within texts, summarising documents, producing chatbots, performing translations into different languages, and undertaking discourse analysis to study complete sentences and paragraphs.

This research is not seeking to advance the field of text analytics but rather to apply established techniques to location-based texts. As presented in chapter two (section 2.2.3), several studies have been published in the decade leading up to this research that have used location-based social media to infer human experiences of a place. However, the focus has been on producing generalisations about a place or analysing individual special events. This research is focused on whether or not the experience of place changes depending on circumstances. The short nature of tweets means that messages are often abbreviated and contain few terms. Thus, this research focuses on the use of terms rather than complete sentences or more in-depth discourse analysis. Table 20 contains a list of terms common to the field of text analysis and used throughout the remainder of this chapter. Words in brackets represent terms that are used interchangeably.

Table 20. Terminology used in text analysis

Term	*Description*
Corpus	A collection of documents
Document (Text)	A semantic unit – spoken or written – consisting of words
Word (Token)	A sequential string of characters that, grouped as a single unit, has meaning
Term	A word (or token) that is used one or more times in a document
Vocabulary	The set of all terms (unique words) found within a document or corpus
Dictionary	A defined set of terms
Tokenisation	The process of converting a document into a series of words (tokens)
Stop words (Noise)	A list of terms to be removed from a tokenised document before analysis
Lemmatisation	Reducing multiple forms of a word to a single dictionary term (its lemma). E.g. 'swim', 'swimming', 'swam', 'swims' all become 'swim'
Stemming	Ignoring word endings, reducing terms to their word stems. E.g. 'swim', 'swimming', swims' become 'swim' but 'swam' would still be 'swam'
Term frequency (TF)	The number of occurrences of a term within a document or corpus
Document frequency matrix (DFM)	A 2-dimensional array identifying each term within a corpus and its frequency for each document within the corpus

Terms in brackets are alternative labels used interchangeably or abbreviations.

The application of a stop word list and lemmatisation or stemming are forms of normalisation for text. As with the normalisation of numeric data, they can introduce assumptions that may affect outcomes. In the study of text, the use of a stop word list potentially creates a modifiable language unit problem (MLUP). The standard stop word list for texts written in English (see Appendix A.2) include modal verbs ('would', 'should', 'might', 'could') and personal and possessive nouns (I, you, we, them). These terms may be relevant when analysing human feelings about a location.

Two open-source NLP toolkits are used to process and analyse texts: NLTK for Python and Quanteda for R. NLTK was created in 2001 at the University of Pennsylvania and is the most popular set of functions used for NLP (Bird, Klein, & Loper, 2009). Quanteda was created in 2013 at the London School of Economics to provide similar capabilities within R, focusing on the quantitative analysis of text (Benoit, et al., 2018).

6.2.1.2 Data retrieval and pre-processing

As explained in chapter four, tweets were retrieved by regularly submitting a query to Twitter's Search API. Code samples for acquiring and pre-processing tweets are detailed in Appendix B.1. For this study, the tweets analysed are those containing geotags that locate them spatially within the QEOP. The query is constructed to retrieve tweets located within a 2.5km radius of the centre of the QEOP. The data returned are a random sample of tweets matching the criteria, assumed to represent 1% of all tweets posted during the period, as detailed in Twitter's official documentation at the time of retrieval (Twitter, n.d.).

The data returned for a query is the full record for each tweet. Only attributes of interest for analysis are retained, including the tweet content, timestamp and spatial coordinates. The full list of attributes is listed in Table 9 in chapter four. The rest of the data is discarded and never stored. The dataset is filtered to only those tweets that fall within the landscape boundary. Retweets (tweets beginning with 'RT') are removed. A retweet is the reposting a tweet posted by somebody else. It is assumed the content is not a direct experience.

Figure 66 describes the process to prepare tweets for all text analyses. The content of a tweet is first 'cleaned' to reduce it to a string of text containing words to analyse. An NLP package is then used to tokenise the string for each tweet. The tokenised tweets are converted into a corpus of documents, where each document is the set of tweets posted on the same day. The corpus can then be used for various term-based analyses, including counting frequency of terms, comparing the similarity of terms between documents, and classifying the documents by comparing their terms against a dictionary, such as scoring for sentiment or emotion expressed.

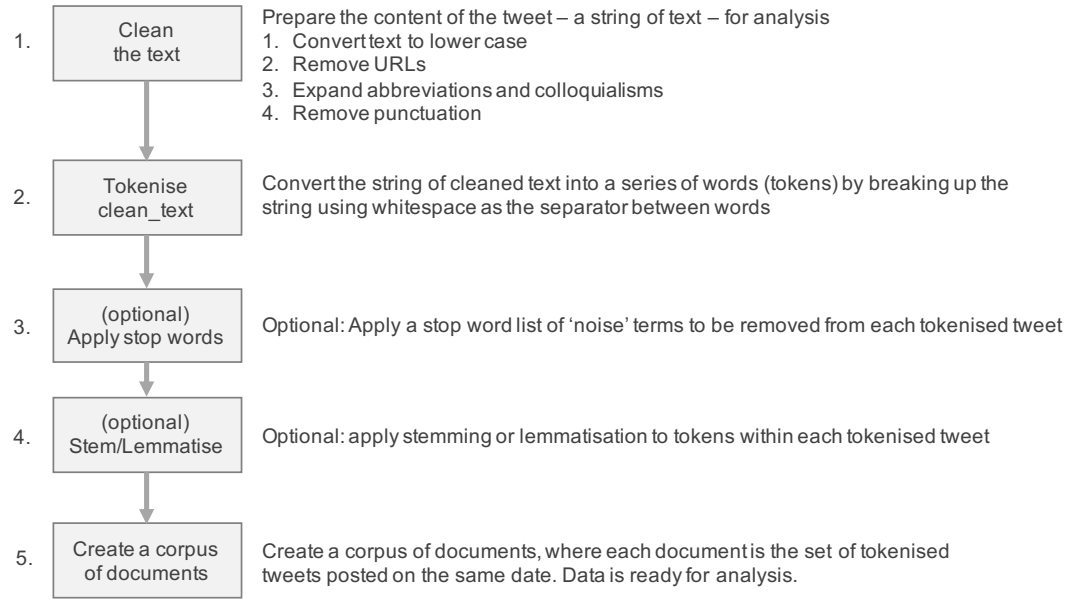


Figure 66. Preparing text for term-based analysis

6.2.2 Results

Two analyses were performed using the geotagged tweets retrieved during March 2016 and from June to August 2016. The first explores the potential to detect different circumstances occurring in the park from the messages posted by people whilst present that may help explain or anticipate changes in population behaviours. The second examines the potential to infer the mood of the population present from emotive terms used within tweets posted on the same date.

6.2.2.1 *Detecting circumstances*

This study performs an analysis of the frequency of terms within sets of tweets posted on different dates to consider if they reveal conditions that may inform variations in population behaviours such as the number of visits to the park. In chapter four, inspecting the daily counts of presence based on the number of unique devices connecting to the free park Wi-Fi revealed dates with anomalies, where readings were either unexpectedly higher or absent. Figure 67 shows the presence counts for the park W-Fi, people posting geotagged tweets, and the number of tweets posted with geotags located within the same boundary as the park Wi-Fi range. Known events are indicated with arrows. The approximate start of the summer school holiday (varied by school from 19th to 22nd July) is indicated with a dotted vertical line and weekends are highlighted in a lighter grey.

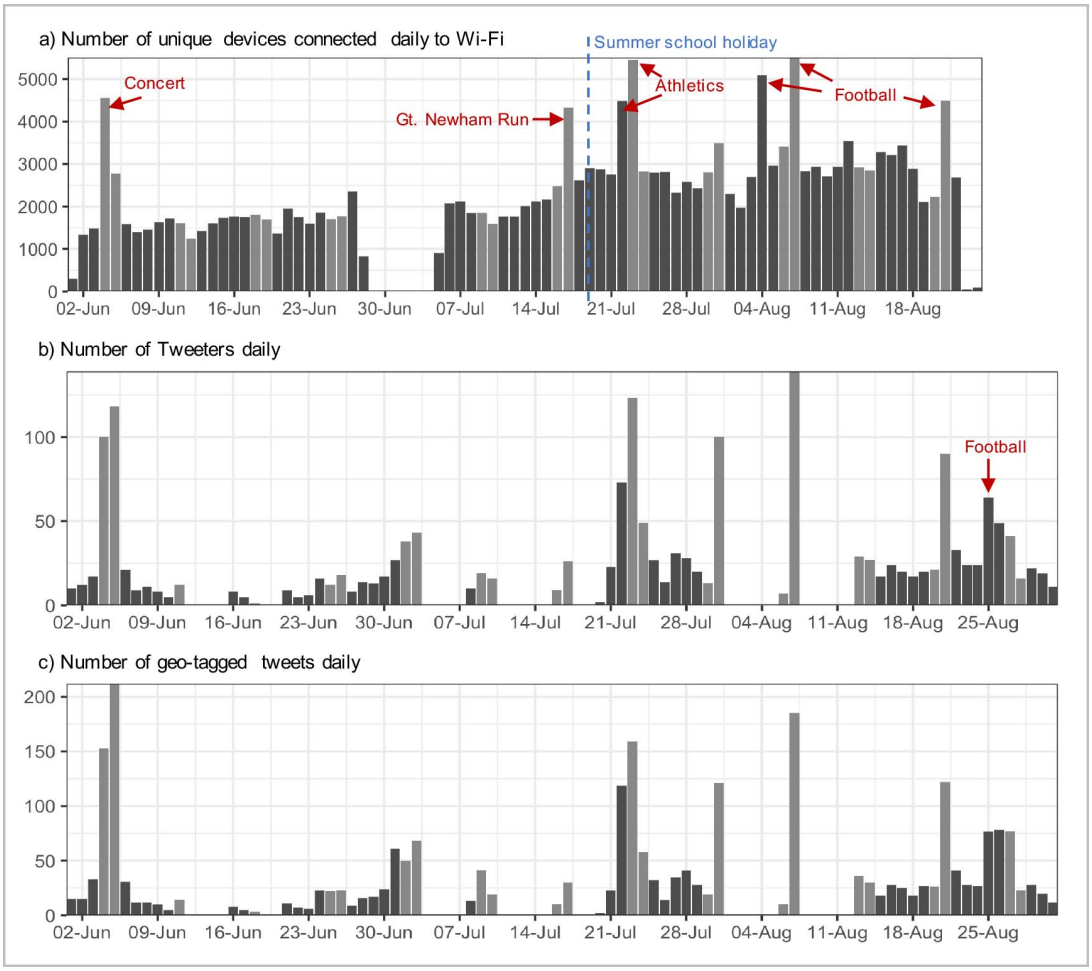


Figure 67. Presence counts within the QEOP, June to August 2016

Weekends are highlighted in a lighter grey.

The Wi-Fi has visible outages on 1 June, 28 June to 5 July and from 23 August onwards. The retrieval of a sample of tweets from the Twitter Search API resulted in several dates with zero counts but did include data on some dates when Wi-Fi readings were absent or unexpectedly high.

For this study, tweets are limited to only those that fall within the Wi-Fi boundary of the QEOP (see blue outline in Figure 20). This is why several dates show zero counts in Figure 67. The majority of

tweets retrieved daily were located outside the Wi-Fi boundary, predominantly within the Westfield retail centre to the east of the park, outside the range of the Wi-Fi network.

An analysis of three dates where Wi-Fi readings were either higher than expected or absent revealed contextual information (see Figure 27 in chapter four). The same study is repeated here and expanded. Figure 68 shows word clouds for dates of interest, where the Wi-Fi readings are either higher or lower than expected, or missing entirely, and where data is available from the Twitter sample. The word clouds show the top 20 terms sized based on the frequency of use.

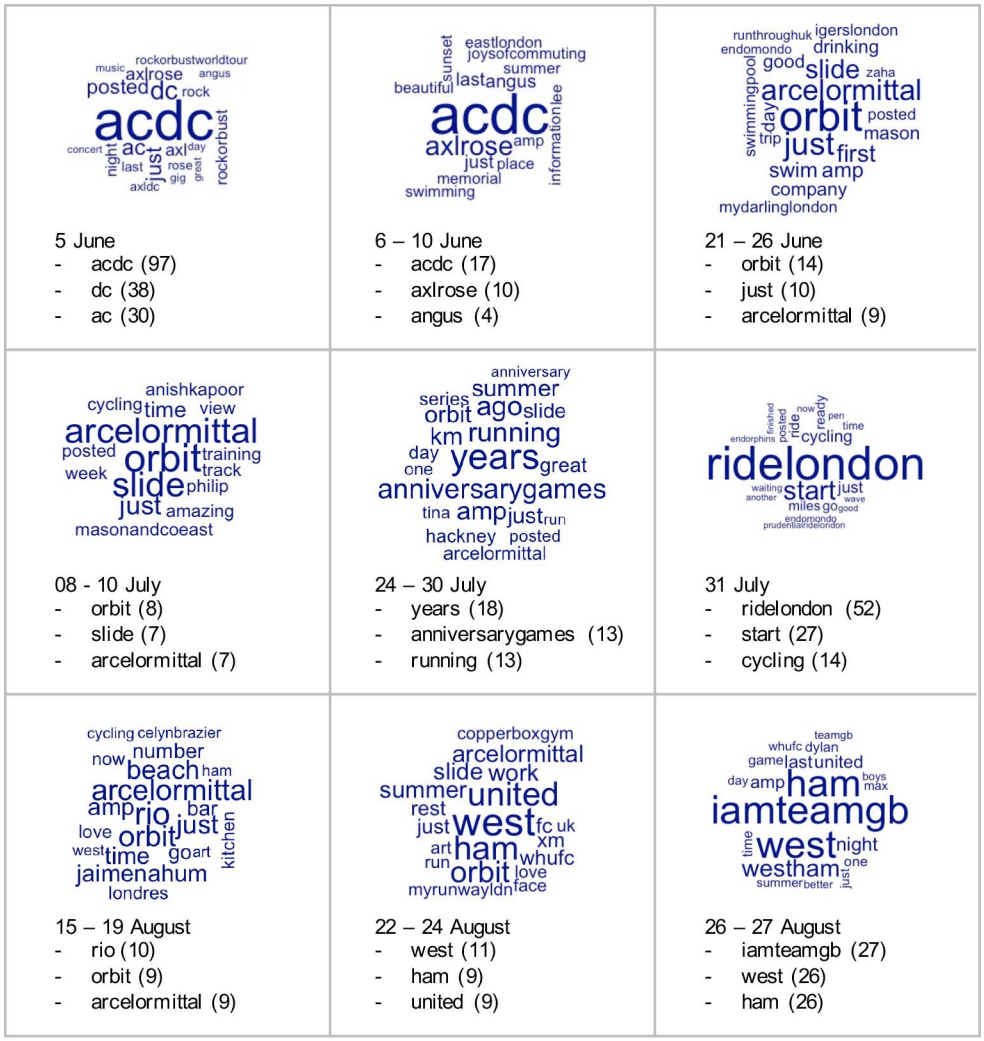


Figure 68. Word clouds for geotagged Twitter activity in the QEOP, June to August 2016

Words are sized by count. Noise words have been excluded by applying a stop word list.

Just plotting as word clouds shows a difference between large events and non-event dates. On event dates, the top terms dominate the word cloud. On non-event days, the counts are much lower and more distributed amongst many terms. Three events are identified. On 31st July, the top term is 'ridelondon'. Performing an internet search query for the phrase 'RideLondon 2016' returned a

news article[5] revealed that a cycling event – 'Prudential RideLondon' – started in the park. It is an annual event with over 30,000 cyclists taking part in a charity ride from London to Surrey.

The word clouds also reveal events that generated national news coverage. On 26th/27th August, the top term is 'iamteamgb'. A news search[6] identified 'I am Team GB' was a national event organised by the National Lottery and ITV to celebrate medallists returning from the Rio 2016 Olympics. It included shutting down ITV's television channel for one hour from 09:30am on Saturday with the switch-off performed by gymnast Matt Whitlock who participated in events at the Copper Box Arena. In late June, the ArcelorMittal Orbit gained a new attraction in the form of the tallest and longest slide in Europe. The term 'slide' is mentioned twice on 5th June and then is mentioned on most days from 23rd June onwards. Consulting news sources, it appears the slide opened to the public on 24th June 2016[7] following media previews on 23rd June. This introduces the possibility of identifying the elasticity of terms within a location-based vocabulary. New words will appear and some words will decline in use whilst others will remain popular. Just as with weights for presence, a location-based dictionary can be self-calibrating through the frequent or continuous sampling of available data sources.

The term 'acdc' remains the top term on the day after the AC/DC music concert and during the week after. This is an important consideration if labelling dates for different contexts. Figure 67 shows that more tweets were posted the day after the concert than on the date itself. That date would be labelled as a non-event day, yet it generated substantially more tweets than would be expected for a non-event day when compared with all other non-event days during June (Figure 63). The same effect was revealed when studying presence, with visits to the park increased above normal on the day after the concert.

This effect is explored further using a dataset generated from tweets posted from within the park during June 2017. Four music concerts were held at the London Stadium in June 2017. Each concert generated a different volume of Tweets both during and after the event. Figure 69 contains the number of tweets, term frequency counts and distribution of usage for each event. For all four concerts, the most frequently used term has the highest count on the day after a concert. This is assumed for Guns N' Roses since the highest count is the day after one performance but also the same day of the second performance, with a higher count on the day after the second performance than on the day of the first performance. All four concerts have the terms 'last' and 'night' in the top twenty most used terms, indicating tweets being posted the day after the concert.

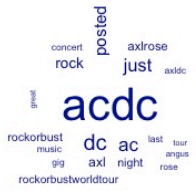
---

[5] Source: https://www.prudentialridelondon.co.uk/news-media/latest-news/prudential-ridelondon-2016-where-to-watch/ accessed 1 November 2018.
[6] Source: https://www.telegraph.co.uk/news/2016/08/27/selfies-and-celebrations-with-team-gb/ accessed 1 Nov 2018.
[7] Source: https://www.theguardian.com/artanddesign/2016/jun/23/carsten-holler-arcelormittal-orbit-slide-first-ride accessed 1 November 2018.

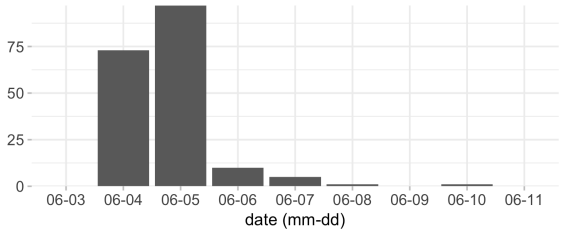
Top terms for the day(s) of the music concert and the day after the concert combined, distribution of top term across dates (geo-tagged tweets within QEOP Wi-Fi boundary)

a) AC/DC on Sat 4th June 2016c

Sat 4 – Sun 5 June 2016
- acdc (170)
- dc (71)
- ac (57)
- just (54)
- posted (48)
- rock (40)
- axl (33)
- axlrose (29)
- night (26)
- rockorbust (26)



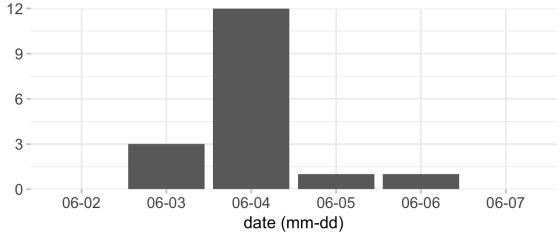
Term: 'acdc'



b) Depeche Mode on Sat 3rd June 2017

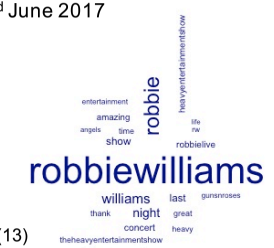
Sat 3 – Sun 4 June 2017
- depechemode(15)
- depeche (7)
- mode (7)
- globalspirittour(3)
- last (3)
- night (3)

Term: 'depechemode'

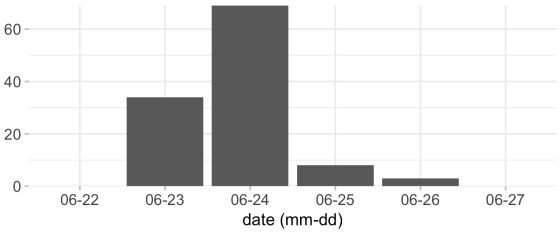


c) Robbie Williams on Fri 23rd June 2017

Fri 23 – Sat 24 June 2017
- robbiewilliams (103)
- robbie (52)
- williams (29)
- night (25)
- show (19)
- last (18)
- concert (15)
- robbielive (14)
- heavyentertainmentshow (13)
- amazing (13)

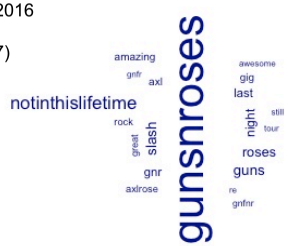


Term: 'robbiewilliams'



a) Guns N' Roses on Fri 16th and Sat 17th June 2017

Fri 16 – Mon 18 June 2016
- gunsnroses (52)
- notinthislifetime (17)
- slash (12)
- guns (12)
- roses (11)
- night (10)
- gnr (9)
- last (9)
- gig (7)
- axl (7)
- amazing (7)



Term: 'gunsnroses'

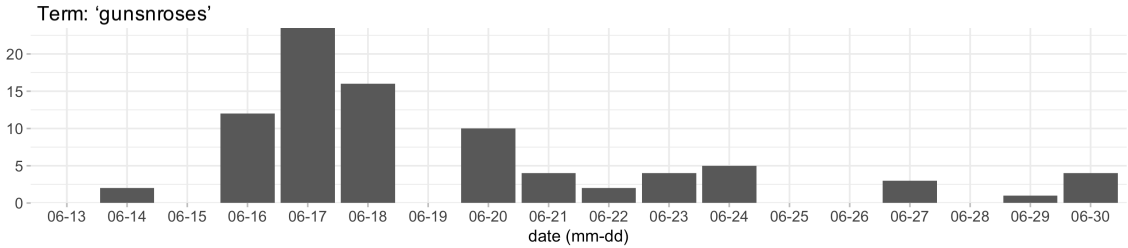


Figure 69. Term frequency for music concerts at the London Stadium, 2017

Whilst not expected, in retrospect this makes sense. All of the concerts took place in the evening and are not environments that involve sitting down waiting for some action during the event, as can be the case with sporting events. The majority of the tweeting on the day of the concert appears to be posted during the build-up to the start of the event. That all concerts generated 'last night' conversations the day after suggests that hosting music concerts for national and global icons will generate surplus tourism for the area. People travelling long distances to attend evening concerts that occur on Fridays or Saturdays are more likely to stay overnight and travel home the next day. For all four concerts, the term count is minimal from two days after the concert, although Guns N' Roses continued to be mentioned for a much longer period. In general, terms referring to the concert disappear within a week of the date of the concert.

The volume of tweets for each concert varied substantially. The AC/DC concert generated the highest volume of tweets but took place a year before the other concerts. It is possible that Twitter adjusted the sampling algorithm and/or integration with other social media channels changed at some point between June 2016 and June 2017. Two of the top ten terms for the AC/DC concert are 'just' and 'posted'. These two terms are default words used when posting to a different social media channel, Instagram, with automatic cross-posting to Twitter. The pair of terms do not appear in any of the 2017 concerts. Inspecting the source data, during June 2016 there were 686 geotagged tweets posted from within the park, of which 54 began with 'Just posted a photo @ …'. In June 2017, there were 531 geotagged tweets posted from within the park, of which just 9 began with 'Just posted a photo @ …' It is an indication that using real-world observations requires frequent recalibration to accommodate changes in the use of digital media within physical settings, and changes to the accessibility of such sources for analysis, as well as changes in vocabulary.

A challenge with analysing terms is that there is variation in the way phrases and abbreviations are used. For example, AC/DC was sometimes represented as ACDC or AC DC after removing punctuation. The athletics championship was officially known as the 'Mueller Anniversary Games', an anniversary event celebrating the stadium's role in the London 2012 Olympics. The promoted hashtag is 'anniversarygames' but multiple other terms are also used, including references to celebrity participants, such as Usain Bolt, and to the activities, such as running. The use of different terms can make it difficult to apply the usual computational linguistic methods, such as stemming or lemmatisation – reducing words to their root form or grouping by a shared lemma – to combine terms referencing the same circumstance within the park. For use in near real-time analytics, text analysis will not have the luxury of preparing a dictionary of expected terms in advance, other than for recurring contexts. Whilst computational text analytics makes it very easy to combine and consolidate terms such as 'ac', 'dc', and 'acdc', such actions require the time to inspect the text. An alternative approach is to examine the amount of similarity in the vocabulary used on different dates. This is the focus of the study presented in section 6.3

The top terms both reveal information about the events taking place and highlight the challenge with using dictionary-based approaches to determine the mood, or sentiment, being expressed

within the landscape. For all four concerts, the top terms include the name of the band and the name of the tour that the concert is part of, respectively:

- AC/DC – Rock Or Bust tour
- Depeche Mode – Global Spirit tour
- Guns N' Roses – Not In This Lifetime tour
- Robbie Williams – Heavy Entertainment Show tour

Axl Rose appears in two of the concerts. He is the lead singer for Guns N' Roses and was the guest lead singer for the AC/DC concert due to the ill health of AC/DC's normal lead. The top terms for the Guns N' Roses concert include 'guns' and 'slash'. Slash is the stage name of Saul Hudson, the lead guitarist for Guns N' Roses. The SentiStrength dictionary (Thelwall, 2017) gives both 'gun' and 'slash' a score of -2 out of a range of -5 (strongly negative) to +5 (strongly positive). Messages about the concert may be incorrectly scored given these terms refer to the band and have nothing to do with the use of such terms in a violent context. This leads us to consider whether or not the mood of a situation can be detected and measured from the content of tweets.

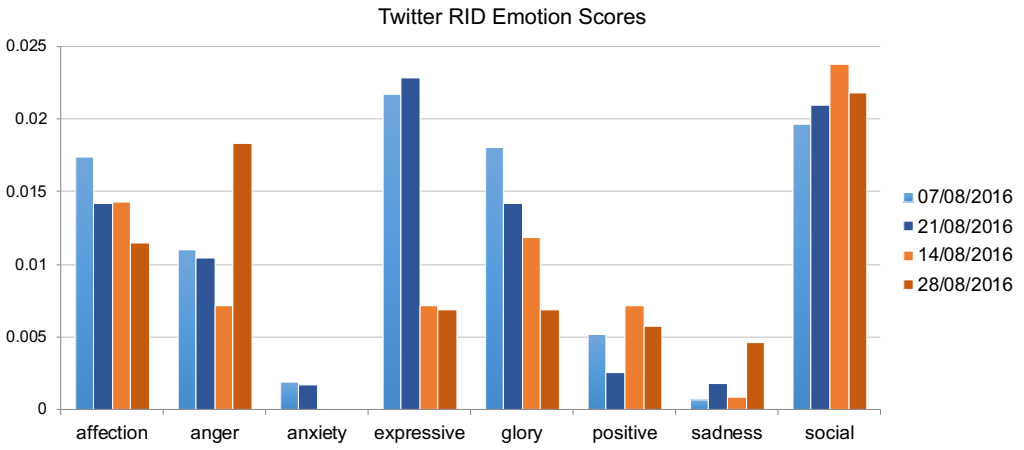
6.2.2.2 Measuring mood

Many studies have been published in the decade since Twitter was launched in 2006 that have explored human emotions being expressed within tweets by scoring terms using a sentiment dictionary, measuring terms as having a positive, negative or neutral valence. Since geotagging was introduced in 2009, a number have explored the potential to infer location-based mood and affect. One of the most prolific was published in 2013 and involved the study of 37 million geotagged tweets to associate happiness with urban mobility. A 2015 study quantified 'at risk' areas in Italy based on the use of hate-related terms (Musto, Semeraro, Lops, & de Gemmis, 2015).

Whilst studies have indicated that location-based mood can be inferred from social media posts, a preliminary exploration of the sentiment in tweets raised concerns that short texts could easily be misclassified. Terms such as 'wicked' and 'dope' could be interpreted as negative using their literal definitions yet are often used as positive indicators in conversation currently. Combinations of terms can invert sentiment, such as comparing 'pretty good' with 'pretty bad'. Linguistic Inquiry and Word Count (LIWC) is considered to be one of the most used text-analytics software tools by social scientists (Settanni & Marengo, 2015). It includes an online tool (https://liwc.wpengine.com). The phrase 'loving the bad ass atmosphere here', is ambiguous with terms that could be considered positive and negative in isolation. It returns a sentiment score of 25.8 out of 100, where above 50 is considered a more positive tone and below 50 is considered a more negative tone.

An alternative approach is to evaluate the words across a range of emotion categories. One of the most comprehensive dictionaries available is the Regressive Imagery Dictionary (RID) that contains over 3,00 words categorised for tones of cognition and affect (Provalis Research, n.d.). A preliminary study analysing geotagged tweets within the QEOP indicated emotive expressions on

different dates (Figure 70). The tweets were pre-processed as described earlier in this chapter and then organised as documents based on date.



Sample of tweets on 28th August, after pre-processing (cleaned text) LIWC

IMAGE REDACTED

Figure 70. Linguistic emotion scores for Sundays in August 2016

LIWC is emotional tone score, out of 100 (higher is more positive, lower is more negative).

Figure 70 shows the emotion scores calculated from a linguistic analysis using the Regressive Imagery Dictionary (RID) for four Sundays in 2016 and lists a sample of tweets from Sunday 28th August with their LIWC emotion tone scores. Across the four Sundays, two were event days (coloured blue), with West Ham United Football Club (WHUFC) playing matches at the London Stadium. The other two Sundays were non-event days (coloured orange). Two noticeable patterns in the RID results are for expressive and anger tones. Expressive terms are much higher on football than on non-football days, whilst one Sunday (28th August) registers a higher anger tone than the other three and could suggest an unexpected incident has occurred. The same day also registers a higher sadness tone. Other emotion tones are less pronounced or differentiated between event and non-event days, although there do appear to be trends, with more affection and glory terms expressed on football days and more social behaviour expressed on non-football days.

Interestingly, positivity is similar between one of the football dates and the two non-event dates. The least positivity is expressed for the winning football match. WHUFC lost 2-3 on Sunday 7th and won 1-0 on Sunday 21st. The loss was against Juventus, one of the top clubs across Europe, whilst the victory was against Bournemouth, a typically lower-performing club within the English Premier League. Perhaps the language reflects that a hard-fought loss is viewed more positively than an easy but perhaps uninspiring win?

A sample of tweets posted on the 28th August, the date displaying the highest anger and sadness RID tones, are displayed in Figure 70 with their associated LIWC emotion tone scores. The tweets expressing the strongest negative emotion are primarily referring to WHUFC and the club's move to the London Stadium. At least one tweet scoring the maximum negativity of 1.0 reads as humour ('that awkward moment when…'). That the LIWC tool only had a range of three scores for the tweets further indicates the difficulty at inferring positive or negative valence from short messages. The emotion scores were also very low. A single tweet using noticeably different language could have a large impact on tones when comparing dates for a local neighbourhood.

These results, combined with other research published at the same time that questioned the underlying foundations for classifying and detecting emotions in micro-expressions (Barrett, 2017) led to the conclusion that sentiment analysis of tweets at this spatial and temporal scale is ineffective and inappropriate. It is too easy to misclassify a tweet as expressing a strong negative sentiment. The decision was taken that further research in emotion detection within language is required, and a more comprehensive source of data needed, before emotions can be inferred with confidence from sampling location-based messages.

A follow-up review in 2019 confirmed that there remain issues with detecting tone from language using dictionary-based approaches. A recently published study found racial bias in measures, with African-Americans twice as likely to be labelled as being offensive because sentiment dictionaries lack cultural awareness (Sap, Card, Gabriel, Choi, & Smith, 2019). Retesting the same phrase 'loving the bad ass atmosphere here' using IBM Watson's Tone Analyzer produced a score of 10% anger, 4% fear, 36% joy and 52% sadness during a visit to IBM's research labs at Hursley Park, England, in September 2019.

For the remainder of this chapter, subjective emotive phrases are treated as part of the same contextual vocabulary as objective descriptive terms. The focus is on whether or not the content of tweets can help inform variations in visits to the park by revealing local conditions in real-time.

6.3 Learning a Contextual Vocabulary

The studies presented in section 6.2 demonstrated that the language expressed in location-based social media is neither general nor random. It varies contextually, revealing different circumstances occurring at different times within the same physical landscape. This section builds on that finding and examines the potential to learn a context-specific vocabulary for a landscape.

6.3.1 Data and methods

The same steps for data acquisition, pre-processing and preparation for text analysis are applied as in the previous section. In this section, additional data was acquired by inferring location based on the content of tweets, and more advanced text analysis methods were performed to model the content of tweets as topics.

6.3.1.1 Inferring location

Geotagged tweets were gathered as a possible indicator of presence within a social landscape and how that presence varies for different contexts. However, as discovered in chapter four, geotagged readings were very sparse when scaling to street-level, particularly for an open outdoor space such as the QEOP. The inclusion of geotags when posting tweets from a mobile device is optional. This issue was noticed early during the study. A second query was introduced, retrieving tweets via the Twitter search API that contained words matching a list of keywords that might indicate the tweet was generated within the landscape of interest. For the QEOP. the list of keywords included the name of the park (in full and abbreviated) and the names of venues and landmarks located within the park. It also included identifiable landmarks within the bounding box of the landscape, specifically Westfields Stratford retail centre, Stratford International rail station, Stratford tube station and Stratford bus station. The list was modified as venue names changed – the Olympic Stadium was renamed as The London Stadium when it reopened in June 2016. The list was also expanded in August to retrieve tweets mentioning West Ham United Football Club (WHUFC) and also mentioning at least one term relating to the park. The full list of terms included in the search query is included in Appendix B.1 code samples.

A visual inspection of the tweets retrieved revealed three types of tweet that were unlikely to be referring to interactions taking place within the landscape at the time of posting:

- tweets referencing specific dates and/or terms relating to the past or future
- tweets containing content referencing activities at other locations
- tweets that appear to be commercial advertising such as real estate and job references

The inspection also revealed that tweets appearing to describe human-environment interactions occurring within the location on a given date typically included at least one of the following:

- terms that indicate the observation is about the present or recent past. For example, ‘just been for a run in…’

- terms that describe some aspect of the environment that needs to be experienced. For example, ‘great atmosphere here…’
- terms that refer to physical interactions. For example, ‘terrible food service at…’

To consider whether or not a tweet is referring to a present experience within the landscape, each tweet is examined programmatically for containing words using the principles above. Figure 71 provides a summary of the workflow and a code sample for the script is provided in Appendix B.6. The first set of rules remove tweets unlikely to be referring to a current situation. For the remaining tweets, the second set of rules are used to identify tweets likely to be referring to a current situation. Any tweets that do not match the second set of rules are also removed.

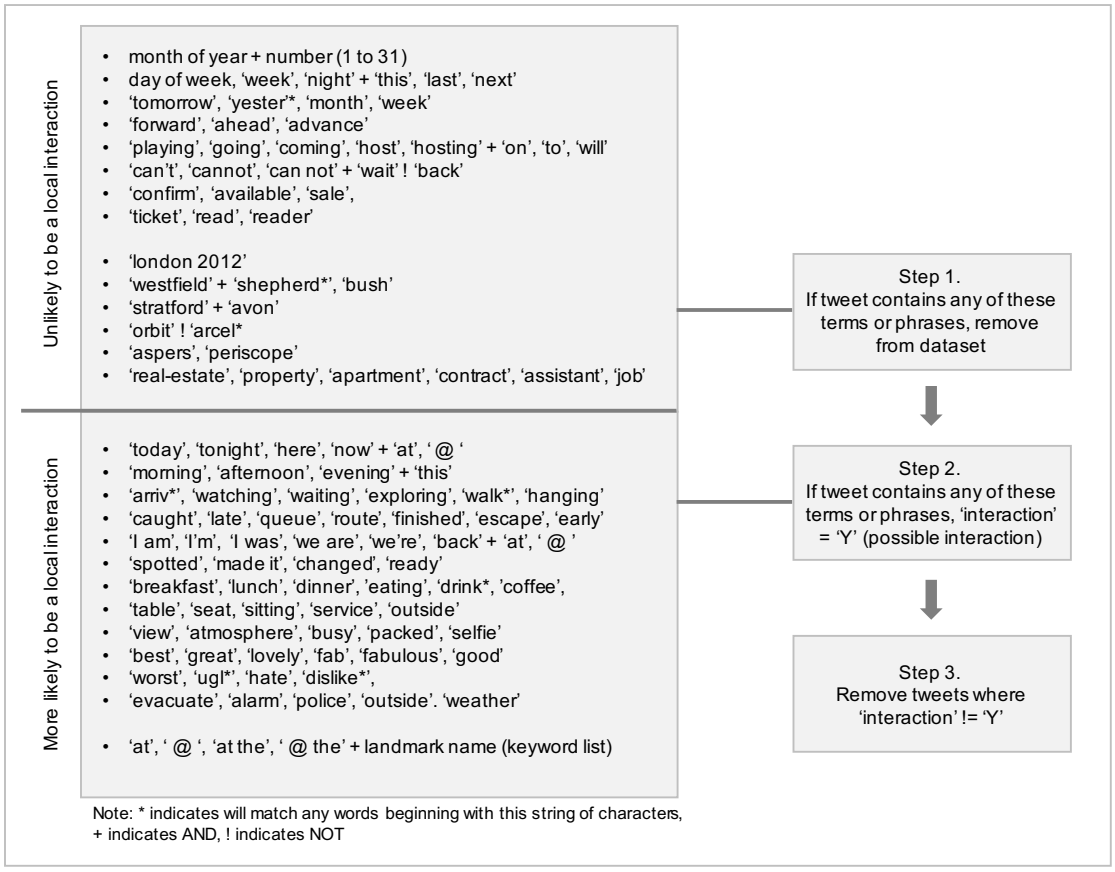


Figure 71. Process for inferring that a tweet was posted from a spatial location

Each row is a separate rule. Terms separated by ‘,’ indicate an OR statement. ‘+’ indicates an AND statement.

The words or word combinations deemed to indicate a local human-environment interaction were developed through trial and error. This was a crude process and intentionally conservative. The goal was to increase the volume of tweets available for analysis without introducing noise that would dilute meaning. At the stage of research when this process was developed, the goal was simply to explore for ways to increase the volume of location-based data available for analysis, not knowing if such analysis would prove beneficial. It was considered preferable to minimise false-positive – tweets estimated as spatially and temporally relevant when they are not. The set of tweets retrieved from May to August was reduced from containing 94,000 tweets matched via keyword to containing fewer than 11,000 tweets after filtering using this method.

Some of the terms are specific to the landscape. In the UK, there are two locations called Stratford: Stratford in East London, home to the QEOP and Stratford on Avon in Warwickshire, approximately 90 miles north of London. Hence tweets containing both 'stratford' and 'avon' are excluded. Also, from visual inspection of the tweets, numerous were posting memories from the London 2012 Olympics, hence anything containing 'london 2012' was automatically excluded.

Figure 72 shows the first 10 tweets before and after space-time filtering, from 11am on Saturday 22nd October, before the start of a football match at the London Stadium, and from 3:29pm, with just over 15 minutes left before full-time. In the sample, the challenge with identifying an interaction from a brief message is evident. The tweets highlighted in pink are those that have been kept as an interaction when they appear not to be. The tweets highlighted in yellow have been filtered out when they should have been kept. The green highlighted tweets are correctly filtered out. The unhighlighted tweets (white or grey background) are correctly retained after filtering. Two sets are shown, the first for tweets posted from 11am and the second for tweets posted from 3:29pm on Saturday 22nd October 2016, when a football match was taking place at the stadium. The majority of tweets are successfully filtered or retained. However, some are open to interpretation as to whether the tweeter was physically present or not. It would be difficult to form agreement amongst humans let alone to train a computer algorithm to automate.

Sampling through the remaining 11,000 tweets, a visual inspection estimated that over 90% of the tweets would be considered a real-time message from within the QEOP landscape. Adding further rules produced an over-aggressive result that substantially increased the number of false negatives by removing too many tweets that, from a visual scan, appear to be relevant. Reducing the rules had the opposite effect and substantially increased false positives by retaining too many tweets that were unlikely to be referring to local human-environment interactions. This was an imprecise approach involving several rounds of trial and error. This exercise revealed the potential for developing a language-based approach to identifying location-based interactions from within texts when a data source lacks spatial coordinates. The method developed is applied within this research to increase the volume of tweets for this study. However, further research is needed to create and evaluate a more robust approach. Such a development was beyond the scope or aims of this thesis and is discussed as a future research direction in chapter eight.

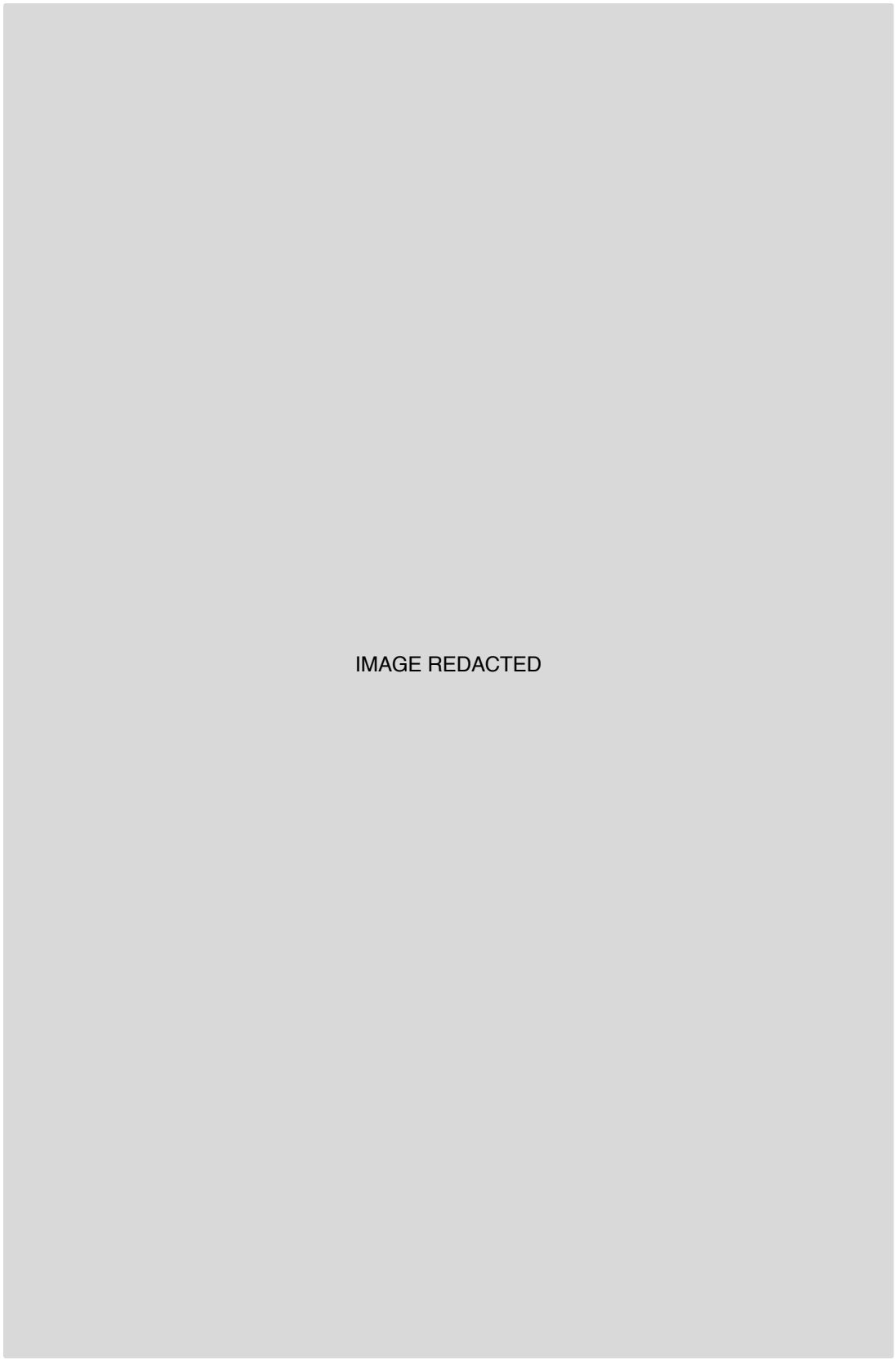


Figure 72. Twitter list before and after filtering to infer space-time relevance

Pink highlight indicates false positive (kept as an interaction when should have been filtered out). Yellow indicates false negative (filtered out when it should have been kept). Green indicates true negative (correctly filtered out). Remainder are true positives (correctly kept as an interaction).

6.3.1.2 Vocabulary analysis

Whereas the analyses in section 6.2 focused on the occurrence and frequency of individual terms within documents, the analyses here focus on sets of terms, the vocabulary of documents. Three techniques are explored: document similarity, collocation and topic modelling. In all approaches, each tweet is a document and a set of tweets is the corpus.

Document similarity calculates the 'distance' between documents within a corpus based on the occurrences of all terms common to each document regardless of the location of the terms within each document. Documents with a higher similarity score share a higher number of terms and have fewer terms unique to each document. If language is independent of different circumstances occurring within the same physical space on different dates than the similarity between documents should be high regardless of date. If language is random, then the similarity between documents should be either uniform or randomly varied. If language is contextual, then the similarity between dates where different circumstances are known to have occurred should be low whilst the language between dates where similar circumstances occur should be high.

Collocation studies combinations of words that appear together in sequence within a document. It is based on the assumption that a term is more related to its neighbours than to other terms within a text, and thus the occurrence of collocated terms may reveal more information than the occurrence of individual terms. The set of terms collocated are referred to as an 'n-gram', where n represents the number of adjacent words. Bi-gram and tri-gram analysis are the studies of collocated word pairs or triplets respectively. The expectation is that combinations of collocated terms may reveal more about the circumstances occurring on different dates and identify phrases that are context-specific versus general for the landscape.

Topic modelling is the process of analysing relationships between terms used within documents in a corpus to identify one or more topics – sets of terms that are likely to occur in a document without requiring them to be collocated. It is the most advanced of the three methods for analysing sets of terms within documents. Whilst document similarity provides a measure for comparing language on different dates, it compares all terms used on each date. Topic modelling offers the potential to identify clusters of terms that may indicate different situations occurring on the same date.

Each of these methods is described in more detail within the corresponding analysis presented in the following results section.

6.3.2 Results

Three analyses are presented. The first uses document similarity to compare the language used on different dates. The second considers if collocation can reveal additional contextual details not known or knowable in advance. The third examines the potential to identify different situations occurring on the same date within the same landscape through topic modelling.

6.3.2.1 Comparing contexts (similarity)

To compare the language expressed on dates known to have different contexts: the occurrence of a large-scale event at the London Stadium or not, document similarity is calculated where each document is the set of tweets posted on the same date.

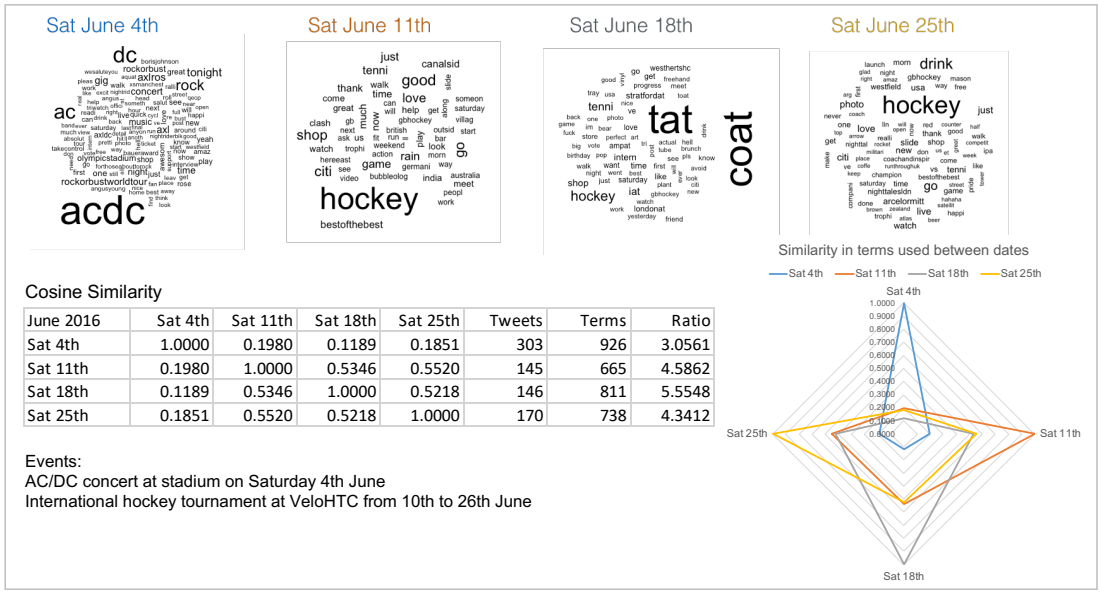
Cosine similarity [4] is used as the distance measure due to its simplicity and popularity for calculating the similarity between documents (Provost & Fawcett, 2013).

$$d_{cosine}(X,Y) = 1 - \frac{X \cdot Y}{||X||_2 \cdot ||Y||_2} \quad [4]$$

Cosine distance measures the occurrences of each term in each document, represented as vector counts. The score returned is a measure between 0 and 1 indicating the similarity between two documents. Cosine distance ignores scale, such as the different lengths of documents. This may have an impact when comparing documents if substantially different volumes of tweets are generated on different dates. Given the exploratory nature of this study, the method was considered sufficient. However, before making any recommendations based on findings, further research would be needed to evaluate alternatives that do incorporate the length of documents.

Two sets of results are presented. The first compares a day of the week for each month from June to August 2016, the same period analysed in the first study presented in chapter four. During June, there was a single music concert on the first Saturday. During July, there was an athletics championship on the third Saturday. During August, there were football matches on the first and third Sundays. The second study compares dates on which football matches occurred at the London Stadium from September to December 2016.

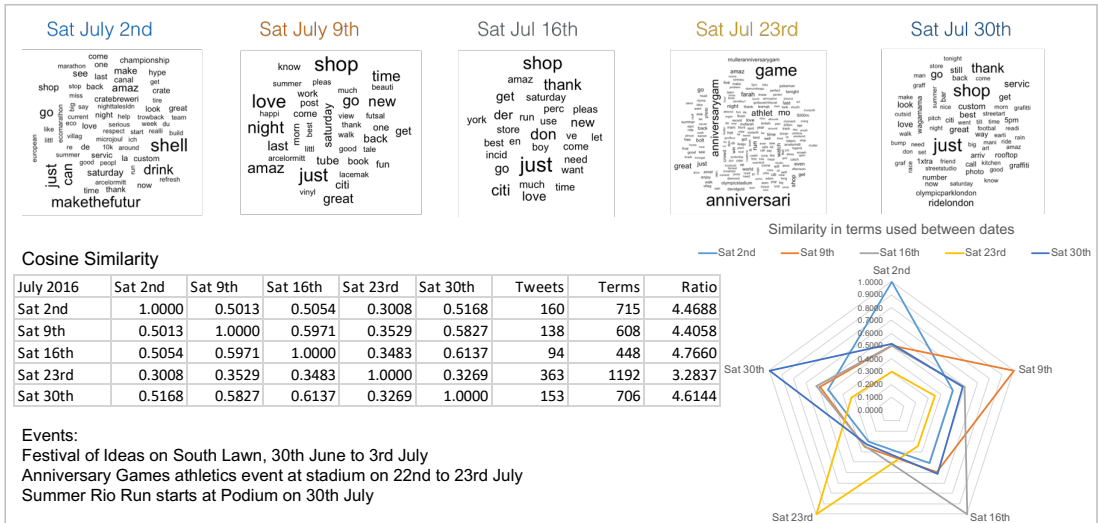
Figure 73 shows results for June, July and August 2016, comparing event days and their non-event equivalents before and/or after. The word clouds show the top terms across both geotagged tweets and keyword matched tweets on each date, using the same method in the previous study. The table presents the cosine similarity scores when comparing each document with each other, where each document is a date during the month. The scores are also visualised as a radar diagram to indicate the amount of similarity between dates. The tables also show the number of tweets included in each document, the number of terms within the document, and the term/tweet ratio. The two football dates had the highest counts, each at over 700. However, it is likely that this due to the recent move of the home football team – West Ham United (WHUFC) to the London Stadium. The move was considered controversial and generated numerous headlines locally and nationally. Likely, the volume of tweets referring to the move from the old stadium will decline over time. The other large events each generated over 300 tweets. Tweets on non-event days ranged from 94 to 180. This does suggest that using a similarity measure that incorporates document size, such as the TF-IDF algorithm (Provost & Fawcett, 2013) may be a more appropriate similarity distance measure for future research.



Cosine Similarity

June 2016	Sat 4th	Sat 11th	Sat 18th	Sat 25th	Tweets	Terms	Ratio
Sat 4th	1.0000	0.1980	0.1189	0.1851	303	926	3.0561
Sat 11th	0.1980	1.0000	0.5346	0.5520	145	665	4.5862
Sat 18th	0.1189	0.5346	1.0000	0.5218	146	811	5.5548
Sat 25th	0.1851	0.5520	0.5218	1.0000	170	738	4.3412

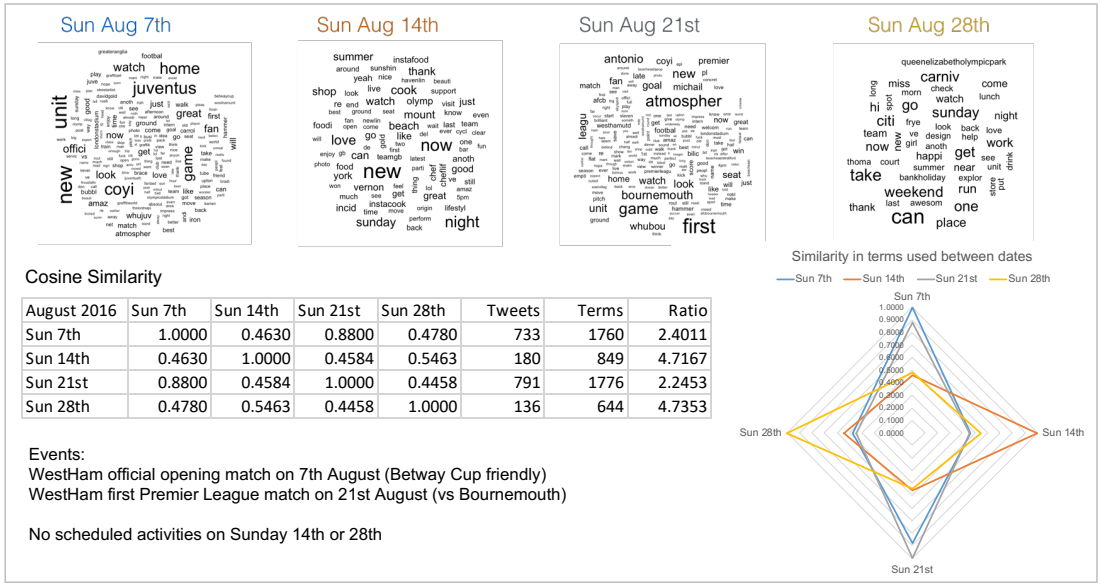
Events:
AC/DC concert at stadium on Saturday 4th June
International hockey tournament at VeloHTC from 10th to 26th June



Cosine Similarity

July 2016	Sat 2nd	Sat 9th	Sat 16th	Sat 23rd	Sat 30th	Tweets	Terms	Ratio
Sat 2nd	1.0000	0.5013	0.5054	0.3008	0.5168	160	715	4.4688
Sat 9th	0.5013	1.0000	0.5971	0.3529	0.5827	138	608	4.4058
Sat 16th	0.5054	0.5971	1.0000	0.3483	0.6137	94	448	4.7660
Sat 23rd	0.3008	0.3529	0.3483	1.0000	0.3269	363	1192	3.2837
Sat 30th	0.5168	0.5827	0.6137	0.3269	1.0000	153	706	4.6144

Events:
Festival of Ideas on South Lawn, 30th June to 3rd July
Anniversary Games athletics event at stadium on 22nd to 23rd July
Summer Rio Run starts at Podium on 30th July



Cosine Similarity

August 2016	Sun 7th	Sun 14th	Sun 21st	Sun 28th	Tweets	Terms	Ratio
Sun 7th	1.0000	0.4630	0.8800	0.4780	733	1760	2.4011
Sun 14th	0.4630	1.0000	0.4584	0.5463	180	849	4.7167
Sun 21st	0.8800	0.4584	1.0000	0.4458	791	1776	2.2453
Sun 28th	0.4780	0.5463	0.4458	1.0000	136	644	4.7353

Events:
WestHam official opening match on 7th August (Betway Cup friendly)
WestHam first Premier League match on 21st August (vs Bournemouth)

No scheduled activities on Sunday 14th or 28th

Figure 73. Term sets and similarity scores for tweets in the QEOP, Jun to Aug 2016

Results based on combination of geotagged tweets and keyword matches after filtering for interactions and de-duplicating. Stop list of words applied. For each image, the same day of week is compared throughout the month. From the top, a) Saturdays during June 2016, b) Saturdays during July 2016, c) Sundays during August 2016.

During June, one large event was held at the London Stadium – the AC/DC music concert on 4th June. It has a noticeably different vocabulary to the other Saturdays during the month, sharing less than 20% vocabulary, whilst the three Saturdays each share approximately 50% of terms with each other. There was another event taking place during this period – an international hockey tournament in the Lee Valley HTC at the northernmost area of the park.

During July, the Mueller Anniversary Athletics Games were held at the stadium. Other smaller scheduled events were held throughout the month. As with June, the date of the stadium event has a vocabulary least in common with all other dates, sharing approximately one-third of terms. The other dates, as with June, share approximately 50% of terms with each other.

During August, two football matches were held on two Sundays, with no large events on the other Sundays, although various small-scale local school holiday activities were taking place, including screening the Rio 2016 Olympic Games. The two football dates share nearly 90% vocabulary whilst having less than 50% in common with the other two Sundays. The two non-event Sundays share just over 50% in common, following the trend visible in the other two months.

The results indicate that different contexts will have distinctly different vocabularies and that there is a core vocabulary shared amongst dates. To explore these findings further, tweets were collected for the remainder of 2016, to enable an analysis of Premier League football matches taking place at the stadium. Figure 74 shows similarity scores and unique terms for four matches that each took place on a Saturday with a 15:00 kick-off.

Saturdays with Premier League Football Matches, 15:00 kick-off

Cosine Similarity

Cosine similarity	Sat 9th Sep	Sat 1st Oct	Sat 22nd Oct	Sat 5h Nov
Sat 9th Sep	1.0000	0.8681	0.8047	0.8160
Sat 1st Oct	0.8681	1.0000	0.7955	0.7856
Sat 22nd Oct	0.8047	0.7955	1.0000	0.7693
Sat 5th Nov	0.8160	0.7856	0.7693	1.0000

Top unique terms

Sat 9th Sep	Sat 1st Oct	Sat 22nd Oct	Sat 5h Nov
watfordfc	whumid	whusun	stoke
hornets	grows	dizzeerascal	scfc
whuwat	link	giants	soccersixbigbang
threaten	mutual	boyindacornder	bojan
deeney	benefits	released	soccersix
comeback	provide	dizzee	whustk
joel	solo	blow	hughes
promise	stranded	fellow	stokecity

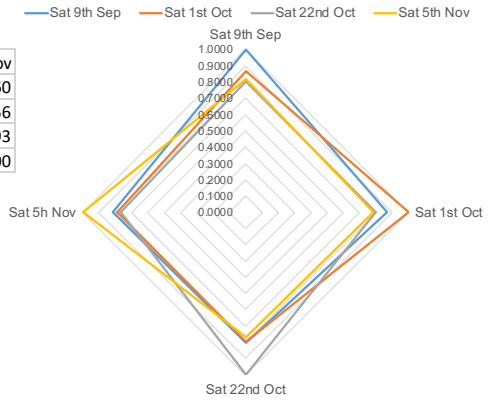


Figure 74. Similarity scores and unique terms per football match at the London Stadium

The four football Saturdays all have similarity scores above 75% when compared with each other, with most scoring over 80%. When examining the top unique terms for each date, the first term identifies the away team, either directly or as a hybrid term using a three-letter abbreviation for each, 'whuxyz' where 'xyz' is the away team: Watford (wat), Middlesbrough (mid), Sunderland (sun) and Stoke City (stk) respectively.

Interestingly, within the top unique terms, three dates have what could be considered emotive terms: 9th September, 'threaten' and 'promise'; 1st October, 'benefits and 'stranded'; 22nd October, 'released' and 'blow'; and, 5th November, 'stoke'. Sentiment analysis was excluded from this research after preliminary studies, due to concerns with the underlying dictionaries and theories used to classify the emotive use of language. This concern is reinforced when studying the language of the music concerts. As explained earlier, the terms 'gun' and 'slash' used when Guns N' Roses are performing a music concert would be incorrectly classified as negative. However, expressive terms do appear to reveal different atmospheres. Whether or not they can be quantified with sufficient rigour for scientific studies is a consideration for further work.

The event with the weakest similarity to the other event days is 22nd October. Examining the top unique terms for this date, the word 'dizzeerascal' is number two. Dizzee Rascal is a grime music artist within the UK. It indicates that there may have been another event also taking place within the park – a music concert. Summer 2016 was the first year when the park venues reopened for hosting special events. Throughout the summer, only one main event was hosted on any day. It is possible that, as the park becomes more developed, multiple larger events will be held simultaneously at different venues. The similarity between all tweets posted on the same date may become less effective as a measure to differentiate contexts. Instead, a method is needed to identify when messages are being posted about different situations occurring within the same physical space. This is the focus of the next study.

6.3.2.2 Learning about contexts (collocation)

The previous section measured the similarity between all terms used on different dates, assuming each date contained only one context. The final test revealed that on at least one football date, a music event was taking place at one of the other venues within the park. On Saturday, 22nd October 2016. West Ham United football match (WHUFC) at the London Stadium and a music concert by grime artist Dizzee Rascal was held at the Copper Box Arena. This date is analysed using two methods: collocation and topic modelling. For both studies, the dataset contains both geotagged and keyword matched tweets with the latter filtered to those inferred to posted from within the landscape, as described earlier in this chapter.

Collocation refers to the study of word sequences that appear unusually often within a text (Bird, Klein, & Loper, 2009). The sequence of words is referred to as a 'n-gram', where n represents the number of adjacent words included in the sequence. The expectation is that sequences of words may reveal more information about the landscape than the frequency of individual terms. Collocation analysis is less likely to be used in the study of large texts because the construction of the feature set increases exponentially with the number of terms in the text. However, this research is working with small texts and computational performance is not an issue.

Collocations are scored using an association measure. The simplest measure is frequency, the number of times a collocation appears in the text. A popular alternative measure is to calculate the pointwise mutual information (PMI) score of each n-gram (Medhat, Hassan, & Korashy, 2014). The

PMI score represents the conditional probability that the terms will occur together given their joint distribution and their individual distributions, assuming independence. PMI is expressed mathematically as equation [5] if performing a bi-gram analysis for terms x and y.

$$pmi(x;y) = log\frac{p(y|x)}{p(y)} \quad [5]$$

PMI will identify combinations of terms that are highly collocated but their occurrence may be infrequent. It is considered useful to require a minimum frequency to exclude outliers.

Figure 75 contains the collocation analysis for words in tweets posted on 22nd October 2016, as bi-grams and tri-grams, shown as the top twenty PMI scores and frequency counts.

```
highest scoring bi-grams: (PMI)

(('stewards', 'released'), 8.678222848845582)
(('dad', 'son'), 8.485577770903186)
(('son', 'attacked'), 8.485577770903186)
(('manage', 'overcome'), 8.398114929652847)
(('men', 'manage'), 8.398114929652847)
(('released', 'away'), 8.163649676015824)
(('attacked', 'outside'), 8.093260348124428)
(('bilic', 'men'), 7.93868331101555)
(('slaven', 'bilic'), 7.894901618987637)
(('nets', 'late'), 7.817625905511001)
(('after', 'stewards'), 7.571307644929071)
(('home', 'crowd'), 7.424466256599798)
(('reid', 'nets'), 7.415188443011793)
(('winston', 'reid'), 7.382484277939085)
(('late', 'winner'), 7.377053314125021)
(('winner', 'slaven'), 7.225050220679968)
(('outside', 'after'), 6.986345144207913)
(('away', 'fans'), 6.723077084629844)
(('1', '0'), 6.098496996490573)
(('fans', 'home'), 5.983893665213818)

most frequently occurring bi-grams:

[(('1', '0'), 46),
 (('0', 'sunderland'), 38),
 (('winston', 'reid'), 27),
 (('slaven', 'bilic'), 21),
 (('late', 'winner'), 20),
 (('reid', 'nets'), 20),
 (('nets', 'late'), 19),
 (('sunderland', 'winston'), 18),
 (('winner', 'slaven'), 18),
 (('men', 'manage'), 17),
 (('attacked', 'outside'), 16),
 (('bilic', 'men'), 16),
 (('away', 'fans'), 14),
 (('after', 'stewards'), 13),
 (('fans', 'home'), 13),
 (('home', 'crowd'), 13),
 (('outside', 'after'), 13),
 (('released', 'away'), 13),
 (('stewards', 'released'), 13),
 (('dad', 'son'), 12)]
```

```
highest scoring tri-grams: (PMI)

(('hunter', 'player', 'ratings'), 19.386193041085217)
(('james', 'hunter', 'player'), 18.80123054036406)
(('york', 'giants', 'la'), 18.51172392316908)
(('boy', 'da', 'corner'), 17.967403406945266)
(('suffer', 'blow', 'police'), 17.77148319697001)
(('pensioner', 'dad', 'son'), 17.386193041085214)
(('system', 'inside', 'delayed'), 17.31580371319382)
(('manage', 'overcome', 'fellow'), 17.29873019983488)
(('blow', 'police', 'radio'), 17.186520696248852)
(('radio', 'system', 'inside'), 17.145878711751507)
(('giants', 'la', 'rams'), 17.13321229991535)
(('new', 'york', 'giants'), 17.11940650039032)
(('police', 'radio', 'system'), 17.046343038200597)
(('dad', 'son', 'attacked'), 16.971155541806375)
(('stewards', 'released', 'away'), 16.841872524861408)
(('men', 'manage', 'overcome'), 16.796229859305697)
(('twickenham', '23', 'oct'), 16.730841212472665)
(('rams', 'twickenham', '23'), 16.615363995052725)
(('trouble', 'outside', 'again'), 16.615363995052725)
(('nfl', 'new', 'york'), 16.578838119027616)

most frequently occurring tri-grams:

[(('1', '0', 'sunderland'), 37),
 (('winston', 'reid', 'nets'), 20),
 (('nets', 'late', 'winner'), 19),
 (('0', 'sunderland', 'winston'), 18),
 (('late', 'winner', 'slaven'), 18),
 (('reid', 'nets', 'late'), 18),
 (('sunderland', 'winston', 'reid'), 18),
 (('winner', 'slaven', 'bilic'), 17),
 (('bilic', 'men', 'manage'), 16),
 (('slaven', 'bilic', 'men'), 16),
 (('after', 'stewards', 'released'), 13),
 (('attacked', 'outside', 'after'), 13),
 (('away', 'fans', 'home'), 13),
 (('fans', 'home', 'crowd'), 13),
 (('outside', 'after', 'stewards'), 13),
 (('released', 'away', 'fans'), 13),
 (('stewards', 'released', 'away'), 13),
 (('dad', 'son', 'attacked'), 12),
 (('men', 'manage', 'overcome'), 12),
 (('son', 'attacked', 'outside'), 12)]
```

Figure 75. Collocation analysis for words in tweets on 22nd October 2016

Reviewing the results for bi-grams, the most frequently occurring bi-grams produce the final score of the match ('1-0'), two names associated with the match ('winston reid' and 'slaven bilic'), that the goal was a late winner, and that there was an incident outside ('attacked outside'). Whilst the score of 1-0 was the most frequently occurring bi-gram, it only just made the top 20 PMI scores. Instead, the PMI scores emphasise the incident ('son attacked') and hints that 'winston reid' was the goal scorer ('reid nets') as well as also confirming it was a late winner ('nets late', 'late winner').

The tri-gram analysis provides clarity for some collocated terms but not others. For example, '1, 0, sunderland' could be interpreted as announcing the score was '1-0 to Sunderland' when in fact it was West Ham who won the match, 'winston, reid, nets' and 'nets, late, winner' confirms that the goal was late in the match and scored by Winston Reid. The tri-grams confirm the attack occurred outside and probably after the match ('attacked, outside, after'), and that a father and son were involved ('dad, 'son', 'attacked). The frequency counts for tri-grams all appear to refer to the football match. The highest PMI scores for tri-grams focus on the attack that occurred and other situations unrelated to the football match. There is more detail about the incident ('pensioner dad son', 'trouble outside again'). Only one references the music concert – 'boy da corner' is the name of the music concert. Others indicate that the data set contains tweets unrelated to the landscape: 'new york giants', 'york giants la' and 'nfl new york' suggest tweets describing a sporting event based in America whilst 'twickenham 23 oct' and 'rams Twickenham 23' indicated that a rugby match may be taking place the following day at the Twickenham stadium in west London. In fact, the event taking place at Twickenham on Sunday 23rd October 2016 was an American National Football League (NFL) match with the New York Giants playing the LA Rams. It was the first-ever non-rugby sporting event held at Twickenham[8]. There were geotagged tweets that would also have benefitted from the rules developed to infer location-based tweets from keyword matched tweets that do not contain coordinates. It had not been anticipated that geotagged tweets should also be filtered if the content can be implied to not be referring to current conditions or experiences.

The results from collocation are mixed. The technique does reveal additional contextual information with varying results if using frequency counts versus PMI to identify the top n-grams. Frequency counting weights towards the most popular collocated terms whilst PMI weights towards the most unexpected frequently occurring collocated terms. PMI can highlight interesting phrases. However, the collocation analysis does not provide is any indication of whether or not the n-grams refer to the same situation or something else. Was the attack outside the stadium or elsewhere? Was it to do with the football match? An alternative approach is to explore the clustering of terms within documents regardless of their sequential order, referred to as topic modelling.

6.3.2.3 Discovering contexts (topic modelling)

Topic modelling is the process of analysing relationships between terms within a document to identify one or more topics. Latent Dirichlet Allocation (LDA), a generative probabilistic model that was introduced in 2003, is considered to be the most popular method for modelling topics in texts (Blei, 2012). It was designed on the assumption that a document will contain a small set of topics and that each topic will use only a small set of terms frequently. Each term is assigned a probability that the term is attributable to the topic. Figure 76 shows the results from performing LDA on geotagged tweets posted on the 22nd October located within the QEOP landscape. Three tests were performed. The first two include all nouns and adjectives organised as 3 and 5 topics

[8] Source: BBC News https://www.bbc.co.uk/sport/american-football/37745803 accessed 19 Dec 2019

respectively. The third uses only hashtags, organised as three topics. The top ten words and their probabilities for being attributed to the topic are listed for each.

```
a) Nouns and adjectives only. 3 topics, 20 passes, top 10 words

[(0, '0.023*"west" + 0.021*"ham" + 0.019*"sunderland" + 0.018*"london" + 0.018*"stratford" +
0.015*"winston" + 0.014*"winner" + 0.014*"stadium" + 0.014*"reid" + 0.013*"late"'),
(1, '0.049*"london" + 0.045*"stadium" + 0.034*"olympic" + 0.024*"park" + 0.020*"ham" + 0.019*"west" +
0.018*"arena" + 0.016*"copperbox" + 0.015*"stratford" + 0.011*"dizzeerascal"'),
(2, '0.068*"stadium" + 0.056*"london" + 0.028*"ham" + 0.028*"west" + 0.016*"sunderland" + 0.015*"fans" +
0.014*"today" + 0.013*"stratford" + 0.011*"safc" + 0.008*"outside"')]

b) Nouns and adjectives, 5 topics, 20 passes, top 10 words

[(0, '0.040*"london" + 0.033*"stadium" + 0.013*"corner" + 0.011*"boy" + 0.009*"dizzeerascal" + 0.009*"da" +
0.007*"boyindacorner" + 0.007*"day" + 0.006*"empty" + 0.006*"live"'),
(1, '0.072*"stadium" + 0.071*"london" + 0.033*"stratford" + 0.032*"ham" + 0.030*"west" + 0.013*"today" +
0.013*"fans" + 0.011*"safc" + 0.009*"station" + 0.009*"sunderland"'),
(2, '0.033*"stadium" + 0.030*"london" + 0.015*"west" + 0.013*"ham" + 0.010*"police" + 0.009*"hackney" +
0.008*"safc" + 0.008*"stratford" + 0.007*"wick" + 0.006*"player"'),
(3, '0.042*"stadium" + 0.030*"london" + 0.011*"today" + 0.007*"whufc" + 0.007*"brewery" + 0.007*"crate" +
0.007*"trouble" + 0.006*"twickenham" + 0.006*"rugby" + 0.006*"hackneywick"'),
(4, '0.046*"olympic" + 0.044*"west" + 0.043*"ham" + 0.032*"park" + 0.031*"sunderland" + 0.026*"stadium" +
0.024*"arena" + 0.022*"london" + 0.021*"copperbox" + 0.018*"winner"')]

c) Hashtags only. 3 topics, 20 passes, top 10 words

[(0, '0.039*"whufc" + 0.023*"coyi" + 0.020*"whusun" + 0.019*"boyindacorner" + 0.008*"london" +
0.006*"rbmauktour" + 0.005*"dizzeerascal" + 0.004*"boyinthecorner" + 0.004*"rbma" + 0.004*"stratford"'),
(1, '0.023*"westham" + 0.015*"london" + 0.004*"londonlife" + 0.004*"westhamfamily" + 0.004*"stadium" +
0.003*"football" + 0.003*"uk" + 0.003*"redbull" + 0.003*"girl" + 0.003*"londongirl"'),
(2, '0.048*"safc" + 0.011*"whufc" + 0.009*"sunderland" + 0.007*"hammers" + 0.006*"hawaythelads" +
0.005*"dizzeerascal" + 0.005*"premierleague" + 0.004*"epl" + 0.004*"dizzee" + 0.004*"catsofinstagram"')]
```

Figure 76. Topic modelling from tweets using LDA

For each list, top 10 words produced for each topic along with probability of it being attributed to the topic

LDA requires specifying the number of topics to be discovered. This can be problematic if the number is unknown. Furthermore, LDA will assign all terms to a topic. As can be seen in Figure 76, many terms appear in all three topics for each test and, whilst it is expected that probabilities will be very low, they decline at varying rates. Focusing on hashtags produces topics that are closest to the two known contexts but with noise. Terms specific to the football match and to the music concert appear in all three topics. For these reasons, LDA is discarded for this research and an alternative is considered, focusing on the hashtags. The alternative method is based on a concept referred to as 'Centering Resonance Analysis' (CRA).

CRA was developed around the same time as LDA, as a method to classify documents by using network analysis and measures of network centrality to identify the terms at the 'centre' of a document and considered to contribute most to the meaning of the text (Cornan, Kuhn, McPhee, & Dooley, 2002). The 'resonance' of the text is then defined as a measure of similarity between two or more texts in terms of their respective centres. Intended for discourse analysis, a similar approach could be adapted to the study of tweets. For this approach, only hashtags extracted from each tweet are used. Using the terminology of network analysis, each hashtag is a node in the network. Links can be created between a hashtag and all other hashtags that appear in all tweets containing this hashtag. The next step is to describe the network using a centrality measure.

Figure 77 describes four common network centrality measures for reference: degree, closeness, betweenness and eigenvector (Kolaczyk, 2009).

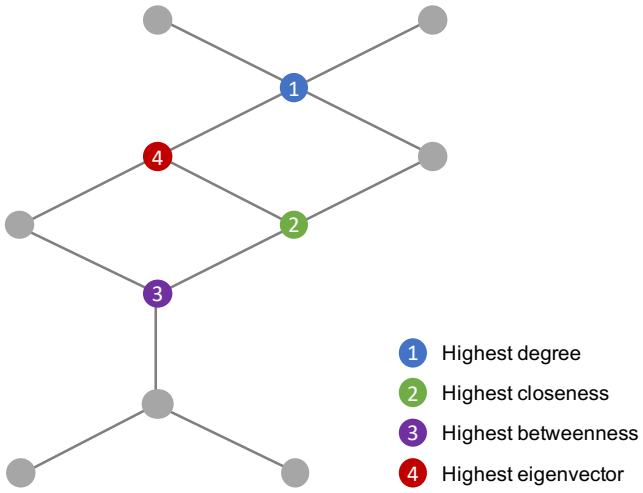


Figure 77. Measures of network centrality, adapted from Kolaczyk, 2009

Figure 4,4 (Kolaczyk, 2009). Degree indicates node with the most connections. Closeness indicates the node with the shortest paths to other nodes. Betweenness indicates the highest intermediary, has the most nodes dependent on it for access to the rest of the network. Eigenvector indicates the node with the most influence – it is closest to both the most connected and the intermediary.

Figure 78 shows the top nodes by the four different network centrality measures applied to the network of hashtags from tweets posted on 22nd October 2016. Of the four measures, Betweenness Centrality appears to provide the most information. It contains terms for both football teams involved in the football match ('westham', 'whufc' and 'sunderland, 'safc'), and the name of the music concert ('boyindacorner') and performer ('dizzeerascal').

Degree Centrality	Betweenness Centrality	Closeness Centrality	Eigenvector Centrality
london (0.2542)	london (0.1748)	london (0.3254)	london (0.4683)
whufc (0.1299)	whufc (0.0602)	whufc (0.2861)	football (0.2856)
westham (0.1243)	westham (0.0478)	football (0.2784)	westham (0.2522)
football (0.113)	stratford (0.0456)	westham (0.2712)	stadium (0.2388)
safc (0.1073)	safc (0.0429)	epl (0.2643)	whufc (0.2324)
stadium (0.0904)	eastlondon (0.0347)	whusun (0.2540)	olympicstadium (0.1855)
boyindacorner (0.0791)	football (0.0252)	olympicstadium (0.2515)	epl (0.1736)
stratford (0.0734)	boyindacorner (0.0245)	stratford (0.2503)	safc (0.1649)
grime (0.0678)	sunderland (0.0114)	stadium (0.2411)	sitting (0.1532)
eastlondon (0.0621)	dizzeerascal (0.0112)	hackneywick (0.2335)	waiting (0.1532)

Hashtag classification:

Whufc event: whufc, westham, football, safc, sunderland, epl, whusun
Music event: boyindacorner, grime, dizzeerascal
Amibiguous: stadium, olympicstadium, sitting, waiting
Location: london, stratford, eastlondon, hackneywick

Figure 78. Scoring hashtags based on network centrality measures

Hashtag classification notes included to indicate different meanings associated with terms.

Betweenness measures a node based on its role as an intermediary, it has the most nodes dependent on it for access to the rest of the network. The removal of intermediaries can lead to the break-up of a network into separate smaller networks, called communities. Figure 79 shows the network produced using betweenness as the centrality measure for hashtags in tweets posted on 22nd October 2016. Hashtags that indicate intermediaries between different topics are circled in red, and terms relating to the two known situations – the West Ham football match at the London Stadium and the Dizzee Rascal music concert at the Copper Box Arena – are shown with links coloured green and purple respectively. Also visible are several other isolated potential topics, for example, one referring to terms relating to Halloween and one referring to Middle Eastern food.

The four circled terms represent the names of four areas of London that the QEOP falls within or is close to: London, East London, Hackney Wick and Stratford. The visible formation of clusters within the network suggests that community detection to partition the network may be a feasible approach to discovering topics. The potential benefit is not requiring any prior knowledge about the topics or the number of topics expected to exist. Furthermore, no dictionary is required. The use of hashtags is language-agnostic.

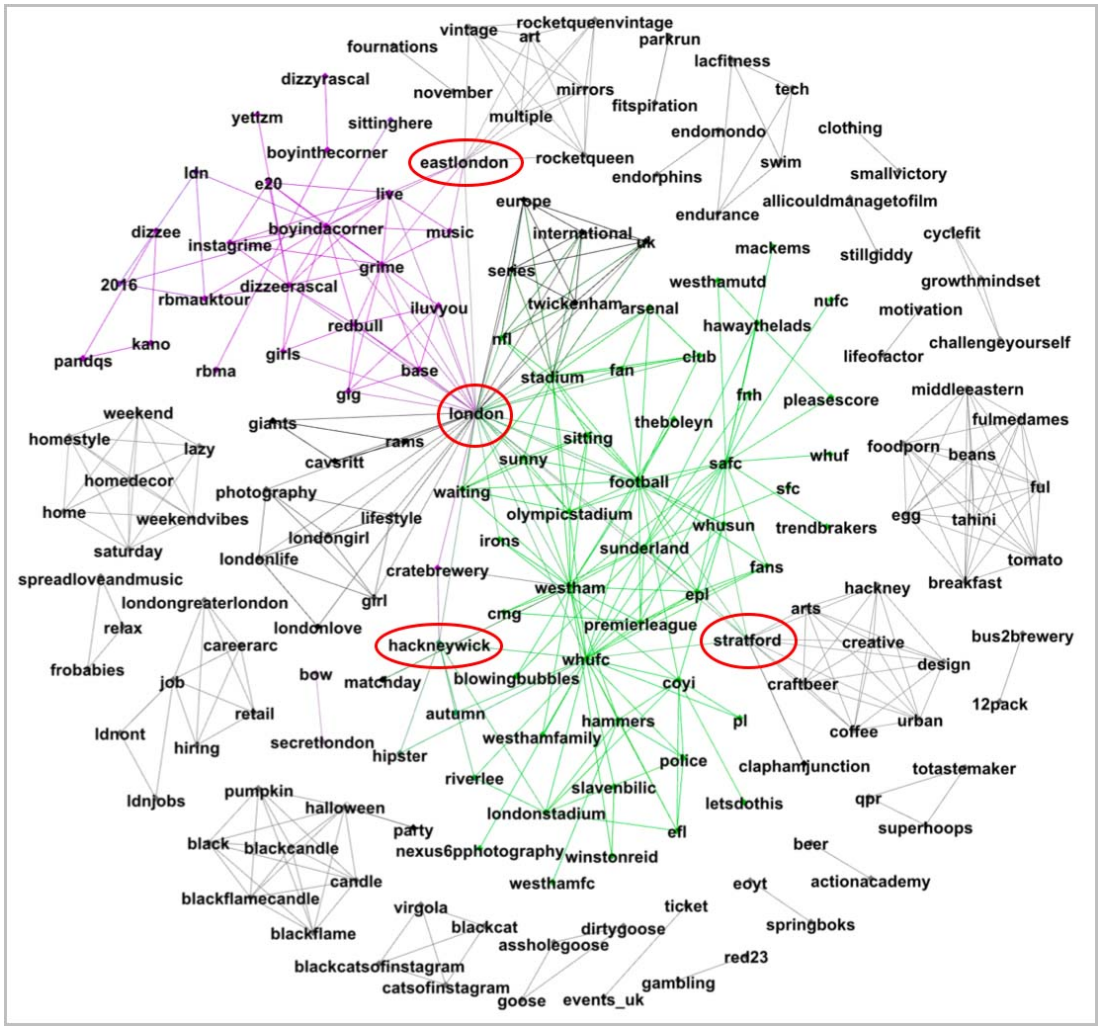


Figure 79. Network of hashtags as nodes for tweets posted 22nd October 2016 in QEOP

Colouring applied to two clusters for emphasis. Location-based betweenness nodes circled in red.

Whilst CRA identifies the terms with the highest betweenness score to indicate the centre of a text, inverting the approach could potentially identify when a text contains multiple topics. Within network analysis, this is referred to as community detection. A community detection algorithm partitions a network into smaller networks, referred to as communities. There are many different approaches and an evaluation of them is beyond the scope of this thesis. However, it is possible that different techniques will produce substantially different results. On this basis, five algorithms have been compared: Fast-and-Greedy by Clauset et al, Girvan-and-Newman, Infomap, Louvain by Blondel et al, and Pons & Latapy. Each adopts a different method for partitioning the network and maximising its modularity. The modularity score is a measure of performance by comparing the

density of connections within clusters compared to the density of connections between clusters, with higher scores indicating better performance (Fortunato, 2010).

Figure 80 shows the results. The five algorithms have each detected between 27 and 30 communities within the network of hashtags with modularity scores of between 0.742 and 0.776. For this network, the Louvain method is arguably the most performant based on its modularity score (highlighted in yellow), closely followed by the Pons & Lapaty method.

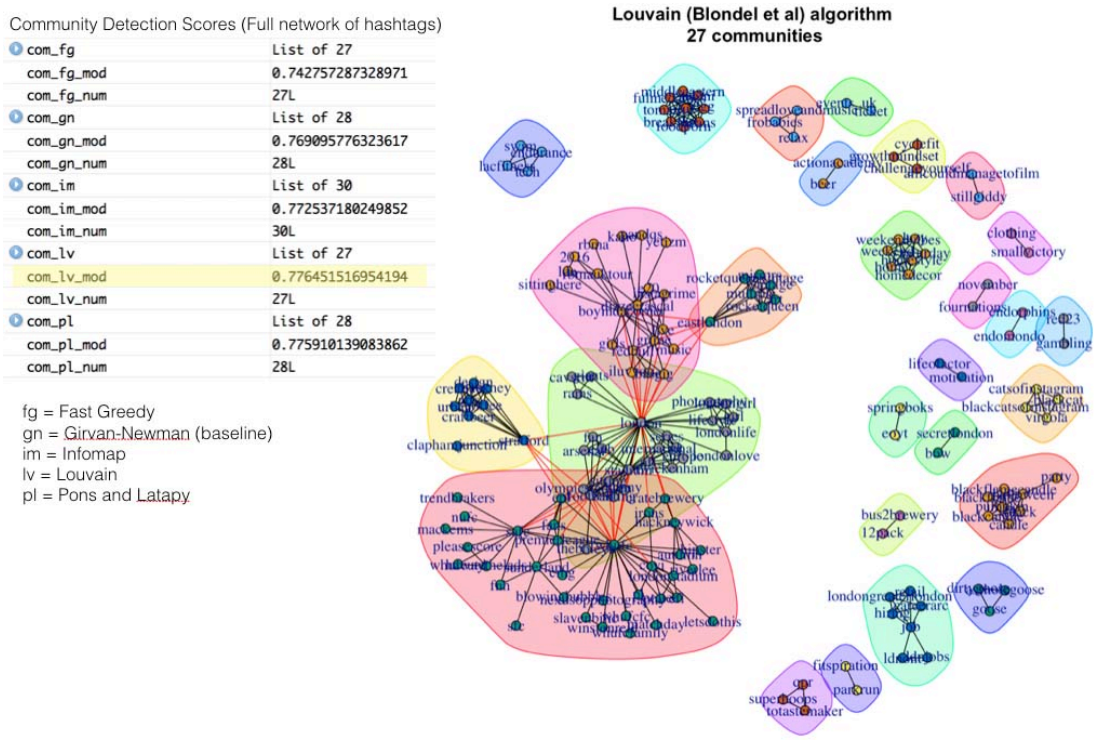


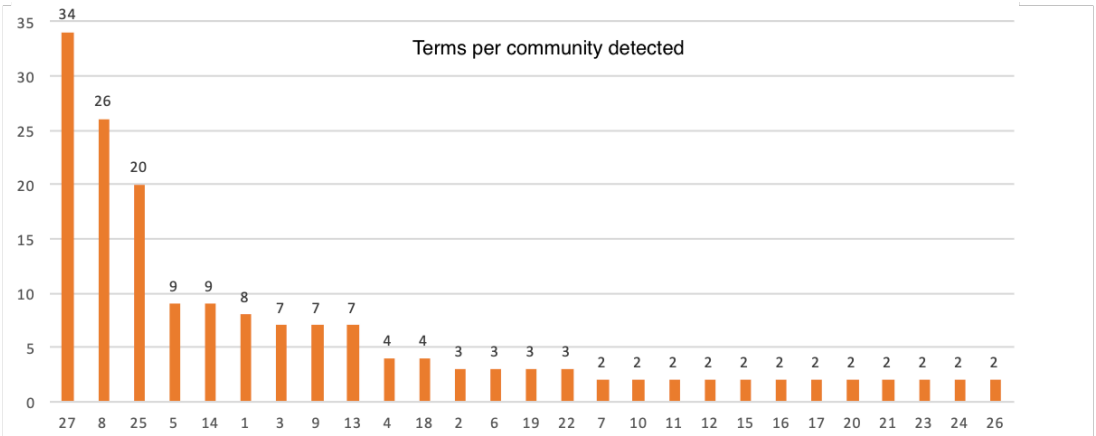
Figure 80. Network community detection of topics within tweets

Table shows modularity scores and number of communities detected for each algorithm. Higher modularity indicates a better community detection score.

Figure 81 shows the distribution of terms per community detected using the Louvain method. Three of the 27 communities have substantially more terms– communities numbered 27, 8 and 25. Community 27 is dominated by terms referring to the football match, community 25 is dominated by terms referring to the music concert, and community 8 is a mix of terms that refer to the London Stadium, some football references but also references to non-football activities.

The results of performing community detection are encouraging. Unlike LDA, it reveals the language of the music concert that took place on the same day as the football match, including the name of the performer ‘dizzeerascal’, the name of the tour ‘boyindacorner’, the music genre ‘grime’, the tour sponsor ‘redbull’ and the official tour hashtag ‘rbmauktour’. All are combined in a single topic. The method is very performant and does not require any prior assumption about the number of topics to be detected. A cut-off could be introduced to ignore any communities containing fewer than a minimum number of terms, either as a fixed number or as a ratio based on the distribution of hashtags for each date analysed. However, as discovered in chapter four, the volume of tweets for non-event days when focusing on the QEOP is low. Whereas it was possible to rapidly generate

a temporal dynamic of presence from the Wi-Fi data traces, and also from the OpenSignal readings in chapter five using just one month of data, a longer period is needed to gather and analyse tweets contextually. It may be possible to develop a contextual vocabulary for the landscape, but not based on a single month. A longitudinal study is required to increase the volume of tweets for studying different everyday contexts. This is discussed in chapter eight.



1	"black" "pumpkin" "candle" "blackcandle" "halloween" "blackflamecandle" "party" "blackflame"
2	"spreadloveandmusic" "frobabies" "relax"
3	"art" "rocketqueen" "rocketqueenvintage" "mirrors" "vintage" "eastlondon" "multiple"
4	"virgola" "blackcatsofinstagram" "catsofinstagram" "blackcat"
5	"coffee" "design" "creative" "stratford" "arts" "craftbeer" "claphamjunction" "urban" "hackney"
6	"cyclefit" "challengeyourself" "growthmindset"
7	"12pack" "bus2brewery"
8	"london" "giants" "sitting" "photography" "sunny" "fan" "nfl" "girl" "waiting" "international" "europe" "lifestyle" "cavsritt" "arsenal" "londonlove" "londongirl" "londonlife" "stadium" "series" "uk" "twickenham" "rams" "olympicstadium" "club" "football" "theboleyn"
9	"saturday" "lazy" "homedecor" "weekendvibes" "home" "homestyle" "weekend"
10	"ticket" "events_uk"
11	"secretlondon" "bow"
12	"eoyt" "springboks"
13	"hiring" "careerarc" "londongreaterlondon" "ldnont" "job" "retail" "ldnjobs"
14	"beans" "tomato" "ful" "middleeastern" "breakfast" "foodporn" "egg" "fulmedames" "tahini"
15	"endorphins" "endomondo"
16	"gambling" "red23"
17	"beer" "actionacademy"
18	"lacfitness" "swim" "endurance" "tech"
19	"goose" "assholegoose" "dirtygoose"
20	"fitspiration" "parkrun"
21	"motivation" "lifeofactor"
22	"totastemaker" "qpr" "superhoops"
23	"smallvictory" "clothing"
24	"november" "fournations"
25	"sittinghere" "ldn" "2016" "dizzeerascal" "iluvyou" "kano" "instagrime" "boyindacorner" "grime" "base" "girls" "live" "gig" "music" "rbmauktour" "redbull" "yetizm" "rbma" "e20" "pandqs"
26	"stillgiddy" "allicouldmanagetofilm"
27	"police" "fans" "cratebrewery" "londonstadium" "safc" "whufcfc" "cmg" "sunderland" "irons" "hipster" "trendbrakers" "mackems" "whufc" "nufc" "fnh" "whufcutd" "hawaythelads" "pl" "blowingbubbles" "riverlee" "autumn" "premierleague" "slavenbilic" "whufcfamily" "nexus6pphotography" "letsdothis" "matchday" "efl" "winstonreid" "sfc" "coyi" "epl" "hackneywick" "pleasescore"

Figure 81. List of terms per topic from community detection

27 communities detected using the Louvain algorithm. Chart shows number of terms per community. List shows terms contained within each community.

6.4 Research Outcomes III

6.4.1 Summary findings

Tweets were used as an available data source to explore the potential for text analysis of short location-based messages to reveal subjective experiences of place and how they vary during different contexts occurring within the same landscape. The results were mixed. The volume of data was too sparse to produce reliable findings for the day-to-day atmosphere of a landscape. Relying only on geotagged tweets also produced small volumes for most situations. This was in part overcome by increasing the dataset. Tweets were retrieved that contained keywords matching the names of features within the landscape. An experimental algorithm was developed to see if the location could be inferred from the content. The method was crude but did expand the data set for studying language expressed within or about the landscape. The method demonstrated potential and is an area for further work, discussed in chapter eight.

The simple term-based analysis provided a quick method for revealing terms of interest but assumes only one situation is occurring within the same landscape. Topic modelling was the more effective approach, able to detect and separate multiple different situations. Whilst LDA is the most popular method for topic modelling within a text, it is ill-suited to this research due to requiring advance knowledge of how many topics to discover. Instead, community detection within networks was applied, using hashtags. The discoverability of hashtags incentivises correct spelling and the re-use existing hashtags rather than creating multiple variations. Furthermore, they are language-agnostic. The results were promising and revealed the potential for creating a contextual vocabulary for a landscape. A longitudinal study could potentially offer further insight. Due to time constraints, it is beyond the scope of this research and a potential direction for future research.

Sentiment analysis and emotion detection had been the original focus at the start of the research, to reveal the social ‘atmosphere’ of place. As introduced in chapter two, many studies have been published using text-based approaches to associate landscapes with subjective well-being that showed promise. However, in researching both the computational techniques and the underlying theories, serious concerns have been raised about the viability of such methods at their current stage of development. A major concern is that interpretation of sentiment analysis is directed towards the landscape or people within it when it is unrelated to physical, environmental or social conditions.

6.4.2 Contextual framework update

Whilst the results have been mixed across the studies presented in this chapter, the studies did show promise for combining both signals and semantics to produce a richer understanding of how context can affect population behaviours within a landscape. Furthermore, similar techniques and findings applied to both approaches. Clustering can be applied both to signals to discover active spaces within the landscape and how they vary for different contexts, and to semantics to discover groups of terms applied to the landscape and how they vary.

The contextual framework is reviewed in terms of the potential to incorporate language as a measure to profile a landscape (Figure 82).

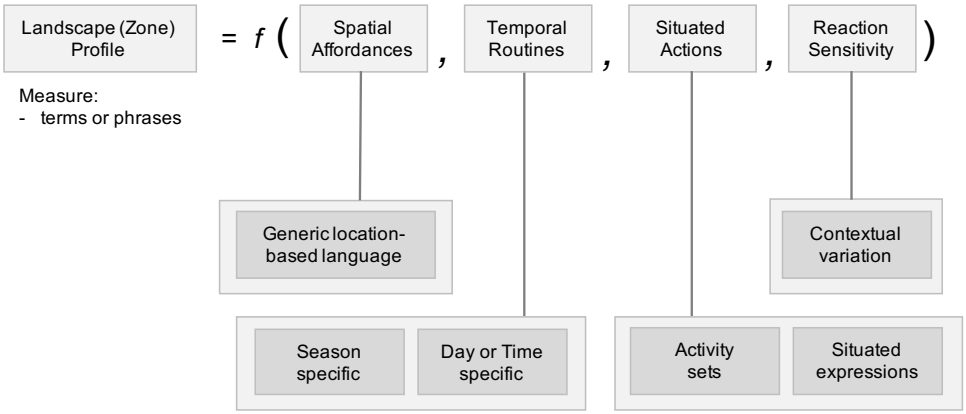


Figure 82. Language version of the contextual framework

For each element of the P-STAR formula, a set of terms is developed instead of a presence weight.

At the first level of context, spatial affordances, a generic set of terms or phrases could be developed to indicate the language used to describe the location independent of any situation, both for the landscape as a whole and for zones within the landscape although the latter is likely to be skewed towards active spaces only. The phrase 'generic local' may seem contradictory. It simply means terms that are generic to any situation but localised to the landscape. The term 'london' is likely to be generic within the QEOP landscape however it would also be generic to any other landscape within London and could be applied both to London in England, and London, Ontario in Canada. The phrase 'london stadium' is more likely to be local to the QEOP as it is the official name of the stadium within the park. However, there may be other stadiums in the world with the same name, so some uncertainty should be incorporated in the likelihood of the phrase being relevant to the landscape.

At the second level of context, different sets of terms could be developed for recurring temporal situations, such as seasonal changes, words specific to a day of the work, and words or phrases specific to the day status, such as Easter, Christmas, school holidays etc. Whilst it may seem odd to consider that the day of the week may have a specific language, the popularity of social media has resulted in recurring hashtags being used for days of the week. For example, '#throwbackthursday' is used to encourage people to post memories each Thursday. Note: this hashtag is also a strong indicator that the content of the tweet is not referring to a present situation.

At the third level of context, sets of terms could be generated for situations as and when they occur. Some situations may have re-use value, such as developing a set of terms likely to appear when music concerts are hosted in the QEOP. More importantly, learning situated language may help identify when trigger words such as 'gun' and 'slash' have a different meaning due to context, such as when a guitar player called Slash is performing with Guns N' Roses at a venue within the landscape. It may also be possible to detect changes in the use of emotive terms to indicate whether the mood of a landscape varies.

The use of community detection based on hashtags within tweets introduces the potential for creating a self-supervised algorithm for learning the contextual language of a landscape that could be applied to any source generating short status updates about people-place interactions. Furthermore, by performing such an analysis automatically, the contextual vocabulary of a landscape can be continually or frequently recalibrated. For example, one of the studies revealed the opening of a new slide attraction within the QEOP. Beforehand, the term ‘slide’ would have been an unexpected occurrence. Afterwards, it has become part of the landscape’s features. Just as recalibration can pick up cultural changes to presence and actions by detecting changes to the volume of signals generated by mobile devices present within the landscape, so recalibration can pick up physical and cultural changes that affect the language used to describe conditions and experiences whilst present within the landscape.

A second shared finding across both signals and semantics is that, when working with small and noisy samples, more can be learned with confidence by focusing on the differences. Even when the samples are too weak to produce robust statistics, they help isolate what matters in a cluttered scene, and to direct questions for further investigation, both quantitatively and qualitatively. Furthermore, combining both signals and semantics can help establish the case for such an investigation. Semantics alone may have little relevance if the presence is very small. Similarly, an unexpected variation in presence is difficult to explain without knowledge about conditions.

PAGE INTENTIONALLY LEFT BLANK

7 Modelling Behaviour Change

This chapter applies the findings and techniques developed from analysing the Queen Elizabeth Olympic Park (QEOP) to three landscapes that each experienced an unexpected major incident during 2017. All three landscapes are located within Central London: Westminster Bridge, London Bridge and Oxford Circus. For this study, the bounding box for each landscape is just under 2 kilometres squared, and are located within an area approximately 3.5km tall (south to north) by 5.5km wide (west to east) in central London (Figure 83).

Figure 83. Annotated satellite map of central London

Map © 2019 Google. Imagery © 2019 Bluesky, Getmapping Plc, Infoterra Ltd & Bluesky, Landsat / Copernicus, Maxar Technologies, The GeoInformation Group. Map data © 2019 Google. Annotations added to show the three landscape map tiles with incident locations labelled.

This chapter revisits the hypothesis that a static population measure, such as residential statistics or an ambient average such as the LandScan estimate, could be a poor indicator of the street population at risk or affected by a real-world phenomenon. It uses the same data source as chapter five – mobile app readings provided by OpenSignal and comprises of two studies. First, a single dataset – June 2017 readings across Greater London – is used to learn the socio-spatial dynamics for each landscape: how the population varies spatially and temporally. The second study uses the socio-spatial dynamics to produce an active population estimate for the hour up to when each incident occurs, comparing it with the best available static alternative. For two landscapes, real-time data is also used to study what, if any, impact a major trauma has on the social dynamic of the landscape. As with earlier studies, all readings are anonymised, and data are aggregated spatially and temporally for all analyses.

7.1 Data and Methods

The analyses presented in this chapter apply the techniques developed for the QEOP landscape. A uniform grid of cells is used for areal aggregation and zones. This is in part due to the weak results from using a data-driven approach to detect active spaces but also to enable comparisons between landscapes. The code for this chapter was developed as a general function that could be applied to any landscape of similar size by supplying coordinates for the bounding box of the landscape. Code samples are included in Appendix B.7.

7.1.1 Grid-based modelling

A LandScan grid of uniform cells is used for areal aggregation. LandScan cells are also sub-divided into 'pixels' 1/16th the size of a LandScan cell to explore for localised clustering within a single LandScan cell. The grid is programmatically generated based on coordinates for a bounding box (minimum and maximum values for the x-axis and y-axis), specified as NN.NNN5, NN.NNN833 or NN.NNN167, where NN.NNN are the degrees and first three decimals of each coordinate. The script will then generate cells sized at LandScan increments and Pixel increments starting from the south-west corner (xMin, yMin). LandScan cells have a width and height of 30 Arc seconds which produces a fraction when converting to degrees Decimal, an increment of 0.0083333. Pixel cells have an increment of 0.0020833. Rounding to a maximum of 7 decimal places means a small error will be introduced. For landscapes in this study, the distance error will be less than 1 metre across the bounding box and is considered acceptable. A correction should be introduced for larger landscapes or if a spatial inaccuracy of 1 metre is problematic.

Each cell is automatically labelled to provide a unique identifier based on an integer value representing the south-west corner of the cell (Figure 84). For a grid containing M points on the x-axis and N points on the y-axis, the maximum grid ID will be x(M-1)y(N-1). Readings within data sets will be tagged with the label for the cell that the reading falls within. Data points that fall outside the grid will be excluded. Cells will be referred to using this labelling system throughout this chapter for both LandScan and Pixel cells.

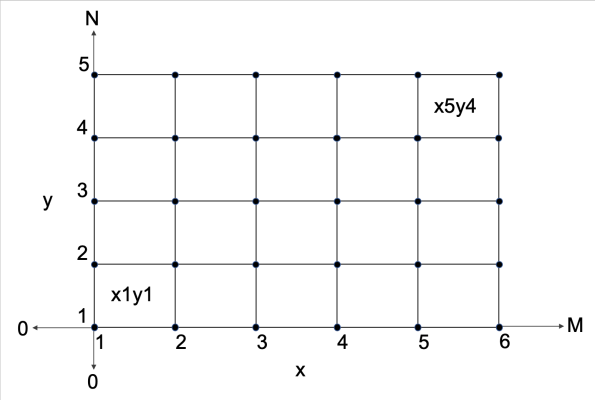


Figure 84. Grid referencing system used in uniform grid-based analysis

Grids are generated from the south-west corner. Cells are labelled as xMyN where M is the western point of the cell on the x-axis and N is the southern point of the cell on the y-axis.

7.1.2 Data preparation

The OpenSignal dataset introduced in chapter five is used here for the three new landscapes to establish their ambient socio-spatial dynamic: hourly and daily variations from a spatial baseline under assumed normal conditions. The OpenSignal June 2017 dataset was released for academic research and provides readings across the whole of the Greater London area (Figure 85).

Figure 85. Mobile device readings across Greater London on 1 June 2017

OpenSignal June 2017 dataset. Coordinates rounded to three decimal places.

Coordinates in this dataset are rounded to three decimal places, creating spatial uncertainty of approximately +/- 55 metres for each reading. Whilst at the city scale, such uncertainty is not apparent, with readings revealing the morphology of the landscape including major roadways (Figure 85), at neighbourhood scales it can mask spatial clustering when viewed in two dimensions (Figure 86a). To aid analysis and visualisation, coordinates are 'jittered' by adding a random value of +/- 0.000499 to each coordinate. Figure 86b shows the results of jittering the coordinates. With the coordinates jittered, clustering around landmarks becomes evident in two dimensions.

For one landscape – Westminster Bridge – a full raw dataset was also provided to enable a more detailed incident analysis. Readings have coordinates of up to 7 decimal places with a location accuracy estimate for each reading. Figure 86c shows the plot of these readings for comparison with the jittered coordinates. The morphology of the landscape is much more clearly defined. Occasional access to a granular dataset such of this could potentially be used to build an algorithm for improved jittering of more opaque datasets. However, for this study, randomised jittering is considered acceptable. The focus is on active spaces at the pixel scale.

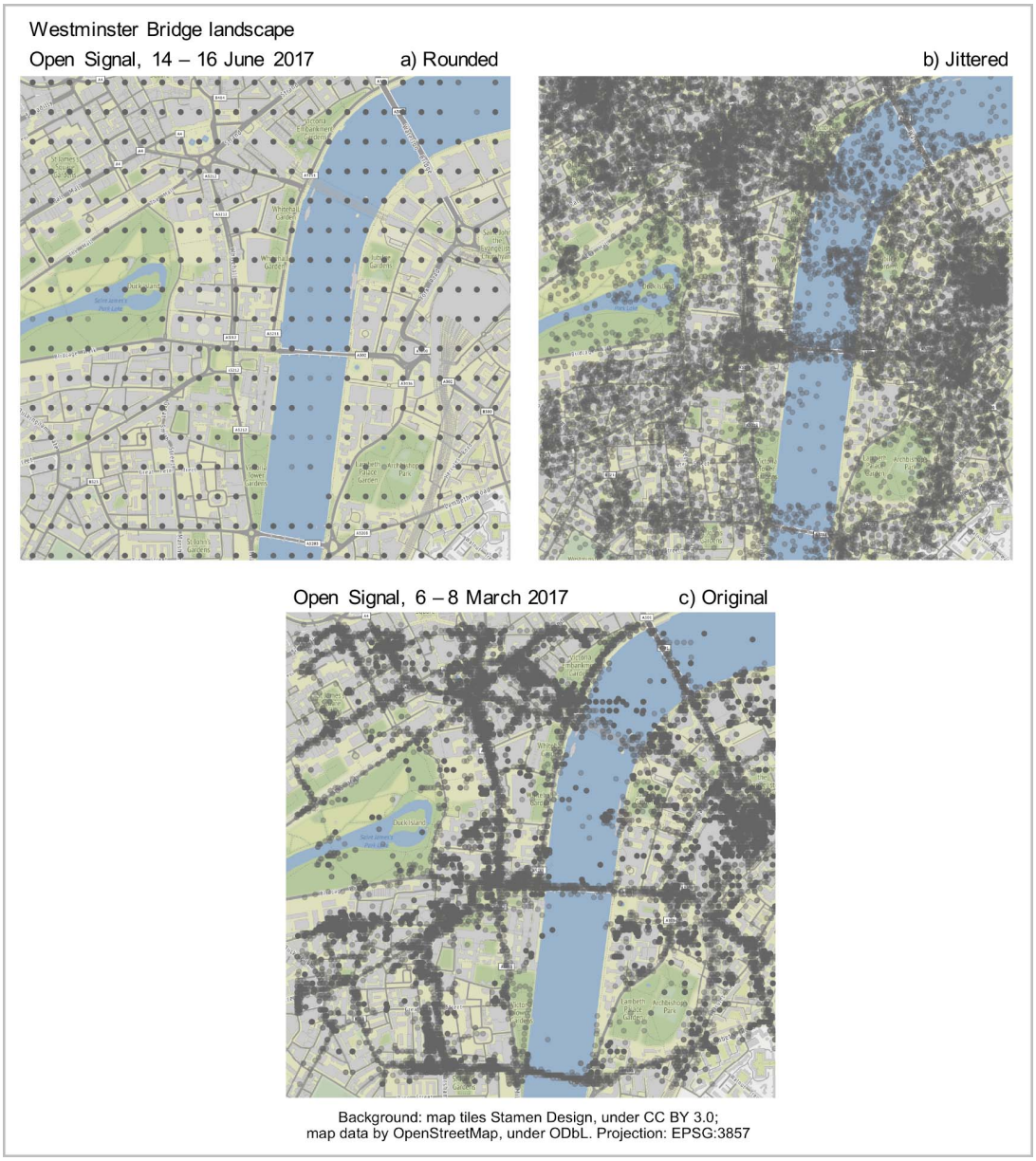


Figure 86. Mobile device readings with coordinates rounded, jittered and original

Clockwise from top left, a) coordinates plotted rounded to three decimal places; b) coordinates 'jittered' to 7 decimal places, randomized adjustment of +/- 0.000499; c) original coordinates to 7 decimal places. Figure a) and b) for 3 days in June 2017, figure c) for 2 days in March 2017.

Figure 87 shows the frequency distribution for location accuracy estimates for the March dataset. 77% of the readings have a location accuracy estimate of 50 metres or less. 82% have an estimate of 100 metres or less, below the 110-metre gap created when plotting coordinates rounded to three decimal places. The June 2017 dataset has a further limitation. Timestamps have been cut off at the hour and only a sample of readings retained across the hour for each device. This makes it difficult to calculate movements. The March 2017 dataset is complete for the Westminster landscape and has nearly five times more readings from half the number of devices compared with the June dataset. Table 21 summarises the data. To enable grid-based surface modelling, each reading is coded with the Pixel and LandScan cell where it is located. Counts are then aggregated and averaged per cell for spatial comparison.

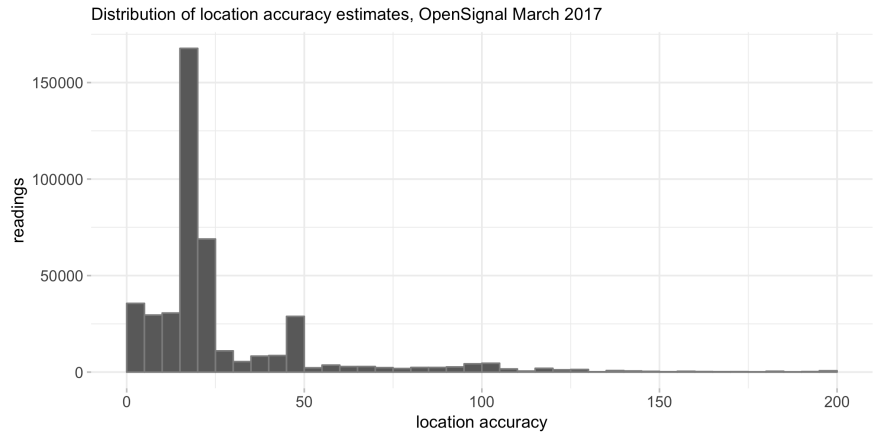


Figure 87. Distribution of location accuracy estimates, March 2017 data

Bins set at 5 metres for the frequency distribution of readings.

Table 21. Mobile data summary for modelling behaviour change

OpenSignal data	*Time period*	*Readings*	*Devices*
Greater London	1 – 28 June 2017	6,077,001	17,578
London Bridge landscape	1 – 28 June 2017	115,836	2,769
Oxford Circus landscape	1 – 28 June 2017	101,081	3,185
Westminster landscape	1 – 28 June 2017	105,093	3,497
Westminster landscape	2 – 29 March 2017	513,370	1,697

In all OpenSignal datasets, the device IDs are fully anonymised, with a different hash applied to each dataset. It is not possible to de-anonymise the device ID or compare presence across different months. However, the June 2017 dataset spans the Greater London area, making it possible to identify what proportion of devices are present in more than one of the three landscapes during that month, summarised as follows based on readings from 1 to 28 June 2017:

- Number of devices across all three landscapes: 4,896
- Devices present in all three landscapes: 749 (15%)
- Devices present in two landscapes: 1,422 (29%)
- Devices present in one landscape: 2,725 (56%)

Just over half of the devices visited only one of the three landscapes compared with 15% present in all three of the landscapes.

Each reading within each OpenSignal data set is given behaviour attributes for duration, trip and stage, as described in chapter three with code samples in Appendix B.2. As with previous studies, a new trip is assumed if two consecutive readings for the same device have a time difference of at least three hours. Trips are divided into stages based on movements between pixel cells. When the next consecutive reading for the same trip for a device is in a different pixel, a new stage of the

trip has begun. Stage durations can then be calculated to indicate if the reading represents an activity that is likely to be dwelling ‘D’, milling ‘M’, or moving ‘R’ whilst present in a pixel cell within the landscape. The method is simplified compared with chapter five. Figure 88 shows the method used to assign a behaviour to a stage based on its duration, depending on the timestamp detail.

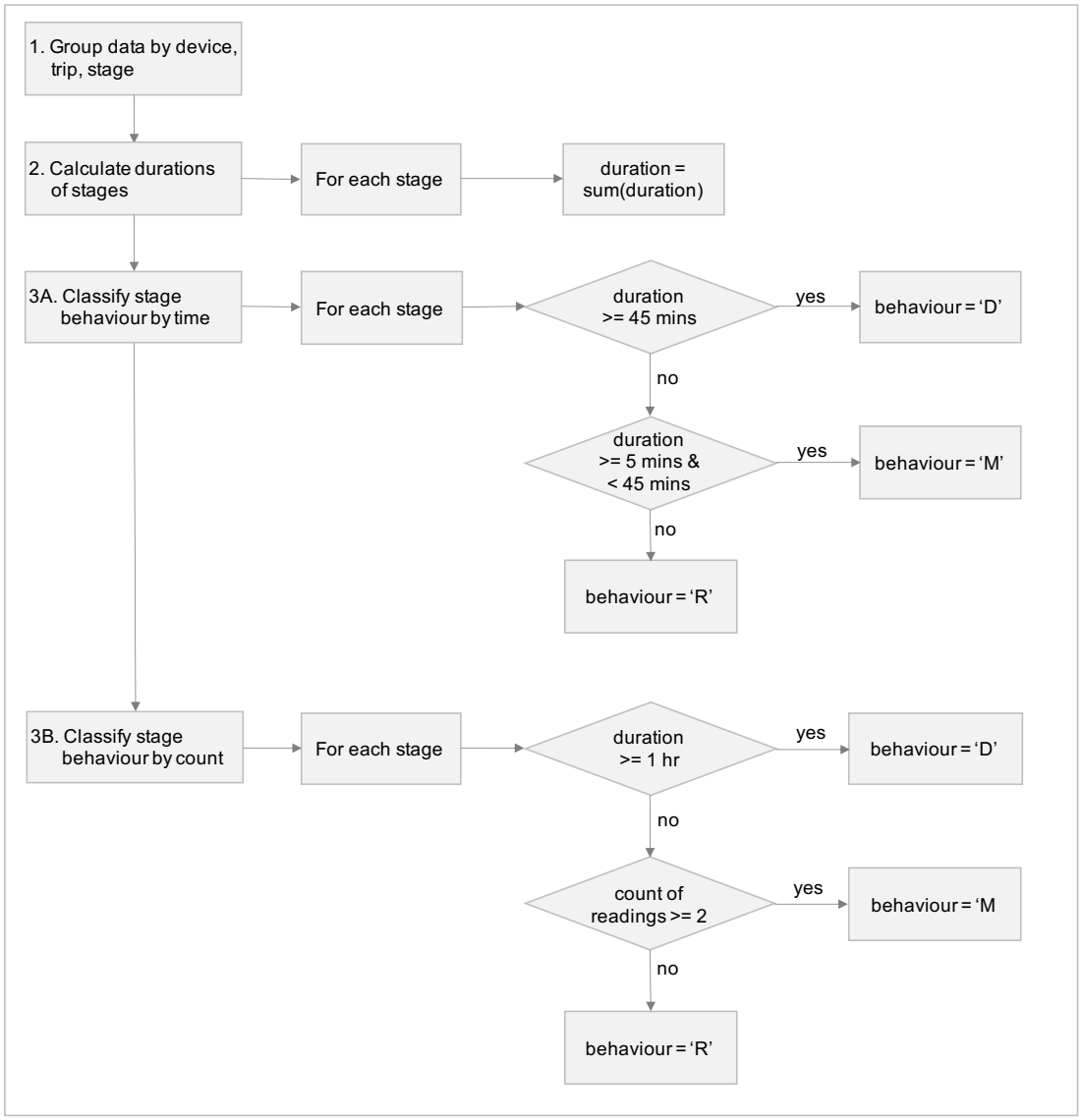


Figure 88. Classifying presence during a trip stage in London landscapes

Assumes individual readings have already been tagged with a trip and stage ID.

For this study, two data sources are analysed, one with a full timestamp and one with a timestamp shortened to the hour. For data with detailed timestamps, a device is assumed to be dwelling in the area if present for more than 45 minutes. If the device is present for more than 5 minutes and less than 45 minutes, it is assumed to be milling. If present for less than 5 minutes, the device is assumed to be moving. For datasets with the time stamp limited to the hour only, the device is assumed to be moving if the duration is more than 1 hour (i.e. the device has readings across two or more hours). If there are 2 or more readings within the stage, the device is assumed to be milling. Otherwise, the device is moving. This is a very crude and somewhat arbitrary method but provides some indication of different stage durations within a dataset containing limited time information.

For trip durations, the behaviour is simplified when the timestamp is limited to hourly and a slightly different classification is used when the full timestamp is provided (Figure 89).

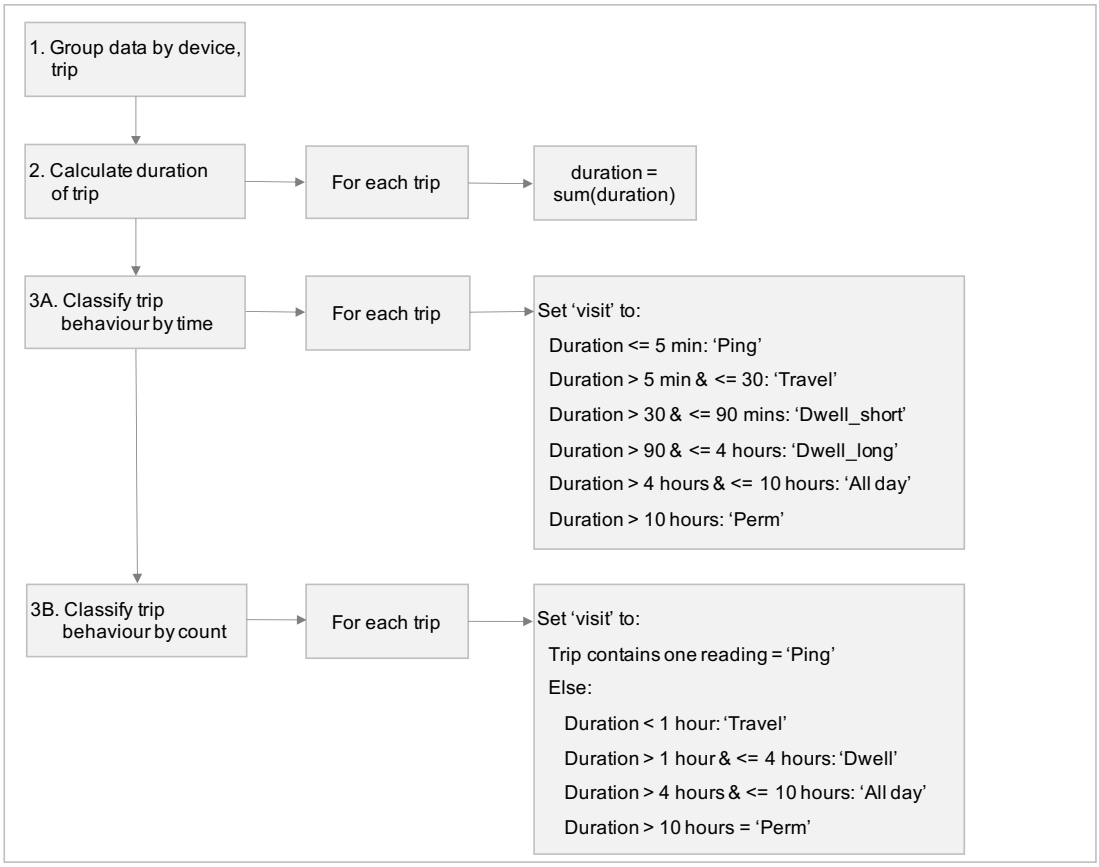


Figure 89. Classifying visit to landscape based on trip duration

Assumes data has already been tagged with trip ID.

For datasets with detailed timestamps, trips lasting 5 to 30 minutes are tagged as 'travel'. They are likely to be moving across the landscape but may also be milling briefly in places. Trips that last 30 to 90 minutes are short dwell visits, likely to be ad-hoc activities such as lunch breaks. Trips that last 90 minutes to 4 hours are longer dwell visits, likely to be due to a specific activity or event taking place in the landscape. Trips that last longer than 4 hours are likely to be due to habitual activities, such as visiting a place of work. However, they could also apply to all-day events that occasionally take place in urban landscapes. A simplified criterion is applied to data with timestamps limited to hour readings. Devices that have only a single reading or are present for too brief an amount of time to analyse presence and/or movement across the landscape (classified as 'ping' visits) are removed from the dataset. The same applies to devices that are present for too long to be visiting the landscape (classified as 'perm' visits).

7.2 Learning Socio-Spatial Dynamics

To provide an active population estimate at the time of a real-world phenomenon, the first step is to learn the socio-spatial dynamic of the landscape This requires defining the landscape and then using a sample of real-world observations to measure presence variations over time.

7.2.1 Landscape definitions

To compare three landscapes that each experienced a traumatic incident in the same year, a row of three LandScan cells is drawn across each landscape, with the centre cell containing the location of the incident. A Pixel grid is overlaid onto the LandScan grid but may also be extended beyond the LandScan boundary to study Pixel cells relevant to specific situations. Background map tiles are retrieved and labelled to aid orientation. The OpenSignal June data set is plotted, with coordinates randomly jittered, increasing the reading from three to seven decimal places.

7.2.1.1 Westminster Bridge

The first landscape is centred on Westminster Bridge and Westminster palace in Central London (Figure 90), where a terrorist attack occurred on Wednesday 22nd March 2017.

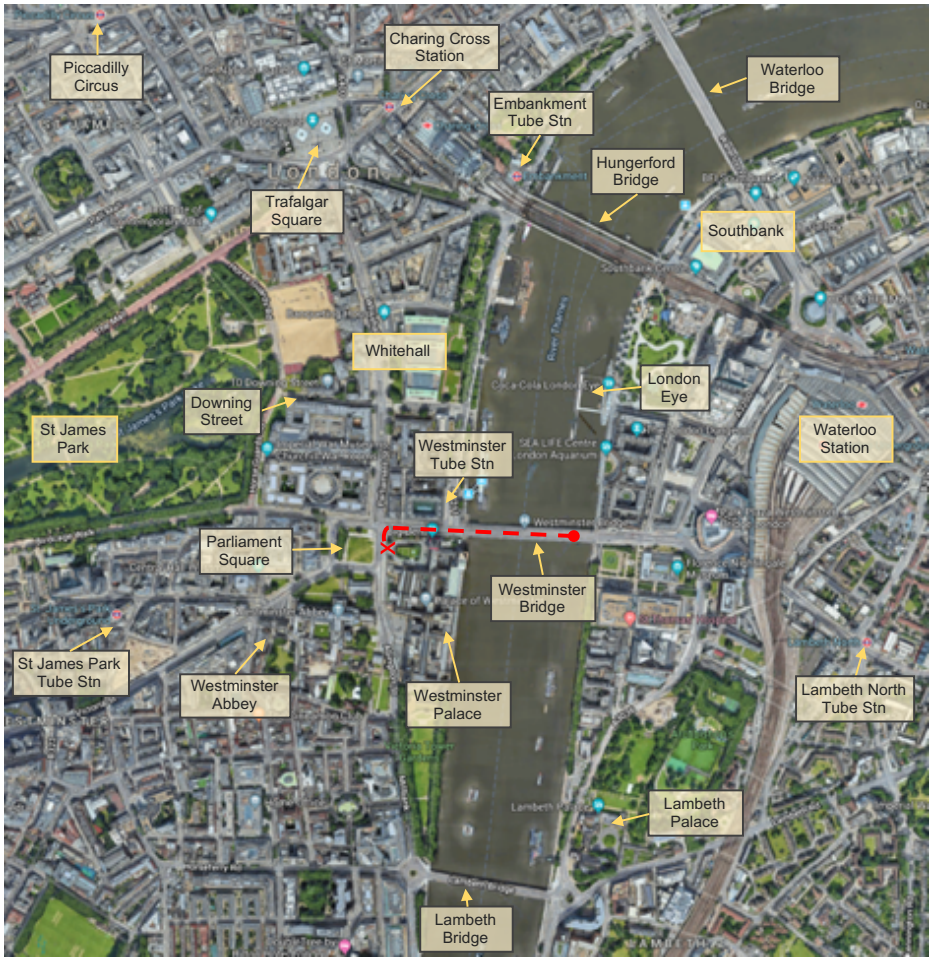


Figure 90. Westminster Bridge landscape, satellite map

Imagery © 2019 Google. Imagery © 2019 Bluesky, Getmapping Plc, Infoterra Ltd & Bluesky, Maxar Technologies, The GeoInformation Group. Map data © 2019 Google. Annotated with labels for key locations. Dotted red line indicates location of major incident. Circle indicates the start. Cross indicates the end of the incident.

The landscape includes several ‘underground’ (London Underground) and ‘mainline’ (London Overground and National Rail) stations. Waterloo station, providing both underground and mainline services, is located on the eastern side of the River Thames that flows from south to the north-east through the map tile. The mainline station is the busiest railway station in the UK, as of 2018, with an estimated 94 million entries/exits during 2017-2018 (Office of Rail and Road, 2018), an average of over 250,000 movements per day. On the eastern side of the river is an arts complex that is home to Festival Hall, various food, drink and entertainment establishments, and the London Eye tourist attraction. Several government buildings and tourist attractions are located on the western side of the river including Trafalgar Square and Westminster Palace, containing the Houses of Parliament and Big Ben. Various historic royal and government buildings are located near the two landmarks including Whitehall, home to Downing Street, Parliament Square and Westminster Abbey. The area is not high density in terms of building height but would be expected to have high footfall throughout the year, both as a result of government and commercial activities at the various buildings and from tourist interest in the area. Four bridges cross the River Thames within the map tile: Lambeth, Westminster, Hungerford and Waterloo (from south to north). Hungerford Bridge comprises three separate bridges, one for a railway and one either side of the railway for pedestrians. The other bridges each contain a roadway and footpaths as a single bridge.

Figure 91 shows the LandScan cells that cover the incident location. The Pixel grid is extended above and below the LandScan grid to enable a comparison of activity around bridges either side of the incident. The pixel cells containing the incident location are highlighted in purple (x6y4, x7y4 and x8y4 spanning London Bridge and Parliament Square). Three locations are highlighted in pink due to exhibiting clusters of activity: x5y7 (Trafalgar Square), x8y7 (Embankment) and x12y5 (Waterloo station). To study the incident impact, 4 further pixels are highlighted: cell x9y4 (eastern side of Westminster Bridge, cells x6y1 and x8y1 (western and eastern sides of Lambeth Bridge to the south); and, x10y6 (eastern side of Hungerford Bridge to the north).

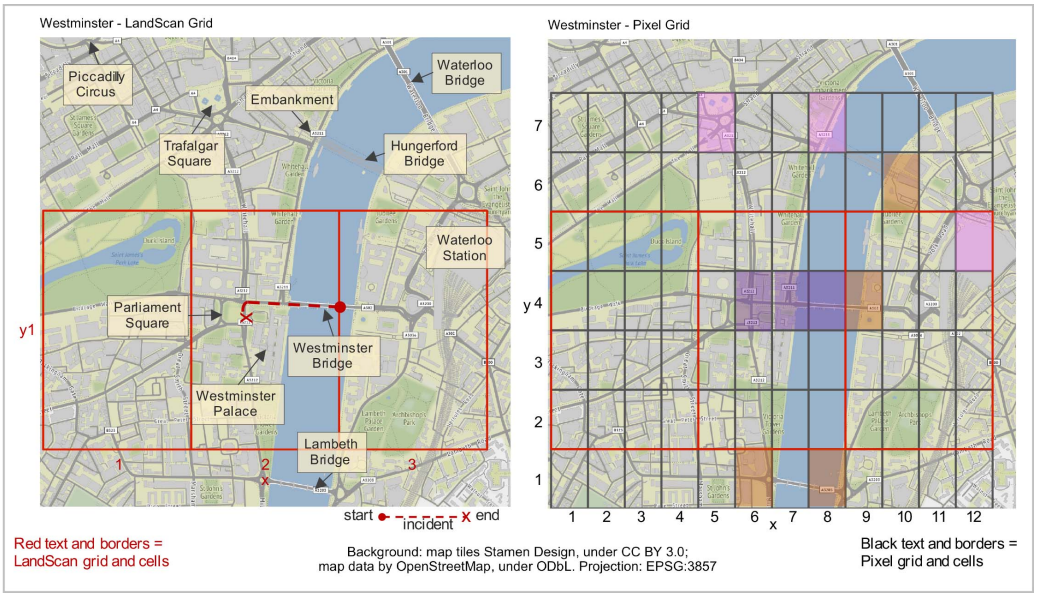


Figure 91. Westminster Bridge landscape data grids

a) The landscape contains three LandScan cells; b) Pixel grid extended above and below LandScan cells.

7.2.1.2 London Bridge

The second landscape is centred on London Bridge (Figure 92) where a terrorist attack took place on Saturday 3rd June 2017, beginning on the bridge and ending at Borough Market.

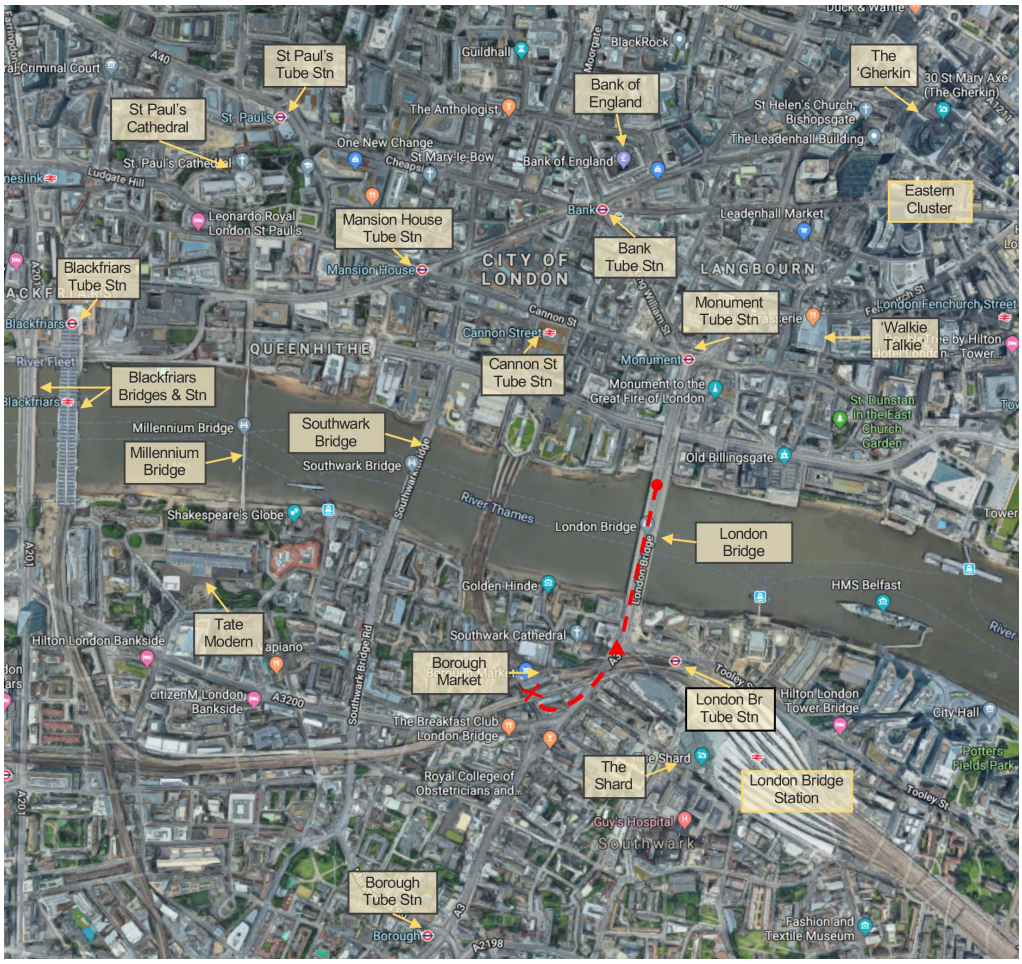


Figure 92. London Bridge landscape, satellite map

Imagery © 2019 Google. Imagery © 2019 Bluesky, Getmapping Plc, Infoterra Ltd & Bluesky, Maxar Technologies, The GeoInformation Group. Map data © 2019 Google. Annotated with labels for key locations. Dotted red line indicates location of major incident.

The north side of London Bridge is home to the City of London, also referred to as the 'Square Mile'. It is the original centre of London and home to its financial district. Several tourist attractions are within proximity on both the north and south sides of the bridge. South-east of London Bridge is London Bridge Station, providing both mainline and underground rail services. The mainline station is the fourth busiest railway station in Great Britain, behind London Waterloo, London Victoria, and London Liverpool Street. In 2017-18, there were over 48 million entrances to and exits from the station, an average of 131,500 per day (Office of Rail and Road, 2018). Above the station is The Shard, the tallest building within the UK, at the time of writing.

Figure 93 contains maps showing the LandScan cells that span the incident, also subdivided into pixels, with pixels of interest highlighted: Borough Market (x7y2), Borough High Street (x8y2) and the south side of London Bridge station (x8y3) that contain the incident location; London Bridge

tube station (x9y2) which appears to contain high footfall close to the incident; and, London Bridge north side (x9y4) for incident impact effects.

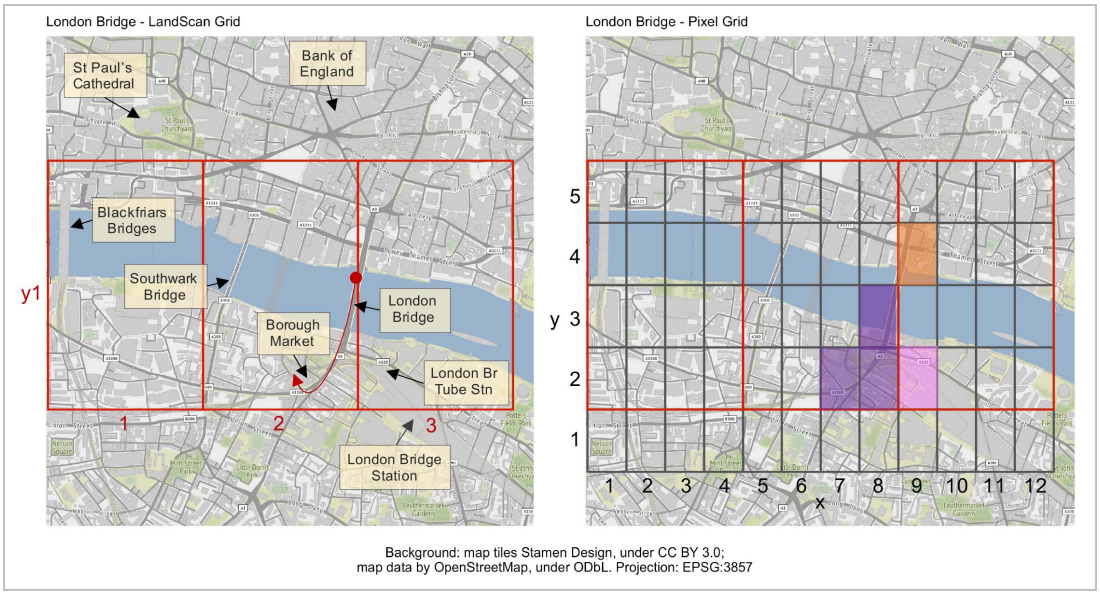


Figure 93. London Bridge landscape data grids

a) The landscape contains three LandScan cells; b) Three landmarks are identified (shaded cells) for comparison analysis at pixel scale. Red grid lines indicate LandScan cells. Black grid lines indicate Pixel cells. Incident indicated by red dotted line with arrow showing direction.

7.2.1.3 Oxford Circus

The third landscape is centred on Oxford Circus (Figure 94). A possible major incident was reported at the Oxford Circus tube station on Friday 24th November but was later found to be false alarm.

Oxford Circus is the junction between Oxford Street and Regent Street. The tube station has entrances on all four corners of the junction and is also an interchange between three different underground lines. Oxford Street is considered the busiest shopping street in Europe, according to a 2017 study that counted 13,500 pedestrians per hour from 14:00 to 16:00 on 10 June 2017 (BNP Paribas Real Estate, 2017). Standard shop opening hours differ from the typical 9am to 5.30pm. Some stores open one hour later at 10am and many stay open into the evening up to 8pm. Regent Street, which crosses Oxford Street at Oxford Circus, is the 9th busiest street in Europe, with approximately 9,000 pedestrians an hour according to the same 2017 study. Oxford Street runs West to East, with the Central underground line running beneath it with stations at the junction between Oxford Street and Tottenham Court Road (Tottenham Court Road station), and the junction with Bond Street (Bond Street station). Regent Street runs north from the Euston Road to south past Oxford Street before curving to the east at Piccadilly Circus and connecting with Leicester Square, the location of numerous theatres, clubs and restaurants.

Figure 95 contains a map tile of the landscape with LandScan cells drawn and the incident zone circled. As with the previous landscapes, pixels are highlighted in purple for the location of the incident (x6y2, x7y2). A neighbouring pixel containing Regent Street and the area behind Oxford Circus (x7y1) is highlighted in orange.

Figure 94. Oxford Circus landscape, satellite map

Imagery © 2019 Google. Imagery © 2019 Bluesky, Getmapping Plc, Infoterra Ltd & Bluesky, Maxar Technologies, The GeoInformation Group. Map data © 2019 Google.

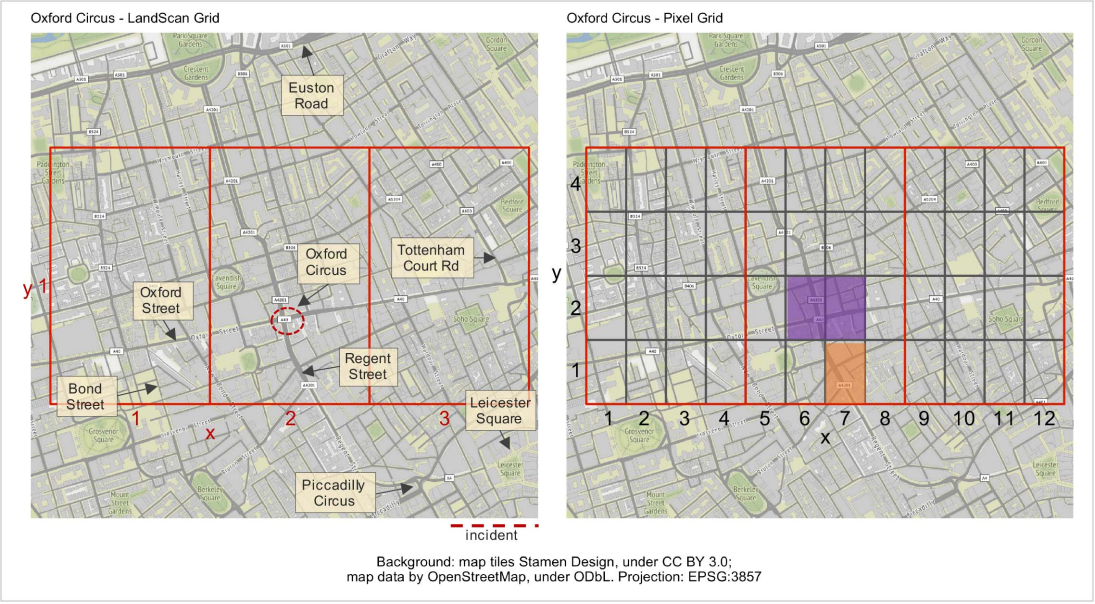


Figure 95. Oxford Circus landscape data grids

a) The landscape contains three LandScan cells; b) Three landmarks are identified (shaded cells) covering the incident location; Red grid lines indicate LandScan cells. Black grid lines indicate Pixel cells.

7.2.1.4 Spatial Affordances

The analyses for each landscape will focus only on the LandScan grid and, where relevant, Pixel cells above and/or below the LandScan grid. Figure 96 shows the OpenSignal June 2017 data points plotted for each map tile with the LandScan cells highlighted. They indicate the variation between LandScan cells both within and between landscapes. In two of the landscapes, the River Thames acts as a barrier. For Westminster Bridge, it is a barrier between cells x2y1 and x3y1, whereas, for London Bridge, it dissects each LandScan cell. Westminster Bridge has substantial green spaces visible whereas Oxford Circus has the least green space but has a higher proportion of space taken up by residential accommodation. These differences in physical environments affect the distribution of readings. All landscapes have visible clustering, emphasising that presence is not randomly or uniformly distributed in space. The next section explores how presence varies in space and time by aggregating readings within LandScan and Pixel cells.

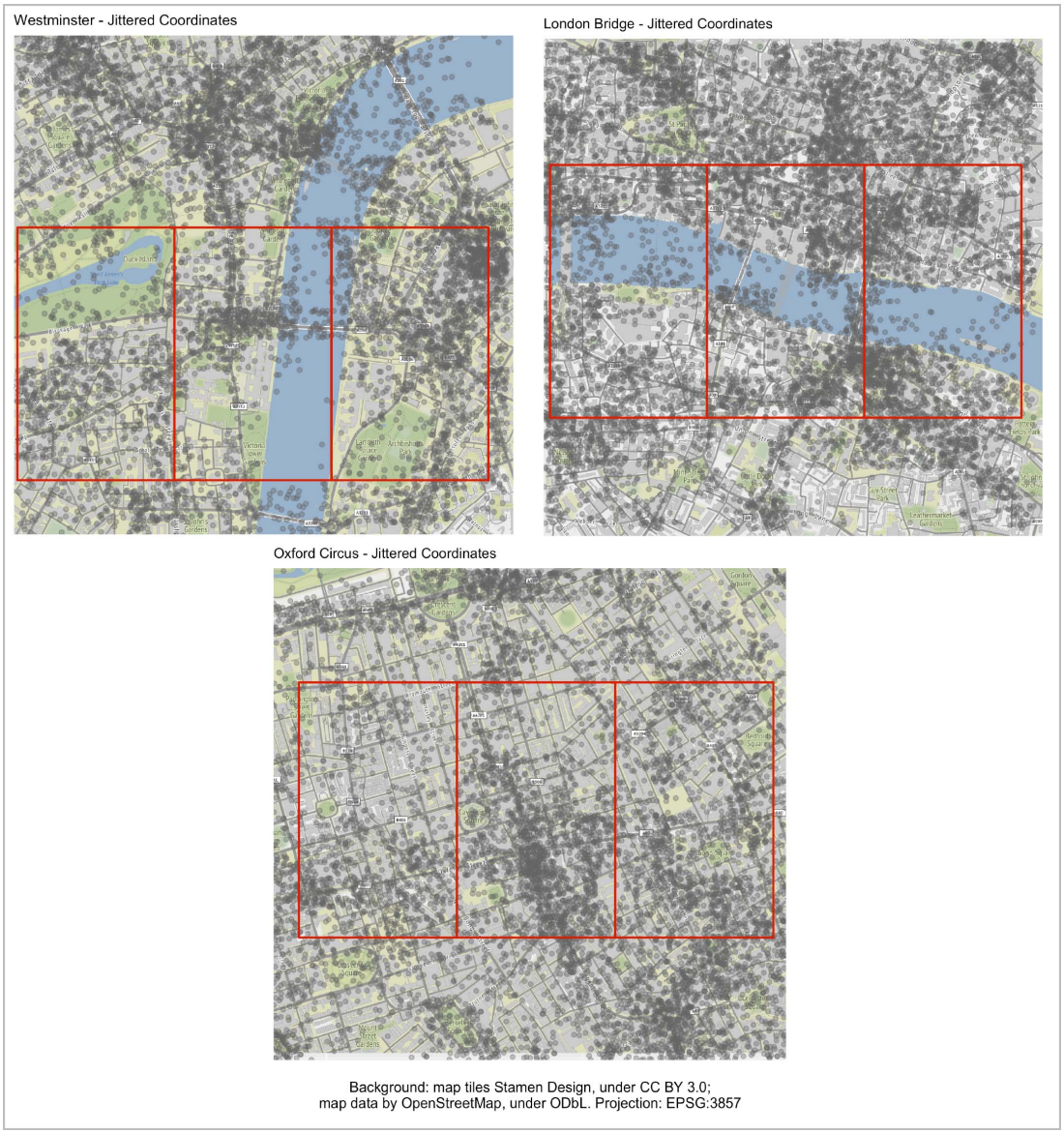


Figure 96. Points-based plots for three landscapes, June 2017

Points are coordinates with jittering applied to extend from 3 decimal to 7 decimal places, and semi-opaque to indicate density of readings.

7.2.2 Space-Time dynamics

To establish the ambient dynamic for each landscape, considered to be variations in population behaviours over time assuming normal conditions, the OpenSignal data set for June 2017 is used to measure presence and movement. Summary counts are provided in Table 22.

Table 22. June 2017 mobile data summary per landscape

Heading 1	*devices*				*readings (to nearest 1,000)*			
1 to 28 June	*tile*	*x1y1*	*x2y1*	*x3y1*	*tile*	*x1y1*	*x2y1*	*x3y1*
Westminster Bridge	3,497	1,021	1,165	1,489	105,000	11,000	12,000	20,000
London Bridge	2,769	1,259	1,400	1,490	116,000	16,000	21,000	21,000
Oxford Circus	3,185	895	1,304	1,358	101,000	12,000	18,000	21,000

xNyN references are for LandScan cells. Counts are for OpenSignal data. Tile = map tile (entire landscape).

Each landscape was visited by over 2,500 devices during the month with from 895 to 1,490 devices per LandScan cell. Each landscape has over 100,000 readings with counts per LandScan cell varying from under 12,000 to over 21,000. Figure 97 shows the daily counts of devices present from 1 to 28 June. Each landscape shows a noticeable variation between weekdays and weekends with the most visible difference in the London Bridge landscape and the least visible for Oxford Circus. A major incident occurred within the London Bridge landscape on 3 June 2017. To compare the landscapes, the temporal variations will be analysed using readings from 8 to 28 June 2017. This will be referred to as the 'ambient context'. No other incidents, abnormal conditions or large-scale events are known to have occurred during this time.

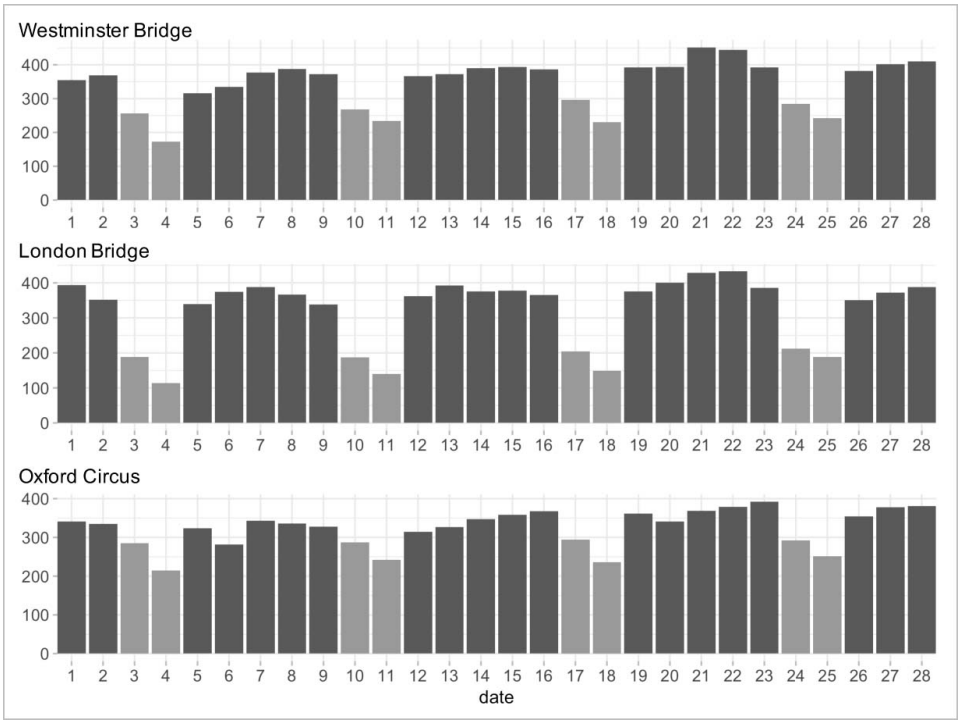


Figure 97. Day counts of mobile devices across each landscape, June 2017

Counts of unique devices present per date. Weekends highlighted.

7.2.2.1 Temporal variations

Variations in presence over time are explored at two scales: daily variations, comparing weekdays with weekends, and hourly variations through the day. For the latter, plots are focused on from 05:00 to 23:59 each day to incorporate daytime and evening activities in public space.

Daily changes

Figure 98 contains box plots showing the mean counts and variations for the number of devices present in each Landscape's LandScan cell by day of the week, and the mean counts scaled from the weekday average plotted as line charts. The counts and scaled weights are listed in Table 23. The weekday average is used as the baseline for comparing weekdays versus weekends and days of the week. This is to align with the LandScan ambient count that is based on weekdays.

Table 23. Daily device counts per landscape and LandScan cell, ambient context

Weekday average	*Westminster Bridge*			*London Bridge*			*Oxford Circus*		
	x1y1	*x2y1*	*x3y1*	*x1y1*	*x2y1*	*x3y1*	*x1y1*	*x2y1*	*x3y1*
Device count	86.13	85.00	145.93	124.53	154.47	161.07	70.33	108.75	124.13
Device Variation	*x1y1*	*x2y1*	*x3y1*	*x1y1*	*x2y1*	*x3y1*	*x1y1*	*x2y1*	*x3y1*
Wkday	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
Wkend	0.65	0.76	0.58	0.47	0.52	0.51	0.88	0.78	0.73
Mon	0.94	1.02	0.95	0.95	0.95	0.95	1.00	0.98	0.97
Tue	1.03	0.95	0.98	1.03	1.04	0.99	0.94	0.93	0.98
Wed	1.08	1.11	1.03	1.06	1.02	1.06	0.98	1.00	1.00
Thu	0.99	0.98	1.04	1.03	1.06	1.03	1.06	1.02	1.00
Fri	0.97	0.95	1.00	0.93	0.94	0.96	1.02	1.07	1.04
Sat	0.67	0.79	0.60	0.53	0.57	0.59	0.94	0.82	0.77
Sun	0.60	0.74	0.56	0.41	0.46	0.43	0.82	0.74	0.68

Weekday devices shows mean weekday device count for each LandScan cell. All other values are scaled by dividing the mean count for the category by the weekday mean (hence weekday is 1.0).

Referring to Figure 98, within the Westminster Bridge landscape, cells x1y1 and x2y1 have similar maximum daily counts although the distribution varies, with a bigger difference between weekdays and weekends in cell x1y1. Cell x3y1, containing Waterloo station, has a much higher count and a much more noticeable difference between weekdays and weekends. For the London Bridge landscape, the three cells exhibit similar distributions across the days of the week. Oxford Circus shows more variation on each day than the other landscapes. All have peak readings during the middle of the week, with lower readings on Mondays and Fridays and substantial drops in activity on weekends. The weekday counts may be reflecting recent cultural changes to office-based working practices in the UK. Organisations are now required to offer flexible working hours, with Friday most likely to be a day when people choose to work from home, followed by Mondays, particularly for those commuting long distances. It indicates the need to regularly recalibrate models and assumptions about urban behaviours.

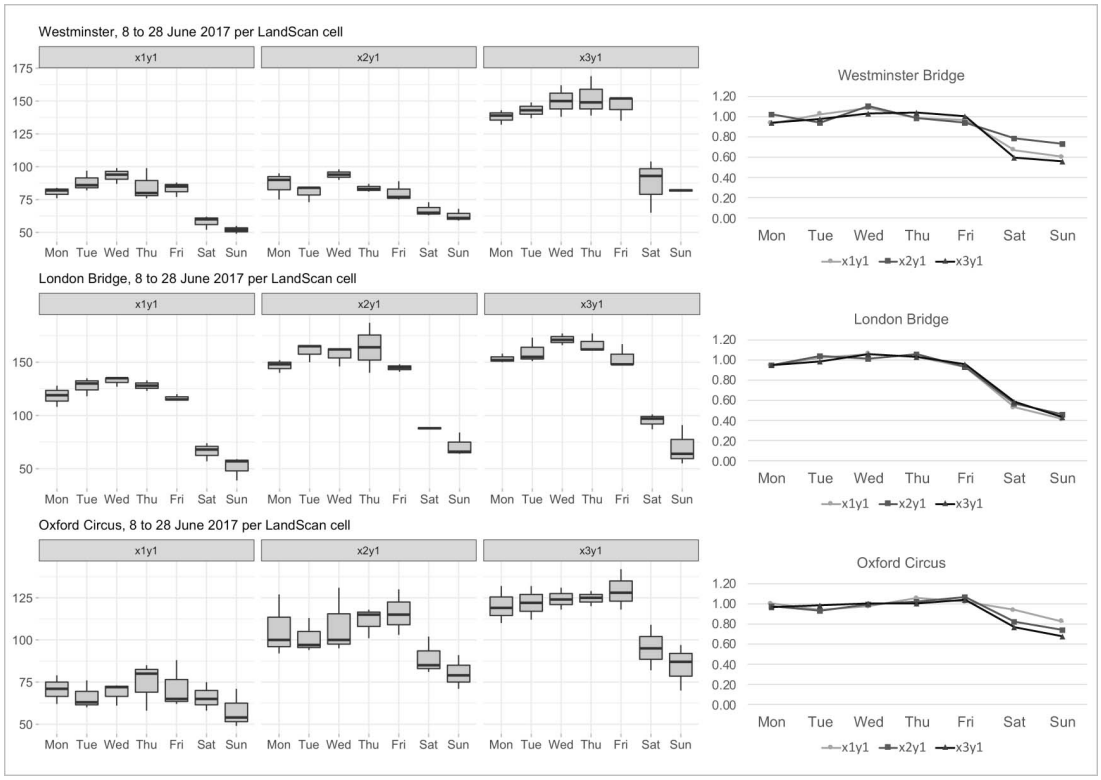


Figure 98. Plots showing device counts per day of week, ambient context

Mean device counts per LandScan cell in each landscape, as box plots showing range, and as line charts showing values scaled from the weekday average.

Cell x3y1 in the Westminster Bridge landscape has a closer match to cell x3y1 in the London Bridge landscape than with its neighbouring LandScan cell, both in terms of the mean weekday device count and the variation across the days of the week. Both cells contain a major transport hub. Similarly, cell x2y1 in the Westminster Bridge landscape has a closer match to cell x2y1 in the Oxford Circus landscape. Both contain a mix of residential and work locations. The results indicate that learning variations in behaviour for landscapes categorised by land-use potential may make it possible to infer population dynamics for an unobserved location if it shares sufficient physical features with an observed location. Whilst beyond the scope of this thesis, the potential is discussed in chapter eight under future research directions.

Another finding is the effect of a river passing through the landscape on neighbouring cells. In the Westminster Bridge landscape, the bridge appears to act as a barrier – readings are very different when comparing cells x2y1 and x3y1. In the London Bridge landscape, the bridge cuts across each cell and thus does not affect an aggregated scale when comparing the two neighbouring cells.

LandScan is an ambient estimate of the number of people present on a working weekday. There is no mention within the LandScan documentation as to whether or not people who briefly enter and leave a landscape are counted or not. As discovered in chapter five, producing day weights will likely result in conservative estimates of active presence at any given time for landscapes due to variation through the day. Hourly weights produced more realistic variations in activity levels. The range in readings potentially indicates the number of trips that are necessary versus optional, adopting Gehl's definition of different reasons for visiting a location (Gehl, 1987).

Hourly changes

Figure 99 shows the mean counts per hour per LandScan cell for each landscape for the ambient context, separated by weekday versus weekend and scaled using min-max normalisation from the weekday hour average to enable a comparison between cells and weekdays versus weekends for each landscape. The red line indicates the weekday hour average. Figure 100 shows the actual counts for individual dates (grey bars) and actual mean counts across all dates (red line) for each landscape, separating and comparing weekend dates with weekend dates to give an indication of how much variation there is above and below the mean for each hour of the day.

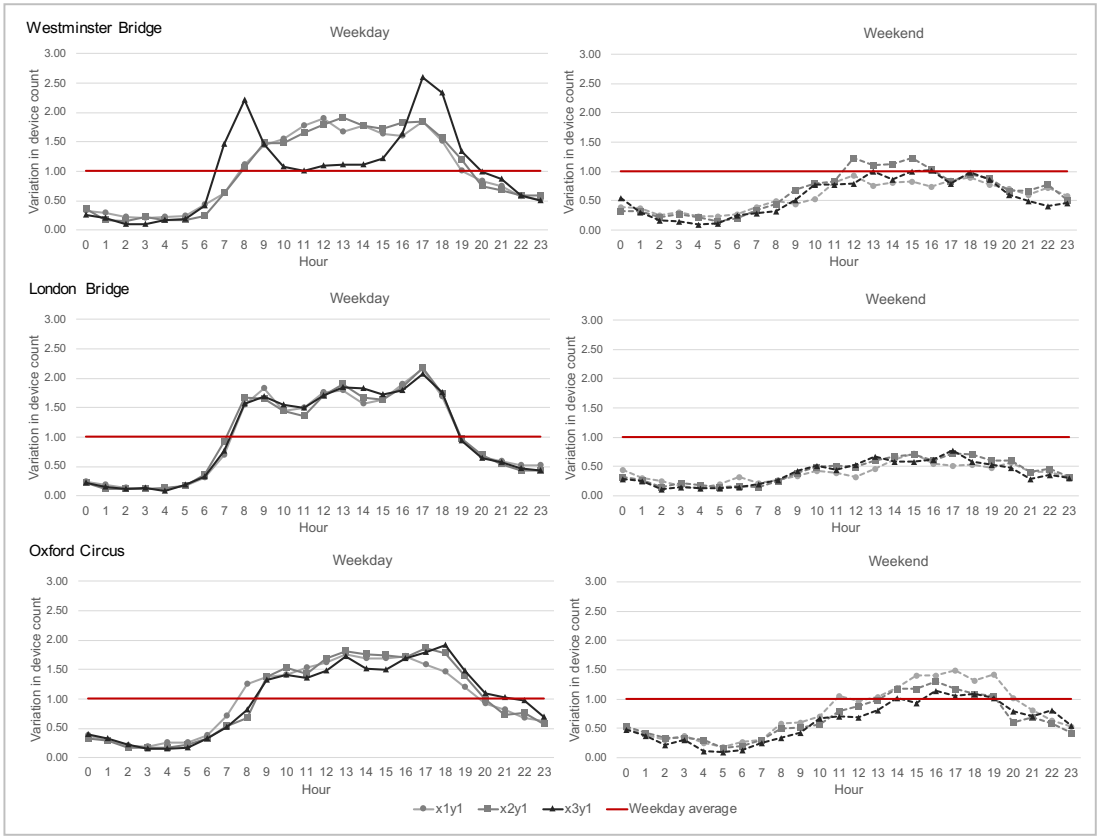


Figure 99. Weighted mean device counts per hour, ambient context

Horizontal red line indicates weekday mean per hour. All plots show hour variations from the weekday mean per landscape. All landscapes have the same y-axis scale for comparison.

From viewing the mean hourly counts per LandScan cell in each Landscape, it is immediately visible that weekends have very different profiles to weekdays both in hourly volumes and in the distribution of presence through the day. Two of the landscapes exhibit similarities across all three cells, with London Bridge showing the most similarity between cells. The Westminster Bridge landscape has a very different profile for one LandScan cell compared with the others. Both Westminster Bridge and Oxford Circus have at least one hour of the weekend that is equivalent to the weekday mean whilst London Bridge is visibly much quieter at weekends. It is a strong indication that reliance on any static measure generalised over time is likely to produce ineffective population estimates.

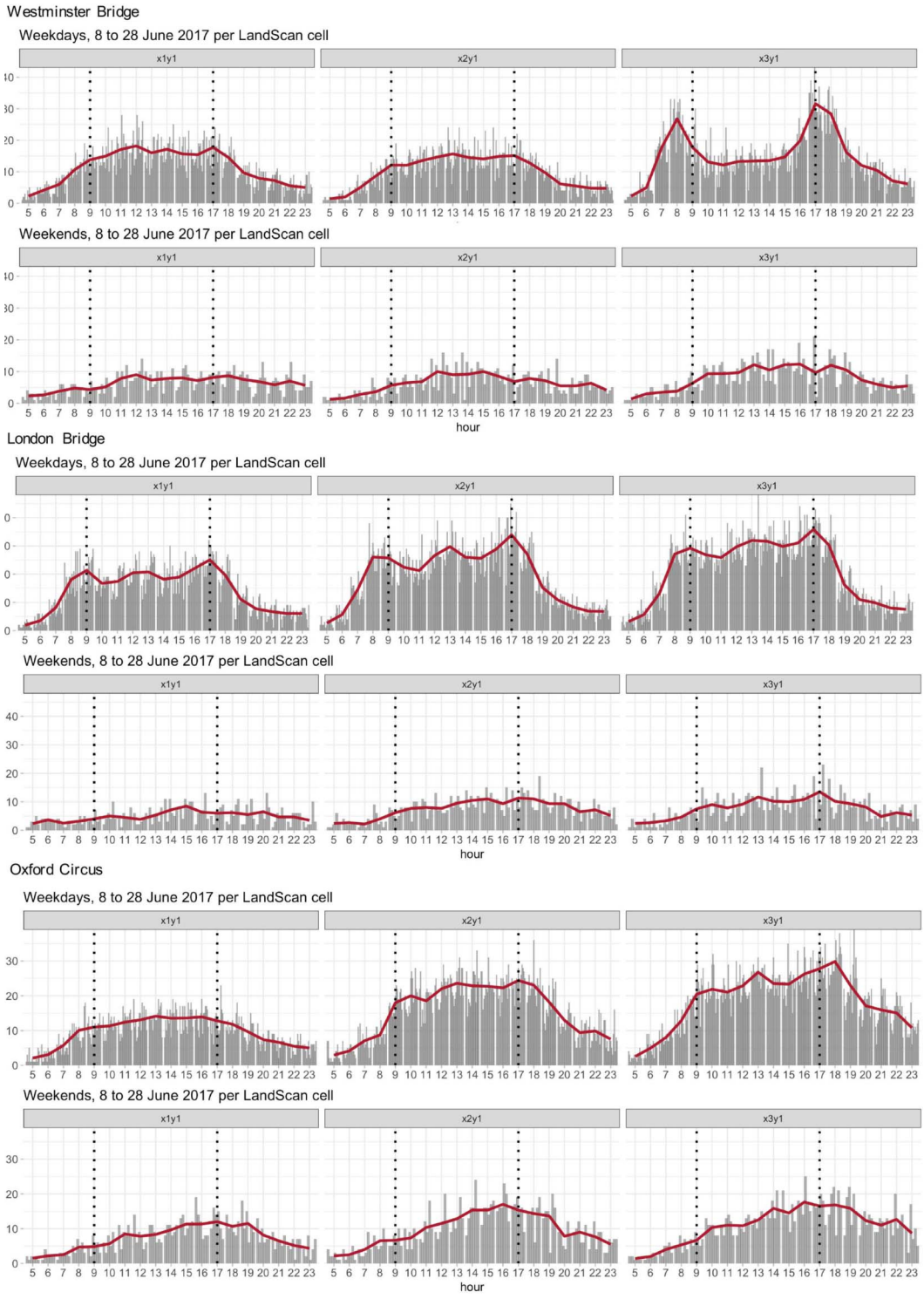


Figure 100. Mean device counts per hour per LandScan cell, ambient context

Grey bars are actual hour counts per date; red line is mean across all dates per hour; vertical dotted lines indicate 9am and 5pm. Hours plotted from 5am to 11pm.

Table 24 summarises the mean device counts and hourly weights for weekdays and weekends across each landscape and LandScan cell. These values are used to estimate an active population for a specific time of day, identifying whether or not it is a weekday or weekend. This is for the ambient context established during June. It assumes no abnormal conditions and does not consider seasonal variations.

Table 24. Ambient context hourly device counts per landscape and LandScan cell

	Westminster Bridge			*London Bridge*			*Oxford Circus*		
Wkday	*x1y1*	*x2y1*	*x3y1*	*x1y1*	*x2y1*	*x3y1*	*x1y1*	*x2y1*	*x3y1*
Device	9.62	8.17	12.15	11.64	15.65	17.34	8.06	13.05	15.56
Scaled	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
Hour weight	*x1y1*	*x2y1*	*x3y1*	*x1y1*	*x2y1*	*x3y1*	*x1y1*	*x2y1*	*x3y1*
05:00	0.24	0.18	0.19	0.17	0.17	0.19	0.26	0.22	0.16
07:00	0.62	0.62	1.46	0.70	0.92	0.76	0.71	0.54	0.51
09:00	1.43	1.48	1.46	1.83	1.64	1.68	1.36	1.38	1.32
11:00	1.77	1.66	1.00	1.50	1.36	1.49	1.54	1.42	1.35
13:00	1.66	1.92	1.10	1.79	1.90	1.84	1.75	1.81	1.72
15:00	1.63	1.72	1.21	1.63	1.64	1.72	1.69	1.74	1.50
17:00	1.84	1.85	2.60	2.15	2.17	2.07	1.59	1.87	1.79
19:00	1.00	1.19	1.33	0.96	0.98	0.94	1.20	1.39	1.48
21:00	0.75	0.67	0.86	0.58	0.53	0.57	0.82	0.72	1.02
23:00	0.52	0.58	0.50	0.52	0.43	0.43	0.62	0.58	0.69
Wkend	*x1y1*	*x2y1*	*x3y1*	*x1y1*	*x2y1*	*x3y1*	*x1y1*	*x2y1*	*x3y1*
Device	5.63	5.39	6.87	4.52	6.61	6.78	6.31	8.75	9.54
Scaled	0.58	0.66	0.57	0.39	0.42	0.39	0.78	0.67	0.61
Hour weight	*x1y1*	*x2y1*	*x3y1*	*x1y1*	*x2y1*	*x3y1*	*x1y1*	*x2y1*	*x3y1*
05:00	0.24	0.15	0.12	0.20	0.15	0.14	0.19	0.17	0.09
07:00	0.40	0.35	0.29	0.21	0.14	0.19	0.31	0.31	0.26
09:00	0.45	0.69	0.52	0.34	0.39	0.43	0.60	0.51	0.43
11:00	0.81	0.84	0.77	0.39	0.51	0.45	1.05	0.79	0.71
13:00	0.76	1.10	1.00	0.46	0.61	0.67	1.03	0.98	0.80
15:00	0.83	1.22	1.00	0.73	0.70	0.58	1.41	1.18	0.93
17:00	0.85	0.84	0.80	0.52	0.72	0.78	1.49	1.18	1.06
19:00	0.78	0.88	0.86	0.47	0.60	0.54	1.43	1.05	1.02
21:00	0.61	0.67	0.49	0.40	0.42	0.28	0.81	0.69	0.71
23:00	0.59	0.51	0.45	0.30	0.33	0.31	0.54	0.42	0.56

Device is the mean hour count of devices across 24 hours. All other values are scaled from the Weekday device count (i.e. weekends are scaled by the weekday average). Every second hour from 05:00 listed.

A final observation is that there appear to be a limited range of curves (Figure 99) such as two peaks for commuting periods, three less-intense peaks when there is also a lunchtime effect, and sustained daytime activity, albeit reduced, at weekends. It is beyond the scope of this thesis but a potential future research direction would be to evaluate whether there is a finite range of time curves and whether or not they are associated with land-use attributes. This could enable a context-specific framework to be applied to an unobserved landscape if its land-use potential is known.

7.2.2.2 Active spaces

The LandScan scale shows localised variations in temporal presence but is too large to study street-level spatial behaviours and potentially masks localised clusters, as discovered in chapter five. To examine for active spaces, a pixel grid is used. Each pixel is 250 metres tall by 155 metres wide. This approach has been selected in preference to detecting data-driven clusters to compare the three landscapes using a standard scale.

Spatial presence

Figure 101 shows the spatial variation in device counts on weekdays and weekends across the pixel grid spanning the LandScan cells for each landscape. For each image (weekdays and weekends per landscape), the presence counts per cell have been scaled across the LandScan grid using z-score standardisation to emphasise high (red) and low (blue) presence. White indicates mean values. Approximate locations of underground and/or mainline stations (yellow circles) are indicated for orientation.

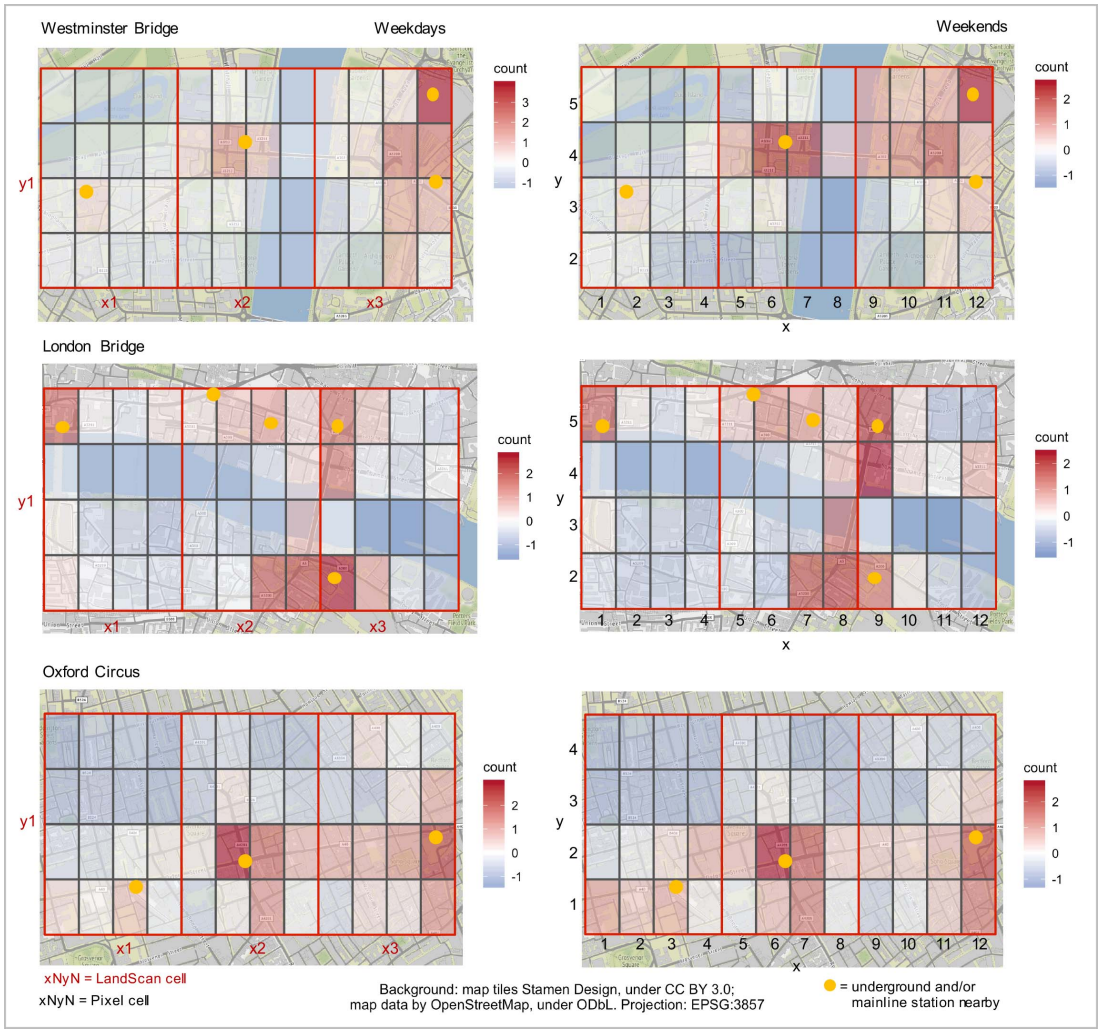


Figure 101. Spatial variation in daily count of unique devices, ambient context

Pixel counts scaled using z-score standardisation for each grid. 0 (white) indicates mean. Red indicates standard deviations above mean. Blue indicates below mean. Scaling applied across each grid (i.e. all three LandScan cells). Note: Westminster Bridge and London Bridge pixel refs start from y2 (first row is outside LandScan grid).

Clustering at pixel-level within the LandScan cells is visible, as is differentiation in cluster formation from weekdays to weekends. For example, within the Westminster Landscape, the majority of readings are located on the eastern side of cell x3y1. In cell x3y1 on weekdays, the pixel containing the entrance to Waterloo station, pixel x12y5, accounts for 17% of presence. The combination of pixels x12y5 whilst cells x12y4 and x11y14 each account for 11% of presence across the LandScan cell. Combined, 39% of presence is concentrated within those three pixels on weekdays, declining to 30% of the weekend count. In LandScan cell x2y1 on weekdays, 18% of readings are located within pixel x6y4, 17% are within pixel x7y4 and 11% are within pixel x5y4. Combined, 46% of weekday readings are concentrated in the three pixels. At weekends, the percentage of presence remains the same in pixels x5y4 and x6y4 with readings in x7y4 increasing to 17%. However, the actual counts are of more interest. Mean device counts in pixel x12y5 drop from 50.87 on weekdays to 25.67 on weekends. Counts in pixel x6y4 also decline, but from 43.87 to 30.00. The weekend count in pixel x6y4 is higher than in pixel x12y5. In all three landscapes, clustering is most likely to occur in a pixel containing a tube or railway station. This is not surprising but indicates that knowledge about entries and exits to tube stations may also provide data about population dynamics. This possibility is revisited in the analysis of the Oxford Circus incident later in this chapter. Another observation is that a change in travel preferences away from using tube stations would have a substantial effect on the street-level population. This could occur if the vision of autonomous electric vehicles is realised within cities, enabling a shift towards personalised and demand-responsive road-based public transport.

The Oxford Circus landscape perhaps most highlights the need to consider public space movements in population estimates. Activity is concentrated in the cells that span Oxford Street as well as the roads that intersect with it including Bond Street, Regent Street and Tottenham Court Road (see Figure 94 and Figure 95 for landscape details). The pixels of the LandScan cells that are dominated by residential and low-density non-residential buildings have the lowest activity, despite likely registering the highest readings if relying on residential population statistics.

Spatial behaviours

Figure 102 shows readings aggregated per Pixel by behaviour: dwelling, moving and milling with readings scaled across each set of LandScan cells. It enables a comparison between behaviours and between weekdays and weekends for each behaviour. In the Westminster landscape, the pixel containing Waterloo station (x12y5) dominates on weekdays for all three behaviours. However, at weekends, the dominant location for dwell behaviour is on the eastern riverbank (x9y5). In the London Bridge landscape, Pixel x9y2, containing the Shard, has the highest dwell count whilst pixels x7y2 and x8y2 have the highest count for movement. In the Oxford Circus landscape, whilst the cells containing the Oxford Circus junction (x6y2 and x7y2) have the highest readings in total, the readings are dominated by movement and milling. Pixel x9y2 has the highest dwell count.

Whilst the spatial variations based on different behaviours could be interesting to explore further, the counts for the dataset become very small at this level of granularity. This direction of analysis would benefit from access to more comprehensive data. However, it does indicate the possibility

not only of providing a real-time population estimate but also an indication of the types of activity taking place. A population that is predominantly in motion, travelling between locations, may react differently to an intervention compared with a population dwelling at an attraction.

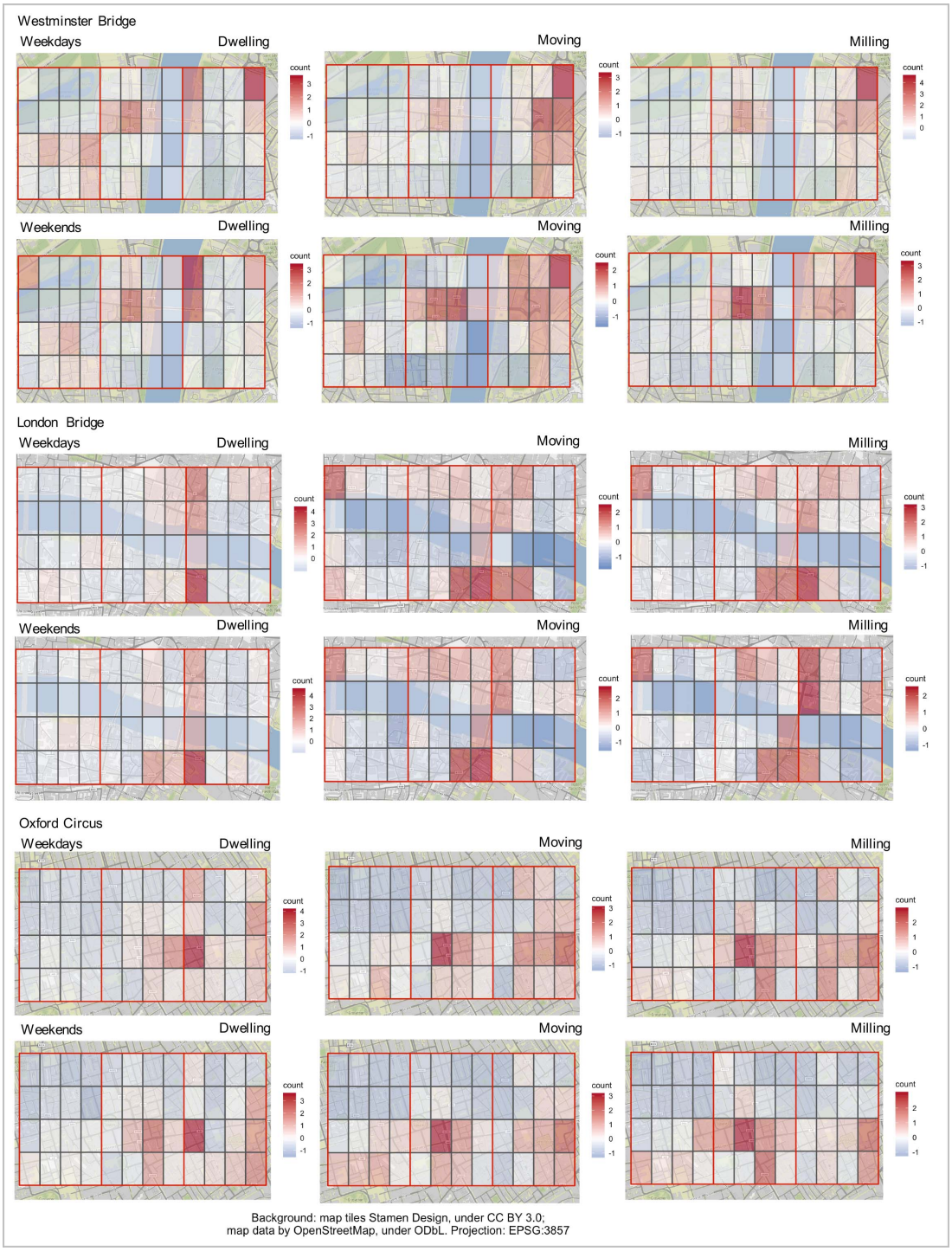


Figure 102. Spatial variation in presence, ambient context

Values scaled using z-score standardization. Red indicates above mean. Blue indicates below mean.

The visualisations of hourly variations in presence within each LandScan cell and the distribution at the Pixel scale within LandScan cells indicate the limitations of using a static population measure. In no LandScan cell are readings dispersed evenly and each exhibits a different clustering pattern,

mimicking the findings of the QEOP landscape in chapter five. Furthermore, there is a variation between weekdays and weekends that is not reflected in an ambient average such as LandScan, even before considering seasonal or situation contexts.

7.2.2.3 Seasonal variation

For one landscape – Westminster Bridge – a second OpenSignal dataset was provided, covering the month of March in 2017 when the incident being analysed occurred. The dataset also allowed evaluation of potential seasonal effects. Earlier research revealed a substantial seasonal variation in presence within open public space at the QEOP, based on the number of devices connecting to the Wi-Fi network over 12 months (see chapter four).

The March dataset was provided for a square mile spanning Westminster Bridge. The first LandScan cell – x1y1 has data cut-off prematurely on the western side that will affect readings slightly. This is visible in Figure 86c. This dataset retained full GPS coordinates to 7 decimal places. The morphology of the landscape is much more clearly defined than when jittering coordinates that have been rounded to 3 decimal places. Also, the full dataset is provided with readings to sub-second level. The result is a much larger dataset for the month – over 500,000 readings across the map tile compared with 85,000 within the landscape for the June dataset.

Daily changes

Plotting the daily count of devices per day in March (Figure 103) highlights possible data retrieval issues. Readings on Thursday 30 are substantially lower than the equivalent day of the week throughout the rest of the month and there are no readings for Friday 31. Wednesday 1 exhibits a much lower reading than all other Wednesdays. These dates are excluded from analysis, leaving four weeks from 2 to 29 March 2017. Readings for Wednesday 29 March are also visibly lower than previous Wednesdays which goes against the pattern in previous weeks. However, the 28th also goes against trend whilst the Monday has the highest count and Saturday 25th is unusually high. These dates are retained for analysis.

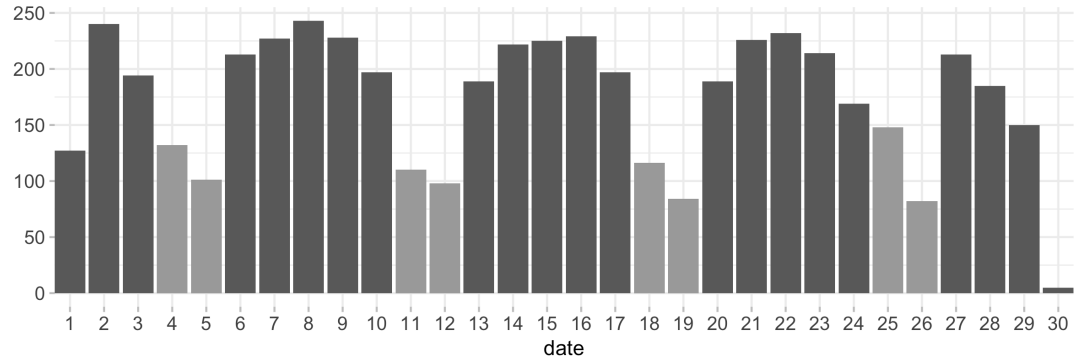


Figure 103. Daily OpenSignal device counts in Westminster landscape, March 2017

Count of devices present on each date within the landscape. Weekends highlighted.

There were no school, public or religious holidays during this period, although the Easter school holiday did begin on Friday 31st March. The nearest weather sensor with archived data available

for March 2017 is located at London City Airport[9] and registered ambient weather conditions. There were no extreme weather events and the temperature remained at or above historical averages. It was unseasonably warm at the middle and end of the month, and mostly dry throughout. Researching news archives for activities on Saturday 25th March revealed that a 'Unite for Europe' march, protesting against the UK vote to leave the European Union took place in central London, beginning at Park Lane and ending at Parliament Square in front of where the attack concluded just three days previously. It was estimated between 25,000 and 100,000 people took part in the march. A minute's silence was observed for the people killed in the attack[10].

Trimming the dataset to 2 to 29 March provides 503,911 readings generated by 1,672 devices in the landscape. Whilst the readings are much higher than the June data set, the number of devices present is nearly half in March. One of the challenges of using secondary data sources is that the apps may undergo updates or promotional campaigns to increase participants. It is not known if the difference is seasonal or due to the mobile app. One of the findings already revealed in earlier chapters has been that any reality data set will need to be regularly recalibrated if it is to be converted into an actual estimation of presence. It seems unlikely for such a difference to be just seasonal variation. However, the study of the QEOP landscape, based on connections to its Wi-Fi network, also registered a near 50% increase in devices present in June compared with March 2017. That two completely different and unrelated data sources for two different landscapes within the same city both register such a change when comparing the same two time periods suggests they may be revealing pedestrian, tourist and (planned) event sensitivity to seasons, activities that are more likely to occur during the summer months. However, it could be a coincidence and there is no data available to verify independently of the two sources.

A summary of the data for each LandScan cell is included in Table 25. This data set is used both for a seasonal evaluation and the incident analysis. To avoid the impact of the incident affecting seasonal analysis, an ambient context period is set for three weeks from 2 to 21 March.

Table 25. Mobile App data summary for Westminster landscape, March 2017

LandScan cells	*x1y1*	*x2y1*	*x3y1*
Dataset (2 to 29 March 2017)			
OpenSignal Readings	87,317	63,798	90,765
Open Signal Devices	463	562	764
Incident date (22 March 2017)			
Open Signal Readings	4,735	2,750	3,762
OpenSignal Devices	50	54	84
Mean device count, Ambient context (2 to 21 March)			
Day average	40.14	36.67	73.79

[9] Weather source: https://www.wunderground.com/history/monthly/gb/london/EGLC/date/2017-3
[10] News source: https://www.rt.com/uk/382309-thousands-london-brexit-protest/

LandScan cells	*x1y1*	*x2y1*	*x3y1*
Weekdays	47.47	40.47	88.37
Weekends	21.83	27.17	37.33

Figure 104 contains box plots for devices present in each LandScan cell by day of the week for 2 to 21 March. Even though the device count is half the size, the variations between weekdays and weekends mostly match the pattern established in the ambient context during June 2017. The most significant difference is that Mondays and Fridays are visibly lower than the rest of the week in all cells and that the cell containing tourist attractions (cell x2y1) is lower than both its neighbours during the week. Scaling the values based on the weekday average shows the difference in the weekday versus weekend for each cell. The overall curve of the chart matches the ambient context but with reduced readings on Mondays and Fridays. It is possible that the difference in device counts is seasonal and that the increase of tourism and/or outdoor activities in the summer affects the daily variation in presence for locations with outdoor actions and attractions, from pedestrian commuting to outdoor leisure activities, such as visiting establishments and attractions on the waterfront on the eastern side of the river in cell x2y1. Culturally, the latter are more likely to occur from Thursday to Saturday. Similarly, the opportunity to work from home could be more popular during the cooler months. Both could explain the reduced readings on Monday and Fridays during March compared with June. One consideration for future work would be to use the findings to inform a qualitative study, prompting questions about behaviours observed rather than assumed. Referring back to the literature, Gehl proposed that spatial interactions can be necessary or optional (Gehl, 1987). The range evidenced on each day could be an indication of when and where substantial optional activities occur.

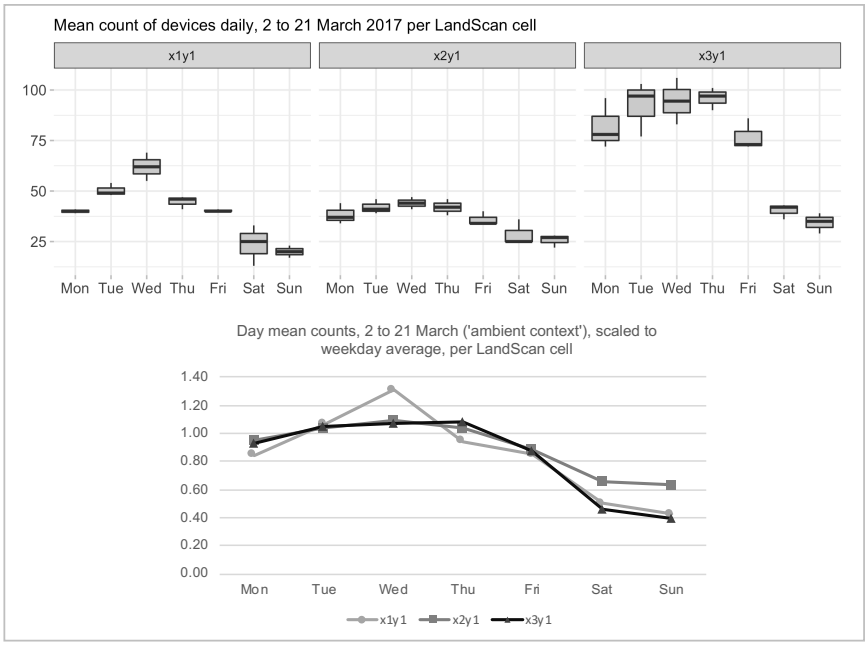


Figure 104. Weekday device counts per LandScan cell, Westminster, March 2017

Counts are for means for readings from 2 – 21 March 2017 to avoid incident effects. Box plots are of actual device counts. Line chart shows mean count per day scaled from weekday average.

Hourly distributions

Plotting the LandScan cell hourly variations (Figure 105) produces plots that are mostly similar to those produced for the ambient context. However, there is a noticeable daytime difference in cells x1y1 and x2y1. During March the three peaks during the morning commute, lunchtime and afternoon commute are more pronounced than for June. For cell x3y1, the increases during peak am and peak pm are higher than in June. It suggests the reduction in volume during March is more noticeable during off-peak hours. This further reinforces the appearance that there is a seasonal and tourism effect having some impact on population behaviours in each cell.

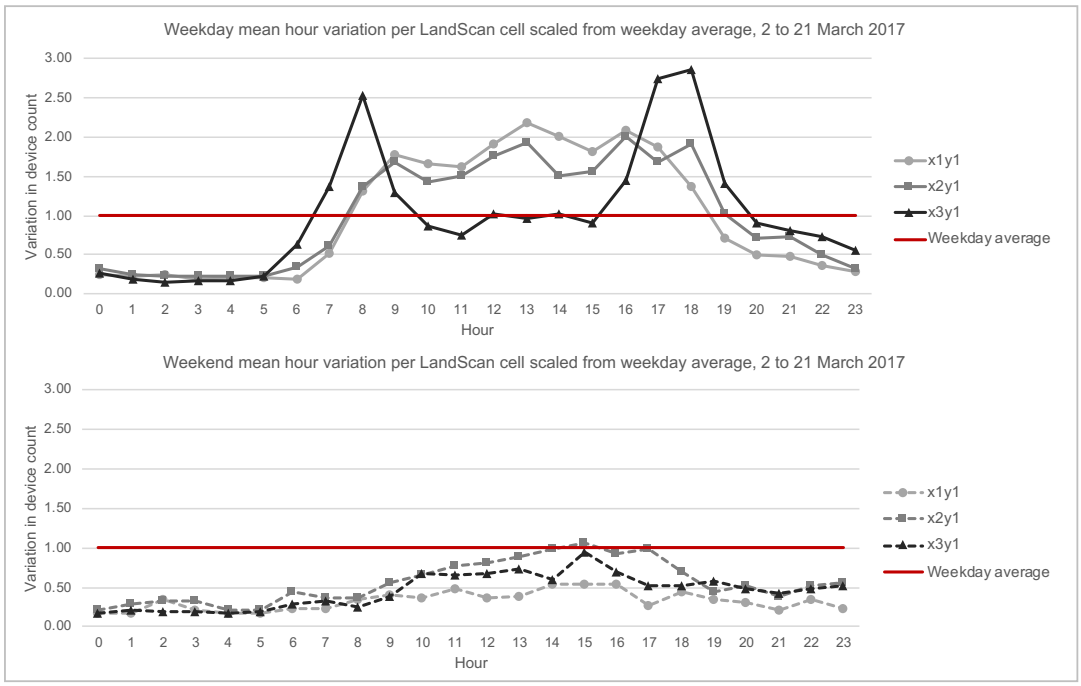


Figure 105. Hourly variations from weekday average, Westminster, March 2017

OpenSignal devices present per hour, weighted for comparison with the weekday hour average, calculated from readings from 2 to 21 March 2017 per LandScan cell.

Spatial distributions

The spatial distribution of presence in March (Figure 106) mostly mimics the June distribution (Figure 101). Waterloo station (x12y5) has the highest actual count on weekdays but is overtaken by Parliament Square (x6y4 and x7y4) at weekends. Parliament Square is also more pronounced at weekends in March than in June. However, this is due to lower readings elsewhere. Within the two Parliament Square pixels, the actual counts are 16.5 and 15.5 respectively, half the count in June. The cells on the eastern bank of the Thames are less prominent at weekends in March than in June (pixels x9y4 and x9y5). These two pixels contain large outdoor seating areas in front of bars and restaurants that would be expected to be more populated during warmer weather. It is further indication that the presence data contains sensitivity to temporal habits that are affected by seasonal expectations and potential differences caused by whether activities being undertaken are necessary or optional. Whilst the actual presence counts are reduced, the locations that attract presence mostly remain the same with outdoor dwell spots for food and drink seeing larger reductions than tourist attractions and landmarks.

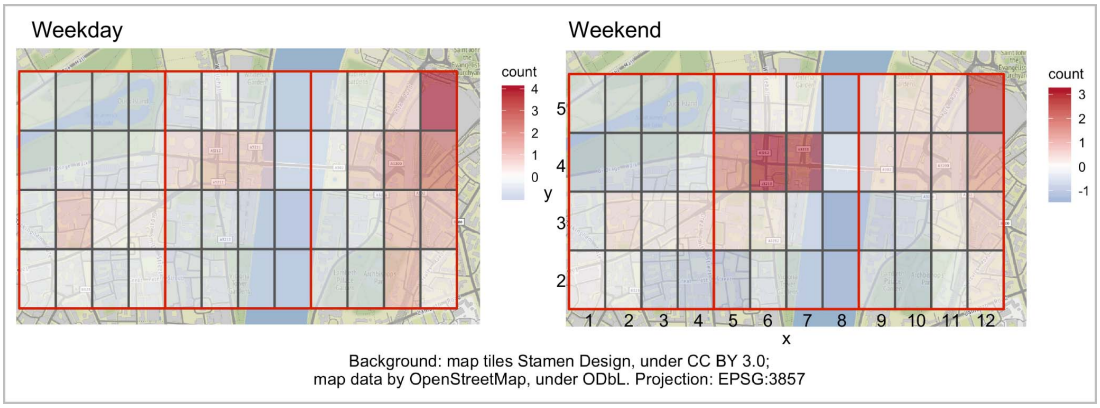


Figure 106. Average device count per pixel, Westminster, March 2017

Mean device counts per weekday and weekend, scaled using z-score standardization across the Pixel grid. Scaling is per dataset (weekend readings are lower than weekdays).

7.2.3 Urban rhythms

Referring back to the context hierarchy in chapter three (Figure 13), physical features provide the first level of variation in behaviour when comparing landscapes and cells within landscapes. The affordances of a landscape determine the capabilities for interactions to occur. One consideration is whether or not a location is more similar to its neighbour or a remote location sharing similar land-use potential. To summarise, the Westminster landscape has one LandScan cell that is likely to be dominated by travel behaviour (x3y1 containing Waterloo station), one cell that is likely to be influenced by tourism (x2y1 containing Westminster Palace and Whitehall), and one that is likely to exhibit diluted daytime behaviours due to mixed land use potential (cell x1y1, partly covered by a large park). The London Bridge has two cells that appear to share a cluster of activity associated with the London Bridge tube station that sits on the border of cells x2y1 and x3y1. The two cells also each contain a portion of the City of London district including several recently developed skyscrapers. The Oxford Circus landscape has the least amount of open space although it also has few high-density buildings. The landscape also has the widest variety of land-use potential across the LandScan cells, with residential properties to the north, entertainment to the south, retail and leisure in between as well as numerous offices, and several tube stations.

The Oxford Circus landscape highlights an issue with cultural assumptions about behaviour. It is anticipated that more people will be active at street-level during warmer months. Tourism within the UK is at its highest during the summer month of July (see Appendix A for details). However, public spaces that attract retail activity, such as Oxford Street, will experience increased footfall during holiday periods, such as the build-up to Christmas, when temperatures are typically much cooler and would normally deter pedestrian activity.

As well as defining the anticipated motivations for visiting a landscape, the spatial context of the framework provides the static baseline for estimating the population size. Table 26 contains the LandScan 2017 population estimates for the LandScan cells in each landscape. Before considering temporal variations, at least one cell looks problematic. Cell x3y1 in the Westminster Bridge landscape has the lowest population estimate yet contains the busiest railway station in the UK.

Even though the duration of presence may be short for the majority of visitors, at any moment during the daytime it is likely to contain substantially more people than indicated by the LandScan estimate. As with the findings from studying the QEOP landscape in chapter five, the estimates appear to be biased towards built locations containing either residential or office buildings compared with open public spaces, retail locations and transport hubs. However, OpenSignal may be biased in the reverse given public Wi-Fi networks are more likely to be operational in these locations. It highlights the difficulties in measuring real-world phenomena.

Table 26. LandScan 2017 ambient population estimates

Landscape	*cell x1y1*	*cell x2y1*	*cell x3y1*
Westminster Bridge	8,222	7,790	1,135
London Bridge	10,708	12,046	6,040
Oxford Circus	7,337	7,495	8,883

Analysing three different landscapes using the same data source revealed variations in presence over space and time for ambient conditions. For the one landscape with a second dataset for a different time of year, a seasonal effect was evidenced similar to that found in the seasonal study of the QEOP in chapter four. In March, the number of devices visiting the landscape was half the number visiting in June the same year.

Table 27 converts the LandScan ambient estimate into active estimates for a range of hours for the LandScan cells containing each incident. The LandScan population estimates are provided in Table 26. The weights are presented in Table 24. To emphasise the uncertainty in these estimates, they are rounded to the nearest 500. The active population can vary substantially based on the day of the week and the time of day, before considering effects such as seasonal variation and environmental conditions, with each cell having a different distribution through the day. The potential to produce even a crude estimate of the active population could help inform responses to unexpected conditions as they emerge.

Table 27. Landscape active population estimates at different hours, ambient context

Landscape LandScan cell	*7am*	*9am*	*1pm*	*3pm*	*5pm*	*7pm*
Westminster x2y1						
Weekday average	5,000	11,500	15,000	13,500	14,500	9,000
Weekend average	2,500	5,500	8,500	9,500	6,500	7,000
London Bridge x2y1						
Weekday average	11,000	20,000	23,000	20,000	26,000	12,000
Weekend average	1,500	4,500	7,500	8,500	8,500	7,000
Oxford Circus x2y1						
Weekday average	4,000	10,500	13,500	13,000	14,000	10,500
Weekend average	2,500	4,000	7,500	9,000	9,000	8,000

Estimates rounded to nearest 500, created from hour weights generated from OpenSignal data (Table 24) multiplied by LandScan baseline for the cell (Table 26).

7.3 Analysing Unexpected Incidents

As mentioned in the chapter introduction, each of the three landscapes in this study experienced major incidents during 2017. This section will consider whether or not knowing population dynamics could have provided information to assist those responding to, at-risk from, or affected by the incident. The second consideration is whether anything can be learned from the impact the incident had on population movements afterwards. Whilst the term is unfortunate in this context, continuous generation of real-world observations enables 'opportunistic data science' (Miller, 2017), the study of unexpected and unpredictable real-world phenomena.

The incidents detailed here were traumatic events in which people lost their lives. The purpose of this study is to consider whether knowledge about socio-spatial dynamics could inform immediate interventions. For example, in such an event, routes may be closed and people may need to be sheltered or directed to move away from the area. Knowledge of the likely population present could help inform such instructions. Whilst providing an overview of the incident for reference, the focus is on the environment in which the incident occurred rather than the incident itself.

7.3.1 Westminster Bridge

On Wednesday 22nd March 2017, a terrorist attack occurred on Westminster Bridge (Figure 107) in London. Reported as starting at 14:40 (H.M. Coroner, 2018), the duration of the attack was 82 seconds. Four pedestrians and one police officer died, and 49 other people were injured.

Figure 107. Photo of Westminster Bridge, London, 23 March 2017

Image taken just after the bridge was reopened following the day after the terror attack. View is facing west, with Westminster Palace (containing the Houses of Parliament) and Elizabeth Tower ('Big Ben') visible to the left. Image source: Flicker, shared under CC-BY-ND 2.0 https://www.flickr.com/photos/dgeezer/33229942940.

7.3.1.1 Estimating the active population

Based on the findings from the previous section of this chapter, it is possible to estimate details about the population within the pixels and LandScan cells where the incident occurred. The June 2017 OpenSignal dataset provides an ambient socio-spatial dynamic for the landscape. The March 2017 dataset provides a seasonal adjustment. It also contains the date when the incident occurred, enabling a study of the impact of the incident on population dynamics in the area afterwards.

Table 28 contains an estimate of the active population for the hour before the incident in LandScan cell x2y1y. It shows the best available static estimate, the LandScan 2017 ambient count, and the number of devices within the OpenSignal (OS) dataset that represents the equivalent to LandScan, the weekday hour average. The active estimate for a weekday from 13:00 to 14:00, the hour before the incident occurred, is calculated based on the weighting available for the ambient context established using the June 2017 dataset (Table 24). A seasonal adjustment is then calculated given we also have a March 2017 dataset for this landscape. Finally, the actual readings are taken from the hour before the incident, 13:00 to 14:00 on Wednesday 22nd March to compare.

Table 28. Estimating active population, Westminster Bridge, 22 March 2017

LandScan (LS) cell x2y1	*OS devices*	*Weight*	*Population*
LandScan 2017 ambient count			7,790
OS weekday hour average (June 2017)	8.17	1.00	7,790
OS active population estimate			
weekday, 13:00 to 13:59 (June 2017)	15.67	1.92	15,000
seasonally adjusted for March 2017	8.71	1.07	**8,500**
Actual readings (OS)			
13:00 to 13:59 on Wednesday 22nd March	**12.00**	1.49	**11,500**

OpenSignal (OS) estimates are rounded to nearest 500.

If we used just the LandScan 2017 ambient count as our population estimate, it would be 7,790. However, that count is regardless of the time of day. As shown in the previous section (Figure 99), activity in cell x2y1 in the Westminster landscape peaks at 13:00, declining slightly to 15:00 before increasing again towards 17:00, but remains well above average during the working day. Using the OpenSignal ambient dataset (June 2017) as weightings, the estimate nearly doubles when including the time of day. However, the March 2017 dataset suggested that visits may be nearly half during March, although the temporal distribution remains the same (Figure 105). Applying a seasonal adjustment brings the count close to the LandScan ambient count at 8,500.

The readings generated in the hour before the incident (13:00 to 13:59 on Wednesday 22nd March) produce an actual count of 11,500. The number of devices present was above the average across 2nd to 21st March. Whilst the number of unique devices is small at this scale, meaning just a single addition or subtraction could have a large effect on the count, it does indicate how much variation there can be in the volume of human activities at different times throughout the day, and throughout

the year. Had the incident occurred in June, it is likely a lot more people would have been at risk or affected by the incident. Even so, it occurred at the worst possible time of day for the location.

7.3.1.2 Evaluating the incident impact

Access to a data source that was continuously available both before and after the incident allows evaluating whether or not the incident had an impact on population movements. Given the severity of the attack, it would be reasonable to assume that people may be deterred from visiting the area, at least in the near term. Figure 108 shows plots of mobile device activity on a) the day before and b) the day after the incident. The visual difference in the volume of readings suggests that presence in the area the day after was affected by the incident.

Background: map tiles Stamen Design, under CC BY 3.0;
map data by OpenStreetMap, under ODbL. Projection: EPSG:3857

Figure 108. Westminster Bridge data maps, incident evaluation

Red circle indicates the area of the incident.

Figure 109 shows readings at the Pixel scale for the pixels highlighted in the landscape definition (see Figure 91): the three bridges along with Parliament Square, Trafalgar Square and Waterloo Station (Embankment is included with Hungerford Bridge).

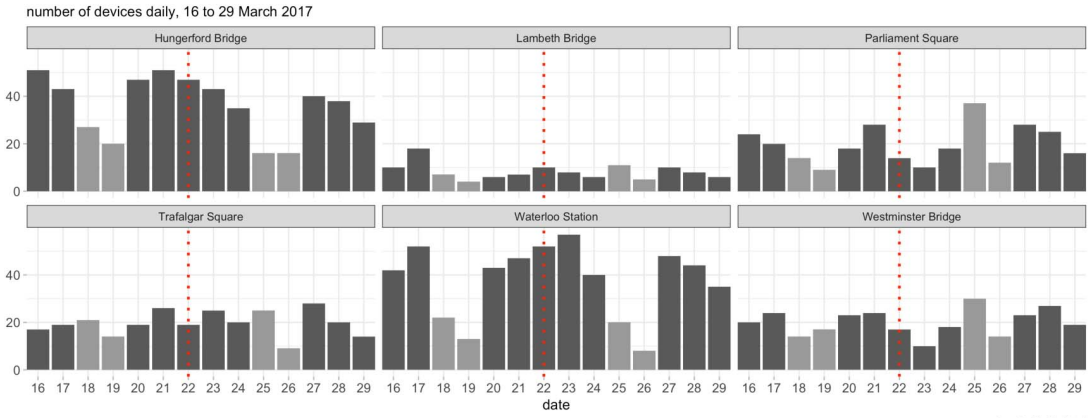


Figure 109. Daily device counts pre- and post-incident, Westminster Bridge, March 2017

The image shows counts from 16 to 28 March. Dotted vertical line indicates incident date. Weekends highlighted.

At Pixel scale, there is a reduction in presence at Parliament Square and on all three bridges. However, the readings are within the range of variation across the month except for Westminster Bridge and the effect lasts just one day. Furthermore, Parliament Square registers its highest count on Saturday 25th, just three days later. This is likely due to the protest march that was scheduled and went ahead. It suggests that people were not deterred from the location after the incident other than the immediate aftermath when routes were closed.

Studying the hourly distribution (Figure 110) shows a substantial increase at 5pm on the day of the incident for Hungerford Bridge, potentially indicating that people switched to it from other routes.

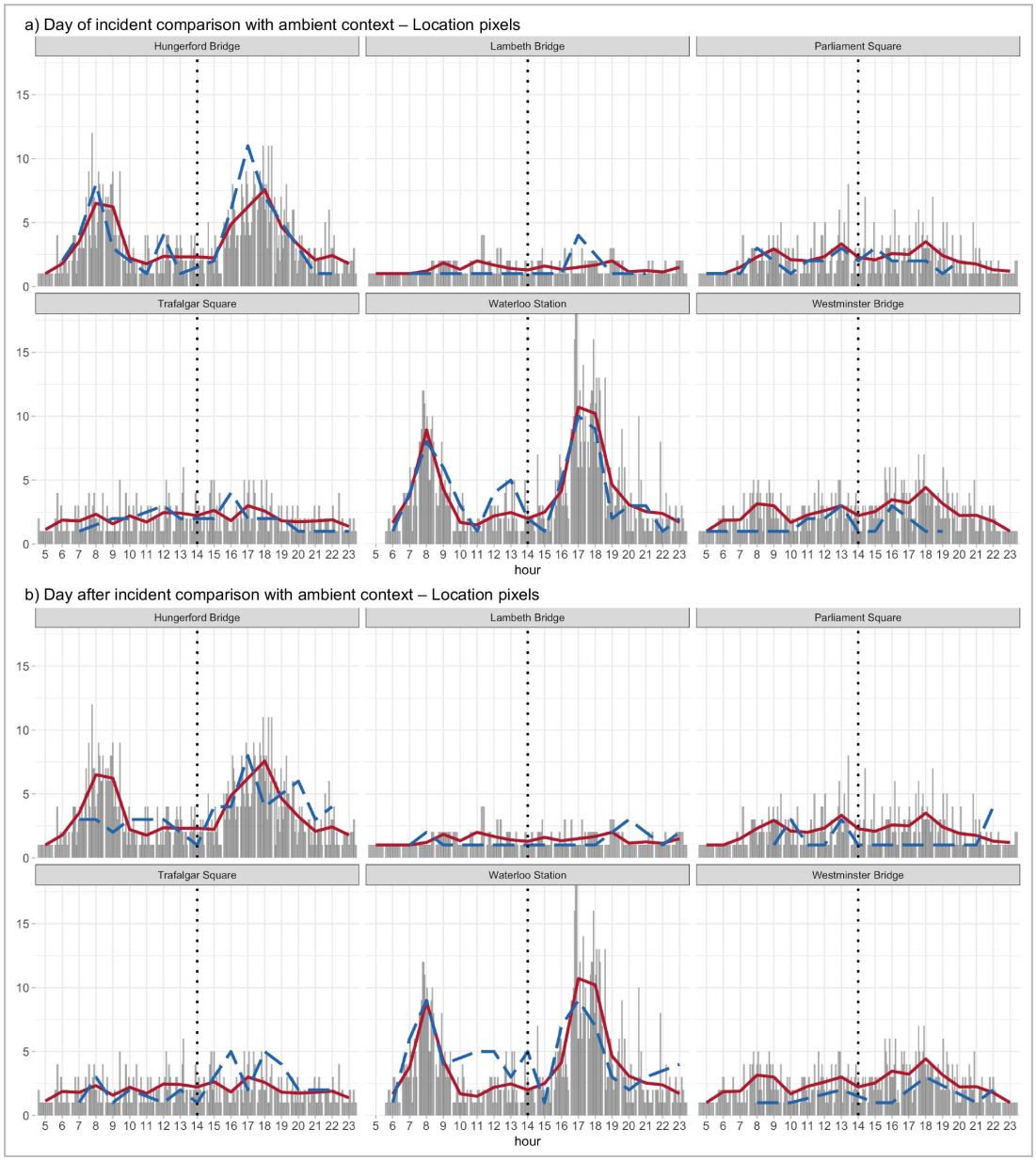


Figure 110. Hourly device counts on day of incident, Westminster Bridge, March 2017

Blue dashed line is the date of interest, red line is ambient mean during March. Grey bars indicate individual readings per day. Dotted vertical line indicates hour when incident occurred. Image a) blue line shows date of incident, counts per location pixel(s); Image b) blue line shows day after incident, counts per location pixel(s).

Hungerford Bridge is a pedestrian-only bridge for traffic within the immediate vicinity that may be affected by road and footpath closures. Lambeth Bridge also has an increase in the same period whilst the pixels at each entrance to Westminster Bridge are understandably lower. Parliament Square shows lower readings after the incident but also has low readings before the incident. Only Hungerford Bridge and Lambeth Bridge register readings higher than on any other date and only Waterloo Station registers readings lower than on any other date. All pixels exhibit a high variety in readings throughout the month. Repeating the pixel scale for the day after the incident (Figure 110b) does indicate an effect in the immediate aftermath, with reduced readings in the morning for Hungerford Bridge, Westminster Bridge and Parliament Square. However, the concern remains that readings at this spatial and temporal scale are very low. Only Hungerford Bridge and Waterloo Station have peaks above 5 unique devices per hour. Waterloo Station appears to be unaffected. The morning peak is similar to the monthly average. The afternoon peak is lower but not outside the range of ambient readings.

7.3.2 London Bridge

On Saturday 3rd June 2017, a second terror attack occurred in London. It started at 22:05 on the north side of London Bridge (Figure 111) within the City of London, also known as the 'Square Mile', the financial centre of London. Saturday 3rd June 2017 was the night of the UEFA Champions League final. According to the inquest into the attack (H.M. Coroner, 2019), the game was being watched on screens in many pubs and other venues in and around the area of the attack. The attack ended Borough Market at 22:16. Eight people were killed and 48 were injured. The duration of the attack was over ten minutes.

Figure 111. Photo of London Bridge, London, June 2017

Viewed from the south side facing north. The image was taken on 9 June 2017 showing security measures implemented following the attack to prevent vehicles mounting the pavement. Shared by ChiralJon under CC-BY-2.0 https://flickr.com/photos/69057297@N04/35039546832.

The two attacks had very different durations and profiles. The Westminster Bridge attack lasted less than 90 seconds during the early afternoon on a weekday in March. It involved a single attacker who targeted a heavily policed area. The London Bridge attack lasted over 10 minutes during late Saturday evening in June. There were three attackers who, after driving a van into pedestrians as they drove across the bridge, continued on foot in a public area with minimal police presence.

As with the Westminster Bridge attack, the purpose of this study is to consider whether or not an active population estimate could provide knowledge about the street-population at the time of the attack with granularity not available when relying on static administrative counts. The second aspect is whether or not a real-time data source can be used to evaluate any lasting impact on population behaviours after the attack.

7.3.2.1 Estimating the active population

The OpenSignal June 2017 dataset is used for this study. The socio-spatial dynamics were generated from readings captured on 8 to 28 June 2017 and the incident occurred on Saturday 3rd June at 22:05. To estimate the active population, the weighting for weekends at 21:00 to 21:59 is used. No seasonal adjustment is required. Table 29 contains the LandScan 2017 ambient estimate, a value that is averaged over 24 hours on working weekdays. The OpenSignal (OS) ambient count mimics LandScan, it is the average number of devices present across 24 hours on weekdays. The active estimate is then calculated for a weekend from 9pm to 10pm and compared with actual readings for the same period on the day of the incident. A seasonal adjustment is not required. There is an event status but it is an unobserved context and cannot be estimated as a weight.

Table 29. Estimating active population, London Bridge, 3 June 2017 at 21:00 to 21:59

LandScan (LS) cell x2y1	*OS devices*	*Weight*	*Population*
LandScan 2017 ambient count			12,046
OS ambient count (June 2017)	15.65	1.0	12,046
OS active population estimate			
weekend, 21:00 to 21:59 (June 2017)	6.5	0.42	**5,000**
seasonal adjustment not required		---	---
event status: major sporting event televised		??	**??**
Actual readings (OS)			
21:00 to 21:59 on Saturday 3rd June	5.0	0.32	**4,000**

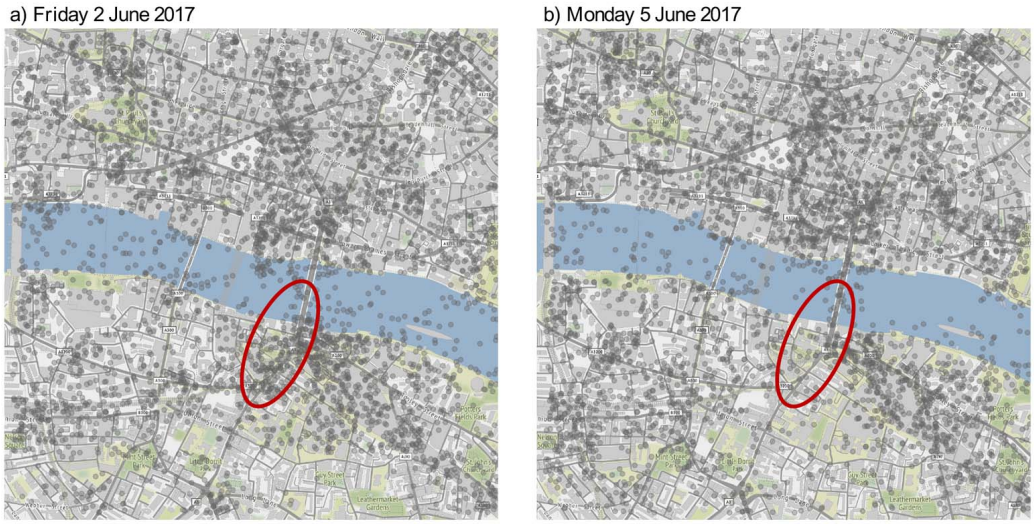
Ambient counts are weekday hour average. OpenSignal (OS)estimates are rounded to nearest 500.

For this estimate, the active population estimate is substantially lower than the LandScan estimate. The LandScan count is averaged across a working weekday. The incident occurred on a late Saturday evening. Daily counts are substantially lower on weekends than weekdays (Figure 98) and during evening hours compared with the daytime hours (Figure 99). Had the incident occurred during the working weekday, between 7am and 7pm, the estimate would have been from 11,000 to 26,000 depending on the hour. Instead, the estimate is 5,000, close to the actual count that produces an active population estimate of 4,000.

As mentioned in estimating for the Westminster Bridge landscape, the device counts are very small when studying individual LandScan cells hourly, meaning small differences will have large effects when using them as weights to multiply and estimate the actual population present. However, having some understanding of population dynamics within the landscape can provide a more informed estimate and an evaluation of how different the impact could be at different times.

7.3.2.2 Evaluating the incident impact

As with the Westminster landscape, having access to a data source that was continuously available allows evaluating whether or not the incident had an impact on population movements. The day after, Sunday, would be quiet under normal conditions. Instead, the working day before (Friday 2nd June) and after (Monday 5th June) are plotted for comparison (Figure 112).



Background: map tiles Stamen Design, under CC BY 3.0;
map data by OpenStreetMap, under ODbL. Projection: EPSG:3857

Figure 112. London Bridge data maps, incident evaluation

Red circle indicates the area of the incident. Coordinates of readings are jittered to 7 decimal places.

The points distribution indicates that the incident may have affected movements on the first weekday after, with visibly fewer readings around the incident area compared to the previous Friday. However, daily device counts through the month at LandScan scale do not reveal any noticeable variation (see Figure 97). Readings on the Sunday after the incident were the lowest of all Sundays throughout June. However, the weather could also have been influential, with below-average temperatures compared with the rest of the month.

Studying at the pixel scale (Figure 113) does indicate a localised impact in the immediate aftermath of the incident. Changes in daily device counts are much more apparent for this situation. All locations within proximity of the incident register substantially lower readings the day after the incident, whilst Monument Station, on the north side of the river, shows minimal effect. Furthermore, the first two days of the week show substantially lower readings in the cells for Borough Market, Borough High Street and London Bridge, with readings up to half those recorded on Wednesday. London Bridge station also recorded reduced readings but to a lesser amount. Monument Station

shows no effect, suggesting that the impact of the incident was localised. From Wednesday onwards, readings return to normal for all pixels. As with the Westminster Bridge landscape, whilst such a traumatic event may terrorise those present and directly affected at the time, it does not appear to have a lasting effect on human activities in the area.

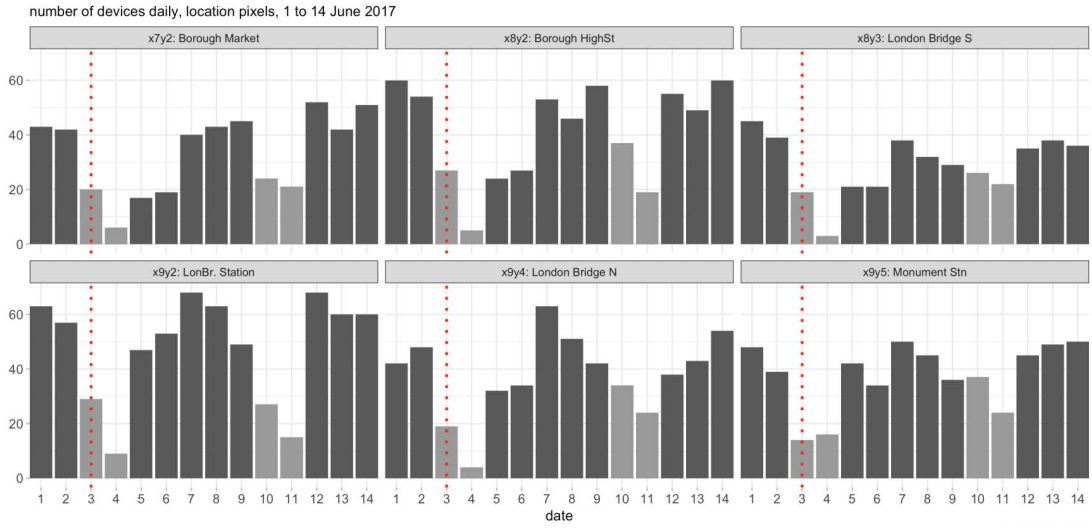


Figure 113. Daily device counts pre- and post-incident, London Bridge, June 2017

Dotted vertical line indicates the incident date. Weekends highlighted.

7.3.3 Oxford Circus

The final incident was initially thought to be a terror attack but proved to be a false alarm. Friday 24th November 2017 was a 'Black Friday' date. 'Black Friday' is a cultural trend imported from the United States of America (USA). The Friday after Thanksgiving is a public holiday in the USA and a time when shops offer heavily discounted sales as part of the build-up to the Christmas festive period, referred to as 'Black Friday'. Shops within the UK have also begun to offer 'Black Friday' sales, attracting large crowds of shoppers although the day is not a public holiday in the UK.

Oxford Circus (Figure 114) is at the junction of Oxford Street and Regent Street. As mentioned in the landscape definition, Oxford Street is the busiest shopping street in Europe, as of 2017. According to live reporting[11], at 16:38 on Friday 24th November, police responded to numerous emergency calls and social media reports of shots fired at Oxford Circus. At 18:09, having found no evidence of gunfire or suspicious packages, the incident was declared to be over. By 18:43, pedestrian activity had returned to normal. It was believed that an altercation between two men on a platform at Oxford Circus tube station had initiated a mass panic. The panic appears to have then been maintained due to misinformation posted on social media.

[11] Timeline of the incident as reported live by The Guardian, collating multiple sources including statements by emergency response teams, other news outlets and social media posts: https://www.theguardian.com/uk-news/live/2017/nov/24/oxford-circus-police-london-tube-gunshots-live accessed 12 February 2019.

Figure 114. Photo of Oxford Circus in November

Oxford Circus, November 2007. Source: Wikimedia Commons, shared under CC BY-SA 3.0. https://commons.wikimedia.org/wiki/File:Multidão_no_Soho_-_panoramio.jpg.

As with the previous two incidents, the purpose of this study is to consider whether or not an active population estimate could provide knowledge about the street-population at the time of the incident. However, there is no mobile data for the period covering this incident to verify the estimate or evaluate any lasting effects. There is, however, transport data spanning the period that is used instead as a comparison for estimating the active population.

7.3.3.1 Estimating the active population

In the study of Westminster, the active population estimate was able to incorporate a seasonal adjustment based on real-world observations. For the study of Oxford Circus, there is only the ambient estimate available, established using readings during June 2017. For this estimate, a conceptual model is used (Figure 115), based on the P-STAR framework, to produce an active population estimate for the period when the incident occurred.

The spatial baseline is the LandScan 2017 estimate of 7,495. It is the best available static measure. The time-based context developed earlier in this chapter indicates that, for this cell, on a weekday between 16:00 and 16:59, the population weight is 1.71, producing an ambient population estimate of 13,000 with the majority of outdoor pedestrians visiting the Oxford Circus junction, in Pixels x6y2 and x7y2. The incident occurred in the winter – November. Based on studies for the QEOP and Westminster landscape, activities can decline by 50% during the winter months. However, Oxford Circus is a retail landmark that is likely to be less sensitive to environmental conditions than other locations. Furthermore, late November is part of the festive build-up to the Christmas holiday period, with late-night shopping hours and increased retail activity. Given the dominance of retail land-use in the bottom half of the LandScan cell, Oxford Circus is likely to be busier than non-retail landscapes and could potentially be as busy as June. Without further information, a seasonal adjustment of 75% is applied, reducing the population count to 10,000. The incident occurred on a working weekday, so no other ambient adjustments are needed.

Spatial baseline			Values
Landscape Zone	Oxford Circus / LandScan x2y1		7,495
Land-use potential	Mixed (non-work indoors, transport hub)		
Time adjustment	**Situation attributes**	**Weights / Updates**	
Day / Hour of day	Weekday, 16:00 to 16:59	1.71	13,000
Season / Sensitivity	Winter / Low (land-use)	75%	10,000
Day status	Term-time	100%	10,000
Action adjustment			
Activity status	Active event (all day, early and late peak)		
Event category	Special event (Retail | Black Friday)		
Event rank	Rank 2	200%	20,000
Weather	Ambient		
Sensitivity	Low (event category + weather conditions)	100%	20,000

Summary

Circumstances:	Land-use potential is mixed, predominantly retail and leisure with a transport hub, Period is Weekday 16:00 to 16:59, Winter, Term-time, Situation is: Active Event, rank 2 (2x population increase) with ambient weather	
Spatial baseline:	7,495	(LandScan cell x2y1 in Oxford Circus Landscape)
Time estimate:	13,000	(Based on OpenSignal readings for Oxford Circus, June 2017)
Seasonal adjustment:	10,000	(Based on OpenSignal readings for Westminster, March 2017)
Situation adjustment:	20,000	(Informal guess that Black Friday could double local population at 4pm)

Figure 115. Estimating a situated active population, Oxford Circus incident

Weights and values in red are estimates for demonstration. Estimates rounded to nearest 500. Situation: Black Friday event. Question: What is the estimated active population present from 16:00 to 16:59 (time of the incident).

The day of the incident was a special event day – 'Black Friday'. The event would likely experience two peaks: a morning peak when shops first open, and a late afternoon/early evening peak for people finishing work who can visit for late-night opening hours. There is no information about how visits to the area vary on and throughout a Black Friday event. An informal estimate is used that the event will be a rank 2 effect, to use the scale considered from the findings in chapter five, and could double the normal number of people present. Based on this assumption, the population size from 16:00 to 16:59 could be double the ambient size. Weather conditions are assumed to be ambient and therefore unlikely to have any effect on a decision to visit the landscape. The combination of factors produces an active population estimate of 20,000 at the time the incident began. This is nearly three times higher than the static LandScan 2017 population estimate.

These are arbitrary values to indicate how a context-aware framework could be applied to a real-world situation, once it is operational with the continuous or frequent sampling of real-time data. There is no mobile data available to validate the estimate. However, two other sources were acquired to compare with the estimate and consider if it improves on the LandScan static count.

7.3.3.2 Comparing real-time data sources

The ambient estimate for the LandScan cell at 16:00 to 16:59 was 13,000 before adding a seasonal adjustment. This is close to the findings of the 2017 study by BNP Paribas Real Estate, that counted 13,500 pedestrians per hour from 14:00 to 16:00 on 10 June 2017 (BNP Paribas Real Estate, 2017). It gives confidence that the mobile data is representative of street-level activities.

Transport for London (TfL) currently release an open data set that contains counts of gate entry/exit across a range of tube stations including Oxford Circus (TfL, 2018). The data retrieved was an average based on gate counts (entries and exits) during five weeks from late September to end of November 2017 excluding the autumn school half-term holiday and any dates when industrial action took place. By chance, this covers the period when the incident occurred. Figure 116 shows hourly counts as entries, exits and the two combined. The number of people exiting the station peak in the morning whilst entries to the station peak in the afternoon, as would be expected for a location that contains work, retail and leisure premises and few residential properties. One interesting observation is the appearance of a two-peak curve when combining entry and exit counts with a small lunchtime bump. This does not concur with the curve from OpenSignal data in June 2017 (see Figure 99) which produced a more defined three-peak curve with a higher reading at midday than the morning commuting peak. Tube statistics only capture trips that require some form of transport and potentially mask short pedestrian trips for local ad-hoc and leisure activities within the local landscape such as lunch breaks. This is similar to the finding that the OpenSignal data had less sensitivity to activities within large open green spaces when compared with the Wi-Fi readings for the QEOP. It is an indication that analysing multiple disparate data sources is likely to produce a more accurate representation of reality than any source alone and caution must be taken when inferring behaviours if only a single source is available.

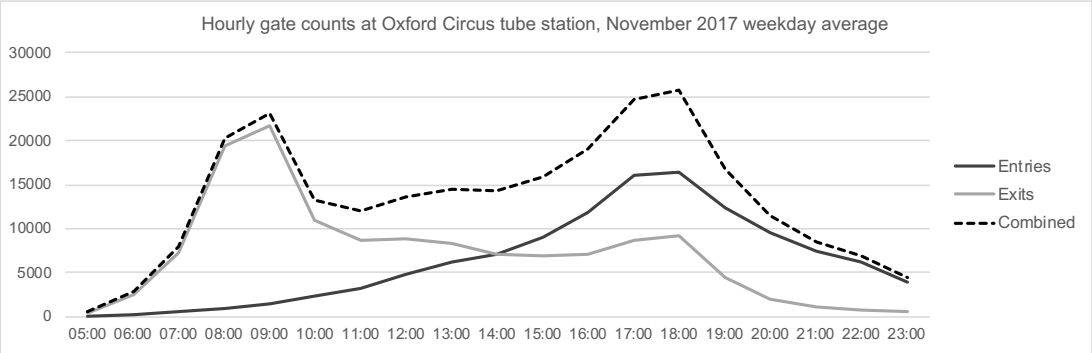


Figure 116. Hourly gate counts at Oxford Circus tube station, November 2017

Averages calculated over five weeks period from late September to end of November 2017, excluding school holidays, industrial action or abnormal conditions (TfL, 2018).

Table 30 shows population estimates generated from each available data source for comparison. Comparing the hourly variation between OpenSignal and TfL highlights that all afternoon hours are substantially above the hour average across the day. However, the distribution is different. TfL weights continue to increase up to a peak at 18:00. OpenSignal has a slight dip in mid-afternoon before peaking at 17:00. It is impossible to say which weighting is closest to reality. However, OpenSignal is not constrained to entry/exit barriers at the station. The LandScan ambient estimate

is substantially lower than the TfL hour average and the BNP Paribas count. It would be beneficial to have access to June statistics for TfL gate counts, or a breakdown of the counts across September to November, to provide an indication for whether or not November does experience above-average visits due to being within a festive period that is expected to generate an increase in activity. However, it is another indication that the LandScan estimate is too low for a public space that can attract high and variable footfall.

Table 30. Hour averages and variations for data about presence at Oxford Circus

		scaled from hour average				
Hourly population weighting (weekday)	*Hour avg*	*14:00*	*15:00*	*16:00*	*17:00*	*18:00*
OpenSignal counts, Jun 2017	1.0	1.75	1.74	1.71	1.87	1.77
TfL Oxford Circus gates, Nov 2017	1.0	1.22	1.36	1.63	2.11	2.20
Hourly population estimate	*Hour avg*	*14:00*	*15:00*	*16:00*	*17:00*	*18:00*
LandScan Ambient Count	7,495	-	-	-	-	-
Situated active population estimate	7,495	20,000	19,500	19,000	21,000	20,000
TfL Oxford Circus gate counts	11,717	14,275	15,892	19,042	24,754	25,762
BNP Paribas count, Sat Jun 10th, 2017 (based on 14:00 to 16:00)	13,500	-	-	-	-	-

Average (avg) is the mean count. For TfL data, excludes 02:00 to 05:00 when the tube station is closed. TfL counts are entries and exits combined into a single count. OpenSignal estimates rounded to nearest 500.

7.3.4 Situated actions

The two terror attacks demonstrated that even the immediate impact of a major incident on people visiting the landscape is very localised, both spatially and temporally. At LandScan scales of aggregation, the impact is barely noticeable. The concentrated effect of such incidents becomes visible at Pixel scales but also reveals their impact is temporary. It suggests that people are resilient to extreme and threatening disruptions to urban environments.

Combining a mobile weight for a given context with the static LandScan estimate does help indicate how much the population can vary in a landscape throughout the day. However, there are several caveats to consider. Assumptions have needed to be made that require verification to improve confidence in this approach. The LandScan ambient count is an average across 24 hours on a working weekday. It is not known if it is an hourly estimate or minute-by-minute estimate or some other time interval. Furthermore, it is independent of time, both during the day and during the year. The OpenSignal data shows that there is great variation in activity both throughout 24 hours and also that there is seasonal variation. This was shown when studying Wi-Fi readings across 12 months for the QEOP and comparing March and June readings for the Westminster landscape. We are assuming that readings during June represent the average across the year. June is the start of the tourist season and it is possible that we should consider that the active population is higher than an ambient average at this time of year.

The OpenSignal data is also not without flaws. It is assumed that the data is biased towards working adults choosing to install the app and make use of public Wi-Fi networks. It may be under-representing family activities, activities involving non-working demographics including children and retired people, residential activities, and non-residential activities taking place in private spaces that do not have a freely available public Wi-Fi network. There potential for error in the readings but it does indicate just how much the size of the population present can vary over a day and at different times of the year when focusing on public spaces.

Referring back to the findings from chapter five, it may be preferable to develop and apply a scale of effect rather than generate precise weights. To provide an estimate of the street population for a given space, time and situation, an approximate percentage used to increase or decrease a static administrative count may be a better approach than a more precise prediction that masks a high degree of uncertainty.

PAGE INTENTIONALLY LEFT BLANK

8 Conclusion

This chapter contains a discussion of the research findings and potential applications. It outlines suggested directions for further research before concluding with closing thoughts.

8.1 Discussion of Findings

The research outcomes have been summarised to conclude each case study (chapters four to seven). The key findings are discussed here from two perspectives. First, the development of the contextual framework that forms the core research contribution. Second, its application to measure changes in population behaviours both as signals of presence and semantics of experience. Data challenges that have impacted the research are considered, before outlining potential applications and recommended directions for further research.

8.1.1 Developing a contextual framework

The hypothesis proposed by this research was that the continuous or frequent sampling of real-time data, analysed programmatically, could be used to learn and model the socio-spatial dynamics of a landscape through the development of a context-aware framework.

8.1.1.1 Context-specific learning

The framework was built on a hierarchy of three place-based contexts: space, time and situation (see Figure 13, chapter three). The spatial layer (S) defines the boundary within which data is collected and aggregated. It enables the creation of a generalised measure that can be aligned with an administrative statistic that does not change other than for a periodic remeasurement such as the ten-year census. It also enables the framework to be integrated with other spatial models.

The temporal layer (T) is the first of two layers introducing time-based variation. It produces a temporal distribution for conditions that occur with some cyclic regularity due to climate and cultural constraints, referred to as the ambient or normal condition. These include diurnal (day-to-night) changes associated with circadian rhythms over a 24-hour day, weekday versus weekend activities and annual seasonal adjustments such as summer versus winter and public holidays. It creates what I refer to as the ‘social heartbeat’ of the landscape, a rhythmic pulse produced by the combined presence and movements of people frequenting the landscape with some regularity.

The situation layer (A) captures situated actions that result from abnormal conditions temporarily disrupting the normal rhythm of the landscape, such as extreme weather and large-scale events that repel or attract a large number of people. Such conditions may be events scheduled or forecast in advance, or they may be incidents that occur without warning such as road traffic accidents. The outcome is a temporary arrhythmic adjustment to the normal routine for the period in question.

8.1.1.2 Profiling landscapes with P-STAR

The framework to profile a landscape, incorporating the contextual hierarchy of space (S), time (T) and situated action (A) is expressed as the P-STAR formula, where P represents a population behaviour of interest, such as presence, movements or language expressed about the landscape, and R represents uncertainty in reactions. The latter incorporates the human element of a person-environment interaction without requiring the study of individual traits. It acknowledges that humans have some degree of agency to react or not to a given set of circumstances and that real-world choices are often ambiguous and can result in arbitrary decisions. Uncertainty can also arise from unknown bias in available data samples. Relying on mobile data sources means the data is only a sample of all interactions taking place and may be generated by a demographic not fully representative of those present in the landscape. For example, mobile app usage in a city centre may be skewed towards adults in work or education and tourists. A benefit of traditional human-observed behaviour maps is that they capture all interactions taking place, albeit for a limited space and time. The use of mobile data sources enables space and time to be scaled, but the trade-off is that not all interactions will be observed and only limited demographic data may be available. Thus, any model using such data needs to consider there will be uncertainty present in results.

Figure 117 visualises the framework in action for a hypothetical landscape, based on readings obtained for the Queen Elizabeth Olympic Park (QEOP) during chapter four.

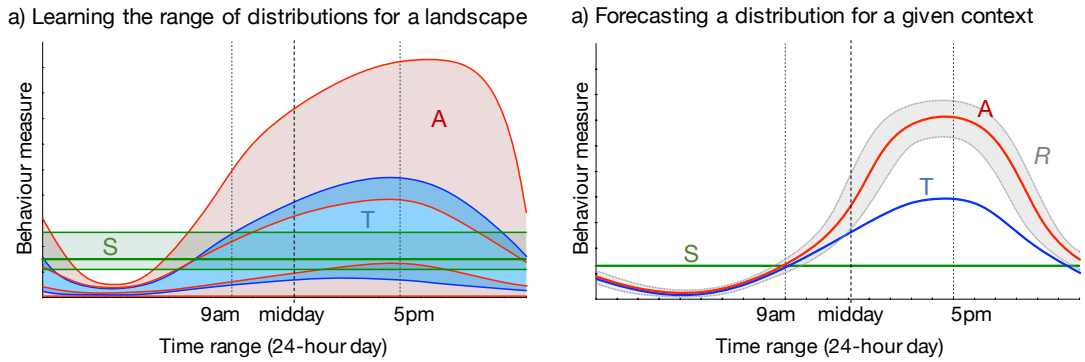


Figure 117. A visual representation of the P-STAR framework in action

First, the ranges of temporal distributions are learned for a behaviour for each tier of the hierarchy – space (S), time (T) and situated action (A). Then a forecast can be made for a given context, with each layer of context adding information to anticipate behaviour changing over a cyclic period, incorporating uncertainty in reactions (R).

The framework comprises two stages. The first stage is to learn how much the temporal distribution of population behaviours vary based on observations (Figure 117a), grouped by context layer. As was discovered during the literature review, whilst samples of mobile data produce diurnal rhythms within urban landscapes, daylight hours produced substantial variations in readings for the same interval (see Figure 9 and Figure 10 in chapter two). This was confirmed by a study of wi-fi device connections in the Queen Elizabeth Olympic Park (QEOP), with both seasonal and situational changes to hourly counts (see Figure 28 in chapter four). Similarly, day counts varied substantially through the course of the year. Whilst the average number of devices connecting daily during 2017 for non-event days was 388, it was just 25 on Christmas day in winter compared with a peak of

over 1,000 during the summer. Thus, the ambient normal distribution (T) is expressed as a range encompassing learned contexts such as day of the week, month of the year, and categories such as term-time and the different religious, public and school holidays that occur on an annual basis.

The normal range (T) is surrounded by two abnormal ranges (A). The upper range represents learned situations that increase the measure of behaviour above the temporal norm, for example attracting more people to the landscape. The lower range represents conditions that decrease the measure below the temporal range, such as repelling people away from the area. It is likely that, for most landscapes, there will be some overlap between the ranges of the normal temporal rhythm (T) and abnormal situations (A). This was evidenced in chapter four in the study of the QEOP. An event day in March generated substantially higher wi-fi readings than on all other dates in March. However, it was within the temporal range for non-event days during the summer months.

The spatial baseline (S) is the average across the time range. It too can have a variation to indicate how much difference there is depending on if the average is calculated across all readings or is separated for normal conditions versus abnormal conditions.

8.1.1.3 Modelling and calibrating population behaviours

Once distribution ranges have been learned, the contextual framework can be used to forecast a behaviour distribution for a given set of circumstances (Figure 117b). Each layer of context adds further information to measure changes in behaviour over a cyclic period. Whilst reaction sensitivity and uncertainty (R) are important to the framework, their quantification is not directly addressed in this research for reasons discussed later under future directions.

The hypothetical scenario presented is based on contexts learned within the QEOP and reflects an afternoon large-scale event taking place on a weekend day in the summer, with a static baseline representing the non-event generalised hour average. As was discovered in chapter four, event days do generate substantial increases in activity, but concentrated around the duration of the event. Up until midday, the readings would be unlikely to be significantly above the range of uncertainty surrounding the routine temporal distribution (T). A similar effect was observed when the weather is hot and sunny during the spring and summer seasons, with a change in presence only occurring from after midday and throughout the afternoon.

The contextual framework was developed and tested using a range of data sources for a single location – the Queen Elizabeth Olympic Park (QEOP) – with results presented in chapter four. The first study compared two data sources embedded within the park – Wi-Fi device connections and hourly headcounts recorded by cameras installed at entrances to the park – with two geo-referenced social media sources, Twitter and Foursquare, analysing daily and hourly counts. It became evident that social media sources are not a suitable measure for analysing presence counts at neighbourhood scales. Whilst sensitive to abnormal situations, readings during normal routine days were sparse. This is one example of how the framework can adapt to different data sources. Social media may not be useful in this landscape for studying presence counts, but it may contain content that can be used to study situated language, discussed in the next section.

Access to daily counts of devices connecting to the park wi-fi over a year demonstrated the need for the three-tier hierarchy of contexts. As shown in Figure 118, the mean count of devices connecting to the park wi-fi on non-event days is substantially different to event days, but there is a large overlap in the range of counts for each context. Producing a generalised count across the two context layers risks creating an ecological fallacy, a measure that is not representative of either normal or abnormal conditions. Such a concern has long been recognised for areal aggregation (Openshaw, 1984) but is demonstrably applicable to temporal aggregation as well. Adopting a contextual framework helps reduce such effects when aggregating real-world observations.

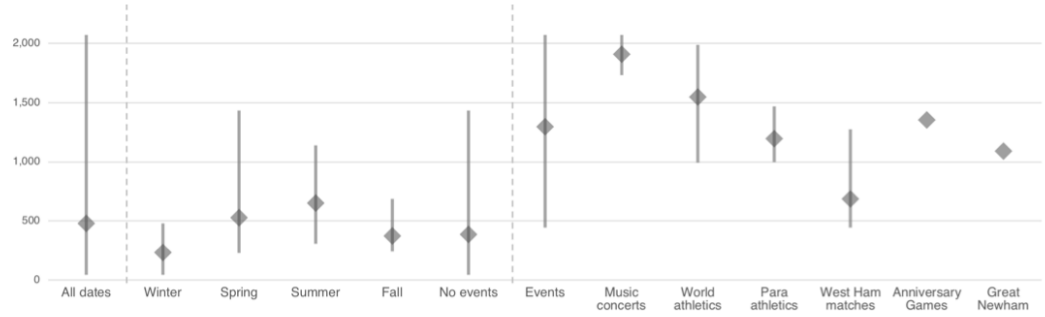


Figure 118. Variation in daily device counts within a context

Number of devices connecting to the QEOP wi-fi per day, grouped by context. Each plot shows the mean and the min-max range. 'No events' is the combination of the four seasons. 'Events' is the combination of all events.

Building a prediction model from the data, using machine learning techniques, emphasised this challenge. Creating a single model across all contexts produced a weaker outcome compared with adopting a modular approach and training the algorithm separately for non-event (representing the temporal layer) and event days (representing the situations layer). The contextual framework replaces the concept of a universal generalised model with a 'many models' approach, each model tuned for different circumstances and able to incorporate different data sets as appropriate.

The variation in readings for different contexts challenges the taxonomy proposed by architect Jans Gehl for human interactions and their relationship with a landscape (Gehl, 1987). Gehl proposed that interactions are necessary or optional, and that necessary interactions are insensitive to the quality of the physical environment whilst optional interactions are highly sensitive to conditions. This appears to be an over-simplification not supported by real-world observations. For example, Westham Premier League football matches can have an attendance of up to 57,000. The club has 52,000 season ticket holders with a waiting list of 55,000 people seeking to purchase season tickets for the 2017/18 season (Inkersole, 2017). Investing in a ticket that grants access to all Premier League matches would suggest a high likelihood of attendance. Yet device counts on match days showed substantial variation throughout 2017 and indicated that low temperatures during winter months have a strong effect on attendance, particularly for weekday matches held in the evening. The commute to work may also be considered a necessary interaction, yet a high variance in counts was exhibited on weekdays across all three city landscapes analysed in chapter seven. Furthermore, the seasonal study of the Westminster landscape indicated sensitivity to climate and flexibility for remote working on Mondays and Fridays. It suggests that whilst the act of work may be considered necessary, it may not translate into a consistent level of necessary human-

environment interactions such as travelling from home to a place of work or study. An alternative approach could be to identify the elasticity of an environment, creating a scale to approximate a measure of its ability to attract or repel people for different contexts. This prospect is considered later in the chapter under future research directions.

8.1.2 Measuring socio-spatial dynamics

Application of the framework was examined for three behaviour measures: counting the number of people present, categorising different actions whilst present (dwelling versus moving), and learning the language used to describe experiences whilst present.

8.1.2.1 Producing a population count

The study of wi-fi readings revealed variations in activity levels between and within normal and abnormal contexts across all the landscapes studied. To inform real-world decisions, they required the means to convert a device count into a population measure. Wi-fi readings are bounded to the area the wi-fi covers and an arbitrary multiplier is used by the provider of the network to estimate the actual number of people present. Analysing daily counts in chapter four indicated that the multiplier was ineffective. Incorporating a spatial baseline context enables the framework to be aligned to alternative sources of information, such as formal population estimates aggregated by output area. Administrative statistics may provide a more robust count of a complete population but lack a temporal distribution. Mobile data has temporal variation but is only a sample and lacks the robustness of administrative statistics. Blending the two can produce a more informed active population estimate than either can provide on its own.

The results presented in both chapters five and seven gave confidence that blending administrative and real-time sources could provide an active population estimate for urban outdoor spaces, areas that cannot be easily measured automatically using methods such as building occupancy. A significant challenge was in deciding on a suitable administrative source for a baseline. As was discussed in chapter two, residential statistics are acknowledged as a poor indicator of presence during daylight hours, particularly on working weekdays when people are at places of work or study. However, static workday population counts are also problematic. The 2011 UK census captured workplace locations for working-age adults in March 2011. However, such a measure only captures a single workplace location. By 2017, how many people would still be working or studying at the same location? Furthermore, it assumes that people spend time away from home at one location. It does not consider people who work at multiple locations, or interactions that occur elsewhere, such as shopping and leisure facilities, rest and recreation periods, community and social activities, tourism and visiting attractions, as well as time spent travelling between locations.

An alternative ambient measure, LandScan was used as a spatial baseline. It also redistributes the residential population according to an ambient working day but incorporates other activities such as travel and tourism. It updates annually using satellite imagery and available data sources. However, when examining individual LandScan cells for landscapes across London, there was concern that the algorithm is a poor indicator for open spaces and transport hubs. The difficulties

sourcing a robust ambient population measure highlighted another aggregation step that can affect the outcomes from the contextual framework: the modifiable interactions unit problem (MIUP). The residential census is limited to a population measure based on where people live, their registered address. The workday census count is limited to an assumption that all working adults are present at a single non-residential location on term-time weekdays. LandScan also incorporates travel and tourism but this is still an incomplete classification for any urban social landscape. One of the more promising ambient measures in development, Population 24/7 (Cockings & Martin, 2018) is incorporating a full set of building attributes to incorporate a more granular time variation, such as considering the different opening hours for categorise such as schools, shops and offices, and weekday to weekend differences. However, it introduces the challenge of maintaining accuracy over time as building use changes, requiring the ongoing maintenance of a centralised database of building descriptors. A key differentiator for the contextual framework is that it does not require any long-term data storage. Instead, it automatically recalibrates over time as new data is analysed to learn and revise weights representing the spatial and temporal distribution of presence.

The modifiable time unit problem (MTUP) presented a challenge for producing a representative population measure. The LandScan count is an ambient population estimate averaged across 24 hours. What isn't known is if an estimate of the people present in a single cell represents the average per minute, per hour or some other interval. For calculation, an hour average was assumed. The results indicated it was a reasonable assumption when alternative data sources were available to verify outcomes, such as attendance at music concerts (chapter five), and gate counts in Oxford Circus (chapter seven). Another time-related challenge was the lack of knowledge about what season such an ambient count represents. The study of device activity in the QEOP demonstrated substantial variation in numbers across the course of 12 months. A key finding from conducting this research was the difficulty in producing a robust and reliable population baseline from which to quantify contextual variations and produce an active population estimate.

8.1.2.2 Classifying presence behaviours

Having assumed that a LandScan ambient estimate represented an average weekday hour during term-time in Spring, it still did not resolve the issue of what the count represents. A count of 10,000 could indicate the same people present simultaneously for the duration of, or part of, the interval, or dispersed across the interval, such as 1,000 people present on average every 6 minutes. Having a data source such as mobile app data, with readings that contained device IDs as well as coordinates and timestamps, creates the potential to quantify both where people are located within the landscape, and whether they are dwelling, milling (pausing briefly) or moving.

The conversion of unclassified mobile data points into classified behaviours (Figure 119) was perhaps one of the most exciting iterations of the framework. It created the possibility that mobile and sensor data could further complement administrative sources by detecting active spaces within an area of aggregation and classifying the duration of presence at different areas across the landscape. This could enable an ambient population measure to be articulated more accurately in terms of the types of interactions taking place. For example, in Figure 119, it is possible to

differentiate a range of different duration lengths for different locations within the landscape. Durations also varied for different contexts. The cluster of red points positioned over the London stadium indicates a duration of between 90 minutes and four hours. These durations only occurred on days when football matches were taking place at the stadium. On non-event dates, the area is dominated by movement and brief dwell times. The area where Westfield Stratford retail centre is located shows a pink cluster representing durations of 20 to 90 minutes, giving an indication of time spent in retail centres. The purple clusters are located at transport hubs and show shorter dwell times of 5 to 20 minutes, indicative of waiting for a bus or train.

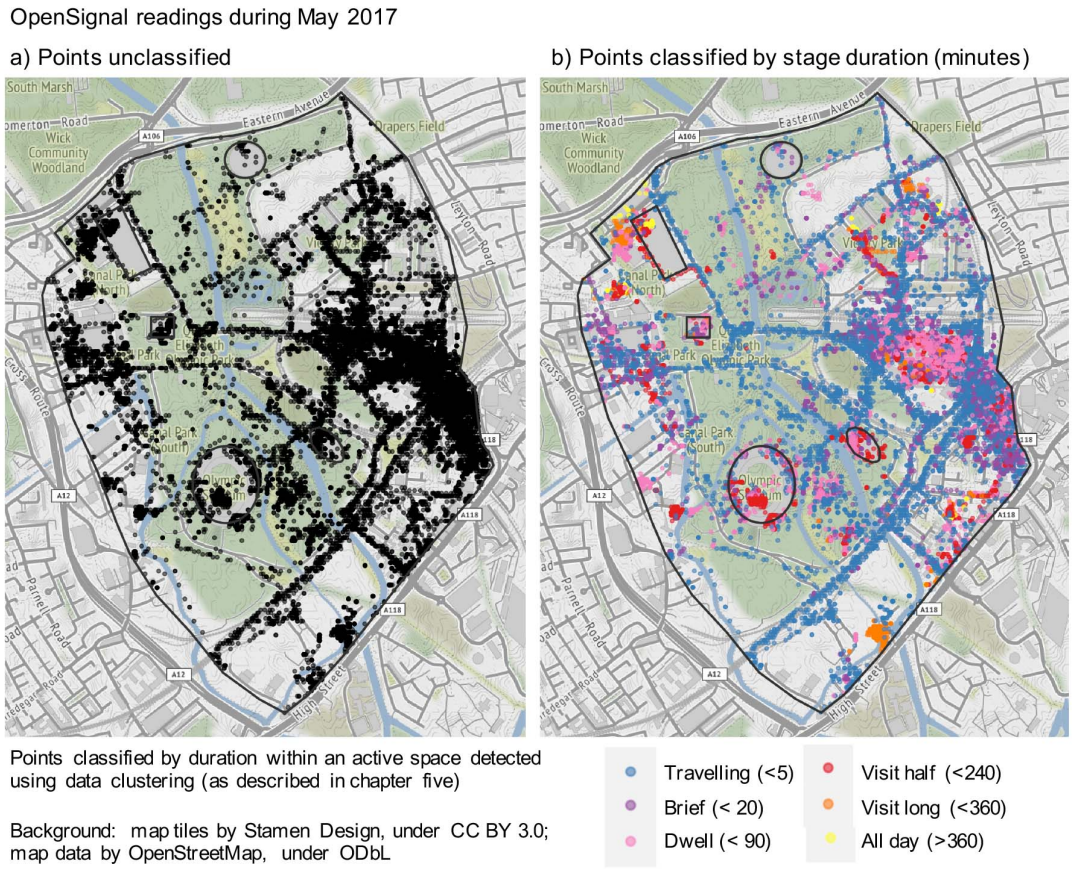


Figure 119. Mapping presence behaviours in a landscape

The image on the right shows the same data points as the image on the left but coloured by presence behaviour of each reading (based on the duration of the stage of a trip the point falls within).

Converting unclassified points into a map of presence behaviours is a capability that would be difficult to achieve with human observers but is relatively simple to perform programmatically with digitised interactions containing a device id, timestamp and spatial coordinates. The result was limited by the small volume of readings provided by the OpenSignal app. However, the approach showed promise and further research is recommended with a more comprehensive data set. Such an outcome could benefit not only population estimates but also provide richer representations for pedestrian modelling, such as the study of trip movements through landscapes and identifying where administrative output areas may unintentionally create ecological fallacies. The clusters detected using mobile data sources could be used to propose alternative aggregation methods such as adopting a nodal approach to focus on active spaces within a landscape.

8.1.2.3 Studying the language of place

The third behaviour considered, following variations in presence counts and movement categories, was the concept that the content of messages could be used to sense people-place experiences. As described in chapter two, several studies have undertaken research using social media sources such as Twitter, Foursquare and Flickr. However, they have typically focused on generalised outcomes that lack temporal variation. Also, whilst the source data contain spatial coordinates and can be aggregated at small output area scales, comparisons are often made across entire cities or regions. However, they have produced measurements such as walkability (Quercia, Aiello, Schifanella, & Davies, 2015) and moods expressed (Musto, Semeraro, Lops, & de Gemmis, 2015), as well as proposing alternative city population measures (Birkin & Malleson, 2013).

This research found that social media sources are not robust enough to act as a population measure at finer spatial resolutions, such as studying variations within a neighbourhood and at street-level. However, the sources were sensitive to abnormal situations that attract people to the landscape and thus could potentially be used to reveal the language of different circumstances occurring within the same physical landscape. This could include learning how the meaning of terms can be drastically changed by context. The example given in chapter six was the language used during a music concert by Guns N' Roses and their lead guitar player Slash. In isolation, the terms 'gun' and 'slash' may be assumed to refer to some act of violence.

More problematic was the use of social media to detect mood or emotion. Not only were there demonstrable difficulties in making a human judgement about positive or negative affect expressed in short messages, but there is also concern regarding the use of both dictionary-based and trained algorithms to quantify emotions. The Facial Action Coding System and its choice of discrete emotion categories dominate cognitive algorithms used to study and interact with humans on an emotional level. Yet the theory is being challenged for lack of reproducibility without priming (Barrett, 2017). Many expressions appear to be ambiguous unless a context is provided. The same is also being demonstrated in language. The prevalent method for calculating the sentiment of a text is to score terms using a dictionary. Testing during research highlighted that positive phrases could be scored negatively and vice versa. Other research published has also raised concerns with generalised dictionaries that, despite being considered universal for a given language, such as English, can contain cultural biases that affect outcomes (Sap, Card, Gabriel, Choi, & Smith, 2019).

Arguably, this demonstrates the need for developing a context-aware approach, creating contextual location-based dictionaries tuned to their environment. As with other sources in a contextual framework, such dictionaries can be self-calibrating, incorporating new terms that become popular for recurring behaviours, and decaying older language that falls out of use. The same three-tiered hierarchy would apply, producing three dictionaries for a landscape: terms used consistently regardless of time (S), terms for recurring periods (T) and situated terms (A).

Whilst such an approach could be used for studying linguistic signatures of place, I believe that sentiment should not be interpreted using emotive tones due to the issues raised. There are moral

and ethical concerns when making judgements about feelings or motivations based on results that may be theoretically unsound or highly sensitive to small changes in the data. To incorporate subjective experiences requires the ability to recognise whether expressions represent the place and situation, and are externally communicated or internalised feelings. Applying categories with emotional labels requires careful communication of findings.

Despite these limitations, there were novel findings from using social media to study contextual variations within a landscape. The use of network analysis and community detection was able to identify different situations occurring on the same date, and even a basic similarity measure comparing terms demonstrated different elasticity of language, with some terms being consistent across situations whilst others were unique to event categories and individual incidents. Studying language revealed the same need for regular recalibration as recognised for presence counts and movement behaviours. For example, the Olympic Stadium within the QEOP was renamed the London Stadium in August 2016. Over time, the descriptor 'Olympic Stadium' will decay in value. Analysing the content of tweets in near real-time uncovered incidents that the park management may otherwise have been unaware of, such as detecting an evacuation that occurred in the Westfield Stratford retail centre, leading people to be temporarily moved into the park.

The research demonstrated that the same contextual framework could be applied to both signals of presence as measured as counts and semantics of experience expressed through language.

8.1.3 Data challenges

Central to this research was the application of data-intensive methods to study population behaviours by aggregating and analysing individual mobile data traces across a landscape. Many challenges were experienced throughout the case studies presented.

An immediate concern with the mobile data sources being used for this research was the size of the samples available. Would they be sufficient for studying local population behaviours? Whilst smaller samples produce results with weaker confidence, the findings confirmed that some data is preferable to none (Norman, 1999). A small number of participants producing readings daily can reveal the dynamics of the landscape provided a sufficient period is studied, preferably at least one month but longer if the samples are very sparse, as was the case for social media. The contextual framework could be made operational for a landscape with a minimum of three weeks of data but would improve with continued or frequent sampling and recalibration over time. The three biggest challenges for this research were accessibility of, bias within, and aggregation of, data sources.

8.1.3.1 Accessibility

Data is being generated about our decisions and actions at an unprecedented rate, referred to as a 'digital shadow' of self (Bentley, O'Brien, & Brock, 2014). However, much of the data is proprietary, owned by private companies and not available for public use or academic research. Furthermore, data can be made available for viewing but remain inaccessible for analysis. It is

possible to view individual tweets from the full archive posted publicly at http://www.twitter.com but it is only possible to query for a sample of tweets posted in the previous seven days.

When relying on secondary and indirect sources, it is unknown how long access will be granted, or how access will change over time. In the period since when this research was first proposed in late 2015, the following changes occurred on the platforms being targeted for reality data:

- Instagram withdrew its public API for accessing publicly-shared images.
- Twitter revised its sampling algorithm, reducing the number of geotagged tweets.
- Intel switched off its trial IoT platform without notice (the platform was used to publish readings from the weather stations installed within the QEOP and intended for use in deploying further embedded sensors for behaviour analysis).
- Google Maps introduced a charging structure for map queries and visualisations.
- Weather Underground, following acquisition by IBM, withdrew free access to its API, that supplied data from weather stations within the QEOP after Intel shut-off its IoT platform.
- Twitter introduced a new charging structure and removed all third-party API access to archived data (after acquiring the most popular method, gnip).
- Foursquare announced a move away from public data sharing.
- Twitter announced the removal of geotagging functionality within Tweets.

A recommendation is to consider new methods for crowd-sourcing data about people-place interactions in open and transparent ways for conducting research in the public interest.

8.1.3.2 Bias

It is no surprise that there is bias in samples of data about real-world situations. The concern has been in the variety of different biases that can affect outcomes from data-intensive studies.

Sampling bias

The bias that was anticipated was that of sampling – is the sample sufficiently representative of the population behaviours being studied and predicted? The initial hypothesis was that the demographics of the sample wouldn't matter if the landscape and/or population behaviours being studied were generic, or if the limitations were acknowledged in the application of resultant algorithms. In this case, the assumption was that outcomes would be focused on the interactions in urban public spaces and thus likely to be weighted towards working adults. The data sources would not be suitable for the study of specialist landscapes used by a non-working demographic, such as areas with a high concentration of schools or care homes for the elderly.

The first study in chapter four included an additional data source that should not suffer any bias – headcounts generated from webcams installed at the main entrances to the QEOP. When comparing sources, the Wi-Fi data, assumed to contain bias, had a robust correlation with the webcams suggesting it is representative of the population. The social media sources had a weak

correlation and were considered unreliable for inferring population dynamics. Instead, their use was restricted to providing additional contextual information about circumstances within the landscape.

A sampling bias that is difficult to measure but warrants consideration is the missing data – what interactions are not being captured digitally and what interactions do not occur because of the conditions of the landscape. For example, reliance on social media also risks over-emphasising extroverted activities and under-representing the mundane or activities within a social landscape that do not encourage self-promotion. Mobile data enables analysis at new spatial and temporal scales but is only a sample of real-world interactions. Many of the findings presented here concluded with a recommendation that more data is needed. The contextual framework could help direct further data collection, both quantitatively and qualitatively, acting as the stepping stone envisaged by middle-range theories, as described in chapter two.

Influencer bias

A relatively new form of bias in spatial data that wasn't anticipated is that of influencers. An influencer is someone who affects the opinions of others. The concept is not new. It is based on a theory developed in the mid 20th century known as the 'two-step flow of communication' (Katz, Lazersfeld, & Roper, 2005). The theory is that people form their opinions under the influence of opinion leaders, who in turn are influenced by mass media. The role of influencer has changed dramatically as a result of global online social networks and this could impact the use of data sources such as social media in research. People who build a large following on online social networks such as Twitter and Instagram are becoming funded as influencers either through direct payments from the providers of products or services being promoted through their social media channel and/or through receiving a share of the advertising revenue generated by the social network platform hosting their channel. Increasingly, such promotions are focused on physical locations. In 2019, several influencers were criticised for promoting an event for the government of Saudi Arabia (Figure 120). Of particular relevance to this research is the comment that people were being offered substantial sums specifically for providing geotagged posts.

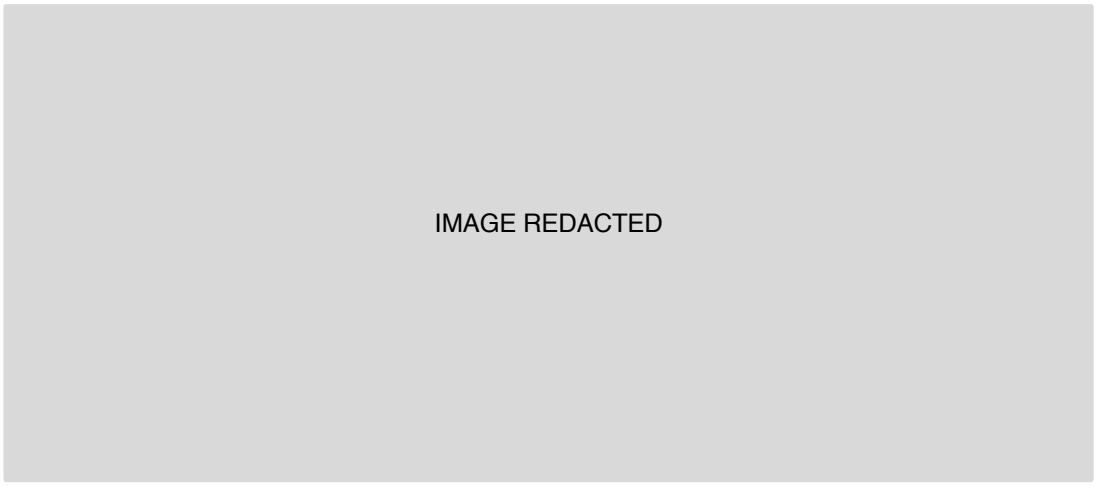


Figure 120. Influencers being paid to promote locations

Image retrieved from author's Twitter timeline, December 2019.

Bot accounts were anticipated during the research – the use of multiple artificial accounts to promote content – and accommodated by de-duplicating identical tweets posted within a short timeframe that were near identical in content. However, influencer actions were not and their growing occurrence in location-based social media lowers the trust in such data as being representative of real-world experiences. Influencers are unlikely to post controversial comments about real-world experiences if the comments might impact their earning potential from sponsorship and advertising revenue. An influencer's popularity is developed and maintained through positive and encouraging messages that may not represent reality. These are substantial concerns for the use of social media to study spatial and social behaviours in real-world situations.

8.1.3.3 Aggregation, scaling and transformation

The modifiable unit problem

A challenge when studying population behaviours is unintentionally introducing error through the necessary acts of aggregation and scaling. These effects can be amplified for small data samples. As has been discussed previously, the modifiable area unit problem (MAUP) is well known to spatial analysis. There has been less acknowledgement about the challenges with space-time analyses and the potential influence of the chosen time interval, referred to as the modifiable time unit problem (MTUP). The appropriate time aggregation may depend on the source of data. Whilst an hourly interval was primarily used in this research, for sparser data sets a larger period may be preferable, such as differentiating only between peak and off-peak periods across a 24-hour day. Blending with an administrative statistic adds a complication in that it requires an assumption regarding what interval the statistic represents as well as the change that can occur across the period between measurements being taken. For example, the LandScan ambient population measure states only that is the average ambient population across a 24-hour workday. It does not specify if the count is the average for any moment in time or the aggregated average for an interval such as hourly. The counting of people based on a single home and work location introduces a further modifiable unit problem: that of interactions (MIUP). Residential statistics are easily identified as being limited in representing where people are present during a working day. However, a workday population estimate requires assuming what interactions to include when generalising where people are present when not at home. A single non-residential location per person is unlikely to be a satisfactory estimate of non-residential populations. LandScan incorporates some element of tourism and business travel. However, it does not consider how much of modern working routines involve multiple locations, remote and flexible working, as well as 'third space' activities involving rest, retail and recreation. Mobile data offers the potential to overcome this challenge.

Incorporating semantics can create a modifiable language unit problem (MLUP) such as whether or not to apply a 'stopwords' list to exclude terms, and whether or not to use techniques such as stemming and lemmatisation to group words spelt differently with shared meanings. All such decisions can be influential and need to be considered when interpreting results.

In addition to language and interactions, this research introduced a further modifiable unit problem: context (MCUP). Throughout the case studies, as the framework was developed, the layer at which time-based attributes were assigned varied between routine behaviours situated actions. In the end, the defining criterion was whether or not the attribute represented a condition that would recur with certainty from one year to the next. It is an imprecise definition. Seasonal variations associated with climate were made part of the temporal routine whilst actual weather conditions were made part of the situated action even though only unexpected and extreme weather conditions are likely to produce a noticeable change in behaviour. Within the QEOP, football matches take place with some regularity as part of an annual football league. They could be argued to be part of the temporal routine. However, it was decided to keep all events within the situation layer. At every stage of the process to record, analyse and measure socio-spatial dynamics there is potential to introduce error through the choices of categorisation and aggregation. It highlights the need to be able to communicate the uncertainty present within studies of real-world phenomena.

Big data analytics

The data-intensive nature of this research required embracing big data analytics. As defined in chapter two, ‘big data’ is considered different to just ‘data’ because it contains features that defy traditional statistical techniques. It requires advanced computational methods, such as machine learning, that can process data that is big in volume, velocity and variety (Laney, 2001). There have been breakthroughs over the past decade in the development of computer algorithms able to outperform humans completing human-like tasks under the same conditions, such as object recognition (He, Zhang, Ren, & Sun, 2015) and 3D multiplayer games (Jaderberg, et al., 2019). However, such developments remain within controlled environments. Yet the foundation of data analytics as a field, as defined by mathematician John Tukey (Tukey, 1962), was in recognition that real-world phenomena occur in uncontrolled environments full of ambiguity and uncertainty:

> *The most important maxim for data analysis to heed, and one which many statisticians seem to have shunned, is this: "Far better an approximate answer to the right question, which is often vague, than an exact answer to the wrong question, which can always be made precise." (Tukey, 1962)*

There have been calls that the results from big data analysis need to be presented with the statistical rigour of traditional methods (National Research Council, 2013). The emphasis on precision may be inappropriate for answers to questions that are often vague. Much of the focus on the so-called ‘V’s of big data (see Table 1 in chapter two) has been on volume – the size of a dataset, whether in terms of the number of records or number of attributes. It has overshadowed the need to consider challenges posed by the other attributes. Data with high velocity may decay rapidly, limiting the usability of results beyond a certain interval. Data with high variety requires transformation – all machine learning algorithms require attributes to be converted into a numerical format – as well as scaling when the range of values for different attributes vary by orders of magnitude. The choices deployed in transformation and scaling can affect outcomes. This was demonstrated in chapter four when using machine learning to predict variations in presence counts. Months of the year are typically expressed numerically from 1 to 12, but the increase from 1 to 12

has no association with changes in presence. Rather, visitor levels increase from Winter to Summer and then decrease from Summer to Winter. Veracity is also problematic. Big data has arisen from the ability to capture real-world observations at scale through digital technology. Yet the quality of readings can vary substantially both within and across data sets in terms of authenticity and accuracy. Social media posts may not represent an actual lived experience but rather a curated experience designed for sharing with an online social network. Mobile data coordinates captured using GPS can have location accuracy estimates varying from a few metres to several hundred.

For every lens through which a real-world phenomenon is measured, there will be some level of aggregation, scaling and transformation that can affect outcomes. It is not sufficient to state that big data defies traditional statistical methods. There needs to be a technique for incorporating the level of approximation present within results and its relationship with the vagueness of the question.

8.1.4 Potential applications

Referring back to the conceptual model in chapter three (Figure 14), the practical application of this research is in the sensing and visualisation of live data feeds to learn and anticipate behaviour. Figure 121 outlines how the framework could be implemented in live operation. The primary focus of this research has been to produce a model that can be initialised with at least three weeks of data to establish the ambient context and begin to learn patterns for different sets of circumstances. The model can then function both to support real-time decision systems as circumstances unfold and to aid planning by providing learned real-world contexts to improve simulations of behaviour.

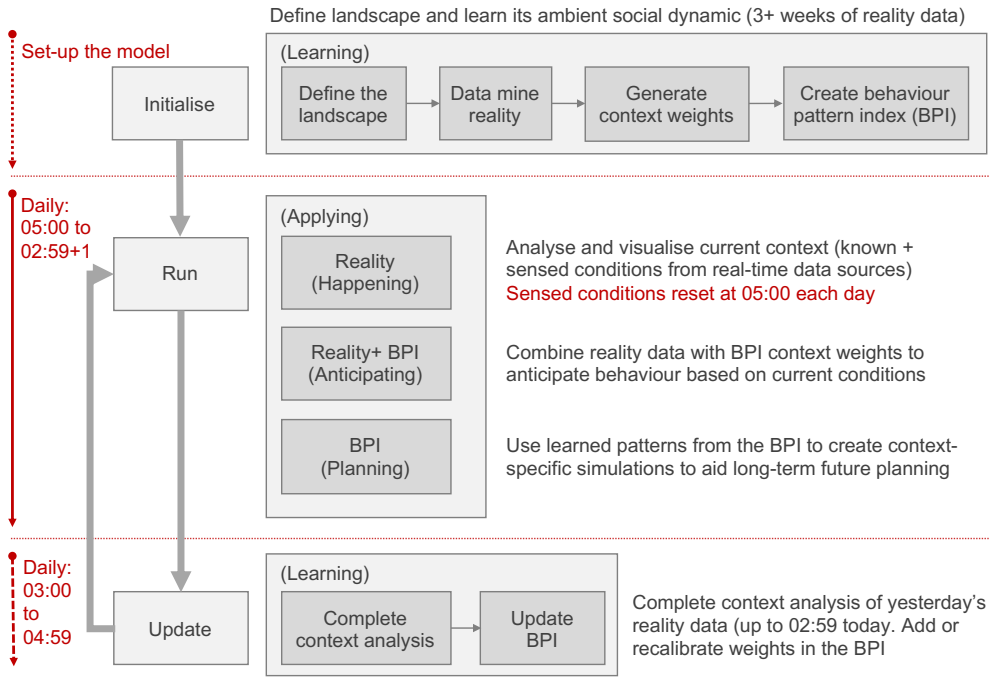


Figure 121. Conceptual workflow for implementing the contextual framework

8.1.4.1 Context-specific research

As introduced in chapter one, all models are simplified representations of reality. Producing a framework to measure how behaviour varies for different contexts within a landscape creates two opportunities. First, it can provide a richer set of initial conditions to inform forecasting and simulation models. For example, an agent-based model could incorporate a more complex set of rules for agents to follow and contain multiple different populations of agents, with each population exhibiting different presence behaviours. Such details would be based on learned contexts rather than theoretical assumptions. Furthermore, the ability to frequently recalibrate the framework enables models to revise recommendations and accommodate longer-term changes in behaviour, such as shifts in preference for office-based versus remote working as evidenced in chapter seven.

A second benefit from learning context-specific behaviours from real-time data traces is to inform and assist qualitative studies and/or more comprehensive quantitative studies using human observers to capture a complete set of interactions. A data-intensive analysis alone is unlikely to be sufficient to fully comprehend an environment and the behaviours it generates for two reasons. First, it will always be possible for a sample to fail to adequately represent the population present. This has been discussed throughout the thesis. The mobile data traces are assumed to be biased towards working adults and under-represent family activities, or non-working demographics. Thus, the outcomes are focused on general urban public spaces. Second, data traces are only produced for what does happen, not what should happen. The optimal approach to study urban landscapes would be to combine multiple and mixed methods. For example, a survey or interview can produce rich individual stories. However, without prior knowledge, an interview risks being skewed by the interviewee's most recent experiences and/or the interviewer's assumptions. Behaviour mapping using human observers is time- and resource-intensive, and typically conducted over small spatial and temporal scales. Conducting a preliminary landscape profile using mobile data traces, as described in this thesis, could enable an evidence-based approach to qualitative research and direct where more comprehensive analyses involving human observers, or the use of embedded sensors would be beneficial to capture a wider range of interactions. Referring back to the concept of middle-range theories introduced in chapter two (Merton, 1967), a data-intensive analysis using real-time data traces can act as a stepping stone towards new theories and improved models by identifying where time and resources should be focused, producing an over-hypothesis to initiate a more comprehensive study, as envisaged by philosopher Nelson Goodman (Goodman, 1990).

8.1.4.2 Real-time interventions

Performing an analysis of an urban environment in real-time within a contextual framework can assist both those responsible for a location or situations occurring within it and those directly affected. There are three approaches: predictive analytics for oversight of an environment; adaptive interfaces embedded within an environment; and, situated intelligence provided to those present.

Cities deploying operations centres and dashboards to monitor urban environments in real-time could benefit from integrating the contextual framework into dashboards. The dashboard could be extended from displaying a live feed of what is currently happening to incorporate predictive

analytics, producing a forecast about what might happen in the near future given learned contexts and/or enabling the development of a recommendation engine to guide interventions. Much of predictive analytics to date has focused on identifying a protagonist and the location for their actions, such as tracking the path of a hurricane or crime patterns. This framework is about providing an understanding of the social environment in which such an action is predicted to occur.

The number of Internet of Things (IoT) connected devices is forecast to grow from 15 billion when this research was first proposed in 2015 to 30 billion at its conclusion in 2020 and is forecast to reach over 75 billion by 2025 (Statista, 2019), a five-fold increase over a decade. Whilst the first decade of smart technology (2007 – 2017) has focused on sensing and analysis, the next decade is anticipated to shift to predicting and actuating. Automation within the physical environment is not new. However, the change forecast by this research is a move towards cognitive automation. Currently, embedded devices enable environments to adapt to changeable conditions through simplistic 'if this… then that…' rules. By incorporating context-specific learning with self-calibration, responses can be adapted to different situations and personalised to individual needs.

The same technology is also advancing the capabilities of mobile and wearable devices. It creates the potential to move beyond adaptive environments to augmenting the spatial intelligence of individuals, creating a 'digital sixth sense' by providing information about the environment that is beyond the reaches of the person's natural senses and combining it with prior experiences and responses. Such a system would behave more like an artificial intuition, tuned not only to the landscape but also to the individual and their learned preferences within the landscape.

8.1.5 Future research directions

Contextual modelling of real-world phenomena is an immature and rapidly growing field. Many of the outcomes from this research produced more questions than answers and warrant further investigation. Three themes have emerged from the research contribution: the potential for new population measures, the creation of a behaviour patterns index, and the need to incorporate uncertainty in results.

8.1.5.1 New population measures

The case studies presented in chapter five uncovered challenges in establishing a measure for quantifying the population targeted, at-risk or affected by a dynamic phenomenon. These included the modifiable unit problems for both space and time, and also for interactions. Counts focused on building occupancy risk under-representing urban mobility and 'third space' activities that occur outside home or work whilst mobile data sources risk the opposite, under-representing people present for long periods within single locations. However, mobile data sources presented the opportunity to develop, or contribute, to new population measures in two ways. First, by converting a static measure into a dynamic social heartbeat representing the temporal variations in presence within a landscape. Second, by identifying the clustering or dispersal of presence both spatially across a landscape, and temporally by calculating and categorising different duration times.

A rating system for measuring population impact

Converting mobile device counts into an hourly weighting to identify variation from the average produced daytime weights ranging from 0.28 to 1.69 on non-event days within the QEOP. On an event day, they ranged from 0.85 to 11.86. However, rather than attempt to produce a precise multiplier for scenarios that are expected to contain uncertainty, a proposed alternative is to develop a rating system, similar to those used to forecast the impact of extreme weather such as hurricanes. The concept was explored briefly in chapter seven, to estimate the effect that the 'Black Friday' event may have on visits to Oxford Circus. However, it was a somewhat arbitrary value based on the weights developed from studying the QEOP. A future study would focus on identifying what would be an acceptable range to adjust an active population estimate for a given set of circumstances, and how well it could generalise across landscapes sharing similar attributes.

From presence counts to presence behaviours

It was demonstrated in chapter five that cluster patterns within mobile data can be used to detect active spaces within a landscape and then convert readings into indications of presence behaviours: moving, milling or dwelling whilst present at different locations across a landscape. This creates two opportunities for population measures. First, adopting a nodal approach instead of a grid-based or polygon-based approach to aggregate readings may reduce the issues of the MAUP by drawing output areas based on where people are detected. Second, a population measure could be broken down to proportions representing movement versus dwelling. A count of 100,000 people dwelling simultaneously within a single location of a landscape for the duration of a given interval will produce a very different social atmosphere to 100,000 people briefly traversing the landscape with presence dispersed across the interval.

Inferring locations in space and time from text

A further finding arose from experiments to acquire location-based tweets for analysis in chapter six. The number of tweets containing geotags positioning within the boundary of the QEOP was very small on non-event days. An alternative approach was considered: to infer location in space and time based on the content of a tweet. For example, certain phrases indicate presence such as, 'loving the great atmosphere here today', whilst others do not, such as 'can't wait for the concert next Saturday'. A set of rules was developed and showed promise but to refine them into a robust approach was beyond the scope of this thesis. It is, however, recommended as a separate future research project and could be of substantial value in many fields seeking to unlock location-based information from language-based sources that lack coordinates or time records.

8.1.5.2 Behaviour patterns index

Chapter seven applied the findings from the studies of the QEOP to three new landscapes, enabling a comparison of areas within a landscape and between landscapes. Each landscape was defined using the same grid, aligned to LandScan to enable an active population estimate. When comparing counts daily, a LandScan cell (x3y1: Waterloo station) within the Westminster landscape had more in common with cells in the London Bridge landscape than with its neighbours (see Figure 98). It suggests that land-use attributes may be an indicator for different population behaviours by

identifying the potential for different interactions to occur. However, the cells did not share the same hourly distribution throughout the day (see Figure 100), suggesting there is a high sensitivity to the specific combination of attributes located within a landscape.

When studying hourly distributions, the nine cells spanning three landscapes in chapter seven, and the cells studied within the QEOP landscape, all produced one or more of four broad distribution curves for temporal routine behaviours, visualised conceptually in Figure 122:

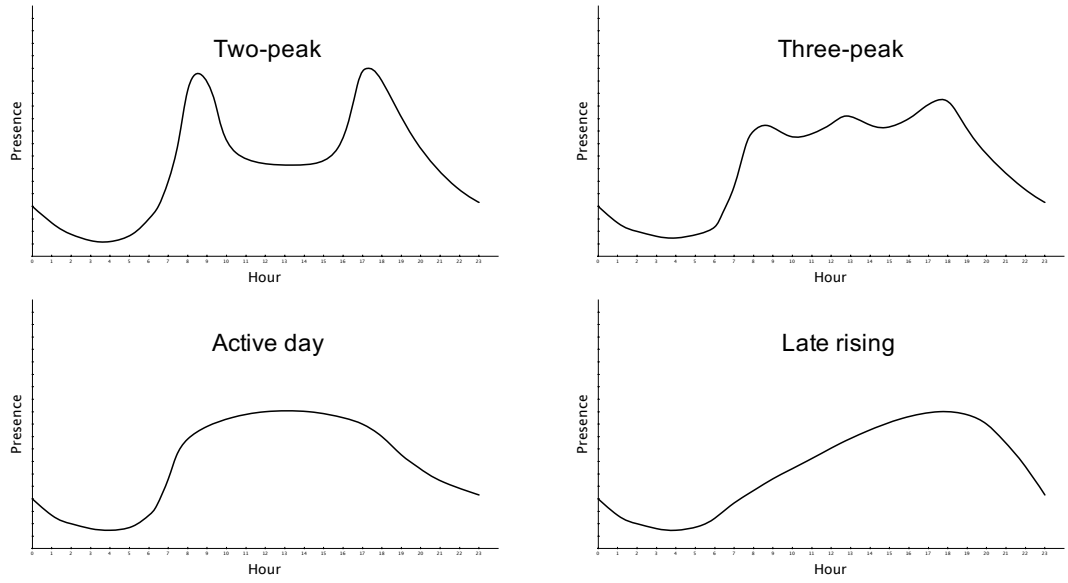


Figure 122. Presence variation over 24 hours as conceptual curves

The hourly distribution of presence across nine cells in three landscapes in London appears to be classified as four broad curves, with a size adjustment for weekday versus weekends (conceptual drawing, not based on actual data).

Most landscapes produced different curves for weekdays versus weekends, in shape and size, but all curves for normal temporal routines can be broadly described as one of these categories:

- A two-peak curve occurring in the morning and afternoon commuting periods, with a substantial dip between the peaks during daytime hours.
- A three-peak curve that aligns with morning and afternoon commutes and lunchtime periods, with dips between each peak.
- An active day curve that rises during the morning commute period, declines during the afternoon commute period but maintains a shallow curve between the two periods.
- A late-rising curve that continues rising through the day peaking late afternoon into evening, indicating the presence of evening/night-life activities within the area.

The dataset used was provided for the whole of Greater London. A logical next step would be to expand the analysis to a LandScan and Pixel grid covering London, to explore for a full set of temporal patterns and whether or not they can be associated with specific land-use attributes and situated circumstances. It creates the potential to estimate population dynamics for unobserved landscapes if they share sufficient similar attributes to an observed landscape, their ‘doppelganger’. The concept of forecasting behaviour based on doppelgangers was introduced in 2003 by Nate

Silver who applied the technique to produce the PECOTA system for forecasting player performance in Major League Baseball (Silver, 2012). Instead of studying averages across baseball players, Silver proposed that, to predict the performance of an unknown player, find the player with the most shared attributes in terms of physiology, psychology, socioeconomic status etc. The same method could be applied to landscapes and their layered contextual attributes.

8.1.5.3 The additional 'V' of big data: vagueness

The framework in this thesis proposes that the circumstances that affect human-environment interactions can be described as three interdependent layers of location-based context: space, time and situated action. A fourth layer was added – reaction sensitivity - to acknowledge the variation and uncertainty in real-world observations of human behaviour. Referring back to Lewin's equation for behaviour [1] and the COM-B model for behaviour change (Figure 5), both identify that behaviour is a complex interaction between person and environment. The three contextual layers focus on the properties of the environment. Reaction sensitivity acknowledges that there are attributes of the person that will also influence their spatial choices. For example, the motivations for being present – whether or not a given interaction is considered necessary or optional.

The intention was to produce a formula to reflect the differences in uncertainty at different time intervals and for different circumstances. Forecasting a distribution based on prior knowledge and sensed conditions can be modelled using Bayesian principles to represent uncertainty in outcomes (Halpern, 2017). However, it does not consider the veracity of the observed data. Researching different approaches, a method called 'subjective logic' showed promise. Subjective logic is based on a combination of Bayesian inference, belief theories and fuzzy logic (Jøsang, 2016). It produces probabilities based on incomplete knowledge and when it is unknown how much of behaviour is determined or random (Blastland, 2019). This is a suitable description for measuring reaction sensitivity in population behaviours. It is closely associated with models seeking to introduce quantum theoretic concepts as a new form of social physics (Wendt, 2015). Such methods are not yet established and hence this is a focus for future research.

Within chapter seven, there was evidence of variations in readings across LandScan cells when comparing the same interval, both for the day of the week and the hour of each day. This could be an indication of the volume of necessary versus optional interactions taking place, assuming necessary interactions to have less sensitivity to variable conditions (Gehl, 1987). It is also possible that it is evidence that human behaviour is simply more random and less predictable than has been assumed in some of the previous research presented in chapter two. Incorporating such vagueness as a quantitative measure would benefit real-world forecasts by emphasising how much is not known when modelling from sensed and crowd-sourced data about real-world situations.

The need to incorporate context and uncertainty when using big data in social and spatial analysis is gaining traction. It was the central theme of the annual AAG conference in 2017 and the subject of a special issue of the International Journal of Geographical Information Science in 2019 (Chun, Kwan, & Griffith, 2019). To develop the concept further, a final outcome from this research is to

propose that we also need to consider the attributes of a social phenomenon that make measuring it so difficult: agency, ambiguity and arbitrary decisions (Table 31). These three variables combined present a proposed additional 'V' in big data analytics, the vagueness described by Tukey (Tukey, 1962). It is an indication that, no matter how much is known about the conditions of a situation, there will always be some amount of randomness or unexplainable preferences in cognitive decisions that create uncertainty in predictions about the actions of people.

Table 31. The three 'A's of socio-spatial phenomena

Agency in choice	Recognising that an individual has some amount of freedom to choose their reaction to a set of circumstances and that the choice may differ for the same person experiencing the same conditions at different times, and for different people experiencing the same conditions at the same time.
Ambiguity in options	Recognising that a given situation and/or the available options to choose from may contain insufficient or missing information to make a reasoned choice based on logic. This relates to the concept of satisficing versus optimising in decision theory (Kahneman, 2003).
Arbitrary decisions	Recognising that there may be more than one equally suitable or valid choice and the selection of one of the other can only be represented as a random choice between equal options given the available information.

Theories are grounded in time and culture (Smith, 1980) and thus, may be inappropriate when applied at a different time. However, data is also not without its limitations. To advance the use of data-intensive methods in behavioural research and produce outcomes that satisfy Brewer's 3 Rs of robustness, relevance, and representativeness (Brewer, 2000) requires a language that can incorporate both uncertainty and vagueness in big data and social phenomena, not just in measurements but in communicating outcomes effectively to inform decisions and interventions that carry consequences. Computer scientists are also recognising the importance of being able to comprehend the variation in human preferences (Russell S. , 2019) and uncertainty in predicting actions (Clark, 2016) when developing artificial intelligence and computer algorithms for operation in real-world situations. This research has focused on the ability to introduce context-specific learning and regular recalibration of social-spatial models through the continuous or frequent sampling of reality. A next stage is to convert the framework into a quantitative method that better represents the vagueness and uncertainty prevalent in observations of real-world behaviours.

8.2 Closing Thoughts

> *"The greatest contribution science can make to the humanities is to demonstrate how bizarre we are as a species, and why." (Wilson, 2014)*

The above quote, by biologist Edward O. Wilson, encapsulates the ideas around which this thesis has been based, that human behaviour has too much variety to be simplified into a generalised model or universal principle absent of context. The true potential of big data analytics is in enabling a 'many models' approach to the study of real-world phenomena. In the introduction to this thesis, reference was made to the classic quote by statistician George Box, that all models are wrong and that the challenge for modellers is knowing how wrong they can be and still be useful (Box & Draper, 1987). To conclude, this research has demonstrated that the real challenge is in not knowing how wrong the model is when it is absent of context.

The novel contribution of this thesis is a framework that applies context-specific learning to sample data about real-world activities and reveal how much variation occurs in people-place experiences over time, from cultural and seasonal routines that create the 'social heartbeat' of a landscape to the arrhythmic impact of abnormal conditions. It introduces a data-driven approach to learning behaviour patterns within a landscape as an alternative to theoretical assumptions. The framework is modular and fully automated through programmatic analytics, enabling it to function standalone or be incorporated within other systems and models. It can be used to visualise the social atmosphere generated by people moving across and dwelling at different locations within the landscape, anticipate behaviour change to assist the design of interventions, and to function as a digital sixth sense, providing situated intelligence to inform and assist those present.

The premise for this research is that we have entered an era when more than half of the global population resides within urban settlements and are anticipating a future where much of the built environment will contain masses of embedded and connected sensors and actuators. Such environments will incorporate artificial intelligence (AI) 'at the edge' operating in real-time. The built environment itself is an artificial construct where social, cultural and political expectations promote and constrain human behaviour. As described by Herbert Simon, social phenomena do not have the necessity of natural phenomena (Simon, 1996). Instead, they have a contingency. Decisions and behaviours are malleable to different circumstances and as such are difficult to generalise. The availability of real-time data affords new modelling techniques that can incorporate the attributes of real-world observations (the 'Vs' of big data) and the attributes of the phenomenon being observed (the 'As' of socio-spatial behaviour).

To enable urban AIs to anticipate, adapt and respond effectively to human behaviour requires a framework that can model the complexity of situated actions and embrace uncertainty. Recent developments in the computational analysis of behaviour have focused on increasing precision by learning from massive amounts of historical data using advanced machine learning algorithms such as neural networks. However, they lack the means to represent the agency, ambiguity and arbitrary choices prevalent in everyday decisions and actions. An informed approximation or heuristic may

outperform a precise measurement when deciding an appropriate response in uncertain and changeable conditions. Thus, the novel contribution from this research is a modular framework within which to structure such heuristics, incorporating multiple inter-dependent attributes across the three axes of space, time and situation (Figure 123). The complexity of the relationships between these attributes, both within and across the contextual layers means that such a framework cannot be easily represented a single mathematic equation. Rather, it is a toolkit that can incorporate different algorithms and data sources with which to model real-world phenomena.

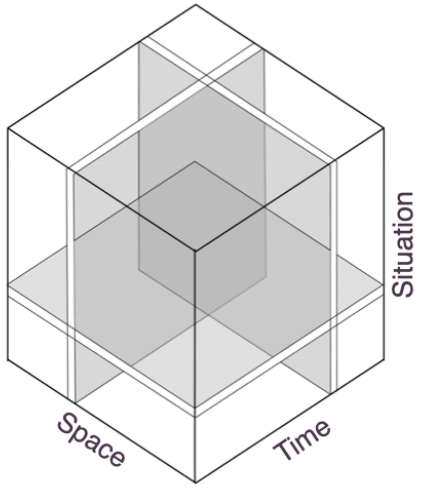


Figure 123. Representing the interdependent axes of space, time and situation

Each axis represents multiple dimensions to study human-environment interactions, based on the schematic cube for representing the dilemma of how to aggregate when studying descriptions of place by people (Canter, 1977).

Human spatial and social behaviour changes over different timescales, from the transformational change created by the physical development of a landscape, to the incremental changes due to evolving cultural habits and climate change, to the temporary adaptions created in response to a specific set of circumstances. The benefit of a continuous source of real-time data combined with a contextual framework is in being able to not only learn about the different population behaviours that occur within a landscape but also to automatically recalibrate as social routines change over time, removing the potential for historical bias to develop or linger. The model is tuned to the place and time in which it is applied and learns from situations as and when they occur.

This research began with a question that was ambitious in scope: can we make use of the data being generated by mobile devices to better understand how context affects behaviour within urban social landscapes? Just as context influences real-world decisions and actions so it should inform models evaluating or predicting real-world choices. The case studies presented in this thesis have opened up several new routes to advance data-driven theories of behaviour. It has become clear that there is not currently the scientific language to quantify social phenomena to the same level of acceptance as modelling natural phenomena. This thesis is a step towards that goal, advancing the computational analysis and visualisation of human behaviour in urban social landscapes.

Epilogue

As the writing of this thesis drew to a close, on 11 March 2020, the World Health Organisation (WHO) declared a pandemic that has caused extreme disruption to the global economy and society (Gopinath, 2020). COVID-19 is an infectious disease caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). As of 1 May 2020, there is no vaccine or cure. To contain the spread of the virus, numerous countries have implemented strict and severe social distancing measures. In an immediate response, lockdowns were initiated that mandated all but essential facilities to close and people to work from home where possible. The WHO recommends that people maintain a distance of at least 1 metre apart and anyone suspected of being infected to wear a mask whenever they leave their home (WHO, 2020). As lockdowns are eased, it is expected that social distancing measures will be required for many aspects of everyday urban life until the virus is either eradicated or a vaccine or cure developed and distributed. It is an unprecedented disruption to the socio-spatial dynamics of an urban landscape.

A finding of this research was that socio-spatial behaviour changes across three timescales. People-place interactions that arise from spatial affordances can remain unchanged for decades but be periodically transformed through developments that alter land-use, building stock and the potential for actions to occur. Temporal routines adjust incrementally in response to evolving changes in culture, society and climate whilst situated actions are temporary adaptions to infrequent disruptive abnormalities. COVID-19 demonstrates that there is a fourth timescale. Temporal routines can experience an acute shock similar to a situated action but with permanent or long-lasting effects. Research has indicated that to maintain social distancing measures according to current WHO guidelines, the effective capacity of public transport could be reduced by up to 90 per cent (Davenport & Levell, 2020) dramatically changing commuting patterns. It is anticipated that the sudden shift from office-based to remote working may become a permanent change for many roles, resulting in fewer trips for work or study and altering demand for commuter transport and associated location-based services.

The findings of this research could be applied in two ways in this unprecedented era. First, as demonstrated in chapter seven, the framework could indicate variable demand for access to the public realm across London. It could identify how streets are currently used, or have been used in the recent past, and recommend reconfigurations to increase the occupancy of under-utilised areas around hot spots, such as transport hubs and tourist attractions. As alluded to in chapter seven, it could help identify methods to reduce crowding and alter behaviour, such as simulating a transition away from mass transit to smaller vehicles able to pick-up and drop-off people at any location. Second, it demonstrates the need for models used in real-world decisions to reflect current conditions. The research has shown that a small number of participants carrying a mobile device with an app installed that emits periodic space-time readings, and/or that prompts for subjective experiences whilst present, could help recalibrate urban behaviour models for the immediate and lasting impact of COVID-19.

THE END

Bibliography

AECbytes. (2018). *LEGION and the Technology of Pedestrian Simulation.* Retrieved September 2019, from Architecture Engineering Construction: http://www.aecbytes.com/feature/2018/Legion-PedestrianSimulation.html

Ahas, R., Aasa, A., Silm, S., & Tiru, M. (2010). Daily rhythms of suburban commuters' movements in the Tallin metropolitan area: Case study with mobile positioning data. *Transportation Research Part C, 18*, 45-54.

Ashour, W., & Sunoallah, S. (2011). Multi Density DBSCAN. *Intelligent Data Engineering and Automated Learning - IDEAL 2011* (pp. 446-453). Norwich: Vol. 6936.

Balduini, M., Bozzon, A., Valle, E. D., Huang, Y., & Houben, G.-J. (2014). Recommending Venues Using Continuous Predictive Social Media Analytics. *IEEE Internet Computing, 18*(5), 28-35.

Balduini, M., Valle, E. D., Dell'Aglio, D., Tsytsarau, M., Palpanas, T., & Confalonieri, C. (2013). Social Listening of City Scale Events Using the Streaming Linked Data Framework. *The Semantic Web - ISWC 2013. 8219.* Lecture Notes in Computer Science.

Ball, P. (2015). *Describing People As Particles Isn't Always a Bad Idea: Using physics to describe social phenomena can work - if it's the right physics*. Retrieved from Nautilus: http://nautil.us/issue/33/attraction/describing-people-as-particles-isnt-always-a-bad-idea

Barker, R. G. (1968). *Ecological psychology: Concepts and methods for studying the environment of human behavior.* Standard, CA: Stanford University Press.

Barrett, L. F. (2017). *How Emotions Are Made: The Secret Life of the Brain.* New York: Pan Books (2018 electronic edition).

Batista e Silva, F., Craglia, M., Freire, S., Rosina, K., Lavalle, C., Marin, M., & Schiavina, M. (2016). *Enhancing activity and population mapping: Exploratory research project interim report.* JRC Technical Reports, Publications Office of the European Union.

Batista e Silva, F., Marin Herrera, M. A., Rosina, K., Barranco, R. R., Freire, S., & Schiavina, M. (2018). Analysing spatiotemporal patterns of tourism in Europe at high-resolution with conventional and big data sources. *Tourism Management, 68*, 101-115.

BBC News. (2020). *Climate change: Australian summers 'twice as long as winters'*. Retrieved March 2020, from BBC News: https://www.bbc.co.uk/news/world-australia-51697803

BBC News. (2020). *Storm Dennis: Met Office issues warnings for more rain and wind*. Retrieved February 2020, from BBC News: https://www.bbc.co.uk/news/uk-51501392

Benoit, K., Watanabe, K., Wang, H., Nulty, P., Obeng, A., Müller, S., & Matsuo, A. (2018). quanteda: An R package for the quantitative analysis of textual data. *The Journal of Open Source Software, 3*(30), 774.

Bentley, R. A., O'Brien, M. J., & Brock, W. A. (2014). Mapping collective behaviour in the big-data era. *Behavioral and Brain Sciences, 37*(1), 63-119.

Betchel, R. B. (1967). Human movement in architecture. *Trans-action, 4*(6), 53-56.

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

Bille, M., Bjerregaard, P., & Sørensen, T. F. (2015). Staging atmospheres: Materiality, culture, and the texture of the in-between. *Emotion, Space and Society 15*, 31-38.

Bird, S., Klein, E., & Loper, E. (2009). *Natural Language Processing with Python.* Sebastopol, CA: O'Reilly Books.

Birkin, M., & Malleson, N. (2013). *Investigating the behaviour of Twitter users to construct an individual-level model of metropolitan dynamics*. Retrieved from National Centre for Research Methods: http://eprints.ncrm.ac.uk/3166/1/complex_city_twitter_Birkin_Malleson.pdf

Blastland, M. (2019). *The Hidden Half: How the World Conceals Its Secrets.* London: Atlantic Books.

Blei, D. M. (2012). Probabilistic Topic Models. *Communications of the ACM, 55*(4), 77-84.

BNP Paribas Real Estate. (2017). *Pan-European Footfall Analysis: Key global and lifestyle cities 2017-2018.* Retrieved from https://f.datasrvr.com/fr1/717/49497/Pan_European_Footfall_Analysis_2017_2018.pdf?cbcachex=736935

Box, G. E., & Draper, N. R. (1987). *Empirical Model-Building and Response Surfaces.* Oxford: John Wiley & Sons.

Bradford Franklin, S. (2019). *AI and Government: Assessing Risks and How to Mitigate Them*. Retrieved December 2019, from Stanford Fall Conference 2019: AI Ethics, Policy and Governance: https://hai.stanford.edu/events/2019-fall-conference/agenda

Brewer, M. B. (2000). Research design and issues of validity. In H. T. Reis, & C. M. Judd. (Eds.), *Handbook of research methods in social and personality psychology* (pp. 3-16). New York: Cambridge University Press.

Burnap, P., Rana, O., Williams, M., Housley, W., Edwards, A., Morgan, J., . . . Conejero, J. (2015). COSMOS: Towards an integrated and scalable service for analysing social media on demand. *International Journal of Parallel, Emergent and Distributed Systems, 30*(2), 80-100.

Cöltekin, A., De Sabbata, S., Willi, C., Vontobel, I., Pfister, S., Kuhn, M., & Lacayo, M. (2011). Modifiable Temporal Unit Problem. *Persistent problems in geographic visualization; ICC2011 Workshop.* Paris.

Canter, D. (1977). *The Psychology of Place.* London: The Architectural Press Ltd.

Carr, S., Francis, M., Rivlin, L. G., & Stone, A. M. (1992). *Public Space.* Cambridge: Cambridge University Press.

Charles-Edwards, E., & Bell, M. (2013). Estimating the Service Population of a Large Metropolitan University Campus. *Applied Spatial Analysis and Policy, 6*(3), 209-228.

Cheng, T., Bowers, K., Longley, P., Shawe-Taylor, J., Davies, T., Rosser, G., . . . Skarlatidou, A. (2016). *CPC: Crime, Policing, Citizenship - Intelligent Policing and Big Data*. Retrieved from ResearchGate: https://www.researchgate.net/publication/303539780_CPC_Crime_Policing_and_Citizenship_-_Intelligent_policing_and_big_data

Chun, Y., Kwan, M.-P., & Griffith, D. A. (2019). Uncertainty and context in GIScience and geography: challenges in the era of geospatial big data. *International Journal of Geographical Information Science, 33*(6), 1131-1134.

Clark, A. (2016). *Surfing Uncertainty: Prediction, Action and the Embodied Mind.* Oxford: Oxford University Press.

Cockings, S., & Martin, D. (2018). Using new and emerging forms of data to produce enhanced spatiotemporal population estimates. *26th GIScience Research UK Conference, University of Leicester.* Leicester: GISRUK.

Cornan, S. R., Kuhn, T., McPhee, R. D., & Dooley, K. J. (2002). Studying Complex Discursive Systems. *Human Communications Research, 28*(2), 157-206.

Cranshaw, J., Schwartz, R., Hong, J. I., & Sadeh, N. (2012). The Livehoods Project: Utilizing Social Media to Understand the Dynamics of a City. *6th International AAAI Conference on Weblogs and Social Media.* Dublin.

Davenport, A., & Levell, P. (2020, May). *Changes down the line: flattening the curve of public transport use*. Retrieved from Institute for Fiscal Studies: https://www.ifs.org.uk/publications/14844

Davies, N., & Clinch, S. (2017). Pervasive Data Science: New Challenges at the Intersection of Data Science and Pervasive Computing. *IEEE Pervasive Computing, 16*(3), 50-58.

de Jong, R., & de Bruin, S. (2012). Linear trends in seasonal vegetation time series and the modifiable temporal unit problem. *Biogeosciences, 9*(1), 71-77.

de Montjoye, Y.-A., Hidalgo, C. A., Verleysen, M., & Blondel, V. D. (2013). Unique in the Crowd: The privacy bounds of human mobility. *Scientific Reports, 3*, 1376.

Deville, P., Linard, C., Martin, S., Gilbert, M., Stevens, F., Gaughan, A., . . . Tatem, A. J. (2014). Dynamic population mapping using mobile phone data. *Proceedings of the National Academy of Sciences, 111*(45), 15888-15893.

Devlin, A. S. (2012). Environmental Perception: Wayfinding and Spatial Cognition. In S. D. Clayton (Ed.), *The Oxford Handbook of Environmental and Conversation Psychology.* Oxford: Oxford University Press.

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithmic Aversion: People Erroneously Avoid Algorithms After Seeing Them Err. *Journal of Experimental Psychology: General, 144*(1), 114-126.

Duarte, F., & Ratti, C. (2016). Smart Cities, Big Data, and the Internet of Things. *IEEE Standards University Smart City Standards, 6*(4).

Eagle, N., & Pentland, A. (2006). Reality mining: sensing complex social systems. *Personal and Ubiquitous Computing, 10*(4), 255-268.

Erkan, Ì., & Hastemoglu, H. S. (2016). Analyzing Level of Service Through Anthropometric Scale and Its Contribution to Transportation Engineering. *International Journal of Civil Engineering, 14*(8A), 585-593.

Ester, M. (2014). Density-Based Clustering. In C. C. Aggarwal, & C. K. Reddy, *Data Clustering: Algorithms and Applications* (pp. 111-127). New York: Chapman and Hall/CRC.

Ester, M., Kriegel, J.-P., Sander, J., & Xu, X. (1996). A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. *Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96).* Portland, OR.

Fortunato, S. (2010). Community detection in graphs. *Physics Reports, 486*(3-5), 75-174.

Foth, M. (2009). *Handbook of Research on Urban Informatics: The Practice and Promise of the Real-Time City.* London: Information Science Reference.

Foth, M., Choi, J. H.-j., & Satchell, C. (2011). Urban Informatics. *CSCW '11 Proceedings of the ACM 2011 conference on Computer supported cooperative work*, (pp. 1-8).

Frank, M. R., Mitchell, L., Dodds, P. S., & Danforth, C. M. (2013). The Geography of Happiness: Connecting Twitter Sentiment and Expression, Demographics, and Objective Characteristics of Place. *Scientific Reports 3*.

Fruin, J. J. (1971). Designing for Pedestrians: A Level-of-Service concept. *50th Annual Meeting of the Highway Research Board*, (pp. 1-15). Washington District of Columbia.

GDRP.EU. (n.d.). *Complete guide to GDPR compliance*. Retrieved December 2018, from GDRP.EU: https://gdpr.eu

Gehike, C. E., & Biehl, K. (1934). Certain Effects of Grouping Upon the Size of the Correlation Coefficient in Census Tract Material. *Journal of the American Statistical Association, 29*(185A), 169-170.

Gehl, J. (1987). *Life Between Buildings: Using Public Space, translation by Jo Koch.* Washington, DC: Island Press, 2011 edition.

Géron, A. (2017). *Hands-on Machine Learning with Scikit-Learn and TensorFlow.* Sebastopol, CA: O'Reilly Media.

Gibson, J. J. (1979). *The Ecological Approach to Visual Perception.* Boston: Routledge edition, published 1986.

Gifford (Ed.), R. (2016). *Research Methods for Environnmental Psychology.* Chichester: Wiley Blackwell.

Glaeser, E. L., Gottlieb, J. D., & Ziv, O. (2016). Unhappy Cities. *Journal of Labor Economics, 34*(2), S129-S182.

González, M., Hidalgo, C. A., & Barabási, A.-L. (2008). Understanding individual human mobility patterns. *Nature, 453*, 779-782.

Goodman, N. (1990). *Fact, Fiction and Forecast: Fourth Edition Paperback.* Cambridge, MA: Harvard University Press.

Gopinath, G. (2020, April). *The Great Lockdown: Worst Economic Downturn Since the Great Depression*. Retrieved April 2020, from International Monetary Fund: https://blogs.imf.org/2020/04/14/the-great-lockdown-worst-economic-downturn-since-the-great-depression/

Graham, L. T., & Gosling, S. D. (2011). Can the Ambiance of a Place be Determined by the User Profiles of the People Who Visit It? *Fifth International AAAI Conference on Weblogs and Social Media.* Barcelona.

Gray, J. (2009). eScience: A Transformed Scientific Method. In T. Hey, S. Tansley, & K. Tolle (Eds.), *The Fourth Paradigm: Data-Intensive Scientific Discovery (Second printing, version 1.1).* Redmond, WA: Microsoft Research.

Greger, K. (2015). Spatio-Temporal Building Population Estimation for Highly Urbanized Areas Using GIS. *Transactions in GIS, 19*(1), 129-150.

Guszcza, J. (2015). *The Last-Mile problem: How data science and behavioural science can work together.* Retrieved December 2015, from Deloitte Review Issue 16: http://dupress.com/articles/behavioral-economics-predictive-analytics/

H.M. Coroner. (2018). *Documents, Submissions, and Rulings.* Retrieved February 2019, from Westminster Bridge Inquests: https://westminsterbridgeinquests.independent.gov.uk/documents-and-rulings/

H.M. Coroner. (2019). *Hearing Transcripts.* Retrieved from London Bridge Inquests: https://londonbridgeinquests.independent.gov.uk/hearing-transcripts/

Hägerstrand, T. (1989). Reflections on 'What about people in regional science?'. *Papers in Regional Science, 66*(1), 1-6.

Halpern, J. Y. (2017). *Reasoning about Uncertainty: Second Edition.* Cambridge, MA: The MIT Press.

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Delving Deep into Rectifiers: Surpassing. Human-Level Performance on ImageNet Classification. *arXiv: 1502.0185*.

Hill, M. R. (1984). Stalking the Urban Pedestrian: A Comparison of Questionnaire and Tracking Methodologies for Behavioural Mapping in Large-Scale Environments. *Environment and Behaviour, 16*(5), 539-550.

Hillier, B., & Hanson, J. (1984). *The Social Logic of Space.* Cambridge: Cambridge University Press.

Inkersole, S. (2017, March 14). *West Ham fans have already renewed a record number of season tickets - after just 24 hours on sale*. Retrieved September 2018, from football.london: https://www.football.london/west-ham-united-fc/news/west-ham-fans-already-renewed-12737605

Innes, M. (2007). Investigation order and major crime inquiries. In T. Newburn, T. Williamson, & A. Wright (Eds), *Handbook of Criminal Investigation* (pp. 255-276). Abingdon, Oxon: Willan Publishing.

Izard, C. E. (2010). The Many Meanings/Aspects of Emotion: Definitions, Functions, Activation and Regulation. *Emotion Review, 2*(4), 363-370.

Jøsang, A. (2016). *Subjective Logic: A Formalism for Reasoning Under Uncertainty.* Cham, Switzerland: Springer International.

Jacobs, J. (1961). *The Death and Life of Great American Cities.* New York: Vintage Books Edition, December 1992.

Jaderberg, M., Czarnecki, W. M., Dunning, I., Marris, L., Lever, G., Castañdea, A. G., . . . Silver, D. (2019). Human-level performance in 3D multiplayer games with population-based reinforcement learning. *Science, 364*(6443), 859-865.

Jayarajah, K., Balan, R. K., Radhakrishnan, M., Misra, A., & Lee, Y. (2016). LiveLabs: Building In-Situ Mobile Sensing & Behavioural Experimentation TestBeds. *MobiSys '16.* Singapore: ACM.

Kahneman, D. (2003). Maps of Bounded Rationality: Psychology for Behavioural Econonmics. *American Economic Review, 93*(5), 1449-1475.

Katz, E., Lazersfeld, P. F., & Roper, E. (2005). *Personal Influence: The Part Played by People in the Flow of Mass Communications (re-release).* New York: Routledge.

Kitchen, R., & Freundschuh, S. (2000). *Cognitive Mapping: Past, Present and Future.* Abingdon, Oxon: Routledge.

Kolaczyk, E. D. (2009). *Statistical analysis of network data: methods and models.* New York: Springer.

Konner, M. (2003). *The Tangled Wing.* New York: First Owl Books edition.

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

Laney, D. (2001). *3D Data Management: Controlling Data Volume, Velocity and Variety.* Retrieved January 2016, from https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabasi, A. L., Brewer, D., . . . Van Alstyne, M. (2009). Life in the network: the coming age of computational social science. *Science, 323*(5915), 721-723.

LeDoux, J. (1998). *The Emotional Brain.* London: Phoenix Books.

Lewin, K. (1936). *Principles of Topological Psychology.* New York: McGraw-Hill.

Leyden, K. M., Goldberg, A., & Michelbach, P. (2011). Understanding the Pursuit of Happiness in Ten Major Cities. *Urban Affairs Review, 47*(6), 861-888.

Li, S., Dragicevic, S., Castro, F. A., Sester, M., Winter, S., Coltekin, A., . . . Cheng, T. (2016). Geospatial Big Data Handling Theory and Methods: A Review and Research Challenges. *ISPRS Journal of Photogrammetry and Remote Sensing, 115*, 119-133.

Li, X., Zhang, C., Li, W., Ricard, R., Meng, Q., & Zhang, W. (2015). Assessing street-level urban greenery using Google Street View and a modified green view index. *Urban Forestry & Urban Greening, 14*, 675-685.

Liu, B. (2015). *Sentiment Analysis.* Cambridge, UK: Cambridge University Press.

Liu, Y., Gao, S., Kang, C., Zhi, Y., Guanghua, C., & Shi, L. (2015). Social Sensing: A New Approach to Understanding Our Socioeconomic Environments. *Annals of the Association of American Geographers, 105*(3), 512-530.

LLDC. (2012). *Our Story.* Retrieved Dec 2018, from Queen Elizabeth Olympic Park: https://www.queenelizabetholympicpark.co.uk/our-story

LLDC. (2016). *Ten Year Plan: 2015/16 - 2024/25*. Retrieved December 2018, from Queen Elizabeth Olympic Park: https://www.queenelizabetholympicpark.co.uk/~/media/lldc/ten year plan.pdf

Longley, P. A., & Adnan, M. (2016). Geo-temporal Twitter demographics. *International Journal of Geographic Information Science, 30*(2), 369-389.

Malleson, N., & Andresen, M. A. (2016). Exploring the impact of ambient population measures on London crime spots. *Journal of Criminal Justice, 46*, 52-63.

Martí, P., Serrano-Estrada, L., & Nolasco-Cirugeda, A. (2017). Using locative social media and urban cartographies to identiy and locate successful urban plazas. *Cities, 64*, 66-78.

Martin, D., Cockings, S., & Leung, S. (2009). Population 24/7: building time-specific population grid models. *European Forum for Geostatistics Conference 2009.* Netherlands.

Martin, D., Cockings, S., & Leung, S. (2015). Developing a Flexible Framework for Spatiotemporal Population Modeling. *Annals of the Association of American Geographers, 105*(4), 754-772.

Martino, M. (2019). *Art + AI = Tool for Artists.* Retrieved 2019 September, from MIT-IBM Research AI: http://ganocracy.csail.mit.edu/slides/Martino_MIT_GANocracy_May_31_2019.pdf

Martino, M., Strobelt, H., Cornec, O., & Phibbs, E. (2017). *Forma Fluens - abstraction, simultaneity and symbolization in drawings*. Retrieved November 2017, from Forma Fluens: http://formafluens.io

Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications. *Ain Shams Engineering Journal, 5*, 1093-1113.

Merton, R. K. (1967). On Sociological Theories of the Middle Range. In R. K. Merton, *On Theoretical Sociology: Five Essays, Old and New* (pp. 39-72). New York: The Free Press.

Met Office. (n.d.). *Daily Weather Summary*. Retrieved November 2016, from Met Office: https://www.metoffice.gov.uk/research/library-and-archive/publications/daily-weather-summary

Michie, S., van Stralen, M. M., & West, R. (2011). The behaviour change wheel: A new method for characterising and designing behaviour change interventions. *Implementation Science, 6*(42).

Miller, H. J. (2010). The Data Avalanche is Here. Shouldn't We Be Digging? *Journal of Regional Science, 50*(1), 181-201.

Miller, H. J. (2017). Geographic information science I: Geographic information observatories and opportunitistic GIS science. *Progress in Human Geography, 4*(4), 489-500.

Miller, H. J., & Goodchild, M. F. (2015). Data-driven geography. *GeoJournal, 80*, 449-461.

Mohan, P., & Singh, M. (2013). Formal Models for Context Aware Computing. *International Journal of Computer Applications Technology and Research, 2*(1), 53-58.

Musto, C., Semeraro, G., Lops, P., & de Gemmis, M. (2015). CrowdPulse: A framework for real-time semantic analysis of social streams. *Information Systems, 54*, 127-146.

National Research Council. (2013). *Frontiers in Massive Data Analysis.* Washington, DC: The National Academies Press.

Nettle, D. (2005). *Happiness: The Science Behind Your Smile.* Oxford: Oxford University Press.

Neutens, T., Witlox, F., & De Maeyer, P. (2007). Individual accessibility and travel possibilities: A literature review on time geography. *European Journal of Transport and Infrastructure Research, 7*, 335-352.

NOAA. (n.d.). *Saffir-Simpson Hurrican Wind Scale*. Retrieved December 2019, from National Hurricane Center and Central Pacific Hurricane Center: https://www.nhc.noaa.gov/aboutsshws.php

NOAA. (n.d.). *The Enhanced Fujita Scale*. Retrieved December 2019, from National Oceanic and Atmospheric Adminsitration: National Weather Service: https://www.weather.gov/oun/efscale

nomis. (2014, May 23). *Census 2011 > Workday Population > Population density*. Retrieved May 2020, from nomis official labour market statistics: https://www.nomisweb.co.uk/census/2011/wd102ew

Norman, D. A. (1999). Affordance, conventions, and design. *Interactions, 6*(3), 38-43.

Noulas, A., Scellato, S., Lambiotte, R., Pontil, M., & Mascolo, C. (2012). A Tale of Many Cities: Universal Patterns in Human Mobility. *PLoS ONE, 7*(5).

Nyhan, M., Grauwin, S., Britter, R., Misstear, B., McNabola, A., Laden, F., . . . Ratti, C. (2016). "Exposure Track" - The Impact of Mobile-Device-Based Mobility Patterns on Quantifying Population Exposure to Air Pollution. *Environmental Science and Technology, 50*, 9671-9681.

OECD. (2013). *OECD Guidelines on Measuring Subjective Well-being.* Retrieved June 2017, from OECD iLibrary: https://www.oecd-ilibrary.org/economics/oecd-guidelines-on-measuring-subjective-well-being_9789264191655-en

OED. (2004). *Concise Oxford English Dictionary, Eleventh Edition.* Oxford: Oxford University Press.

Ofcom. (2011). *A nation addicted to smartphones.* Retrieved July 15, 2015, from Ofcom: http://media.ofcom.org.uk/news/2011/a-nation-addicted-to-smartphones/

Ofcom. (2017). *Fast facts: Proportion of adults with a smartphone.* Retrieved December 27, 2017, from Ofcom: https://www.ofcom.org.uk/about-ofcom/latest/media/facts

Office of Rail and Road. (2018). *Estimates of Station Usage 2017-2018.* Retrieved February 2019, from ORR Data Portal: https://dataportal.orr.gov.uk/media/1214/estimates-of-station-usage-2017-18-key-facts.pdf

O'Neill, E., Kostakos, V., Kindberg, T., Schiek, A. F., Penn, A., Fraser, D. S., & Jones, T. (2006). Instrumenting the City: Developing Methods for Observing and Understanding the Digital Cityscape. *Ubicomp 2006* (pp. 315-332). Lecture Notes in Computer Science number 4206.

ONS. (2015). *Nearly one in five people had some form of disability in England and Wales.* Retrieved November 2016, from Office for National Statistics: https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/disability/articles/nearlyoneinfivepeoplehadsomeformofdisabilityinenglandandwales/2015-07-13

ONS. (2016, January 18). *Output geography*. Retrieved May 2017, from Office for National Statistics: https://www.ons.gov.uk/census/2011census/howourcensusworks/howwetookthe2011census/howweplannedfordatadelivery/outputgeography

Openshaw, S. (1984). Ecological fallacies and the analysis of areal census data. *Environment and planning. A, 16*(1), 17-31.

Openshaw, S., & Taylor, P. J. (1981). The modifiable areal unit problem. In N. Wrigley, & R. J. Bennett (Eds.), *Quantitative Geography: A British View* (pp. 60-70). London: Routledge and Kegan Paul.

ORNL. (n.d.). *Documentation*. Retrieved August 2017, from LandScan: https://landscan.ornl.gov/index.php/documentation

Orr, S., Paskins, J., & Chaytor, S. (2014). *UCL Policy Briefing - October 2014: Valuing Urban Green Space: Challenges and Opportunities*. Retrieved January 2016, from UCL Public Policy: https://www.ucl.ac.uk/public-policy/sites/public-policy/files/migrated-files/urban_green_spaces_briefing_FINAL.pdf

Oxford Poverty & Human Development Initiative. (n.d.). *Bhutan's Gross National Happiness Index*. Retrieved December 2019, from Oxford Poverty & Human Development Initiative: https://ophi.org.uk/policy/national-policy/gross-national-happiness-index/

Paldino, S., Kondor, D., Bojic, I., Sobolevsky, S., González, M. C., & Ratti, C. (2016). Uncovering Urban Temporal Patterns from Geo-Tagged Photography. *PLoS ONE, 11*(12).

Penn, A. (2003). Space Syntax and Spatial Cognition: Or Why the Axial Line? *Environment and Behavior, 35*(1), 30-65.

Pentland, A. (2014). *Social Physics: How Good Ideas Spread.* New York: The Penguin Press.

Provalis Research. (n.d.). *Regressive Imagery Dictionary*. Retrieved December 2017, from WordStat: https://provalisresearch.com/products/content-analysis-software/wordstat-dictionary/regressive-imagery-dictionary/

Provost, F., & Fawcett, T. (2013). *Data Science for Business: What You Need to Know About Data Mining and Data-Analytic Thinking.* Sebastopol, CA: O'Reilly Media.

Pushkarev, B., & Zupan, J. M. (1975). *Urban Space for Pedestrians: A Report of the Regional Plan Association.* Cambridge, MA: MIT Press.

PwC and Demos. (2016). *Good growth for cities 2016: A report on urban economic wellbeing from PwC and Demos*. Retrieved November 2017, from Demos: https://demos.co.uk/project/good-growth-for-cities-2016/

quanteda. (n.d.). *stopwords function*. Retrieved February 2020, from quanteda: https://quanteda.io/reference/stopwords.html

Quercia, D., Aiello, L. M., Schifanella, R., & Davies, A. (2015). The Digital Life of Walkable Streets. *WWW '15: Proceedings of the 24th International Conference on World Wide Web* (pp. 875-884). Florence: ACM.

Quercia, D., Schifanella, R., & Aiello, L. M. (2014). The Shortest Path to Happiness: Recommending Beautiful, Quiet, and Happy Routes in the City. *HT '14: Proceedings of the 25th ACM conference on Hypertext and social media.* Santiago: ACM.

Ratti, C., Pulselli, R. M., & Williams, S. (2006). Mobile Landscapes: using location data from cell phones for urban analysis. *Environment and Planning B: Planning and Design, 33*, 727-748.

Richardson, S. (2015). Measuring Mobile Digital Footprints: A Modern Index of Urban Interaction, Masters dissertation. University College London.

Richardson, S. (2019). Predicting Presence in Urban Outdoor Spaces. *IEEE Pervasive Computing, 8*(3), 21-30.

Richardson, S., & Hudson-Smith, A. (2018). Estimating the Active Population of an Urban Outdoor Space in London from a Sample of Mobile Data Traces. *26th GIScience Research UK Conference, University of Leicester.* Leicester: GISRUK.

Russell, M. A. (2013). *Mining the Social Web: Facebook, Twitter, Linkedin, Google+, Github, And More, Second Edition.* Sebastopol, CA: O'Reilly Media.

Russell, S. (2019). *Human Compatible: AI and the Problem of Control.* UK: Allen Lane.

Russell, S. J., & Norvig, P. (2014). *Artificial Intelligence: A Modern Approach. Third Edition.* Harlow: Peason Education Limited.

Sap, M., Card, D., Gabriel, S., Choi, Y., & Smith, N. A. (2019). The Risk of Racial Bias in Hate Speech Detection. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 1668-1678). Florence, Italy: Association for Computational Linguistics.

Satyanarayanan, M. (2001). Pervasive Computing: Vision and Challenges. *IEEE, 8*, 10-17.

Seresinhe, C., Moat, S. H., & Preis, T. (2015). Quantifying the impact of scenic environments on health. *Scientific Reports*(5:16899).

Settanni, M., & Marengo, D. (2015). Sharing feelings online: studying emotional well-being via automated text analysis of Facebook posts. *Frontiers in Psychology, 6*, Article 1045.

Silver, N. (2012). *The Signal and the Noise: The Art and Science of Prediction.* London: Allen Lane.

Simon, H. A. (1996). *The Sciences of the Artificial. Third Edition.* Cambridge, MA: MIT Press.

Smart London Board. (2013). *Smart London Plan.* Retrieved July 2015, from https://www.london.gov.uk/sites/default/files/smart_london_plan.pdf

Smith, M. P. (1980). *The City and Social Theory.* Oxford: Blackwell.

Statista. (2019). *Internet of Things (IoT) connected devices installed base worlwide from 2015 to 2025.* Retrieved Dec 16, 2019, from Statista: https://www.statista.com/statistics/471264/iot-number-of-connected-devices-worldwide/

Still, G. K. (2000). Crowd Dynamics. *PhD Thesis. University of Warwick*.

Sui, D. (2012). Looking through Hägerstrand's dual vistas: towards a unifying framework for time geography. *Journal of Transport Geography, 23*, 5-16.

Sutton, P. C., Elvidge, C., & Obremski, T. (2003). Building and Evaluating Models to Estimate Ambient Population Density. *American Society for Photogrammetry and Remote Sensing, 69*(5), 545-553.

TfL. (2007). *Gender Equality Scheme: 2007 - 2010*. Retrieved June 2019, from Transport for London: content.tfl.gov.uk/gender-equality-scheme-2007-2010.pdf

TfL. (2018). *Our open data - Network statistics*. Retrieved September 2018, from Transport for London: https://tfl.gov.uk/info-for/open-data-users/our-open-data?intcmp=3671

The Editors of Encyclopaedia Britannica. (2015). *Nuclear family*. Retrieved February 2020, from Encyclopædia Britannica: https://www.britannica.com/topic/nuclear-family

Thelwall, M. (2017). Heart and Soul: Sentiment Strength Detection in the Social Web with SentiStrength. In H. Janusz (Ed.), *Cyberemotions: Collective Emotions in Cyberspace* (pp. 119-134). Cham, Switzerland: Springer International.

Tukey, J. W. (1962). The Future of Data Analysis. *Mathematical Statistics, 33*(1), 1-67.

Twitter. (n.d.). *Search Tweets*. Retrieved March 2016, from Twitter API Documentation: https://developer.twitter.com/en/docs/tweets/search/overview

Vanderbilt, T. (2012). *Sidewalk Science: The peculiar habits of the pedestrian explained*. Retrieved January 2016, from Slate Magazine: http://www.slate.com/articles/life/walking/2012/04/walking_in_america_what_scientists_know_about_how_pedestrians_really_behave_.html?via=gdpr-consent

Webb, W. (2017, April 3rd). Meeting between Sharon Richardson and William Webb to discuss OpenSignal dataset.

Weiser, M. (1991). The Computer for the 21st Century. *Scientific American, 265*(3), 94-105.

Weisner, M., & Seely Brown, J. (1995). *Designing Calm Technology*. Retrieved December 2018, from Archive of ubiq.com (no longer online): https://www.karlstechnology.com/blog/designing-calm-technology/

Wendt, A. (2015). *Quantum Mind and Social Science: Unifying physical and social ontology.* Cambridge: Cambridge University Press.

WHO. (2020, April 29). *Coronavirus disease (COVID-19) advice for the public*. Retrieved May 2020, from World Health Organisation: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/advice-for-public

Whyte, W. (1980). The Social Life of Small Urban Spaces (Documentary written, directed and narrated by William H. Whyte). New York: Project for Public Spaces Inc.

WiFiSpark. (2016). Estimating the range of WiFi Access Points. Discussion via email with Alex Howe, Network Operations Engineer.

Williams, M. L., Burnap, P., & Sloan, L. (2016). Crime sensing with big data: The affordances and limiations of using open-source communications to estimate crime patterns. *British Journal of Criminology. Advanced access, published March 2016*.

Wilson, E. O. (2014). *The Meaning of Human Existence.* New York: Liveright Publishing Corporation.

Wortley, R. (2012). Exploring the Person-Situation Interaction in Situational Crime Prevention. In N. Tilley, G. Farrell, & (Eds), *The Reasoning Criminologist: Essays in Honour of Ronald V. Clarke.* London: Routledge.

Zhou, Y., De, S., & Moessner, K. (2016). Real world city event extraction from Twitter data streams. *International Workshop on Data Mining on IoT Systems. 98*, pp. 443-338. Procedia Computer Science.

Appendix A: Supplemental Information

A.1 UK climate summary

Source: https://weatherspark.com/y/45062/Average-Weather-in-London-United-Kingdom-Year-Round accessed 18 August 2019.

A.3 English stop word list

The following are the lists of English stop words that are removed when applied to tokenized texts using NLP tools such as NLTK (in Python) and Quanteda (in R).

This version is from Quanteda as of 1 February 2020.

```
 [1] "i"   "me"   "my"   "myself"  "we"   "our"
 [7] "ours"  "ourselves" "you"  "your"  "yours"  "yourself"
 [13] "yourselves" "he"   "him"  "his"  "himself" "she"
 [19] "her"  "hers"  "herself" "it"   "its"  "itself"
 [25] "they"  "them"  "their"  "theirs"  "themselves" "what"
 [31] "which"  "who"  "whom"  "this"  "that"  "these"
 [37] "those"  "am"   "is"   "are"  "was"  "were"
 [43] "be"   "been"  "being"  "have"  "has"  "had"
 [49] "having"  "do"   "does"  "did"  "doing"  "would"
 [55] "should"  "could"  "ought"  "i'm"  "you're"  "he's"
 [61] "she's"  "it's"  "we're"  "they're" "i've"  "you've"
 [67] "we've"  "they've" "i'd"  "you'd"  "he'd"  "she'd"
 [73] "we'd"  "they'd"  "i'll"  "you'll"  "he'll"  "she'll"
 [79] "we'll"  "they'll" "isn't"  "aren't"  "wasn't"  "weren't"
 [85] "hasn't"  "haven't" "hadn't"  "doesn't" "don't"  "didn't"
 [91] "won't"  "wouldn't" "shan't"  "shouldn't" "can't"  "cannot"
 [97] "couldn't" "mustn't" "let's"  "that's"  "who's"  "what's"
[103] "here's"  "there's" "when's"  "where's" "why's"  "how's"
[109] "a"   "an"   "the"  "and"  "but"  "if"
[115] "or"   "because" "as"   "until"  "while"  "of"
[121] "at"   "by"   "for"  "with"  "about"  "against"
[127] "between" "into"  "through" "during"  "before"  "after"
[133] "above"  "below"  "to"   "from"  "up"   "down"
[139] "in"   "out"  "on"   "off"  "over"  "under"
[145] "again"  "further" "then"  "once"  "here"  "there"
[151] "when"  "where"  "why"  "how"  "all"  "any"
[157] "both"  "each"  "few"  "more"  "most"  "other"
[163] "some"  "such"  "no"   "nor"  "not"  "only"
[169] "own"  "same"  "so"   "than"  "too"  "very"
[175] "will"
```

The Quanteda English stopwords list is built from The SMART information retrieval system (obtained from Lewis, David D., et al. "Rcv1: A new benchmark collection for text categorization research." Journal of machine learning research (2004, 5 April): 361-397. (quanteda, n.d.)

The following is the list of NLTK English stopwords as of July 2018:

```
['i', 'me', 'my', 'myself', 'we', 'our', 'ours', 'ourselves', 'you', "you're",
"you've", "you'll", "you'd", 'your', 'yours', 'yourself', 'yourselves', 'he',
'him', 'his', 'himself', 'she', "she's", 'her', 'hers', 'herself', 'it', "it's",
'its', 'itself', 'they', 'them', 'their', 'theirs', 'themselves', 'what', 'which',
'who', 'whom', 'this', 'that', "that'll", 'these', 'those', 'am', 'is', 'are',
'was', 'were', 'be', 'been', 'being', 'have', 'has', 'had', 'having', 'do',
'does', 'did', 'doing', 'a', 'an', 'the', 'and', 'but', 'if', 'or', 'because',
'as', 'until', 'while', 'of', 'at', 'by', 'for', 'with', 'about', 'against',
'between', 'into', 'through', 'during', 'before', 'after', 'above', 'below', 'to',
'from', 'up', 'down', 'in', 'out', 'on', 'off', 'over', 'under', 'again',
'further', 'then', 'once', 'here', 'there', 'when', 'where', 'why', 'how', 'all',
'any', 'both', 'each', 'few', 'more', 'most', 'other', 'some', 'such', 'no',
'nor', 'not', 'only', 'own', 'same', 'so', 'than', 'too', 'very', 's', 't', 'can',
'will', 'just', 'don', "don't", 'should', "should've", 'now', 'd', 'll', 'm', 'o',
're', 've', 'y', 'ain', 'aren', "aren't", 'couldn', "couldn't", 'didn', "didn't",
'doesn', "doesn't", 'hadn', "hadn't", 'hasn', "hasn't", 'haven', "haven't", 'isn',
"isn't", 'ma', 'mightn', "mightn't", 'mustn', "mustn't", 'needn', "needn't",
'shan', "shan't", 'shouldn', "shouldn't", 'wasn', "wasn't", 'weren', "weren't",
'won', "won't", 'wouldn', "wouldn't"]
```

PAGE INTENTIONALLY LEFT BLANK

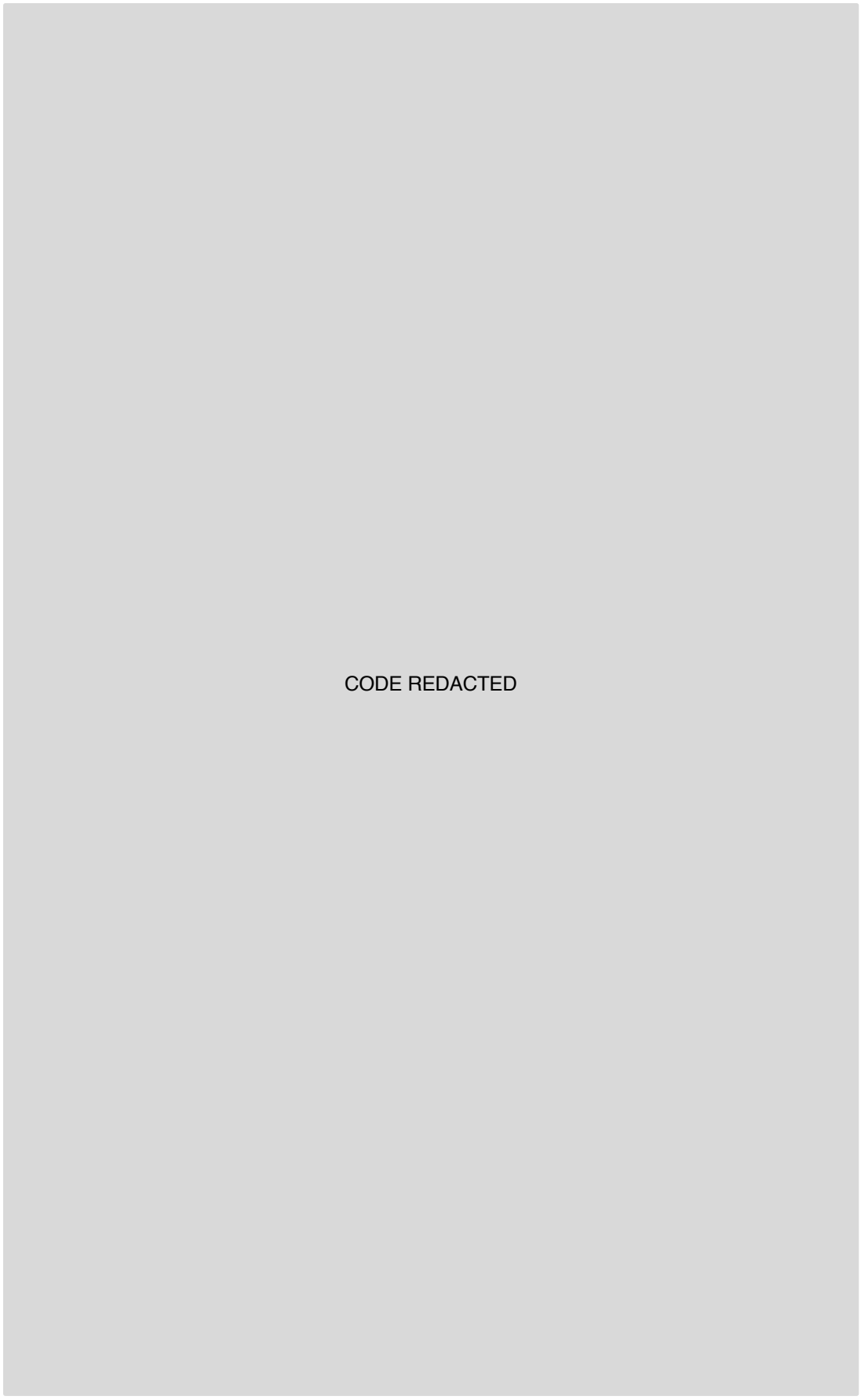
Appendix B: Code Samples

All studies involved the use of scripts developed iteratively. Code samples are reproduced here for key stages for reference. The objective of this research was not to create a production-ready model and approaches taken may not represent the most efficient coding techniques.

B.1 Data acquisition and pre-processing

Mobile data: Wi-Fi network readings from QEOP

CODE REDACTED


CODE REDACTED

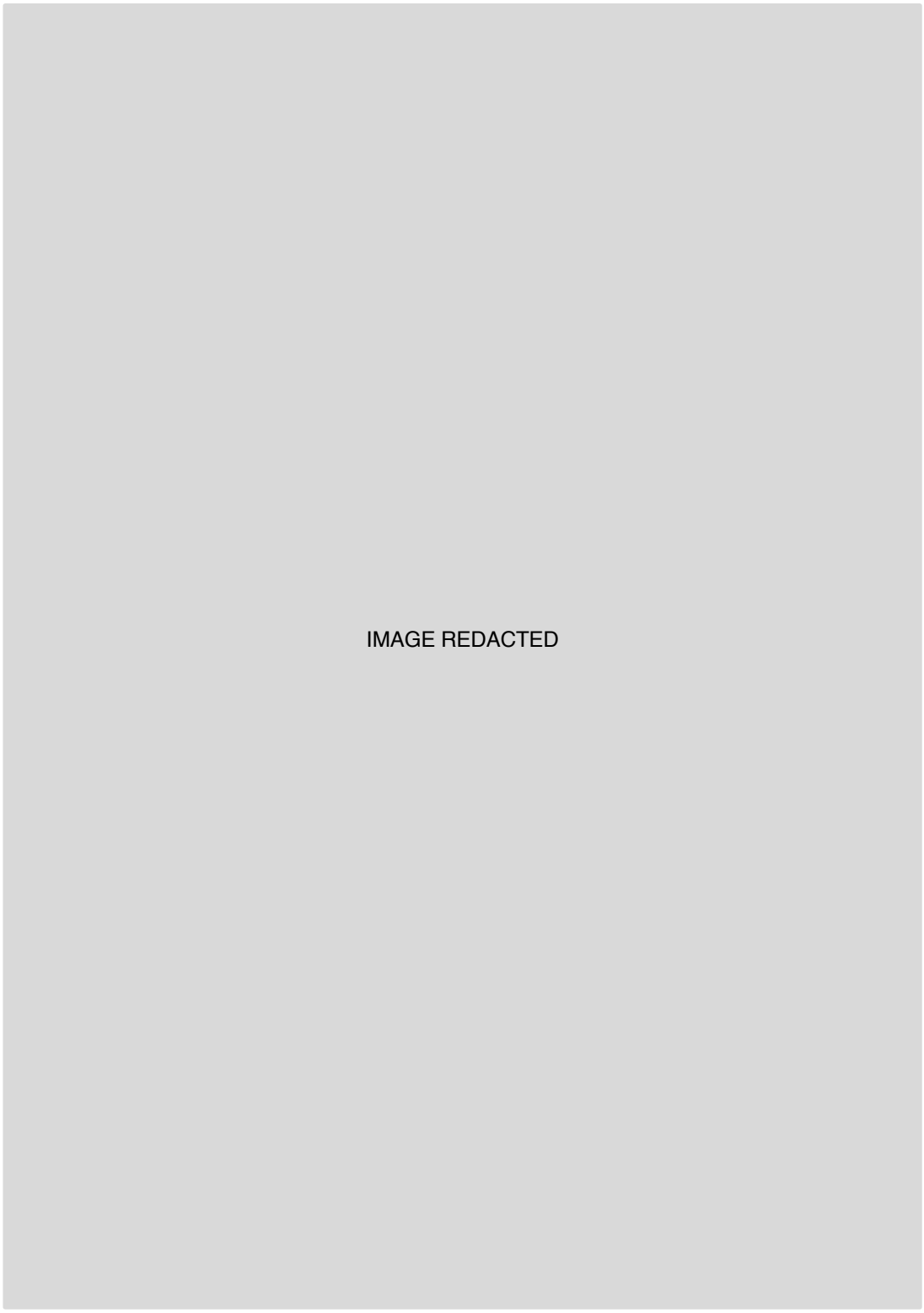


Figure 124. Results from extracting Wi-Fi device activity from a system log

Mobile data: OpenSignal app

CODE REDACTED

Social media: Twitter and Foursquare

Social media data were acquired by querying application programming interfaces (APIs) for publicly shared data. The files were then uploaded to a MySQL database to combine and produce monthly datasets for comparative analysis with the mobile data sources.

Acquiring and pre-processing data via Twitter Search API

Twitter data was retrieved regularly from the free and publicly available search API. The method went through several iterations during the research. For the initial pilot study (March 2016, QEOP landscape), the data was acquired by manually executing a script weekly. The script was then scheduled to run automatically overnight on a local computer from May to August 2016. From September 2016, the script was scheduled to run hourly on a virtual machine hosted online and export the results as a file to my personal Dropbox account.

The code to retrieve tweets was developed based on an existing example (Russell M. A., 2013) and then modified. Two queries are performed: To retrieve geotagged tweets, and to retrieve keyword matches. In both cases, pre-processing is performed on the results returned from the query to only store data required for analysis and to drop tweets if they are retweets (the content begins with 'RT…'. A Twitter bot (@bigben) that posts a tweet automatically every hour is used to define the time range for queries. Private keys have been removed from the code.

```
#!/usr/bin/python

# Twitter Live Scraping of QEOP
# - schedule will be created to run script every hour
# - will use BigBen status_ID to start and end search

# ---------
# initalise
# ---------
# import required packages
import csv
import codecs
import unicodecsv
import sys
import time
import json
import io
from datetime import datetime
from collections import Counter

from urllib2 import URLError
from httplib import BadStatusLine

import twitter
import dropbox

# DropBox config
dropbox_client = dropbox.client.DropboxClient('#### - enter private key - ####')

# ----------------------
# set variable parameters
# ----------------------
```

```
# Storage folder location
file_loc = '<FOLDER>' # for tweets
bb_loc = '<FOLDER>' # for bigben file

# DropBox app location
dbx_loc = '<FOLDER>'

# area to scrape
max_range = 2.5  # radius in KM

area_qeop = 'qeop'

lat_qeop = 51.5410 # Centre of the QEOP
lng_qeop = -0.0140

search_qeop = ['QEOP OR "Queen Elizabeth Olympic Park"', 'Olympic Park London OR Stratford',
'Olympic Stadium London OR Stratford', 'London Stadium OR CopperBox OR "Copper Box" OR Velopark
OR Velodrome OR Aquatics OR "Olympic Pool"', 'Lee Valley Velopark OR Velodrome OR Hockey OR
Tennis OR HTC', 'Orbit ArcelorMittal OR AM', 'Stratford Westfield OR Westfields OR "east village"
OR station OR international', 'West Ham stratford OR Westfield OR Olympic', 'whufc stratford OR
Westfield OR Olympic']

# --------------
# load functions
# --------------

# Function: authenticate with Twitter
# -----------------------------------
def oauth_login():
 CONSUMER_KEY = '#### - enter private key - ####'
 CONSUMER_SECRET = '#### - enter private key - ####'
 OAUTH_TOKEN = '#### - enter private key - ####
 OAUTH_TOKEN_SECRET = '#### - enter private key - ####'

 auth = twitter.oauth.OAuth(OAUTH_TOKEN, OAUTH_TOKEN_SECRET, CONSUMER_KEY, CONSUMER_SECRET)

 twitter_api = twitter.Twitter(auth=auth)
 return twitter_api

# Function: make a robust Twitter request (cope with errors and maxing out connection limits)
# -------------------------------------------
def make_twitter_request(twitter_api_func, max_errors=10, *args, **kw):

 def handle_twitter_http_error(e, wait_period=2, sleep_when_rate_limited=True):

  if wait_period > 3600: # Seconds
   print >> sys.stderr, 'Too many retries. Quitting.'
   raise e
  if e.e.code == 401:
   print >> sys.stderr, 'Encountered 401 Error (Not Authorized)'
   return None
  elif e.e.code == 404:
   print >> sys.stderr, 'Encountered 404 Error (Not Found)'
   return None
  elif e.e.code == 429:
   print >> sys.stderr, 'Encountered 429 Error (Rate Limit Exceeded)'
   if sleep_when_rate_limited:
    print >> sys.stderr, "Retrying in 15 minutes...ZzZ..."
    sys.stderr.flush()
    time.sleep(60*15 + 5)
    print >> sys.stderr, '...ZzZ...Awake now and trying again.'
    return 2
   else:
    raise e # Caller must handle the rate limiting issue
  elif e.e.code in (500, 502, 503, 504):
   print >> sys.stderr, 'Encountered %i Error. Retrying in %i seconds' % \
    (e.e.code, wait_period)
   time.sleep(wait_period)
   wait_period *= 1.5
   return wait_period
  else:
   raise e

 # End of nested helper function

 wait_period = 2
 error_count = 0

 while True:
  try:
   return twitter_api_func(*args, **kw)
  except twitter.api.TwitterHTTPError, e:
   error_count = 0
   wait_period = handle_twitter_http_error(e, wait_period)
   if wait_period is None:
    return
  except URLError, e:
```

```
    error_count += 1
    print >> sys.stderr, "URLError encountered. Continuing."
    if error_count > max_errors:
     print >> sys.stderr, "Too many consecutive errors...bailing out."
     raise
   except BadStatusLine, e:
    error_count += 1
    print >> sys.stderr, "BadStatusLine encountered. Continuing."
    if error_count > max_errors:
     print >> sys.stderr, "Too many consecutive errors...bailing out."
     raise

# Function: retrieve geotagged tweets within range
# ------------------------------------------------
def geotagged_tweets(start_id, end_id, file_loc, max_range, location, lat, lng):

 # add current date-time to filename to ensure each file is unique
 start_time = datetime.now().strftime('%Y%m%d-%H%M')

 # create CSV file to receive data
 file_name = "tw_" + location + "_" + start_time + "_geo" + ".csv" # ensures file is unique
 filename = file_loc + file_name # full path including directory
 dbxfile = dbx_loc + location + "/" + file_name # full path for DropBox copy

 csvfile = open(filename, 'wb')
 csvwriter = unicodecsv.writer(csvfile)

 qt = '' # no keywords, i.e. grab all geotagged tweets within spatial range

 since_id = start_id
 last_id = end_id

 result_count = 0 # count number of tweets collected

 while since_id < last_id:

  # query Twitter search API using parameters
  query = make_twitter_request(twitter_api.search.tweets, q = qt, geocode = "%f,%f,%dkm" % (lat,
lng, max_range), count = 100, max_id = last_id)

  for result in query["statuses"]:

   current_id = result["id"]

   # only process results since last scrape
   if current_id >= since_id:

    # only process a result if it has a geolocation
    if result["geo"]:
     status_id = result["id_str"]
     tweet_text = result["text"]
     tweet_text = tweet_text.encode('ascii', 'replace')
     posted = time.strftime('%Y-%m-%d %H:%M:%S', time.strptime(result['created_at'],'%a %b %d
%H:%M:%S +0000 %Y'))
     source = result["source"]
     lat = result["geo"]["coordinates"][0]
     lng = result["geo"]["coordinates"][1]
     user_id = result["user"]["id"]
     user_name = result["user"]["screen_name"]

     # strip HTML tags from source attribute
     head, sep, tail = source.partition('>') # drop opening tag
     head, sep, tail = tail.partition('<') # drop closing tag
     source = head       # rename back to source

     # adding additional fields for use in MySQL
     geo = 'Y'

     # write row to CSV file
     row = [ status_id, tweet_text, posted, source, lat, lng, user_id, user_name, geo]
     csvwriter.writerow(row)
     result_count += 1

   last_id = result["id"]

 # finish up
 csvfile.close()

 print "For %s, %d results written to %s" % (location, result_count, file_name)

 # copy contents of local file to DropBox file
 csvfile = open(filename, 'rb') # open in read mode (can't copy whilst in write mode)
 dropbox_client.put_file(dbxfile, csvfile)
 csvfile.close()

 print "uploaded %s to Dropbox" % (file_name)
```

```
# Function: retrieve tweets matching search keywords
# -------------------------------------------------
def keyword_tweets(start_id, end_id, file_loc, max_range, location, search_terms):

 print "Keyword queries for %s" % location

 # add current date-time to filename to ensure each file is unique
 start_time = datetime.now().strftime('%Y%m%d-%H%M')

 # create CSV file to receive data
 file_name = "tw_" + location + "_" + start_time + "_srch" + ".csv" # ensures each file is unique
 filename = file_loc + file_name # full path including directory
 dbxfile = dbx_loc + location + "/" + file_name # full path for DropBox copy

 csvfile = open(filename, 'wb')
 csvwriter = unicodecsv.writer(csvfile)

 total_results = 0 # count total number of tweets collected for location
 s = 0

 while s < len(search_terms):

  qt = search_terms[s]

  since_id = start_id
  last_id = end_id
  result_count = 0

  while since_id < last_id:

   query = make_twitter_request(twitter_api.search.tweets, q = qt, include_entities = 1, count =
100, max_id = last_id)

   for result in query["statuses"]:
    current_id = result["id"]
    check_text = result["text"]
    check_text = check_text[:3]

    # only process results since last scrape
    if current_id >= since_id:

     # only process if NOT a retweet (don't want RTs)
     if check_text != "RT ":
      status_id = result["id_str"]
      tweet_text = result["text"]
      tweet_text = tweet_text.encode('ascii', 'replace')
      posted = time.strftime('%Y-%m-%d %H:%M:%S', time.strptime(result['created_at'],'%a %b %d
%H:%M:%S +0000 %Y'))
      source = result["source"]
      if result["geo"]:
       lat = result["geo"]["coordinates"][0]
       lng = result["geo"]["coordinates"][1]
      else:
       lat = -998
       lng = -998
      user_id = result["user"]["id"]
      user_name = result["user"]["screen_name"]

      # strip HTML tags from source attribute
      head, sep, tail = source.partition('>') # drop opening tag
      head, sep, tail = tail.partition('<') # drop closing tag
      source = head       # rename back to source

      # adding additional fields for use in MySQL
      geo = 'N'

      # write row to CSV file
      row = [ status_id, tweet_text, posted, source, lat, lng, user_id, user_name, geo ]
      csvwriter.writerow(row)
      result_count += 1

    last_id = result["id"] # to query for next 100 records, until reach since_id (counts back)

  print "%d results for query: %s" % (result_count, qt)

  # aggregate results count and run next search query
  total_results += result_count
  s +=1

 # finish up
 csvfile.close()

 print "Retrieved %d tweets for %s written to %s \n" % (total_results, location, file_name)

 # copy contents of local file to DropBox file
```

```
 csvfile = open(filename, 'rb')
 dropbox_client.put_file(dbxfile, csvfile)
 csvfile.close()

 print "uploaded %s to Dropbox" % (file_name)

# -----------
# run program
# -----------

# authenticate with Twitter
# -------------------------
twitter_api = oauth_login()
print twitter_api

# set startID and endID using BigBen twitter account
# --------------------------------------------------
fileBigBen = bb_loc + "tw_bigben.csv" # (created 1st use manually)

# read status ID stored in file and set as start ID
csvFile = open(fileBigBen, 'rb') # read contents of file
for row in csvFile:
 start_id = row

csvFile.close()

start_id = int(start_id) # will read in as a string
print start_id

# retrieve latest bong from BigBen account, set as End ID and write to file
user = 'big_ben_clock'
query = twitter_api.statuses.user_timeline(screen_name=user, count = 1)

for result in query:
 end_id = result["id"]
print end_id

csvFile = open(fileBigBen, 'wb') # open file to overwrite contents
csvWriter = csv.writer(csvFile)
row = [end_id]
csvWriter.writerow(row)
csvFile.close()     # this end_ID will become the start_ID next time

# scrape data
# -----------
print "\nFiles will be saved in %s \n" % file_loc

std_args = (start_id, end_id, file_loc, max_range)

# scrape QEOP
geotagged_tweets(*std_args, location=area_qeop, lat=lat_qeop, lng=lng_qeop)
print "Geotag scrape complete."

keyword_tweets(*std_args, location=area_qeop, search_terms=search_qeop)
print "Keyword scrape complete."
```

Acquiring Foursquare 'Here Now' venue counts

Querying the Foursquare API followed a similar iterative development process as for Twitter. From September 2016, data collection was automated using an online virtual machine (VM) running on Digital Ocean. Foursquare was queried every 15 minutes for active venues ('Here Now' count > 0) within proximity of the QEOP. Note: two other venues were also queried – Wembley Stadium and Cardiff Stadium – for a research idea that was not followed up and is not included in this thesis.

```
# Foursquare Live Scraping
# - single instance
# - schedule will be created to run script every 15 minutes
# - designed for cloud server (DigitalOcean)
# script to be stored on Droplet at /home/dataminer

# initalise
# ---------

# import required packages
import csv, json
import codecs, unicodecsv
import sys, io, time
from datetime import datetime
```

```
import foursquare
import dropbox

# Foursquare keys (userless mode)
fsq_client_id = '#### - enter private key - ####'
fsq_client_secret = #### - enter private key - ####'
fsq_api = foursquare.Foursquare(fsq_client_id, fsq_client_secret)

# DropBox config
dropbox_client = dropbox.client.DropboxClient('#### - enter private key - ####')

# set parameters
# --------------

# folder location and filename beginning
file_loc = '<FOLDER>'

# DropBox app location
dbx_loc = '<FOLDER>_'


# creating arrays for each type of data needed per area to loop through

# used in file names
area_names = ['qeop', 'wembley', 'cardiff']

# coordinates
area_coords = ['51.5410, -0.0140', '51.5560, -0.2795', '51.4810, -3.1835']

# scrape data
# -----------
fsq_api = foursquare.Foursquare(fsq_client_id, fsq_client_secret)
current_time = datetime.now().strftime('%Y-%m-%d %H:%M') # timestamp record
start_time = datetime.now().strftime('%Y%m%d-%H%M') # format for filename

x = 0
while x < len(area_names):

 # create/open file in write mode (wb)
# file_name = area_names[x] + "_" + str(start_time) + ".csv"
# fileFsq = file_loc + file_name
# csvFile = open(fileFsq, 'wb')
# csvWriter = unicodecsv.writer(csvFile)

 results = 0

 # collect data
 dataFsq = fsq_api.venues.explore(params={'ll':area_coords[x], 'radius':'2500', 'time': 'any',
'day':'any'}) # radius is in metres

 print dataFsq

 # convert from JSON to array to extract properties of interest
 for item in dataFsq['groups'][0]['items']:
  herenow = item['venue']['hereNow']['count']

  if herenow > 0:
   vid = item['venue']['id']
   venue = item['venue']['name']
   lat = item['venue']['location']['lat']
   lng = item['venue']['location']['lng']
   count = item['venue']['hereNow']['count']

   # write to file
   row = [current_time, vid, venue, lat, lng, count]
#   csvWriter.writerow(row)
   results += 1

# csvFile.close()
 print "%d results for %s" % (results, area_names[x])

 # open local file in read mode (can't upload to DropBox in write mode)
# csvFile = open(fileFsq, 'rb')

 # specify filename and location on DropBox (in /scrapes app)
# dbxFile = dbx_loc + area_names[x] + "_" + str(start_time) + ".csv"

 # copy contents of local file into Dropbox file
# dropbox_client.put_file(dbxFile, csvFile)

# csvFile.close()

# print "uploaded %s to Dropbox" % (file_name)

 x += 1 # increment to retrieve next area
```

Bulk importing files to MySQL

Periodically, social media files were bulk imported to a MySQL database table to enable the files to be combined into monthly datasets. The import was performed by running a Bash script at the command line within the folder where files were being stored following retrieval via API.

The code below is a sample (note: using MySQL root account, the password has been removed)

```
#!/bin/bash
for f in *.csv
do
mysql -e "LOAD DATA LOCAL INFILE '"$f"' INTO TABLE raw_twt_201609 FIELDS TERMINATED BY ','
ENCLOSED BY '\"' LINES TERMINATED BY '\r\n'" -u root --password= --local-infile phd
done
```

Converting to datasets for analysis

To enable a comparative analysis with the Wi-Fi data, the social media readings were organised into datasets matching the periods: March 2016 and May to August 2016. The data were queried from MySQL into R and then the output was exported as a text file ready for analysis. Note, some additional steps included that were not used in the analysis (e.g. tagging Tweeters as habitual or explorer based on the number of days posting tweets.)

```
# extract daily volumes - device or user counts - for May to August 2016

# SETUP
# =====

# import libraries
# ----------------
library(RMySQL)
library(dplyr)
library(lubridate)  # for extracting values from dates

# set point at which plot will switch from standard form
options(scipen=9)

# connect to SQL
# --------------
#iMac
db <- dbConnect(MySQL(), user='root', dbname='phd', host='localhost')

# bounding box (LandScan, do <=, >= to snip to this range)
# ------------
N = 51.5541
S = 51.5292
E = 0.0041
W = -0.0291

# EXTRACT DATA
# ============

# Twitter geotagged
# -----------------
dataSelect <- paste("select * from raw_twt_201605 WHERE geotag LIKE 'Y'")
query <- dbSendQuery(db, dataSelect)
df <- fetch(query, n=-1)

# Convert timestamp to date/time format and extract breakdowns
df$date_time <- as.POSIXct(df$date_time, format='%Y-%m-%d %H:%M:%S')
df$mth = month(df$date_time)
df$day = day(df$date_time)
df$hr = hour(df$date_time)
df$date = date(df$date_time)

# drop records outside of 1 May to 31 August
df <- subset(df, df$mth==5 | df$mth==6 | df$mth==7 | df$mth==8 )

# drop records outside bounding box
df <- subset(df, df$lat>=S & df$lat<=N & df$lng>=W & df$lng<=E )
```

```
# write-out raw tweets
outputfile = 'tweets_geo_raw_2016MaytoAug.csv'
write.csv(df, outputfile, row.names = F)


# tag geotagged tweets as habitual (<3 daily presence) or explorer (3+ daily presence)
# - aggregate per tweeter (user id)
agg <- group_by(df, user_id)

subdf <- summarise(
 agg,
 num_days = n_distinct(date), # count number of days each device is present
)

devices <- as.data.frame(subdf)
devices$visit <- 'explorer' # default to explorer
devices$visit[devices$num_days >= 3] <- 'habitual'

# append visit behaviour to raw dataset
df <- merge(df, devices, by.x='user_id', by.y='user_id', all.x = T)

# write-out raw session data with devices tagged
outputfile = 'tweets_geo_raw_tagged_2016MaytoAug.csv'
write.csv(df, outputfile, row.names = F)

# Twitter keywords
# ----------------
# need to remove duplicates
dataSelect <- paste("select * from twt_sch_201605")
query <- dbSendQuery(db, dataSelect)
df <- fetch(query, n=-1)

# Convert timestamp to date/time format and extract breakdowns
df$date_time <- as.POSIXct(df$date_time, format='%Y-%m-%d %H:%M:%S')
df$mth = month(df$date_time)
df$day = day(df$date_time)
df$hr = hour(df$date_time)
df$date = date(df$date_time)

# drop records outside of 1 June to 31 August
df <- subset(df, df$mth==5 | df$mth==6 | df$mth==7 | df$mth==8 )

# want to remove duplicates
# - first remove URLs from text
df$tweet_text <- gsub("htt.*",'', df$tweet_text)

subdf <- distinct(df, tweet_text, .keep_all = TRUE)

outputfile = 'tweets_sch_raw_2016MaytoAug.csv'
write.csv(subdf, outputfile, row.names = F)

# aggregate number of people tweeting and number of tweets per day
daily_tweeters <- group_by(subdf, date) # group by date

subdf <- summarise(
 daily_tweeters,
 twtrs_sch = n_distinct(user_id), # number of people tweeting
 twts_sch = n()      # number of tweets
)

# write-out tweet volumes
outputfile = 'tweets_sch_dailyvols_2016MaytoAug.csv'
write.csv(subdf, outputfile, row.names = F)

# aggregate number of people tweeting and number of tweets per day
hr_tweeters <- group_by(subdf, date, hr) # group by date

subdf <- summarise(
 hr_tweeters,
 twtrs_sch = n_distinct(user_id), # number of people tweeting
 twts_sch = n()      # number of tweets
)

# write-out tweet volumes
outputfile = 'tweets_sch_hourvols_2016MaytoAug.csv'
write.csv(subdf, outputfile, row.names = F)

rm(df, subdf, daily_tweeters, hr_tweeters, outputfile)

# Foursquare
# ----------
dataSelect <- paste("select * from raw_fsq_201605")
query <- dbSendQuery(db, dataSelect)
df <- fetch(query, n=-1)

# Convert timestamp to date/time format and extract breakdowns
df$date_time <- as.POSIXct(df$date_time, format='%Y-%m-%d %H:%M:%S')
```

```
df$mth = month(df$date_time)
df$day = day(df$date_time)
df$hr = hour(df$date_time)
df$date = date(df$date_time)

# drop records outside of 1 June to 31 August
df <- subset(df, df$mth==5 | df$mth==6 | df$mth==7 | df$mth==8 )

# drop records outside bounding box
df <- subset(df, df$lat>=S & df$lat<=N & df$lng>=W & df$lng<=E )

# write-out raw records
outputfile = 'fsquare_raw_2016MaytoAug.csv'
write.csv(df, outputfile, row.names = F)

# sum venues and check-ins per hour per day, and then per day...
# - have to do in stages, calculating average number of check-ins per hour, per venue
# - then adding up the averages across venues for a total check-in count per hour, per day

# 1. calculate average check-ins per venue per hour per day
agg <- group_by(df, date, hr, venue_id) # group by date and venue

subdf <- summarise(
 agg,
 avg_here = mean(here_now), # average number of check-ins per hour of the day, per venue
)

subdf <- as.data.frame(subdf) # convert to data frame for next step

# 2. count number of venues and sum the hourly averages, per hour per day
agg <- group_by(subdf, date, hr)

subdf <- summarise(
 agg,
 num_venues = n_distinct(venue_id), # number of venues with check-in data
 sum_here = sum(avg_here)   # sum the average hourly check-ins
)

subdf <- as.data.frame(subdf) # convert to dataframe

# write-out hourly volumes
outputfile = 'fsquare_hourvols_2016Mar.csv'
write.csv(subdf, outputfile, row.names = F)

subdf <- as.data.frame(subdf) # convert to dataframe

# 3. can now just add up the sums per venue per day
agg <- group_by(subdf, date)

subdf <- summarise(
 agg,
 fsq_venues = sum(num_venues),  # number of venues with check-in data per day
 fsq_checkins = sum(sum_here)  # total daily check-ins
)

# write-out volumes
outputfile = 'fsquare_dailyvols_2016MaytoAug.csv'
write.csv(subdf, outputfile, row.names = F)

rm(df, subdf, agg, outputfile)
```

B.2 Data preparation

Spatial filtering and zone assignment

The simplest filtering is to restrict spatial readings to a rectangular bounding box. Then, readings can be filtered to those whose coordinates fall within the bounding box by ensuring each coordinate is greater than the minimum and less than the maximum on each axis. If a multi-sided polygon is used, then a more complicated process is required to calculate whether or not readings fall within the polygon or outside it, using a shapefile.

Filter/assign readings using a shapefile

Shapefiles were created by manually drawing boundaries on an OpenStreetMap (OSM) set of map tiles, using the opensource GIS application QGIS. The shapefiles can then be loaded as data using specialist GIS packages in R (rgeos and rgdal). To assign readings to a shapefile polygon requires ensuring both the shapefile and readings use the same coordinates system. Shapefile were exported containing both OSM projection (epsg:3857) and WGS84 (lat/lng) projection (epsg: 4326)

The following code sample is for snipping OpenSignal readings to the QEOP Wi-Fi boundary for comparing the two data sources (described in chapter five).

```
# DEVICE
# ======
root_folder <- '<FOLDER>'

# SETUP PART
# =====
library(plyr)   # for hulls and joins
library(dplyr)   # for data manipulation, grouping, mutating
library(rgeos)   # for gIntersects
library(rgdal)   # for readOGR and SpatialPointsDataFrame
library(lubridate)  # date/time calcs

# folders to save plots
plots_folder <- './plots/'
img_title <- 'Rplot_Hours_'

# set point at which plot will switch from standard form
options(scipen=7)

# set coordinate referencing system (for changing CRS but not reprojecting)
crs_wgs84 <- "+init=epsg:4326" # lat/lng
crs_osm <- "+init=epsg:3857" # OSM projection

# Set coordinate systems for reprojecting
proj_wgs84 <- '+proj=longlat +datum=WGS84'
proj_osm <- '+proj=merc +a=6378137 +b=6378137 +lat_ts=0.0 +lon_0=0.0 +x_0=0.0 +y_0=0.0 +k=1.0
+units=m +nadgrids=@null +wktext +no_defs'

# Set outer bounding box for map
LatMax <- 51.5564 # North
LatMin <- 51.5256 # South
LngMax <- 0.0049  # East
LngMin <- -0.0300 # West

# load shapefiles - lldc park boundary, venues and wifi outline
# -------------------------------------------------------------
# folder containing shapefiles data
folder <- paste0(root_folder, 'data/ShapeFiles/')

# function to convert to a dataframe that can be plotted using geom_polygon
map_to_df <- function(map) {

 # convert shapefile to dataset for plotting as geom_polygon
 map@data$id <- rownames(map@data)
 map.points <- fortify(map, region="id")
 df <- join(map.points, map@data, by="id")
```

```
 names(df)[names(df) == "long"] <- "lng" # rename from long to lng

 # reproject map to OSM to also get OSM coordinates
 proj4string(map) <- CRS(crs_wgs84)
 map <- spTransform(map, CRS(proj_osm))

 map@data$id <- rownames(map@data)
 map.points <- fortify(map, region="id")
 osmdf <- join(map.points, map@data, by="id")

 names(osmdf)[names(osmdf) == "long"] <- "osm_x"
 names(osmdf)[names(osmdf) == "lat"] <- "osm_y"

 # append OSM coordinates to DF with lat/lng
 osmdf <- osmdf[, 1:2] # just keep the osm_x and osm_y cols (is in same order as df)
 df <- cbind(df, osmdf)

 return(df)
}

# load shapefiles
park_shapefile <- readOGR(dsn=folder, "qeop_lldc_park_outline")
venues_shapefile <- readOGR(dsn=folder, "qeop_venues")
wifi_shapefile <- readOGR(dsn=folder, "qeop_lldc_park_wifi")

# convert from map to dataset for plotting as geom_polygon
park_outline <- map_to_df(park_shapefile)
park_venues <- map_to_df(venues_shapefile)
park_wifi <- map_to_df(wifi_shapefile)

# only keep the main park venues, drop the rest from the venues shapefile
park_venues <- subset(park_venues, park_venues$id>=0 & park_venues$id<=3 | park_venues$id==8)

# DATA FUNCTIONS
# ==============
# snip data points to within park outline shapefile
# - assumes columns containing coords are 'lat' and 'lng'
# - df = data file, shapefile = shapefile containing boundary/ies to be snipped within
snip_data <- function(df, shapefile) {

 # create SpatialPointsPolygon
 dataset_map_coords <- df[c("lng", "lat")]
 dataset_map_data <- df[ ,c(3:ncol(df))]  # need to keep the data, check fits with dataset

 dataset_map <- SpatialPointsDataFrame(coords=dataset_map_coords, data=dataset_map_data)
 proj4string(dataset_map) <- CRS(crs_wgs84) # set CRS as WGS84 (lat/lng)

 # --- match data points within zone, drop records outside of boundaries
 area_intersect <- gIntersects(dataset_map, shapefile, byid=T)
 clipped <- apply(area_intersect == F, MARGIN = 2, all)
 data_included <- dataset_map[which(!clipped), ]

 # replace points_map and points_data with clipped data (data points within boundary only)
 dataset_map <- data_included

 # write snipped data back to dataframe and include zone tag for each data point
 newdata <- as.data.frame(dataset_map)

 return(newdata)
}

# LOAD DATA
# =========
# Specify shapefile to be used for boundary (park outline or wifi range)
shapefile = wifi_shapefile

# OpenSignal June 2017
folder <- paste0(root_folder, 'data/OpenSignal/')
data_file <- 'opensignal_raw_tagged_2017Jun.csv'
input_data <- paste0(folder, data_file)
df <- read.csv(input_data)
df$date_time <- as.POSIXct(df$date_time, format = '%Y-%m-%d %H:%M:%S')
df <- mutate(df, date_time = date_time + hours(1)) # converting time from GMT to BST
df$date <- date(df$date_time)
df$mth <- month(df$date_time)
df$day <- day(df$date_time)
df$hr <- hour(df$date_time)
df$wday <- ifelse(wday(df$date)==1,7,wday(df$date)-1)
df <- df[, c(7:8, 1:6, 9:12)] # order lat, lng, then data for snip function
df <- snip_data(df, shapefile)
```

Assign readings to zones using data-driven clusters

This technique was developed in chapter five. The code sample is provided in Appendix B.5. Each reading is assigned the cluster ID that it falls within, or 0 if it does not fall within any cluster.

Assign readings to zones within a uniform grid

Generate a grid by providing base coordinates and then count the number of readings per cell of the grid. This code sample was used for the studies presented in chapter five. A refined version was created for the study presented in chapter seven as part of the function in Appendix B.7

```
# plotting data points as a thematic grid (uniform cells)
# aligned to LandScan grid and then scaled to 1/4 and 1/16

# notes:
#
# LandScan visualisation is a forced thematic grid by using LandScan coordinates to plot counts
# counts are pre-prepared and tagged with label for each grid cell

# TinyPixels visualisation is a correctly produced thematic grid - raw data is counted and
# plotted thematically by constructing a grid and then counting occurrences within each cell

# ----------------------------------------------------------

# DEVICE
# ======
root_folder <- '<FOLDER>'

# SETUP
# =====
library(dplyr)    # for mutate
library(ggplot2)  # for plots (and theme)

# set point at which plot will switch from standard form
options(scipen=7)

# base chart theme for all plots
t <- theme_bw() + theme(panel.border = element_rect(colour = "#bdbdbd") +
                          theme(title = element_text(size=10)))

# create grid labels - LandScan cells and Tiny Pixels (1/16th size of LandScan cell)
# ------------------
# set bounding box/grid extent (LandScan grid)
LatMax <- 51.5541666666661    # North
LatMin <- 51.5291666666661    # South
LngMax <- 0.00416666666596077     # East
LngMin <- -0.0291666666673731    # West

# LandScan labels in Lat (Y: values 1-3) /Lng (X: values A-D)
LS_A <- -0.0291667
LS_B <- -0.0208333
LS_C <- -0.0125
LS_D <- -0.0041667
LS_1 <- 51.5458333
LS_2 <- 51.5375
LS_3 <- 51.5291667

# Tiny Pixels - 1/6th size of a LandScan grid cell
increment <- 0.00208333333333333          # 1/4 of height and width (cell will be 1/16th size)
x <- seq(LngMin, LngMax,by=increment)     # West to East coordinates, with increment
y <- seq(LatMin, LatMax, by=increment)    # South to North coordinates, with increment
grid <- expand.grid(x, y)

# provision grid dataframe for results
results <- as.data.frame(grid)
names(results)[names(results) == "Var1"] <- 'x'
names(results)[names(results) == "Var2"] <- 'y'

results$pixel <- paste0(sprintf('%.6f',results$x), 'x_',  sprintf('%.6f',results$y), "y")

# load LandScan data (to draw cells in thematic grid - using its coordinates as base to start)
# ------------------
landscan <- read.csv(paste0(root_folder, 'data/landscan/2015_coords_landscan_qeop.csv'), header =
T, sep = ',')
landscan <- landscan[, c(1:10)]   # don't need ward codes and names here
landscan <- mutate(landscan, count_pct = count/sum(count))  # add count as a percentage
```

```
# ==================== #
# CONFIGURATION PART 1  #
# ==================== #

# date range
year <- '2017'    # yyyy (e.g. 2016, 2017)
month <- 'May'    # Mon (e.g. Mar, Apr, May, Jun + MaytoAug for 2016)

# load data
# =========

# OpenSignal
# ----------
input_file <- paste0(root_folder, 'data/opensignal/opensignal_raw_tagged_', year, month, '.csv' )
df <- read.csv(input_file, header = T, sep = ',')
df <- subset(df, df$lat > LatMin & df$lat < LatMax & df$lng > LngMin & df$lng < LngMax)  # make
sure data is within grid extent

# sort by device (want to count number of devices, not observations, for ambient population)
df <- df[order(df$pseudo_id), ]

# add landscan labels for each data point (based on which Landscan cell its coordinates fall
within)
df$row[df$lat > LS_1] <- '1'
df$row[df$lat > LS_2 & df$lat <= LS_1] <- '2'
df$row[df$lat > LS_3 & df$lat <= LS_2] <- '3'
df$col[df$lng > LS_D] <- 'D'
df$col[df$lng > LS_C & df$lng <= LS_D] <- 'C'
df$col[df$lng > LS_B & df$lng <= LS_C] <- 'B'
df$col[df$lng > LS_A & df$lng <= LS_B] <- 'A'

df <- mutate(df, ls_label = paste0(col, row))   # create label by joining 'row' and 'col'
df <- df[, c(1:11, 14)]   # drop the 'row' and 'col' columns once labels created

# add pixel_grid labels (to enable unique device counts)
binxy <- data.frame(x=findInterval(df[,8], x),  # lng   <- note: LNG for X
                    y=findInterval(df[,7], y))  # lat   <-       LAT for Y

df <- cbind(df, binxy)
df <- mutate(df, pixel=paste0('x', x, 'y', y))
df <- mutate(df, dev_pixel = paste0(pseudo_id, '_', pixel))
df <- df[ , c(8, 7, 1:2, 5:6, 9:12, 15:16)]   # put lng in 1, and lat in 2. Order is important
for viz

# ==================== #
# CONFIGURATION PART 2  #
# ==================== #

# base grid for LandScan visual plots
lscan <- landscan[, c(1:5)]

# set visual title (what data is being plotted)
charttitle <- 'QEOP: OpenSignal May 2017'

# choose data to visualise (using subdf to keep original df intact)
subdf <- df

# group data by LandScan grid cell (pixel) and count unique devices
by_pixel <- group_by(subdf, ls_label)
lscan_df <- summarise(by_pixel, count = n_distinct(pseudo_id))
lscan_df <- as.data.frame(lscan_df)

# de-dup data for TinyPixels grid (counting device once per pixel)
# note: different method to LandScan because count will occur as part of visualisation
tiny_df <- subdf[!duplicated(subdf$dev_pixel),]

# ===============
# Set up viz data
# ===============

# LandScan grid cell size
# -----------------------
# merge with lsgrid for cell coordinates
lsgrid <- merge(lscan, lscan_df, by.x='label', by.y='ls_label', all.x=T)
lsgrid[is.na(lsgrid)] <- 0   # set any NAs to 0

# create distribution summary - averages (check divisor!) and percentages across the grid
lsgrid <- mutate(lsgrid, avg=count/31) # may need to adjust, e.g. /30 for daily average in June
lsgrid <- mutate(lsgrid, pct = count/sum(count))


# Visualise thematic grid
d <- ggplot(aes(x=lng, y=lat, fill=avg), data=lsgrid)    # use pct or count depending on visual

# display
d + geom_tile() +
```

```
  labs(title=charttitle,x="lng", y="lat",
       subtitle="Scale: LandScan pixel (1km tall, approx 0.66km wide)",
       fill="Average",  # can only use if binding fill colour in AES for plot
       colour=NULL,     # can only use if binding border colour in AES for plot
       caption=NULL) +
  theme_bw() + theme(title=element_text(size=10)) +
#  scale_fill_continuous(low="#762a83", high="yellow")
  scale_fill_continuous(low="#313131", high="#efefef")   # grey scale

# Tiny Pixels grid
# ----------------
binxy <- data.frame(x=findInterval(tiny_df[,1], x),
                    y=findInterval(tiny_df[,2], y))
results <- table(binxy)
d2 <- as.data.frame.table(results)
d2 <- mutate(d2, cell = paste0('x', x, 'y', y))

d <- d2  # use first time, then comment out if using hack (run edit at end)
         # make sure first use is for full dataset that will have a count per cell

# hack when needing to ensure 192 rows for data with less
#d <- merge(d, d2, by.x='cell', by.y='cell', all.x=T)
#d <- d[, c(1:3, 6)]
#names(d)[names(d) == 'x.x'] <- 'x'
#names(d)[names(d) == 'y.x'] <- 'y'
#d[is.na(d)] <- 0

# after doing d <- d2 or running the hack above
d <- mutate(d, avg = Freq/31)         # average across dataset
d <- mutate(d, pct = Freq/sum(Freq))   # monthly average as a percentage

# thematic visualisation
v <- ggplot(aes(x=x, y=y, fill=pct), data=d)
v + geom_tile() +
  labs(title=charttitle,x="lng", y="lat",
       subtitle="Scale: 1/16th LandScan pixel (0.25km tall, approx 0.16km wide)",
       fill="Average",  # can only use if binding fill colour in AES for plot
       colour=NULL,     # can only use if binding border colour in AES for plot
       caption="lighter cells indicate higher presence count"
       ) +
  theme_bw() + theme(title=element_text(size=10)) +
#  scale_fill_continuous(low="#762a83", high="yellow")
  scale_fill_continuous(low="#313131", high="#efefef")

#d <- d[, c(1:3)]  # when using hack

# edit: creating a fixed d grid to make sure all 192 cells are plotted with or without data
# keeping a version of d that has all 192 rows. Then using d2 for each dataset after, and
resetting d after each viz
#d <- d[, c(1:2)]
#d <- mutate(d, cell = paste0('x', x, 'y', y))

# using log transform
# ===========================
# note: adding 0.1 to 0 counts to avoid div/0 error on log transform

# Landscan
# --------
# this requires the previous LandScan visualisation to have been run
lsgrid[is.na(lsgrid)] <- 0   # set any NAs to 0
lsgrid$count[lsgrid$count == 0] <- 0.1   # avoid div/0 error on log transform
lsgrid <- mutate(lsgrid, logt = log10(count))

d <- ggplot(aes(x=lng, y=lat, fill=logt), data=lsgrid)
d + geom_tile() +
  labs(title=charttitle,x="lng", y="lat",
       subtitle="Scale: LandScan pixel (1km tall, approx 0.66km wide) | Count is Log10
Transformed",
       fill="Average",  # can only use if binding fill colour in AES for plot
       colour=NULL,     # can only use if binding border colour in AES for plot
       caption="lighter cells indicate higher presence count"
       ) +
  theme_bw() + theme(title=element_text(size=10)) +
  #  scale_fill_continuous(low="#762a83", high="yellow")
  scale_fill_continuous(low="#313131", high="#efefef")

# Pixel grid
# ----------
# this requires the previous TinyPixels visualisation to have been run
# just adds in 2 lines to the data to log transform it

d$Freq[d$Freq == 0] <- 0.1           # avoid div/0 error on log transform
d <- mutate(d, logt = log10(Freq))   # average across dataset

# thematic visualisation
v <- ggplot(aes(x=x, y=y, fill=logt), data=d)
v + geom_tile() +
```

```
  labs(title=charttitle,x="lng", y="lat",
       subtitle="Scale: 1/16th LandScan pixel (0.25km tall, approx 0.16km wide)",
       fill="Average",  # can only use if binding fill colour in AES for plot
       colour=NULL,     # can only use if binding border colour in AES for plot
       caption="lighter cells indicate higher presence count"
  ) +
  theme_bw() + theme(title=element_text(size=10)) +
  #  scale_fill_continuous(low="#762a83", high="yellow")
  scale_fill_continuous(low="#313131", high="#efefef")

d <- d[, c(1:2)]
d <- mutate(d, cell = paste0('x', x, 'y', y))
```

Tagging with visit attributes

Identifying trips

The workflow for tagging each record of a dataset with a trip ID is described in chapter three. The following code shows how a trip ID is calculated and added to each reading. The dataset has to first be ordered by device id, and then by date and time. Each reading is then appended with data from the previous record to calculate whether or not the current reading is the start of a new trip. If it is, the trip counter is incremented by 1 ready for the next reading. The sample included here is for the OpenSignal May 2017 dataset, using Python.

```
# DEVICE PARAMETERS
# -----------------
base_folder = '<FOLDER>'
print(base_folder)

# SETUP
# =====
%matplotlib inline

import pandas as pd
import numpy as np
import matplotlib.pyplot as plt

# May 2017 data
# =============
source_folder = 'data/OpenSignal/'
source_file = 'opensignal_raw_tagged_2017May.csv'
output_file = 'opensignal_2017May_trips.csv'

input_file = base_folder + source_folder + source_file
df = pd.read_csv(input_file, sep=',')

# must be sorted by device, then date_time order
# ----------------------------------------------
df = df.sort_values(by=['pseudo_id', 'date_time'])
df = df.reset_index(drop=True)

# add values from previous row to current row using shift(1)
df['prev_id'] = df['pseudo_id'].shift(1)
df['prev_date'] = df['date'].shift(1)
df['prev_hr'] = df['hr'].shift(1)
df[:3]

# tag each record with a trip number
# ----------------------------------
devices = [] # just capturing as can do quick check of device to trip counts
trips = []
trip = 0  # setting to 0 because need first record in loop to get 1 (prev_dev is NaN)

for index, row in df.iterrows():

 if (np.any(row['pseudo_id'] != row['prev_id']) == True): # if device is different, reset trip to
1
   trip = 1
 elif (np.any(row['date'] != row['prev_date']) == True):  # if date is different to last date,
increase trip
   trip += 1
 elif (np.any(row['hr'] >= (row['prev_hr']+2) ) == True): # if duration between events is larger
than 3 hours, increase trip
   trip += 1
 else: # else nothing changes
```

```
  trip = trip

 # append results to lists
 devices.append(row['pseudo_id'])
 trips.append(trip)

 # read values ready for next row
 device = row['pseudo_id'] # update device and date from this row before reading next row
 date = row['date']

# once done, append results back into df
df['dev_trip'] = trips

# write out results
# -----------------
output_file = 'opensignal_2017May_trips.csv'
df.to_csv(output_file, sep=',', encoding='utf-8', index=False)
print("done")
```

Note: early iterations performing this function took a different approach, appending readings from the following record instead of the previous record. However, the same result is achieved.

Identifying stages

Once records have been given trip IDs and readings have been assigned to zones within a landscape, whether using a manually-drawn shapefile, uniform grid or data-driven clusters, each reading can be assigned a stage ID. A new stage begins each time a device moves between zones within the landscape. The sample below is for assigning where the zones are based on data-driven clusters. The stage ID is assigned and duration is calculated for the reading.

```
# DEVICE PARAMETERS
# -----------------
base_folder = '<folder>'

# SETUP
# =====
%matplotlib inline

import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from datetime import datetime

# Version for data-driven clusters
# ================================
# Specify data source and output filename (time period - month of data)
# ---------------------------------------
source_file = 'osmay17_trips_dataclusters_2tier_alldays.csv'   # input data
output_file = 'osmay17_trips_dataclusters_2tier_alldays_stages.csv' # output file

# run the rest
source_folder = 'data/OpenSignal/'
input_file = base_folder + source_folder + source_file
df = pd.read_csv(input_file, sep=',')

# must be sorted by device, then date_time order
df = df.sort_values(by=['pseudo_id', 'date_time'])
df = df.reset_index(drop=True)

df['date_time'] = pd.to_datetime(df['date_time']) # switch from object to date_time

# add values from previous row to current row using shift(1) and for next row using shift(-1)
df['prev_trip'] = df['pseudo_trip'].shift(1)
df['prev_cluster'] = df['cluster2tier'].shift(1)
df['next_trip'] = df['pseudo_trip'].shift(-1)
df['next_time'] = df['date_time'].shift(-1)

# tag each record with a stage number and duration of the reading
# ---------------------------------------------------------------
# note use of numpy - np.any(statement) == True within if statements. Is needed else get
'ambiguous...'
# np.any() requires any items in the statement to be true, np.all() requires all statements to be
true
# in this case, there is only one argument, so could have used either.
stages = []
durations = []
stage = 0  # setting to 0 because need first record in loop to get 1 (prev_dev is NaN)
```

```
for index, row in df.iterrows():

 # calculate stage of trip
 if (np.any(row['pseudo_trip'] != row['prev_trip']) == True): # if trip is different, reset stage
to 1
   stage = 1
 elif (np.any(row['cluster2tier'] != row['prev_cluster']) == True): # else if zone is different,
increment stage
   stage += 1
 else: # else nothing changes
   stage = stage

 # calculate duration of current reading (set to 60 seconds if next reading is the start of a new
trip)
 duration = (row['next_time'] - row['date_time']).total_seconds() # calculates differnce in
seconds
 if (np.any(row['pseudo_trip'] != row['next_trip']) == True):
   duration = 60.0

 # append results to lists
 stages.append(stage)
 durations.append(duration)

# once done, append results back into df
df['trip_stage'] = stages
df['duration'] = durations

# write out results
# -----------------
output_results = base_folder + source_folder + output_file
df.to_csv(output_results, sep=',', encoding='utf-8', index=False)
```

B.3 Visual exploration

Spatial background preparation

Acquiring map tiles for visual backgrounds

Static spatial visualisations with a map as the background were produced in R using the OpenStreetMap package. For the backgrounds, map tile designs were either built-in to the package or produced by Stamen Designs (https://stamen.com).

Code sample is a function for retrieving map tiles using OpenStreetMap package in R:

```
library(OpenStreetMap) # for map background

 # Map bounding box
 LatMax <- 51.51022 # North
 LatMin <- 51.49384 # South
 LngMax <- -0.111176  # East
 LngMin <- -0.137676 # West

map_background <- function(map_type='osm', map_design='osm', map_zoom=15) {

 if (map_type == 'osm') {
 ul <- c(LatMax,LngMin)
 lr <- c(LatMin,LngMax)
 basemap <- openmap(ul,lr, zoom=map_zoom, type=map_design, mergeTiles=FALSE)
 }
 else {
 tile_server <- paste0("http://tile.stamen.com/", map_design, "/{z}/{x}/{y}.png") # Stamen Design
 basemap <- openmap(c(LatMax,LngMin),c(LatMin,LngMax), zoom=map_zoom, type=tile_server,
mergeTiles=FALSE)
 }

 return(basemap)
}

# to plot just the map background
autoplot(basemap)
```

Reprojection to a common coordinates system

Spatial plots require all data to share the same coordinates system. If data uses a different coordinates system, it requires reprojection. If a map is used as the background for a spatial plot, it too must have the same coordinates system.

All map backgrounds and data points in this thesis use one of three coordinate referencing systems (CRS): WGS84 (epsg:4326) with latitude (‘lat’) and longitude (‘lng’); BNG (British National Grid, epsg:27700) with Eastings and Northings; and, OpenStreetMap (OSM, epsg:3857).

The following code sample is used to either set the CRS or reproject it to a different CRS.

```
library(rgdal)   # for readOGR and SpatialPointsDataFrame

# set coordinate referencing system (for changing CRS but not reprojecting)
crs_wgs84 <- "+init=epsg:4326" # lat/lng
crs_bng <- "+init=epsg:27700"
crs_osm <- "+init=epsg:3857" # OSM projection

# Set coordinate systems for reprojecting
proj_wgs84 <- '+proj=longlat +datum=WGS84'
proj_bng <- '+proj=tmerc +lat_0=49 +lon_0=-2 +k=0.9996012717 +x_0=400000 +y_0=-100000 +ellps=airy
+datum=OSGB36 +units=m +no_defs'
```

```
proj_osm <- '+proj=merc +a=6378137 +b=6378137 +lat_ts=0.0 +lon_0=0.0 +x_0=0.0 +y_0=0.0 +k=1.0
+units=m +nadgrids=@null +wktext +no_defs'

# REPROJECT FROM LAT/LNG TO OSM COORDINATES
# - requires column names for lat and lng to be specified as 'y' and 'x' (default 'lng' and
'lat')
# - a and b will be the names for the OSM coordinate columns (default 'osm_x' and 'osm_y')
# - appends the OSM coordinates as new columns to original dataframe
reproject <- function(df, x='lng', y='lat', a='osm_x', b='osm_y') {
 dataset_map_coords <- df[c(x, y)]
 dataset_map_data <- as.data.frame(df[ ,c(1)]) # doesn't matter which column, is dropped at end
 dataset_map <- SpatialPointsDataFrame(coords=dataset_map_coords, data=dataset_map_data)

 # set CRS and reproject to OSM for OSM number system
 dataset_map@proj4string # check first, should be NA - not yet been set
 proj4string(dataset_map) <- CRS(crs_wgs84) # set the current coordinates system
 dataset_map <- spTransform(dataset_map, CRS(proj_osm)) # reproject to OSM

 # convert back to dataframe with OSM cooordinates and then amend to original dataset
 newdf <- as.data.frame(dataset_map)
 names(newdf)[names(newdf) == x] <- a; names(newdf)[names(newdf) == y] <- b
 newdf <- newdf[, 2:3] # just want to keep the OSM coordinates to append back to dataset
 df <- cbind(df, newdf)
 return(df)
}
```

B.4 Analyses for chapter four

The majority of programmatic work during chapter four was in preparing the data for analysis and then performing descriptive statistics at daily and hourly scales.

The code for the second study is reproduced here: evaluating machine learning algorithms to produce a predictive model for changes to the number of visits to the park for different contexts. The model uses machine learning algorithms included in the Scikit-Learn package for Python. The code was developed based on an existing example (Géron, 2017).

Setup environment

```
# BASE CONFIG
# ============
dropbox = '<FOLDER>'

# define data source
# ------------------
root_folder = <FOLDER>'
source_folder = 'data/qeop/'
source_data = 'qeop_diary_2017_events.csv' # input data

# define project label (will prefix output files)
# ---------------------
project_label = 'MLtest'

path = dropbox + root_folder
source = path + source_folder + source_data
print(source)

# SETUP
# =====
%matplotlib inline

import pandas as pd
import numpy as np
import matplotlib.pyplot as plt

from pandas.plotting import scatter_matrix

from sklearn.model_selection import train_test_split   # for random sampling
from sklearn.model_selection import StratifiedShuffleSplit # for stratified sampling
from sklearn.preprocessing import Imputer     # for handling missing values
from sklearn.preprocessing import OneHotEncoder    # for binary classification of factorised
categories

from sklearn.preprocessing import StandardScaler    # Standardisation method for feature scaling
from sklearn.preprocessing import MinMaxScaler    # Normalisation method for feature scaling

from sklearn.base import BaseEstimator, TransformerMixin  # for writing own transformers
from sklearn.pipeline import Pipeline      # for building pipeline of transformations
from sklearn.pipeline import FeatureUnion     # for joining pipelines together

# the ML algorithms and methods to evalute/fine-tune
from sklearn.linear_model import LinearRegression  # Linear regression
from sklearn.tree import DecisionTreeRegressor   # Decision tree regression
from sklearn.ensemble import RandomForestRegressor  # Random forest regression
from sklearn.svm import SVR        # Support Vector Machine regression
from sklearn.neural_network import MLPRegressor   # Neural Network regression

from sklearn.metrics import mean_squared_error   # classic evaluation of linear regression
(variance)
from sklearn.model_selection import cross_val_score  # cross-validation of model
from sklearn.model_selection import GridSearchCV   # Grid Search for hyperparameter tuning
from sklearn.model_selection import RandomizedSearchCV # Randomised search
from scipy.stats import randint       # for generating random integers used in Randomised Search

# for pseudo-randomisation
np.random.seed(42)

# save figures
plt.rcParams['axes.labelsize'] = 13
plt.rcParams['xtick.labelsize'] = 11
plt.rcParams['ytick.labelsize'] = 11

def save_fig(fig_id, tight_layout=True, fig_extension="png", resolution=300):
 path = os.path.join(path, fig_id + "." + fig_extension)
 print("Saving figure", fig_id)
```

```
 if tight_layout:
  plt.tight_layout()
 plt.savefig(path, format=fig_extension, dpi=resolution)

print('ready')
```

Read and prep source data

```
# import csv as dataframe and prep additional attributes (decided during data discovery)
df = pd.read_csv(source, sep=',')

df['season'] = 2
df['season'][(df['date_mth'] >=4) & (df['date_mth'] <=6)] = 4
df['season'][(df['date_mth'] >=7) & (df['date_mth'] <= 8)] = 6
df['season'][(df['date_mth'] <= 3) & (df['date_mth'] >= 11)] = 0

df['wk_we'] = 1
df['wk_we'][(df['day_of_wk'] > 1) & (df['day_of_wk'] < 7)] = 0

df['stadium_cat'] = df['stadium_cat'].fillna('none') # need to fix NaNs

df['event_rank'] = 0
df['event_rank'][df['stadium_cat'].str.contains('westham')] = 1
df['event_rank'][df['stadium_cat'].str.contains('music')] = 3
df['event_rank'][(df['stadium_cat'].str.contains('athletics')) |
(df['stadium_cat'].str.contains('race'))] = 2

df['temp_high'] = df['temp_high'].astype(int)

df.info()
df.head()

df_original = df.copy() # creating source copy to base each model on

# keep only numeric fields to be quantified + date for checking (will drop it before running
models)
df_original[:3]

# setup each dataset to build model from

# full year of data, excluding precipitation due to NaNs
year = df_original[['date_v2', 'wifi_dev', 'season', 'day_of_wk', 'wk_we', 'stadium_event',
'event_rank' ,
   'holiday', 'temp_high']].copy()

# including precipitation, dropping dates with missing data
rain = df_original[['date_v2', 'wifi_dev', 'season', 'day_of_wk', 'wk_we', 'stadium_event',
'event_rank' ,
   'holiday', 'temp_high', 'pod_prec']].copy()
rain = rain.dropna(subset=['pod_prec']).copy()

# full year excluding stadium events
year_normal = year[(year.event_rank == 0)].copy()

# precipitation excluding stadium events
rain_normal = rain[(rain.event_rank == 0)].copy()

# events only
year_events = year[(year.event_rank > 0)].copy()


print("ready")
```

Ready to build the model

```
# first select the dataset to build model from, then run through the rest...

#df = year.copy()   # full year
#df = rain.copy()   # partial year, due to gaps in rain data
#df = year_normal.copy() # normal days only
#df = rain_normal.copy()
df = year_events.copy() # event days only

df.info()
df = df.reset_index(drop=True) # need to do this when re-running this code with different df to
full set
```

Create test set

```
# NOTE!!!
# - use 'stadium_event' for full dataset (year models)
# - use 'day_of_wk' for dataset that excludes events (events are mostly at weekends)
# - use 'event_rank' for dataset for events only

# random sampling
```

```
# ---------------
# random_state is same as using setseed()
train_set, test_set = train_test_split(df, test_size=0.1, random_state=42)

# stratified sampling
# -------------------
split = StratifiedShuffleSplit(n_splits=1, test_size=0.1, random_state=42)

for train_index, test_index in split.split(df, df['event_rank']):
 strat_train_set = df.loc[train_index]
 strat_test_set = df.loc[test_index]

# can compare both approaches to the overall dataset
# --------------------------------------------------
# calculate proportions for each - the dataset, random sample and stratified sample
def data_prop(data):
 return round(data['event_rank'].value_counts() / len(data) * 100, 4)


compare_props = pd.DataFrame({
 'Dataset': data_prop(df),
 'Random': data_prop(test_set),
 'Stratified': data_prop(strat_test_set),
}).sort_index()

compare_props["R. %error"] = round(100 * compare_props["Random"] / compare_props["Dataset"] -
100, 4)
compare_props["S. %error"] = round(100 * compare_props["Stratified"] / compare_props["Dataset"] -
100, 4)

compare_props # view results (proportions and % difference between each sample and the dataset)
```

Prepare for machine learning

```
# separate predictors and target values
# ----------
df = strat_train_set.drop('wifi_dev', axis=1) # copy of dataset without date or predictor
attribute
df_labels = strat_train_set['wifi_dev'].copy() # copy of the predictor attribute from the dataset

df = df.drop('date_v2', axis=1) # need to drop date_v2 as won't contribute to the model
df[:3]

# creating pipeline - values feature scaled
# -----------------------
num_attribs = list(df)   # creates list of column names to feed into function

pipeline = Pipeline([
  ('std_scaler', StandardScaler()),
 ])

df_prepared = pipeline.fit_transform(df)

df_prepared
```

Select and train a model

```
# Descriptive stats reminder
# --------------------------
# For the predictor variable, so can see how the results compare...
df_labels.describe()

# and creating a sample from within the training data
# to see how predictions look compared with actual values
some_data = df.iloc[:5]
some_labels = df_labels.iloc[:5]
some_data_prepared = pipeline.transform(some_data)

# evaluate using cross validation
# -------------------------------
scores = cross_val_score(tree_reg, df_prepared, df_labels, scoring='neg_mean_squared_error',
cv=10)
tree_rmse_scores = np.sqrt(-scores)

def display_scores(scores):
 print("Cross Validation with 10 folds")
 print("Scores: ", scores)
 print("Mean: ", scores.mean())
 print("St.Dev: ", scores.std())

display_scores(tree_rmse_scores)
```

```
# Linear regression model
```

```
# -----------------------
# create the model
lin_reg = LinearRegression()
lin_reg.fit(df_prepared, df_labels)

print("Predictions ", lin_reg.predict(some_data_prepared))
print("Actual values ", list(some_labels))

# evaluate performance (linear method - mean squared error)
df_predictions = lin_reg.predict(df_prepared)
lin_mse = mean_squared_error(df_labels, df_predictions)
lin_rmse = np.sqrt(lin_mse) # to get variance between prediction and actual (residuals are
squared, need sq_rt)
print("\nprediction variance: " + str(lin_rmse))

lin_scores = cross_val_score(lin_reg, df_prepared, df_labels, scoring='neg_mean_squared_error',
cv=10)
lin_rmse_scores = np.sqrt(-lin_scores)

display_scores(lin_rmse_scores)
```

```
# Decision Tree model
# -------------------
# build model from training data
tree_reg = DecisionTreeRegressor()
tree_reg.fit(df_prepared, df_labels)

# make predictions
df_predictions = tree_reg.predict(df_prepared)

# evaluate performance of prediction (mean squared error)
tree_mse = mean_squared_error(df_labels, df_predictions)
tree_rmse = np.sqrt(tree_mse)

print("Predictions ", tree_reg.predict(some_data_prepared))
print("Actual values ", list(some_labels))
print("\nprediction variance: " + str(tree_rmse))
```

```
# Random Forest ensemble learning method
# -------------
forest_reg = RandomForestRegressor()
forest_reg.fit(df_prepared, df_labels)

print("Predictions ", forest_reg.predict(some_data_prepared))
print("Actual values ", list(some_labels))

df_predictions = forest_reg.predict(df_prepared)
forest_mse = mean_squared_error(df_labels, df_predictions)
forest_rmse = np.sqrt(forest_mse)
print("\nprediction variance: " + str(forest_rmse) + "\n")

forest_scores = cross_val_score(forest_reg, df_prepared, df_labels,
        scoring="neg_mean_squared_error", cv=10)
forest_rmse_scores = np.sqrt(-forest_scores)
display_scores(forest_rmse_scores)
```

```
# Support Vector Machine
# ----------------------
svm_reg = SVR(gamma='auto')
svm_reg.fit(df_prepared, df_labels)

print("Predictions ", svm_reg.predict(some_data_prepared))
print("Actual values ", list(some_labels))

df_predictions = svm_reg.predict(df_prepared)
svm_mse = mean_squared_error(df_labels, df_predictions)
svm_rmse = np.sqrt(svm_mse)
print("\nprediction variance: " + str(svm_rmse) + "\n")

svm_scores = cross_val_score(svm_reg, df_prepared, df_labels,
        scoring="neg_mean_squared_error", cv=10)
svm_rmse_scores = np.sqrt(-svm_scores)
display_scores(svm_rmse_scores)
```

```
# Neural Network
# --------------
```

```
nn_reg = MLPRegressor()
nn_reg.fit(df_prepared, df_labels)

print("Predictions ", nn_reg.predict(some_data_prepared))
print("Actual values ", list(some_labels))

df_predictions = nn_reg.predict(df_prepared)
nn_mse = mean_squared_error(df_labels, df_predictions)
nn_rmse = np.sqrt(nn_mse)
print("\nprediction variance: " + str(nn_rmse) + "\n")

nn_scores = cross_val_score(nn_reg, df_prepared, df_labels,
        scoring="neg_mean_squared_error", cv=10)
nn_rmse_scores = np.sqrt(-nn_scores)
display_scores(nn_rmse_scores)
```

Fine-tune the model (use grid search for best performing parameters

```
# Grid Search for best hyperparameter combination
# -----------------------------------------------
# testing with the Random Forest model, since it was the best performing of the three tested
param_grid = [
 # try 12 (3×3) combinations of hyperparameters
 {'n_estimators': [3, 10, 30], 'max_features': [2, 3, 6]},
 # then try 6 (2×3) combinations with bootstrap set as False
 {'bootstrap': [False], 'n_estimators': [3, 10], 'max_features': [2, 3, 6]},
 ]

# will explore 9 + 6 (ie 15) combinations, training the model 5 times = 75 rounds of training
grid_search = GridSearchCV(forest_reg, param_grid, cv=5, scoring='neg_mean_squared_error')

grid_search.fit(df_prepared, df_labels)

# display best result from the grid search
grid_search.best_params_

# get best estimator directly - this is your best performing model
grid_search.best_estimator_

# evaluate the scores of each combination
cvres = grid_search.cv_results_

for mean_score, params in zip(cvres['mean_test_score'], cvres['params']):
 print(np.sqrt(-mean_score), params)

# evaluate feature importance
# ---------------------------
feature_importances = grid_search.best_estimator_.feature_importances_

# display feature importance values next to their attribute names
attributes = num_attribs
sorted(zip(feature_importances, attributes), reverse=True)

# create and compare tuned model
# ------------------------------
final_model = grid_search.best_estimator_
final_model.fit(df_prepared, df_labels)

print("Predictions ", final_model.predict(some_data_prepared))
print("Actual values ", list(some_labels))

df_predictions = final_model.predict(df_prepared)
final_model_mse = mean_squared_error(df_labels, df_predictions)
final_model_rmse = np.sqrt(final_model_mse)
print("\nprediction variance: " + str(final_model_rmse) + "\n")

final_model_scores = cross_val_score(final_model, df_prepared, df_labels,
        scoring="neg_mean_squared_error", cv=10)
final_model_rmse_scores = np.sqrt(-final_model_scores)
display_scores(final_model_rmse_scores)
```

Evaluate final model (tuned random forest) against best performing untuned models

This uses the test data (held back from training)

```
# final model selection
# ---------------------
# going to pick the best estimator from randomised search of hyperparameters
final_model = grid_search.best_estimator_

test_data = strat_test_set.drop('wifi_dev', axis=1)
test_data = test_data.drop('date_v2', axis=1) # need to drop date_v2 as won't contribute to the
model
```

```
actual_values = strat_test_set['wifi_dev'].copy()
test_dates = strat_test_set['date_v2'].copy()

test_data_prepared = pipeline.transform(test_data)

# pull out sample of the test data to show predictions versus actuals (rather than full set)
sample_data = test_data.iloc[:5]
sample_data_prepared = pipeline.transform(sample_data)
sample_labels = actual_values.iloc[:5]
sample_dates = test_dates.iloc[:5]

predicted_values = final_model.predict(test_data_prepared)

print("Dates   ", list(sample_dates))
print("Predictions ", final_model.predict(sample_data_prepared).astype(int))
print("Actual values ", list(sample_labels))

final_mse = mean_squared_error(actual_values, predicted_values)
final_rmse = np.sqrt(final_mse)

final_rmse

# random forest model
# -------------------
predicted_values = forest_reg.predict(test_data_prepared)

print("Dates   ", list(sample_dates))
print("Predictions ", forest_reg.predict(sample_data_prepared).astype(int))
print("Actual values ", list(sample_labels))

final_mse = mean_squared_error(actual_values, predicted_values)
final_rmse = np.sqrt(final_mse)

final_rmse

# decision tree model
# -------------------
predicted_values = tree_reg.predict(test_data_prepared)

print("Dates   ", list(sample_dates))
print("Predictions ", tree_reg.predict(sample_data_prepared).astype(int))
print("Actual values ", list(sample_labels))

final_mse = mean_squared_error(actual_values, predicted_values)
final_rmse = np.sqrt(final_mse)

final_rmse

# linear regression model
# -----------------------
predicted_values = lin_reg.predict(test_data_prepared)

print("Dates   ", list(sample_dates))
print("Predictions ", lin_reg.predict(sample_data_prepared).astype(int))
print("Actual values ", list(sample_labels))

final_mse = mean_squared_error(actual_values, predicted_values)
final_rmse = np.sqrt(final_mse)

final_rmse
```

B.5 Analyses for chapter five

Comparing population measures

For comparisons within the Wi-Fi boundary of the QEOP and administrative output areas, shapefiles are used to aggregate readings by polygon. For LandScan comparisons, a grid of uniform cells is generated to aggregate readings. See Appendix B.2 for code samples.

Comparing areal scales

```
# Map of Boundaries
#
# Shapefiles for admin areas: MSOA, LSOA and OA, will be snipped to boundary of map
# Shapefile for park boundary (drawn in QGIS)
# Grid generated for LandScan cells
#
############

# DEVICE CONFIG
# =============
# root folder for source data
source <- '<FOLDER>'

# SET-UP
# ======
library(dplyr)
library(ggplot2)
library(scales)
library(OpenStreetMap)   # for map background
library(rgeos)
library(rgdal)          # to load shapefiles
library(maptools)       # to fortify shapefiles
library(lubridate)

# set point at which plot will switch from standard form
options(scipen=7)

# set coordinate referencing system (for changing CRS but not reprojecting)
crs_wgs84 <- "+init=epsg:4326"  # lat/lng
crs_bng <- "+init=epsg:27700"
crs_osm <- "+init=epsg:3857"    # OSM projection

# Set coordinate systems for reprojecting
proj_wgs84 <- '+proj=longlat +datum=WGS84'
proj_bng <- '+proj=tmerc +lat_0=49 +lon_0=-2 +k=0.9996012717 +x_0=400000 +y_0=-100000 +ellps=airy
+datum=OSGB36 +units=m +no_defs'
proj_osm <- '+proj=merc +a=6378137 +b=6378137 +lat_ts=0.0 +lon_0=0.0 +x_0=0.0 +y_0=0.0 +k=1.0
+units=m +nadgrids=@null +wktext +no_defs'

# theme adjustments applied to all plots (for maps only)
t <- theme(axis.text = element_text(colour="white"),
           title = element_text(size=11),
           axis.ticks = element_line(colour="white"),
           axis.title.x = element_text(size=10),
           legend.text = element_text(size=10)
)

# FUNCTIONS
# =========
# reproject from lat/lng to OSM coordinates
# -----------------------------------------
# requires columns labeled 'lat', 'lng' specified as 'y' and 'x'
reproject <- function(df, x, y) {
  dataset_map_coords <- df[c(x, y)]
  dataset_map_data <- as.data.frame(df[ ,c(3)])
  dataset_map <- SpatialPointsDataFrame(coords=dataset_map_coords, data=dataset_map_data)

  # set CRS and reproject to OSM for OSM number system
  dataset_map@proj4string # check first, should be NA - not yet been set
  proj4string(dataset_map) <- CRS(crs_wgs84)  # set the current coordinates system
  dataset_map <- spTransform(dataset_map, CRS(proj_osm))  # reproject to OSM

  # convert back to dataframe with OSM cooordinates and then amend to original dataset
  newdf <- as.data.frame(dataset_map)
  names(newdf)[names(newdf) == x] <- 'osm_x'; names(newdf)[names(newdf) == y] <- 'osm_y'
  newdf <- newdf[, 2:3]   # just want to keep the OSM coordinates to append back to dataset
  df <- cbind(df, newdf)
  return(df)
```

```
}

# LOAD AND PREP OPENSTREETMAP MAP DATA
# ----------------------------
# Set bounding box for OSM map (to cover LSOAs, so this is larger than the normal bounding box
used)
LatMax <- 51.554    # North
LatMin <- 51.529    # South
LngMax <- 0.0040    # East
LngMin <- -0.034    # West

# Retrieve OSM maptiles
tile_server <- "http://tile.stamen.com/terrain/{z}/{x}/{y}.png"   # Stamen Design
basemap <- openmap(c(LatMax,LngMin),c(LatMin,LngMax), zoom=15, type=tile_server,
mergeTiles=FALSE)

#basemap <- openproj(basemap, projection=proj_wgs84)  # if reprojecting

# LOAD AND PREP MSOA, LSOA AND OA SHAPEFILES + WORKPLACE ZONES
# -----------------------------------------------------------
# load shapefile and reproject from default Merc to OSM, then convert shapefile to dataframe for
ggplot2
map_areas <- readOGR(dsn="./maps", "MSOA_2011_London_gen_MHW")
map_areas <- spTransform(map_areas, CRS(proj_osm))
areas_fort <- fortify(map_areas, region='MSOA11CD')
areas_geom <- arrange(areas_fort, group, order)
msoa_map <- areas_geom

# load shapefile and reproject from default Merc to OSM, then convert shapefile to dataframe for
ggplot2
map_areas <- readOGR(dsn="./maps", "LSOA_2011_London_gen_MHW")
map_areas <- spTransform(map_areas, CRS(proj_osm))
areas_fort <- fortify(map_areas, region='LSOA11CD')
areas_geom <- arrange(areas_fort, group, order)
lsoa_map <- areas_geom

# load shapefile and reproject from default Merc to OSM, then convert shapefile to dataframe for
ggplot2
map_areas <- readOGR(dsn="./maps", "OA_2011_London_gen_MHW")
map_areas <- spTransform(map_areas, CRS(proj_osm))
areas_fort <- fortify(map_areas, region='OA11CD')
areas_geom <- arrange(areas_fort, group, order)
oa_map <- areas_geom

# load shapefile and reproject from default Merc to OSM, then convert shapefile to dataframe for
ggplot2
map_areas <- readOGR(dsn="./maps", "Workplace_Zones__December_2011__Boundaries")
map_areas <- spTransform(map_areas, CRS(proj_osm))
areas_fort <- fortify(map_areas, region='wz11cd')
areas_geom <- arrange(areas_fort, group, order)
wz_map <- areas_geom

# count number of workplace zones
wz_map %>% summarize(count=n_distinct(id))

# LOAD AND PREP QEOP SHAPEFILE
# ----------------------------
# load park outline shapefile and project to OSM
folder <- paste0(source, 'ShapeFiles/')

# choose which outline to plot for QEOP
#map_park <- readOGR(dsn=folder, "qeop_lldc_park_outline")   # boundary includes immediate
surrounding area
map_park <- readOGR(dsn=folder, "qeop_lldc_park_wifi")      # boundary is park only - WiFi
coverage

map_park <- spTransform(map_park, CRS(proj_osm))

# convert to dataframe for ggplot
park_fort <- fortify(map_park, region='id')
park_geom <- arrange(park_fort, group, order)

park <- park_geom

# load OpenSignal data
# ----------------------------------------
input_file <- paste0(source, 'OpenSignal/os_london_2017-06_extract.csv' )
df <- read.csv(input_file, header = T, sep = ',')
names(df)[names(df) == "pseudo_id"] <- 'device'

# reproject to OSM for points plot
y <- 'lat'   # rem: y = lat
x <- 'lng'   # rem: x = lng
df <- reproject(df, x, y)

df$date <- as.Date(df$date_time)
```

```
# LOAD AND PREP LANDSCAN GRID
# ---------------------------
# set LandScan bounding box to trim population data to 3x4 grid cells (using their SW corner
coords)
LatMax <- 51.546    # North
LatMin <- 51.528    # South
LngMax <- -0.0040     # East
LngMin <- -0.0300    # West

# load landscan London population data, snip to park 3x4 grid and reproject
landscan <- read.csv(paste0(source, 'LandScan/LandScan2015_London.csv'), header = T, sep = ',')
landscan <- subset(landscan, landscan$lat >= LatMin & landscan$lat <= LatMax &
                     landscan$lng >= LngMin & landscan$lng <= LngMax)

# reprojecting the centroids. geom_tile draws based on centroids
y <- 'centreLat'   # rem: y = lat
x <- 'centreLng'   # rem: x = lng
landscan <- reproject(landscan, x, y)

#####################################

# get LandScan grid boundary in OSM coordinates to snip LSOAs and data to plot
# -----
osm_max_x <- max(landscan$osm_x)
osm_max_y <- max(landscan$osm_y)
osm_min_x <- min(landscan$osm_x)
osm_min_y <- min(landscan$osm_y)

osm_x_diff <- osm_max_x - osm_min_x
osm_y_diff <- osm_max_y - osm_min_y

basemap_min_x <- osm_min_x - (osm_x_diff * 0.4)
basemap_max_x <- osm_max_x + (osm_x_diff * 0.4)
basemap_min_y <- osm_min_y - (osm_y_diff * 0.4)
basemap_max_y <- osm_max_y + (osm_y_diff * 0.4)


# then snip the output area data to the basemap limits
msoa <- subset(msoa_map, (msoa_map$long >= basemap_min_x & msoa_map$long <= basemap_max_x) &
                 (msoa_map$lat >= basemap_min_y & msoa_map$lat <= basemap_max_y))

lsoa <- subset(lsoa_map, (lsoa_map$long >= basemap_min_x & lsoa_map$long <= basemap_max_x) &
                 (lsoa_map$lat >= basemap_min_y & lsoa_map$lat <= basemap_max_y))

oa <- subset(oa_map, (oa_map$long >= basemap_min_x & oa_map$long <= basemap_max_x) &
                 (oa_map$lat >= basemap_min_y & oa_map$lat <= basemap_max_y))

wz <- subset(wz_map, (wz_map$long >= basemap_min_x & wz_map$long <= basemap_max_x) &
               (wz_map$lat >= basemap_min_y & wz_map$lat <= basemap_max_y))

df <- subset(df, (df$osm_x >= basemap_min_x & df$osm_x <= basemap_max_x) &
               (df$osm_y >= basemap_min_y & df$osm_y <= basemap_max_y))


# VISUAL - MSOA, LSOA and OA with Landscan boundaries overlaid on OpenStreetMap
# ========
chart_title <- "Areal boundaries spanning the QEOP"
sub_title <- NULL
labels <- labs(title=NULL,
               subtitle = chart_title,
               x="Background: map tiles Stamen Design, under CC BY 3.0;\nmap data by
OpenStreetMap, under ODbL",
               y=NULL,
               caption = NULL)

landscan_layer <- geom_tile(aes(x=osm_x, y=osm_y, fill=NULL), alpha=0, colour="#555555",
size=0.8, data=landscan)
park_layer <- geom_polygon(aes(x=long, y=lat, fill=NULL, group=group, colour="park"), alpha=0,
size=1.0, colour="#0000ff", data=park)
msoa_layer <- geom_polygon(aes(x=long, y=lat, fill=NULL, group=group, colour="MSOA"), alpha=0,
size=0.8, data=msoa)
lsoa_layer <- geom_polygon(aes(x=long, y=lat, fill=NULL, group=group, colour="LSOA"), alpha=0,
size=0.8, data=lsoa)
oa_layer <- geom_polygon(aes(x=long, y=lat, fill=NULL, group=group, colour="OA"), alpha=0,
size=0.8, data=oa)
wz_layer <- geom_polygon(aes(x=long, y=lat, fill=NULL, group=group, colour="WZ"), alpha=0,
size=0.8, data=wz)

# MSOA version
autoplot(basemap) + msoa_layer + landscan_layer + park_layer + labels + coord_equal() + t
ggsave(filename=paste0('Rplot_MSOABoundaries_QEOP.png'), plot=last_plot(), device='png',
       scale=0.9, width=16, height=16, units='cm', dpi=300, limitsize=TRUE)

# LSOA version
autoplot(basemap) + lsoa_layer + landscan_layer + park_layer + labels + coord_equal() + t
ggsave(filename=paste0('Rplot_LSOABoundaries_QEOP.png'), plot=last_plot(), device='png',
       scale=0.9, width=16, height=16, units='cm', dpi=300, limitsize=TRUE)
```

```
# OA version
autoplot(basemap) + oa_layer + landscan_layer + park_layer + labels + coord_equal() + t
ggsave(filename=paste0('Rplot_OABoundaries_QEOP.png'), plot=last_plot(), device='png',
       scale=0.9, width=16, height=16, units='cm', dpi=300, limitsize=TRUE)

# WZ version
autoplot(basemap) + wz_layer + landscan_layer + park_layer + labels + coord_equal() + t
ggsave(filename=paste0('Rplot_WZBoundaries_QEOP.png'), plot=last_plot(), device='png',
       scale=0.9, width=16, height=16, units='cm', dpi=300, limitsize=TRUE)


# OpenSignal data
os <- subset(df, df$date <= '2017-06-03')
os_layer <- geom_point(aes(x=osm_x, y=osm_y), alpha=0.2, size=4, data=os)

landscan_layer <- geom_tile(aes(x=osm_x, y=osm_y, fill=NULL), alpha=0, colour="#ffff33",
size=0.8, data=landscan)
park_layer <- geom_polygon(aes(x=long, y=lat, fill=NULL, group=group, colour="#000000"), alpha=0,
size=1.0, colour="#33FFFF", data=park)
autoplot(basemap) + os_layer + oa_layer + landscan_layer + park_layer + labels + coord_equal() +
t
ggsave(filename=paste0('Rplot_OSdata_oa_QEOP.png'), plot=last_plot(), device='png',
       scale=0.9, width=16, height=16, units='cm', dpi=300, limitsize=TRUE)

autoplot(basemap) + os_layer + lsoa_layer + landscan_layer + park_layer + labels + coord_equal()
+ t
ggsave(filename=paste0('Rplot_OSdata_lsoa_QEOP.png'), plot=last_plot(), device='png',
       scale=0.9, width=16, height=16, units='cm', dpi=300, limitsize=TRUE)

autoplot(basemap) + os_layer + msoa_layer + landscan_layer + park_layer + labels + coord_equal()
+ t
ggsave(filename=paste0('Rplot_OSdata_msoa_QEOP.png'), plot=last_plot(), device='png',
       scale=0.9, width=16, height=16, units='cm', dpi=300, limitsize=TRUE)

autoplot(basemap) + os_layer + wz_layer + landscan_layer + park_layer + labels + coord_equal() +
t
ggsave(filename=paste0('Rplot_OSdata_wz_QEOP.png'), plot=last_plot(), device='png',
       scale=0.9, width=16, height=16, units='cm', dpi=300, limitsize=TRUE)
```

Comparing LandScan grid of uniform cells

LandScan grid generated manually by specifying coordinates. Code sample includes both LandScan 2017 and LandScan 2015 estimates. 2017 data added to check if there are significant changes in counts (there weren't, so original 2015 visuals used for chapter).

Note: Pixel grid not included here as it essentially follows the same approach but with the grid auto-generated (that process is included in chapter seven).

The same technique was then also applied to produce counts for an active population estimate using just the LandScan and OpenSignal data.

```
# Comparing counts using LandScan grid
#
########################

# DEVICE CONFIG
# =============
# root folder for source data
source <- '<FOLDER>'

# setup
# -----
library(dplyr)
library(ggplot2)
library(maps)
library(OpenStreetMap)
library(RColorBrewer)
library(scales)

# set point at which plot will switch from standard form
options(scipen=5)

# base chart theme
t <- theme_bw() + theme(panel.border = element_rect(colour = "#bdbdbd"),
      axis.text = element_text(colour="white"),
```

```
      axis.ticks = element_line(colour="white"), # hiding axis data because is OSM number system,
not meaningful
      axis.title.x = element_text(size=10),
      legend.title = element_text(colour="white")
)

# ====================== #
# CONFIGURATION PART 1 #
# ====================== #

# date range (yyyy-mm)
period <- '2017-06'
month <- 6

# specify area (will append with date for output filename)
area <- 'qeop'

# set bounding box/grid extent (OSM tile and datasnip)
LatMax <- 51.5541666666661 # North
LatMin <- 51.5291666666661 # South
LngMax <- 0.00416666666596077  # East
LngMin <- -0.0291666666673731 # West

# LandScan labels in Lat/Lng
LS_A <- -0.0291666667
LS_B <- -0.0208333334
LS_C <- -0.0125000001
LS_D <- -0.0041666667
LS_1 <- 51.545833333
LS_2 <- 51.537499995
LS_3 <- 51.529166666

# Function to add LandScan labels (to count within LandScan grid cells)
add_labels <- function(df) {

 df$row[df$lat >= LS_3 & df$lat < LS_2] <- '3'
 df$row[df$lat >= LS_2 & df$lat < LS_1] <- '2'
 df$row[df$lat >= LS_1 & df$lat < LatMax] <- '1'
 df$col[df$lng >= LS_D & df$lng < LngMax] <- 'D'
 df$col[df$lng >= LS_C & df$lng < LS_D] <- 'C'
 df$col[df$lng >= LS_B & df$lng < LS_C] <- 'B'
 df$col[df$lng >= LS_A & df$lng < LS_B] <- 'A'
 df <- mutate(df, ls_label = paste0(col, row))

 num_cols <- ncol(df) # figure out number of columns to clean up after adding labels
 df <- df[, c(1:(num_cols-3), num_cols)]

 return(df)
}

# load data
# =========

# Landscan
# --------
# LandScan 2017
df <- read.csv(paste0(source, 'LandScan/LandScan2017_London.csv'), header = T, sep = ',')

# snip to QEOP grid and setup labels
df <- subset(df, df$lat > 51.5290 & df$lat < 51.5459 & df$lng > -0.030 & df$lng < -0.0040)

# add landscan labels
df <- add_labels(df)
#df <- mutate(df, pct = count/sum(count)) # add count as a percentage
landscan <- df
rm(df)

# LandScan 2015
df <- read.csv(paste0(source, 'LandScan/LandScan2015_London.csv'), header = T, sep = ',')
df <- subset(df, df$lat > 51.5290 & df$lat < 51.5459 & df$lng > -0.030 & df$lng < -0.0040)
df <- add_labels(df)
lscan2015 <- df

# OpenSignal
# ----------
#input_file <- paste0(source, 'OpenSignal/os_', area, '_', period, '_prepped_agg.csv' )
input_file <- paste0(source, 'OpenSignal/opensignal_raw_tagged_2017Jun.csv' )
df <- read.csv(input_file, header = T, sep = ',')
df <- subset(df, df$lat > LatMin & df$lat < LatMax & df$lng > LngMin & df$lng < LngMax) # make
sure data is within grid extent
names(df)[names(df) == "pseudo_id"] <- 'device'

# sort by device (want to count number of devices, not observations, for ambient population)
df <- df[order(df$device), ]

# add landscan labels
df <- add_labels(df)
```

```
pop_os <- df
rm(df)

#pop_all <- pop_os
#pop_os <- subset(pop_os, pop_os$day == day_of_int)

# adding deviceID including grid square (to count unique devices daily per grid square)
#pop_os <- mutate(pop_os, deviceGrid = paste0(device, '_', ls_label))

# Crime Stats (street-level)
# ------------
# have to use the street-level stats that contain jittered coordinates.
get_crimes <- function(period) {

 input_file <- paste0(source, 'DataStore/lmps_', period, '-street.csv' )
 df <- read.csv(input_file, header = T, sep = ',')
 df <- subset(df, df$lat > LatMin & df$lat < LatMax & df$lng > LngMin & df$lng < LngMax) # snip
records to grid extent

 return(df)
}

period <- '2017-03'
df <- get_crimes(period)
crime <- df

period <- '2017-04'
df <- get_crimes(period)
crime <- rbind(crime, df)

period <- '2017-05'
df <- get_crimes(period)
crime <- rbind(crime, df)

# add landscan labels
crime <- add_labels(crime)
rm(df)

# Fire records
# ------------
input_file <- paste0(source, 'DataStore/lfb_fire2017_recoded.csv' )
df <- read.csv(input_file, header = T, sep = ',')
df <- subset(df, df$lat > LatMin & df$lat < LatMax & df$lng > LngMin & df$lng < LngMax) # snip
records to grid extent

# add landscan labels
df <- add_labels(df)

fire <- df
rm(df)

#fire_all <- fire
#fire <- subset(fire, fire$month == month & fire$day == day_of_int)
#fire <- fire_all

# ===================== #
# CONFIGURATION PART 2 #
# ===================== #

# create single table containing population counts and scaled using min-max where min = 0
# ================================================================================

# base grid to plot
lsgrid <- landscan[, c(10, 1:2, 4:7)]
lsgrid <- lsgrid[order(lsgrid$ls_label), ]

# landscan population
# -------------------
# 2017 data
df <- landscan[, c(10, 3)] # already aggregated by ls_label, prepping to merge
names(df)[names(df) == 'count'] <- 'lscan'
df <- mutate(df, lscan_nm = lscan/max(df$lscan, na.rm=T)) # divide count by max value
lsgrid <- merge(lsgrid, df, by.x='ls_label', by.y='ls_label', all.x = T)

# 2015 data
df <- lscan2015[, c(10, 3)] # already aggregated by ls_label, prepping to merge
names(df)[names(df) == 'count'] <- 'ls2015'
df <- mutate(df, ls2015_nm = lscan/max(df$ls2015, na.rm=T)) # divide count by max value
lsgrid <- merge(lsgrid, df, by.x='ls_label', by.y='ls_label', all.x = T)

# OpenSignal counts
# -----------------
# all data
df <- pop_os
df <- as.data.frame(df %>% group_by(ls_label) %>% summarise(osdev = n_distinct(device)))
df <- mutate(df, osdev_nm = osdev/max(df$osdev, na.rm=T)) # divide count by max value
```

```
lsgrid <- merge(lsgrid, df, by.x='ls_label', by.y='ls_label', all.x = T)

# daytime (7am-7pm)
df <- pop_os
df <- subset(df, df$hr >= 7 & df$hr < 19)
df <- as.data.frame(df %>% group_by(ls_label) %>% summarise(osday = n_distinct(device)))
df <- mutate(df, osday_nm = osday/max(df$osday, na.rm=T)) # divide count by max value
lsgrid <- merge(lsgrid, df, by.x='ls_label', by.y='ls_label', all.x = T)

# night (7pm-7am)
df <- pop_os
df <- subset(df, df$hr >= 19 | df$hr < 7)
df <- as.data.frame(df %>% group_by(ls_label) %>% summarise(osnt = n_distinct(device)))
df <- mutate(df, osnt_nm = osnt/max(df$osnt, na.rm=T)) # divide count by max value
lsgrid <- merge(lsgrid, df, by.x='ls_label', by.y='ls_label', all.x = T)

# Crime
# -----
# all crime
df <- crime
df <- as.data.frame(df %>% group_by(ls_label) %>% summarise(crime = n()))
df <- mutate(df, crime_nm = crime/max(df$crime, na.rm=T)) # divide count by max value
lsgrid <- merge(lsgrid, df, by.x='ls_label', by.y='ls_label', all.x = T)

# theft from person
df <- crime
df <- subset(df, df$crime_type == 'Theft from the person')
df <- as.data.frame(df %>% group_by(ls_label) %>% summarise(ctheft = n()))
df <- mutate(df, ctheft_nm = ctheft/max(df$ctheft, na.rm=T)) # divide count by max value
lsgrid <- merge(lsgrid, df, by.x='ls_label', by.y='ls_label', all.x = T)

# Anti-social behaviour
df <- crime
df <- subset(df, df$crime_type == 'Anti-social behaviour')
df <- as.data.frame(df %>% group_by(ls_label) %>% summarise(canti = n()))
df <- mutate(df, canti_nm = canti/max(df$canti, na.rm=T)) # divide count by max value
lsgrid <- merge(lsgrid, df, by.x='ls_label', by.y='ls_label', all.x = T)

# Public order
df <- crime
df <- subset(df, df$crime_type == 'Public order')
df <- as.data.frame(df %>% group_by(ls_label) %>% summarise(cpublic = n()))
df <- mutate(df, cpublic_nm = cpublic/max(df$cpublic, na.rm=T)) # divide count by max value
lsgrid <- merge(lsgrid, df, by.x='ls_label', by.y='ls_label', all.x = T)

# Fire
# ----
df <- fire
df <- as.data.frame(df %>% group_by(ls_label) %>% summarise(fire = n()))
df <- mutate(df, fire_nm = fire/max(df$fire, na.rm=T)) # divide count by max value
lsgrid <- merge(lsgrid, df, by.x='ls_label', by.y='ls_label', all.x = T)

# daytime (7am-7pm)
df <- fire
df <- subset(df, df$hr >= 7 & df$hr < 19)
df <- as.data.frame(df %>% group_by(ls_label) %>% summarise(fireday = n()))
df <- mutate(df, fireday_nm = fireday/max(df$fireday, na.rm=T)) # divide count by max value
lsgrid <- merge(lsgrid, df, by.x='ls_label', by.y='ls_label', all.x = T)

# night (7pm-7am)
df <- fire
df <- subset(df, df$hr >= 19 | df$hr < 7)
df <- as.data.frame(df %>% group_by(ls_label) %>% summarise(firent = n()))
df <- mutate(df, firent_nm = firent/max(df$firent, na.rm=T)) # divide count by max value
lsgrid <- merge(lsgrid, df, by.x='ls_label', by.y='ls_label', all.x = T)

subgrid <- subset(lsgrid, lsgrid$ls_label == 'B1' | lsgrid$ls_label == 'B2' | lsgrid$ls_label ==
'B3' |
     lsgrid$ls_label == 'C2' | lsgrid$ls_label == 'D2')


# VISUALISATION
# =============

# Set any NAs to 0
# ----------------
# ensures all 12 cells always plot
lsgrid[is.na(lsgrid)] <- 0

# function to plot as 3x4 choropleth grid (using LandScan coordinates)
# ---------------------------------------
# rename selected data column to standard label for plot, then rename back after
plot_data <- function(data_col, data_title) {

 names(lsgrid)[names(lsgrid) == data_col] <- 'pop_data' # rename column for plotting

 labels <- labs(title=NULL,
```

```
     subtitle = data_title,
     x=NULL,
     y=NULL,
     caption = NULL)

 layer <- geom_tile(aes(x=lng, y=lat, fill=pop_data), colour='#888888', size=0.2, data=lsgrid)

 # display
 plot <- ggplot() + layer + labels + scale_fill_continuous(limits=c(0, 1), low="#333333",
high="#efefef") +
 scale_x_continuous(expand = c(0,0)) + scale_y_continuous(expand = c(0,0)) + t

 # save
 ggsave(filename=paste0('./Rplot_lscan_qeop/Rplot_qeop_lscan', data_col, '.png'),
plot=last_plot(), device='png',
   scale=0.8, width=10, height=8, units='cm', dpi=300, limitsize=TRUE)

 names(lsgrid)[names(lsgrid) == 'pop_data'] <- data_col # rename column back to original

 return(plot) # defining the plot and returning means it will display in Plots window
}

# Visualise thematic grid
# -----------------------
# LandScan
data_col = 'lscan_nm'
data_title = "a) LandScan Ambient"
plot_data(data_col, data_title)

# OpenSignal devices
data_col = 'osdev_nm'
data_title = "b) OpenSignal Devices"
plot_data(data_col, data_title)

# OpenSignal day
data_col = 'osday_nm'
data_title = "c) OpenSignal - Day"
plot_data(data_col, data_title)

# OpenSignal night
data_col = 'osnt_nm'
data_title = "d) OpenSignal - Night"
plot_data(data_col, data_title)

# Crime
data_col = 'crime_nm'
data_title = "e) Crime Incidents"
plot_data(data_col, data_title)

# Crime - theft
data_col = 'ctheft_nm'
data_title = "f) Crime - Theft"
plot_data(data_col, data_title)

# Crime - anti social behaviour
data_col = 'canti_nm'
data_title = "g) Crime - Antisocial"
plot_data(data_col, data_title)

# Crime - public disorder
data_col = 'cpublic_nm'
data_title = "h) Crime - Disorder"
plot_data(data_col, data_title)
# Fire
data_col = 'fire_nm'
data_title = "i) Fire Incidents"
plot_data(data_col, data_title)

# Fire day
data_col = 'fireday_nm'
data_title = "j) Fire - Day"
plot_data(data_col, data_title)

# Fire night
data_col = 'firent_nm'
data_title = "k) Fire - Night"
plot_data(data_col, data_title)
```

Detecting active spaces and analysing trip behaviours (stages)

Data has already been tagged with trips, see Appendix B.2

Data-driven clustering using DBSCAN (Two-tier approach)

Active spaces identified based on clusters of data points. Script to load data, snip to Shapefile boundary (park outline), detect clusters (in two tiers), and draw convex hulls around points within each cluster. All points that fall within the cluster is then tagged with the ID for that cluster.

```
# Clustering analysis
#
# DEVICE
# ======
root_folder <- '<FOLDER>'

# SETUP
# =====
library(plyr)   # for hulls and joins
library(dplyr)
library(dbscan)   # clustering algorithm
library(ggplot2)   # for plots (and theme)
library(rgeos)   # for gIntersects
library(rgdal)   # for readOGR and SpatialPointsDataFrame
library(OpenStreetMap) # for map background
library(lubridate)  # date/time calcs
library(geosphere)  # to calculate centroids of polygons

# set point at which plot will switch from standard form
options(scipen=7)

# set coordinate referencing system (for changing CRS but not reprojecting)
crs_wgs84 <- "+init=epsg:4326" # lat/lng
crs_bng <- "+init=epsg:27700"
crs_osm <- "+init=epsg:3857" # OSM projection

# Set coordinate systems for reprojecting
proj_wgs84 <- '+proj=longlat +datum=WGS84'
proj_bng <- '+proj=tmerc +lat_0=49 +lon_0=-2 +k=0.9996012717 +x_0=400000 +y_0=-100000 +ellps=airy
+datum=OSGB36 +units=m +no_defs'
proj_osm <- '+proj=merc +a=6378137 +b=6378137 +lat_ts=0.0 +lon_0=0.0 +x_0=0.0 +y_0=0.0 +k=1.0
+units=m +nadgrids=@null +wktext +no_defs'

# theme adjustments applied to all plots (for maps only)
t <- theme(axis.text = element_text(colour="white"),
   title = element_text(size=11),
   axis.ticks = element_line(colour="white"), # hiding axis data because is OSM number system,
not meaningful
   axis.title.x = element_text(size=10),
   legend.text = element_text(size=10)
)

# Set bounding box for map
LatMax <- 51.554167 # North
LatMin <- 51.529167 # South
LngMax <- 0.0041667  # East
LngMin <- -0.0291667 # West

# Retrieve OSM maptiles
tile_server <- "http://tile.stamen.com/toner-lite/{z}/{x}/{y}.png" # Stamen Design
basemap <- openmap(c(LatMax,LngMin),c(LatMin,LngMax), zoom=15, type=tile_server,
mergeTiles=FALSE)

# load shapefiles - lldc park boundary, venues and cluster zones
# --------------------------------------------------------------
folder <- paste0(root_folder, 'data/ShapeFiles/')

# function to convert to a dataframe that can be plotted using geom_polygon
map_to_df <- function(map) {

 # convert shapefile to dataset for plotting as geom_polygon
 map@data$id <- rownames(map@data)
 map.points <- fortify(map, region="id")
 df <- join(map.points, map@data, by="id")

 names(df)[names(df) == "long"] <- "lng" # rename from long to lng

 # reproject map to OSM to also get OSM coordinates
 proj4string(map) <- CRS(crs_wgs84)
 map <- spTransform(map, CRS(proj_osm))

 map@data$id <- rownames(map@data)
 map.points <- fortify(map, region="id")
 osmdf <- join(map.points, map@data, by="id")

 names(osmdf)[names(osmdf) == "long"] <- "osm_x"
 names(osmdf)[names(osmdf) == "lat"] <- "osm_y"
```

```
 # append OSM coordinates to DF with lat/lng
 osmdf <- osmdf[, 1:2] # just keep the osm_x and osm_y cols (is in same order as df)
 df <- cbind(df, osmdf)

 return(df)
}

# load shapefiles
park_shapefile <- readOGR(dsn=folder, "qeop_lldc_park_outline")
venues_shapefile <- readOGR(dsn=folder, "qeop_venues")
zones_shapefile <- readOGR(dsn=folder, "qeop_lldc_park_outline_zoned")

# convert from map to dataset for plotting as geom_polygon
park_outline <- map_to_df(park_shapefile)
park_venues <- map_to_df(venues_shapefile)
park_zones <- map_to_df(zones_shapefile)

# only keep the main park venues, drop the rest from the venues shapefile
park_venues <- subset(park_venues, park_venues$id>=0 & park_venues$id<=3 | park_venues$id==8)

# FUNCTIONS
# =========

# snip data points to within park outline
# ---------------------------------------------
# assumes columns containing coords are 'lat' and 'lng'
# requires park_shapefile to have been loaded (is done in setup)
snip_data <- function(df) {

 # create SpatialPointsPolygon
 dataset_map_coords <- df[c("lng", "lat")]
 dataset_map_data <- df[ ,c(1:9, 12:ncol(df))]  # need to keep the data, check this fits with
dataset

 dataset_map <- SpatialPointsDataFrame(coords=dataset_map_coords, data=dataset_map_data)

 # check coordinates referencing system (CRS) should be NA - not yet been set
 dataset_map@proj4string

 # set CRS as WGS84 (lat/lng)
 proj4string(dataset_map) <- CRS(crs_wgs84)
# dataset_map <- spTransform(dataset_map, CRS(proj_osm))

 # --- match data points within zone, drop records outside of boundaries
 area_intersect <- gIntersects(dataset_map, park_shapefile, byid=T)
 clipped <- apply(area_intersect == F, MARGIN = 2, all)
 data_included <- dataset_map[which(!clipped), ]

 # replace points_map and points_data with clipped data (data points within boundary only)
 dataset_map <- data_included

 # write snipped data back to dataframe and include zone tag for each data point
 newdata <- as.data.frame(dataset_map)

 return(newdata)

}

# run clustering
# --------------
find_clusters <- function(df, eps, min_pts) {

 # convert df to matrix/array for feeding to DBSCAN
 x <- as.matrix(df[, c('lng', 'lat') ])

 # run DBSCAN clustering
 set.seed(42) # for randomising within DBSCAN (KNN)
 res <- dbscan(x, eps = eps, minPts = min_pts, weights = NULL, borderPoints = TRUE)

 # merge cluster results back into dataset
 results <- as.data.frame(res$cluster)  # res$cluster contains number of the cluster the data
point is a member of
 names(results)[names(results) == 'res$cluster'] <- 'cluster'
 df <- cbind(df, results)
 df$cluster <- sprintf("%02d",df$cluster) # to ensure discrete colour selection in plots (is a
category, not a value)

 # note: will append the number of the cluster to each data item
 # cluster '0' means data point is not part of any cluster (i.e. 0 = the set of unclustered data
points)

 return(df)
}

# reproject from lat/lng to OSM coordinates
```

```
# ----------------------------------------
# requires columns labeled 'lat', 'lng', 'pseudo_id', and 'cluster' (latter is created in this
script)
reproject <- function(df) {
 dataset_map_coords <- df[c("lng", "lat")]
 dataset_map_data <- df[ ,c("pseudo_id", "cluster")]
 dataset_map <- SpatialPointsDataFrame(coords=dataset_map_coords, data=dataset_map_data)

 # set CRS and reproject to OSM for OSM number system
 dataset_map@proj4string # check first, should be NA - not yet been set
 proj4string(dataset_map) <- CRS(crs_wgs84) # set the current coordinates system
 dataset_map <- spTransform(dataset_map, CRS(proj_osm)) # reproject to OSM

 # convert back to dataframe with OSM cooordinates and then amend to original dataset
 newdf <- as.data.frame(dataset_map)
 names(newdf)[names(newdf) == "lng"] <- 'osm_x'; names(newdf)[names(newdf) == "lat"] <- 'osm_y'
 newdf <- newdf[, 3:4] # just want to keep the OSM coordinates to append back to dataset
 df <- cbind(df, newdf)
 return(df)
}

# find convex hulls to draw boundaries around clusters in ggplot
# ------------------------------------------------------------
# requires colum,ns 'osm_x' and 'osm_y' (i.e. run this after reprojection)
find_hull <- function(df) df[chull(df$osm_x, df$osm_y), ]

convex_hulls <- function(df) {
 clustered <- subset(df, cluster != '00') # only include points in a cluster
 hulls <- ddply(clustered, "cluster", find_hull)
 return(hulls)
}

# get centroid for each cluster convex hull
findCentroid <- function(x, y, ...){
 centroid(cbind(x, y), ...)
}

# detect and plot clusters on an OSM background, save to file (image and clustered data)
# ------------------------------------------------------------
plot_clusters <- function(df, eps, min_pts, chart_title, pic_num, file_name) {

 df <- find_clusters(df, eps, min_pts)  # find clusters, append cluster ids to dataset
 df <- reproject(df)      # convert lat/lng to OSM coordinates for map plotting
 hulls <- convex_hulls(df)     # find convex hulls to surround each cluster detected

 by_cluster <- group_by(hulls, cluster)
 clusters <- summarise(by_cluster,
      devices = n_distinct(pseudo_id),
      ctr_x = min(osm_x)+(max(osm_x)-min(osm_x))/2,
      ctr_y = min(osm_y)+(max(osm_y)-min(osm_y))/2
 )

 clustered <- subset(df, df$cluster != '00')  # plot clustered data points as a layer (will be
coloured per cluster)
 notclustered <- subset(df, df$cluster == '00') # plot unclustered datapoints as separate layer
(will be coloured grey)

 colour_notcluster <- '#aaaaaa' # greyscale colour for unclustered data points
 num_clusters <- max(df$cluster) # count number of clusters detected

 labels <- labs(title=chart_title,
     subtitle = paste0(pic_num, "DBSCAN with eps: ", eps, " and min pts: ", min_pts, ". Clusters:
", num_clusters),
     x="Background: map tiles by Stamen Design, under CC BY 3.0;\nmap data by OpenStreetMap,
under ODbL",
     x=NULL,
     y=NULL,
     caption = NULL)

 # note: using cluster_id for colour, needs to be a string or will apply a gradient instead of
discrete colour scale
 layer_clustered <- geom_point(aes(x=osm_x, y=osm_y, colour=cluster), data=clustered, size=1,
alpha=0.6)
 layer_notclustered <- geom_point(aes(x=osm_x, y=osm_y), data=notclustered, size=1,
colour=colour_notcluster, alpha=0.4)
 layer_hulls <- geom_polygon(aes(x=osm_x, y=osm_y, group=cluster, colour=cluster, fill=cluster),
data=hulls, alpha=0.2)
 layer_clusterids <- geom_label(aes(x=ctr_x, y=ctr_y, label=cluster, fill=cluster),
data=clusters, alpha=0.4)
 layer_outline <- geom_polygon(aes(x=osm_x, y=osm_y, group=id), data=park_outline,
colour='#333333', fill=NA, size=0.8)
 layer_venues <- geom_polygon(aes(x=osm_x, y=osm_y, group=id), data=park_venues,
colour='#333333', fill=NA, size=0.8)

 #basemap <- openproj(basemap, projection=proj_osm) # not needed, already on OSM projection

# plot <- autoplot(basemap) + layer_clustered + layer_notclustered + layer_hulls +
layer_clusterids +
```

```
# layer_outline + layer_venues + labels + coord_equal() + t

 plot <- autoplot(basemap) + layer_clustered + layer_notclustered + layer_clusterids +
 layer_outline + layer_venues + labels + coord_equal() + t

 # save plot to file
 ggsave(filename=paste0('Rplot_clusteranalysis_', file_name, '.png'), plot=last_plot(),
device='png',
   scale=0.8, width=20, height=20, units='cm', dpi=300, limitsize=TRUE)

 # save cluster results to file
 write.csv(df, paste0('osmay17_trips_dataclusters_', file_name, '.csv'), row.names = F)

 return(plot)
}

# two-tier clustering on an OSM background, save to file (image and clustered data)
# ---------------------------------------------------------
# already have completed find_clusters and reproject
# need a modified convex_hulls to use the column for both sets of clustering
twin_clusters <- function(df, file_name) {

 clustered <- subset(df, cluster != '00' & cluster2tier != paste0(max_cluster, '_00')) # only
include points in a cluster
 hulls <- ddply(clustered, "cluster2tier", find_hull)

 by_cluster <- group_by(hulls, cluster2tier)
 clusters <- summarise(by_cluster,
      devices = n_distinct(pseudo_id),
      ctr_x = min(osm_x)+(max(osm_x)-min(osm_x))/2,
      ctr_y = min(osm_y)+(max(osm_y)-min(osm_y))/2
 )

 clustered <- subset(df, df$cluster != '00' & df$cluster2tier != paste0(max_cluster, '_00'))  #
plot clustered data points as a layer (will be coloured per cluster)
 notclustered <- subset(df, df$cluster == '00' | df$cluster2tier == paste0(max_cluster, '_00')) #
plot unclustered datapoints as separate layer (will be coloured grey)

 colour_notcluster <- '#aaaaaa' # greyscale colour for unclustered data points
 labels <- labs(title='Two-tier DBSCAN.',
     subtitle = 'Adjusted eps and min pts for dividing mega cluster',
     x="Background: map tiles by Stamen Design, under CC BY 3.0;\nmap data by OpenStreetMap,
under ODbL",
     x=NULL,
     y=NULL,
     caption = NULL)

 # going to split cluster labelling up to only show _NN for the mega cluster
 clusters <- as.data.frame(clusters)
 cluster_ids1 <- subset(clusters, clusters$cluster2tier >= '02_00')
 cluster_ids1$cluster2tier <- strtrim(cluster_ids1$cluster2tier, 2)
 cluster_ids2 <- subset(clusters, clusters$cluster2tier >= '01_00' & clusters$cluster2tier <
'02_00')


 # note: using cluster_id for colour, needs to be a string or will apply a gradient instead of
discrete colour scale
 layer_clustered <- geom_point(aes(x=osm_x, y=osm_y, colour=cluster2tier), data=clustered,
size=1, alpha=0.6)
 layer_notclustered <- geom_point(aes(x=osm_x, y=osm_y), data=notclustered, size=1,
colour=colour_notcluster, alpha=0.4)
 layer_hulls <- geom_polygon(aes(x=osm_x, y=osm_y, group=cluster2tier, colour=cluster2tier,
fill=cluster2tier), data=hulls, alpha=0.2)
 layer_clusterids1 <- geom_label(aes(x=ctr_x, y=ctr_y, label=cluster2tier, fill=cluster2tier),
data=cluster_ids1, alpha=0.4)
 layer_clusterids2 <- geom_label(aes(x=ctr_x, y=ctr_y, label=cluster2tier, fill=cluster2tier),
data=cluster_ids2, alpha=0.4)
 layer_outline <- geom_polygon(aes(x=osm_x, y=osm_y, group=id), data=park_outline,
colour='#333333', fill=NA, size=0.8)
 layer_venues <- geom_polygon(aes(x=osm_x, y=osm_y, group=id), data=park_venues,
colour='#333333', fill=NA, size=0.8)

 #basemap <- openproj(basemap, projection=proj_osm) # not needed, already on OSM projection

 # one of the following, depending on whether or not splitting the cluster id labels
# plot <- autoplot(basemap) + layer_clustered + layer_notclustered + layer_hulls +
layer_clusterids +
# layer_outline + layer_venues + labels + coord_equal() + t

 plot <- autoplot(basemap) + layer_clustered + layer_notclustered + layer_hulls +
layer_clusterids1 + layer_clusterids2 +
 layer_outline + layer_venues + labels + coord_equal() + t

 # save plot to file
 ggsave(filename=paste0('Rplot_clusteranalysis_', file_name, '.png'), plot=last_plot(),
device='png',
   scale=0.8, width=20, height=20, units='cm', dpi=300, limitsize=TRUE)

 # save cluster results to file
```

```
 write.csv(df, paste0('osmay17_trips_dataclusters_', file_name, '.csv'), row.names = F)

 return(plot)
}

##########################################

# LOAD DATA - MAY 2017 with trips
# ================================
folder <- paste0(root_folder, 'data/opensignal/')
data_file <- 'opensignal_raw_tagged_2017May_trips.csv'

# load and prep data
input_data <- paste0(folder, data_file)
df <- read.csv(input_data)
df$date_time <- as.POSIXct(df$date_time, format = '%Y-%m-%d %H:%M:%S')
df$date <- as.Date(df$date)
df$wday <- ifelse(wday(df$date)==1,7,wday(df$date)-1)

# add in pseudo_id + trip
df$pseudo_trip <- paste0(df$pseudo_id, '_', df$dev_trip)

# drop surplus columns
df <- subset(df[, c(1:6, 16, 15, 17, 7:11)])

df <- df[order(df$pseudo_id, df$date_time),] # make sure in date_time order
df$mins <- minute(df$date_time) # extract mins as separate column
df$mod <- df$mins %/% 1   # create 1-minute interval (return number of divisions)
df <- mutate(df, dev_hr_mod = paste0(pseudo_id, hr, mod)) # create field for device + hr + mod
interval

# reduce to one device reading per interval (1-second)
df$persec <- paste0(df$pseudo_id, '_', df$date_time)
df <- df[match(unique(df$persec),df$persec),]  # keep only first entry (per device per interval)
df <- df[, c(1:17)] # drop the persec column

# note: if doing per-minute, could just use the mins column. But sticking with mod to keep code
consistent
# mod %% 1 returns number of divisions, e.g. for 5 mins, will be 0 to 11, for 1 mins, will be 0
to 59, for 15 mins, 0 to 3

# reduce to readings from 05am to 11pm (22:59)
#df <- subset(df, df$date <= '2017-05-28')
df <- subset(df, df$hr >= 5)

n_distinct(df$pseudo_id) # count number of unique devices
n_distinct(df$pseudo_trip)
test <- df[match(unique(df$dev_hr_mod),df$dev_hr_mod),] # reduce to one reading per minute per
device


# now snip to datapoints falling within park and surrounding area outline
# ------------------------------------------------------------------------
df <- snip_data(df)

n_distinct(df$pseudo_id) # count number of unique devices
n_distinct(df$pseudo_trip)

snipped <- df

##########################################

# Two-tier DBSCAN clustering
# ==========================
# run individual functions separately, then run twin_clusters function to layer up the two sets

# first pass
subdf <- df
chart_title <- 'a) All days, 1 to 30 May, 2017'
file_name <- 'alldays_2tier'
pic_num <- ''
min_pts <- 200
eps <- 0.0006
plot_clusters(subdf, eps, min_pts, chart_title, pic_num, file_name)

subdf <- find_clusters(subdf, eps, min_pts)  # find clusters, append cluster ids to dataset then
subset largest
subdf$val <- 1 # adding a val of 1 to every record, to add them up in aggregate (give it
something to add)
clusters <- aggregate(x=subdf[,19], by=list(cluster=subdf$cluster), FUN=sum)

# find the largest cluster (this assumes there is only one mega cluster)
clusters <- clusters[2:nrow(clusters), ] # drop first row (cluster 0 = unclustered points)
sum_count <- sum(clusters$x) # total count across clusters
max_cluster <- clusters[which.max(clusters[,2]),1] # find max value in col 2 (x), return val in
col 1 (cluster)

# subset the megacluster and detect its own clusters. Rename cluster id column so can merge
```

```
newdf <- subset(subdf, subdf$cluster == max_cluster) # new subset - the mega cluster
newdf <- newdf[, 1:17] # drop the cluster columns ready to re-run

# now re-run cluster detection just for the points in the mega-cluster
chart_title <- 'b) Pass two: Largest cluster re-evaluated'
pic_num <- ''
min_pts <- 800
eps <- 0.0004
file_name <- 'alldays_2tierb'
plot_clusters(newdf, eps, min_pts, chart_title, pic_num, file_name)

# merge the two tiers together as a single data set and visual
newdf <- find_clusters(newdf, eps, min_pts)  # find clusters, append cluster ids to dataset then
subset largest
names(newdf)[names(newdf) == 'cluster'] <- 'cluster2'
newdf$merge <- paste0(newdf$pseudo_id, '_', newdf$date_time) # creates unique identifier for each
record
subdf$merge <- paste0(subdf$pseudo_id, '_', subdf$date_time) # ditto

newdf <- newdf[, c(19,18)] # just need unique identifier and cluster2 col
results <- merge(subdf, newdf, by.x='merge', by.y='merge', all.x=T)
results$cluster2[is.na(results$cluster2)] <- '00'
results$cluster2tier <- paste0(results$cluster, '_', results$cluster2)
results <- results[, c(2:19, 21:22)]

# now plot and export results for two-tier clustering (snipping to boundary)
results <- reproject(results)      # convert lat/lng to OSM coordinates for map plotting
file_name <- 'alldays_2tier_merge'
twin_clusters(results, file_name)

# note: have two options for plotting twin_clusters within the function
# default will use cluster2tier id, else modify to show shorter versions for park clusters
```

Analysing trip movements and dwell durations

Requires data to have been tagged with trip IDs and stage IDs (see Appendix B.2). Stage IDs for this analysis are based on active spaces identified using data-driven cluster detections.

```
# Analysing dwell times at, and movements between, data-driven clusters
#
####################################

# DEVICE
# ======
root_folder <- '<FOLDER>'

# SETUP
# =====
library(reshape2)  # dcast (pivot as a dataframe, acast for as an array/matrix)
library(dplyr)
library(ggplot2)   # for plots (and theme)
library(lubridate)  # date/time calcs

# set point at which plot will switch from standard form
options(scipen=7)

# theme adjustments applied to all plots
t <- theme(axis.text = element_text(colour="white"),
   title = element_text(size=11),
   axis.ticks = element_line(colour="white"), # hiding axis data because is OSM number system,
not meaningful
   axis.title.x = element_text(size=10),
   legend.text = element_text(size=10)
)

# base output filename
base_output <- 'data-cluster_stage_analysis_osmay17'
base_imgname <- 'Rplot_cluster_stage_analysis_osmay17'

# LOAD DATA - MAY 2017 with trips, clusters and stages
# ====================================================
folder <- paste0(root_folder, 'data/opensignal/')
data_file <- 'osmay17_trips_dataclusters_2tier_alldays_stages.csv'

# load and prep data
input_data <- paste0(folder, data_file)
df <- read.csv(input_data)
df$date_time <- as.POSIXct(df$date_time, format = '%Y-%m-%d %H:%M:%S')
df$date <- as.Date(df$date)
df$wday <- ifelse(wday(df$date)==1,7,wday(df$date)-1)
```

```
# add in pseudo_id + trip + stage (unique identifier for presence for each stage of trip, stage =
presence in a cluster (or in no cluster))
df$pseudo_stage <- paste0(df$pseudo_trip, '_', df$trip_stage)
df$pseudo_trip <- as.character(df$pseudo_trip)

str(df)

# renaming 'cluster' to 'cluster_basic' to focus on the 2-tier clusters (renaming them to
'cluster')
names(df)[names(df) == 'cluster'] <- 'cluster_basic'
names(df)[names(df) == 'cluster2tier'] <- 'cluster'

# making all non cluster readings the same (i.e. 02_00 is a non-cluster from 2nd tier DBSCAN,
00_00 is non-cluster from 1st tier DBSCAN)
df$cluster[df$cluster == '01_00'] <- '00_00'
df$cluster <- as.character(df$cluster)

original <- df

#########################################################

# PREP DATA FOR ANALYSIS

df <- original

# Duration of each stage (i.e. time spent in cluster for each stage of a trip)
# -----------------------
analysis <- df[order(df$pseudo_id, df$date_time), ]
analysis <- analysis %>% group_by(pseudo_stage, pseudo_id, dev_trip, trip_stage, cluster,
trip_stage, date, wday) %>%
 summarise(first_time = min(date_time),
   last_time = max(date_time))

analysis <- as.data.frame(analysis)

# ASSUMPTION: IF ONLY ONE READING IN A STAGE, SET DURATION TO 15 SECONDS (don't know how long
they were present, but they were present.)

# calculating and then adding 15 seconds to all durations
analysis <- mutate(analysis, duration = (last_time - first_time)+15)
analysis <- mutate(analysis, dur_mins = duration/60)

# add trip back in
analysis <- mutate(analysis, pseudo_trip = paste0(pseudo_id, '_', dev_trip))

# how many devices, trips and stages in the analysis
n_distinct(analysis$pseudo_id)
n_distinct(analysis$pseudo_trip)
n_distinct(analysis$pseudo_stage)

# how many devices and trips do not fall into any cluster at least once (i.e. only ever present
in clusters 00 and 01_00)
clustered <- subset(analysis, analysis$cluster != '00_00' & analysis$cluster != '01_00' )
notclustered <- subset(analysis, analysis$cluster == '00_00' | analysis$cluster == '01_00' )

length(setdiff(notclustered$pseudo_id, clustered$pseudo_id)) # number of not clustered devices
length(setdiff(notclustered$pseudo_trip, clustered$pseudo_trip)) # number of not clustered trips

# number of unique devices in each zone
test <- analysis %>% group_by(cluster) %>% summarise(count = n_distinct(pseudo_id))

# create summary that shows which cluster zones are formed on which days of the month
# - count number of devices in cluster on each date, earliest and last timestamp
df <- analysis %>% group_by(cluster, date) %>% summarise(count = n_distinct(pseudo_id)) %>%
ungroup()

df <- dcast(df, date ~ cluster) # pivots to show counts per cluster per date
df[is.na(df)] <- 0   # convert NAs (no cluster) to 0s

# Snipping to zones of interest
# =============================
df <- analysis
df <- subset(df, df$cluster != '00_00' & df$cluster != '01_00' & df$cluster != '01_04' &
    df$cluster != '06_00' & df$cluster != '09_00' & df$cluster != '10_00' &
    df$cluster != '11_00' & df$cluster != '13_00' & df$cluster != '14_00' &
    df$cluster != '17_00' & df$cluster != '12_00' &
    df$cluster != '02_00' & df$cluster!= '16_00')

analysis <- df

# Initial Statistics
# ==================
stats <- analysis %>% group_by(cluster) %>%
 summarise(devices = n_distinct(pseudo_id),
   trips = n_distinct(pseudo_trip),
   visits = n_distinct(pseudo_stage),
   trips_dev = trips/devices,
   visits_trips = visits/trips,
```

```
   min_dur = min(dur_mins),
   max_dur = max(dur_mins),
   mean_dur = mean(dur_mins),
   sd_dur = sd(dur_mins))

# bar chart showing count of devices, trips, visits (stages) at each cluster
# ----------------------------
df <- analysis
df <- as.data.frame(df %>% group_by(cluster) %>% summarise(devices=n_distinct(pseudo_id),
trips=n_distinct(pseudo_trip), stages=n_distinct(pseudo_stage)))
ggplot() +
 geom_bar(aes(x=cluster, y=devices), data=df, stat = "identity") +
 labs(subtitle='Count of devices per cluster: OpenSignal readings within QEOP, May 2017')
ggsave(filename=paste0(base_imgname, '_bar_devices.png'), plot=last_plot(), device='png',
  scale=0.8, width=20, height=14, units='cm', dpi=300, limitsize=TRUE)

ggplot() +
 geom_bar(aes(x=cluster, y=trips), data=df, stat = "identity") +
 labs(subtitle='Count of trips per cluster: OpenSignal readings within QEOP, May 2017')
ggsave(filename=paste0(base_imgname, '_bar_trips.png'), plot=last_plot(), device='png',
  scale=0.8, width=20, height=14, units='cm', dpi=300, limitsize=TRUE)

# histogram of duration of presence within cluster
# ------------------------------------------------
df <- analysis

binsize = 5
ggplot(data=df, aes(df$dur_mins)) +
 geom_histogram(breaks=seq(0, 1200, by=binsize), col='#444444', fill='#565656', alpha=0.8) +
 scale_x_continuous(breaks=seq(0, 1200, by=60), expand=c(0.01, 0.01)) +
 scale_y_continuous(expand=c(0.01, 0.01)) +
 labs(subtitle='Histogram of dwell time within clusters: OpenSignal readings within QEOP, May
2017') +
 labs(x=paste0("duration in minutes (bin width = ", binsize, " minutes)"), y="count")

ggsave(filename=paste0(base_imgname, '_hist.png'), plot=last_plot(), device='png',
  scale=0.8, width=20, height=14, units='cm', dpi=300, limitsize=TRUE)

# categorise and view duration of presence in clusters
# ----------------------------------------------------
df <- analysis
df$dwell_cat <- 'perm' # more than 360 minutes
df$dwell_cat[df$dur_mins <= 1] <- 'd01mins'
df$dwell_cat[df$dur_mins >1 & df$dur_mins <= 5] <- 'd05mins'
df$dwell_cat[df$dur_mins > 5 & df$dur_mins <= 20] <- 'd20mins' # 5 to 20 mins
df$dwell_cat[df$dur_mins > 20 & df$dur_mins <= 90] <- 'd90mins' # 20 to 90 mins
df$dwell_cat[df$dur_mins > 90 & df$dur_mins <= 360] <- 'dhours' # 20 to 90 mins

analysis <- df # adding in cluster dwell_time categories

################################
#
# ANALYSIS AND VISUALISATIONS
#
################################

# FUNCTIONS TO RE-RUN ANALYSIS AND VISUALS WITH DIFFERENT SUBSETS OF DATA

# group by duration of visit to cluster zone, pivot by category
# -------------------------------------------------------------
visit_duration <- function(df) {
 stats <- df %>% group_by(cluster, dwell_cat) %>% summarise(visits = n())
 pivot <- dcast(stats, cluster ~ dwell_cat)
 pivot[is.na(pivot)] <- 0
 return(pivot)
}

# Analyse clusters for different contexts - presence and dwell times
# ==================================================================

# All days (May 1 to 30)
subdf <- analysis
n_distinct(subdf$date) # number of days
res_dwell <- visit_duration(subdf)

# Weekdays no event or holiday (exclude 1, 5, 29, 30)
subdf <- subset(analysis, analysis$wday <= 5)
subdf <- subset(subdf, subdf$date != '2017-05-01' & subdf$date != '2017-05-05' & subdf$date !=
'2017-05-29' & subdf$date != '2017-05-30')
n_distinct(subdf$date) # number of days
res_dwell <- visit_duration(subdf)

# Weekends, no events (exclude 14)
subdf <- subset(analysis, analysis$wday >= 6)
subdf <- subset(subdf, subdf$date != '2017-05-14')
n_distinct(subdf$date) # number of days
res_dwell <- visit_duration(subdf)
```

```
# snip to clusters of interest
# ============================
subdf <- analysis
subdf <- subset(subdf, subdf$cluster=='01_01' | subdf$cluster=='01_02' | subdf$cluster=='01_03' |
     subdf$cluster=='04_00' | subdf$cluster=='05_00' | subdf$cluster=='08_00')

# box plot of min, max and mean dwell time per cluster
# ----------------------------------------------------
chart_title <- "Dwell times at data-driven clusters of interest, OpenSignal data, May 2017"
sub_title <- "all dates"
labels <- labs(title=chart_title,
    subtitle = sub_title,
    x=NULL,
    y=NULL,
    caption = NULL)
scales <- scale_y_continuous(limits=c(0,240)) # snipping outliers with duration longer than 4
hours

# weekdays, excluding bank and school holidays, and events
newdf <- subset(subdf, subdf$wday<6)
newdf <- subset(newdf, newdf$date != '2017-05-01' & newdf$date != '2017-05-05' & newdf$date <=
'2017-05-28')
file_name <- 'wkdays'
sub_title <- "box plot of dwell time (minutes), weekdays, no events or holidays"
labels[4] <- sub_title # sub-title label
ggplot(newdf, aes(cluster, dur_mins)) +
 geom_boxplot(outlier.colour = "#bb0000", outlier.shape = 1, outlier.alpha=0.6) +
 labels + scales

ggsave(filename=paste0('Rplot_data-cluster_boxplot_osmay17_', file_name, '.png'),
plot=last_plot(), device='png',
  scale=0.9, width=20, height=10, units='cm', dpi=300, limitsize=TRUE)

# plot the start times of visits for each cluster
# ===============================================
# weekdays, excluding bank and school holidays, and events
newdf <- subset(subdf, subdf$wday<6)
newdf <- subset(newdf, newdf$date != '2017-05-01' & newdf$date != '2017-05-05' & newdf$date <=
'2017-05-28')
file_name <- 'wkdays'

newdf$start_hr <- hour(newdf$first_time)
newdf$start_mins <- minute(newdf$first_time)
newdf$start_time <- paste0(newdf$start_hr, ":", newdf$start_mins)

# scatter plot
layer_data <- geom_point(aes(x=start_hr, y=cluster), data=newdf, shape=20, col="#dd0000", size=3,
alpha=0.8)
#layer_data2 <- geom_point(aes(x=last_zone, y=lst_hr), data=vizdf, shape=0, col="#0000ee",
size=4, alpha=0.8)
labels <- labs(subtitle = "Hour of first reading, per zone",
    x="zone", y="first hour",
    caption = NULL)
ggplot() + layer_data + labels +
 scale_x_continuous(breaks = seq(5, 23, by=1), limits=c(5,23), expand=c(0.01,0.01))

# plot count of visits as a cumulative distribution over time, based on start hour of visit
labels <- labs(title="Cumulative frequency of visit start times (hour), OS readings May 2017",
    y="percentage",
    x="start_hour of visit"
)
ggplot(newdf, aes(x=start_hr, group=cluster, colour=cluster)) + stat_ecdf() +
 scale_y_continuous(breaks=seq(0, 1, 0.1)) + labels


# plot as a count per hour per cluster (start hour of visit) - ploting with facet_wrap()
# =======================================================
# calculating average daily count, then plotting

# Weekdays versus weekends
# ------------------------
newdf <- subdf
newdf <- subset(newdf, newdf$date != '2017-05-01' & newdf$date != '2017-05-05' & newdf$date !=
'2017-05-14' & newdf$date <= '2017-05-28')
newdf$start_hr <- hour(newdf$first_time)

wkdays <- subset(newdf, newdf$wday <= 5)
num_days <- n_distinct(wkdays$date)
wkdays <- as.data.frame(wkdays %>% group_by(cluster, start_hr) %>% summarise(count =
n()/num_days))
wkends <- subset(newdf, newdf$wday >= 6)
num_days <- n_distinct(wkends$date)
wkends<- as.data.frame(wkends %>% group_by(cluster, start_hr) %>% summarise(count =
n()/num_days))

labels <- labs(subtitle = "Count of visits per start hour, per cluster, OpenSignal, May 2017",
    x="hour", y="count (smoothed)",
```

```
    caption = "Weekdays versus weekends (dotted)")
ggplot() +
 geom_smooth(aes(x=start_hr, y=count, group=cluster, colour=cluster), data=wkdays, size=1,
alpha=0.6, se=F) +
 geom_smooth(aes(x=start_hr, y=count, group=cluster, colour=cluster), data=wkends, size=1,
alpha=0.6, se=F, linetype="dotted") +
 scale_x_continuous(breaks=seq(5, 23, 2), limits=c(5,23), expand=c(0.01,0.01)) +
 facet_wrap(vars(cluster)) +
 labels

ggsave(filename=paste0('Rplot_clusters_starthour_counts_osmay17.png'), plot=last_plot(),
device='png',
  scale=1.0, width=20, height=16, units='cm', dpi=300, limitsize=TRUE)

# DURATION OF VISITS - DOES IT VARY?
# ==================================
# redoing the original box plot and distributions, now that the data has been recalculated with
mega-cluster split

# Box plots and duration statistics
# =================================
chart_title <- "Dwell times at data-driven clusters, OpenSignal data, May 2017"
sub_title <- "all dates"
labels <- labs(title=chart_title,
    subtitle = sub_title,
    x=NULL,
    y=NULL,
    caption = "excluding extreme outliers (0 or >240 mins)")
scales <- scale_y_continuous(limits=c(0,240)) # snipping outliers with duration longer than 4
hours

# exclude the visits with duration time of 0 or longer than 240 mins

subdf <- analysis
subdf$duration <- subdf$duration - 15 # removing the addition
subdf$dur_mins <- subdf$duration/60
subdf <- subset(subdf, subdf$duration > 0 & subdf$dur_mins <= 240)

# function to summarise presence duration
cluster_presence <- function(df) {
 df <- df %>% group_by(cluster) %>%
 summarise(devices = n_distinct(pseudo_id),
    trips = n_distinct(paste0(pseudo_id, dev_trip)),
    stages = n_distinct(paste0(pseudo_id, dev_trip, trip_stage)),
    dwell_time = sum(duration/60),
    min_dwell = min(duration/60),
    max_dwell = max(duration/60),
    mean_dwell = mean(duration/60),
    sd_dwell = sd(duration/60),
    iqr_dwell = IQR(duration/60))

 return(df)
}

# weekdays, excluding bank and school holidays, and events
# ---
newdf <- subset(subdf, subdf$wday<6)
newdf <- subset(newdf, newdf$date != '2017-05-01' & newdf$date != '2017-05-05' & newdf$date <=
'2017-05-28')

# boxplot
file_name <- 'wkdays'
sub_title <- "box plot of dwell time (minutes), weekdays, no events or holidays"
labels[4] <- sub_title # sub-title label
ggplot(newdf, aes(cluster, dur_mins)) +
 geom_boxplot(outlier.colour = "#bb0000", outlier.shape = 1, outlier.alpha=0.6) +
 labels + scales
ggsave(filename=paste0('Rplot_cluster_boxplot_osmay17_', file_name, '.png'), plot=last_plot(),
device='png',
  scale=0.8, width=20, height=10, units='cm', dpi=300, limitsize=TRUE)

# statistics
stats <- cluster_presence(newdf)

# weekends, excluding bank and school holidays, and events
# --------
newdf <- subset(subdf, subdf$wday>=6)
newdf <- subset(newdf, newdf$date != '2017-05-14')

# statistics
stats <- cluster_presence(newdf)

# football weekdays
# -----------------
newdf <- subset(subdf, subdf$date == '2017-05-05')
stats <- cluster_presence(newdf)

# football weekends
```

```
# -----------------
newdf <- subset(subdf, subdf$date == '2017-05-14')
stats <- cluster_presence(newdf)

# Studying the daily formation (or lack of) clusters for clusters of interest
# ============================================================================
# want to see what days there are clusters at the Copper Box, Podium and Stadium - can events be
detected
subdf <- analysis
subdf <- subset(subdf, subdf$cluster != 'O2_OO' & subdf$cluster != '16_OO')
n_distinct(subdf$pseudo_id)

subdf <- as.data.frame(subdf %>%
       group_by(date, wday, cluster) %>%
       summarise(devices = n_distinct(pseudo_id), trips = n_distinct(pseudo_trip), visits =
n_distinct(pseudo_stage)))

wkdays <- subset(subdf, subdf$wday <= 5)
wkends <- subset(subdf, subdf$wday >= 6)
layer_wkdays <- geom_bar(aes(x=date, y=devices, group=cluster, fill=cluster), data=wkdays,
colour='#232323', size=0.1, alpha=0.90, stat='identity')
layer_wkends <- geom_bar(aes(x=date, y=devices, group=cluster, fill=cluster), data=wkends,
colour='#232323', size=0.1, alpha=0.65 ,stat='identity')
ggplot() + layer_wkdays + layer_wkends +
 scale_x_date(date_breaks = "2 days", date_labels = "%d-%b", expand = c(0.01,0.01)) +
 scale_y_continuous(expand=c(0, 0)) +
 scale_fill_brewer(palette="Set1") +
 labs(subtitle = "Count within DBSCAN cluster per day, OpenSignal, May 2017",
  x = "date (weekends highlighted)")
ggsave(filename=paste0('Rplot_cluster_distr_osmay17.png'), plot=last_plot(), device='png',
  scale=1.2, width=20, height=18, units='cm', dpi=300, limitsize=TRUE)

# snipping out clusters for Westfield and Stratford
wkdays <- subset(subdf, subdf$wday <= 5 & subdf$cluster > '01_03')
wkends <- subset(subdf, subdf$wday >= 6 & subdf$cluster > '01_03')
layer_wkdays <- geom_bar(aes(x=date, y=devices, group=cluster, fill=cluster), data=wkdays,
colour='#232323', size=0.1, alpha=0.90, stat='identity')
layer_wkends <- geom_bar(aes(x=date, y=devices, group=cluster, fill=cluster), data=wkends,
colour='#232323', size=0.1, alpha=0.60 ,stat='identity')
ggplot() + layer_wkdays + layer_wkends +
 scale_x_date(date_breaks = "2 days", date_labels = "%d-%b", expand = c(0.01,0.01)) +
 scale_y_continuous(expand=c(0, 0)) +
 scale_fill_brewer(palette="Set1") +
 labs(subtitle = "Count within DBSCAN cluster per day, OpenSignal, May 2017",
  x = "date (weekends highlighted)",
  caption = "Westfield and Stratford clusters (01_01 to 01_03) removed")
ggsave(filename=paste0('Rplot_cluster_distr_v2_osmay17.png'), plot=last_plot(), device='png',
  scale=1.2, width=20, height=10, units='cm', dpi=300, limitsize=TRUE)

# Statistics for daily formation of clusters
# ==========================================
subdf <- analysis
subdf <- subset(subdf, subdf$cluster != 'O2_OO' & subdf$cluster != '16_OO')
subdf <- as.data.frame(subdf %>%
       group_by(date, wday, cluster) %>%
       summarise(devices = n_distinct(pseudo_id), trips = n_distinct(pseudo_trip), visits =
n_distinct(pseudo_stage)))

subdf$context <- 'wkdays'
subdf$context[subdf$wday >= 6 & subdf$date != '2015-05-14'] <- 'wkends'
subdf$context[subdf$date == '2017-05-01' | subdf$date >= '2017-05-29'] <- 'holday'
subdf$context[subdf$date == '2017-05-05' | subdf$date == '2017-05-14'] <- 'footbl'
subdf$context[subdf$date == '2017-05-03' | subdf$date == '2017-05-04' | subdf$date == '2017-05-
09'] <- 'w_cool'
subdf$context[subdf$date == '2017-05-24' | subdf$date == '2017-05-25' | subdf$date == '2017-05-
26'] <- 'w__hot'

# number of days clusters form across the month
stats <- subdf %>% group_by(cluster) %>%
 summarise(days = n_distinct(date))

# number of days that clusters form for each context
stats <- subdf %>% group_by(cluster, context) %>%
 summarise(days = n_distinct(date))
pivot <- dcast(stats, cluster ~ context)

# mean count of each cluster across the month
stats <- subdf %>% group_by(cluster) %>%
 summarise(mean_dev = mean(devices))

# mean count of devices for each context
stats <- subdf %>% group_by(cluster, context) %>%
 summarise(mean_dev = mean(devices))
pivot <- dcast(stats, cluster ~ context)

# mean count of daily devices across all clusters
stats <- subdf %>% group_by(context) %>%
 summarise(mean_dev = mean(devices))
```

```
# Analysing categorised dwell times within clusters
# ================================================

df <- analysis

# view as summary (visits)
df %>% group_by(cluster) %>% summarise(visits = n()) # total visits per cluster

stats <- df %>% group_by(cluster, dwell_cat) %>% summarise(visits = n())
pivot <- dcast(stats, cluster ~ dwell_cat)
pivot[is.na(pivot)] <- 0
visits <- as.data.frame(df %>% group_by(cluster) %>% summarise(visits = n()))
pivot <- merge(pivot, visits, by.x='cluster', by.y='cluster')

# adding percentages
pivot <- mutate(pivot, pct01 = round(d01mins/visits, digits=2), pct05 = round(d05mins/visits,
digits=2),
    pct20 = round(d20mins/visits, digits=2), pct90 = round(d90mins/visits, digits=2),
    pcthr = round(dhours/visits, digits=2), pctpm = round(perm/visits, digits=2))

# Visit stats
n_distinct(analysis$pseudo_id)
n_distinct(analysis$pseudo_trip)
n_distinct(analysis$pseudo_stage)
sub <- subset(analysis, analysis$duration == 15)
sub <- subset(analysis, analysis$dur_mins > 240)
df %>% group_by(cluster) %>% summarise(visits = n())

# weekdays
subdf <- subset(df, df$wday <= 5 & df$date != '2017-05-01' & df$date!= '2017-05-05' & df$date <=
'2017-05-28')
visits <- as.data.frame(subdf %>% group_by(cluster) %>% summarise(visits = n()))
stats <- subdf %>% group_by(cluster, dwell_cat) %>% summarise(count = n())
pivot <- dcast(stats, cluster ~ dwell_cat)
pivot[is.na(pivot)] <- 0
pivot <- merge(pivot, visits, by.x='cluster', by.y='cluster')

# stats - how many devices present at a cluster for longer than 1 minute, and longer than 5
minutes
stats <- df %>% group_by(dwell_cat) %>% summarise(devices = n_distinct(pseudo_id), trips =
n_distinct(pseudo_trip))

perms <- subset(df, dwell_cat == 'perm')
shorts <- subset(df, dwell_cat == '01mins')

# number of devices that are only present for a very long or short time
subdf <- subset(df, df$dwell_cat != 'perm')
length(setdiff(perms$pseudo_id, subdf$pseudo_id)) # number of devices only present for > 360 mins
(perms)

subdf <- subset(df, df$dwell_cat != '01mins')
length(setdiff(shorts$pseudo_id, subdf$pseudo_id)) # number of devices only present for < 1 min
(shorts)
```

B.6 Analyses for chapter six

The following code samples are specific to studies presented in chapter six of this thesis.

Term frequencies within sets of tweets

Figure 65 in chapter six shows the frequencies of terms used in tweets posted on four consecutive days, from Friday 3rd to Monday 6th June. In the following code sample, the tweets have already been loaded into memory as a dataframe called 'df' and filtered to within the QEOP Wi-Fi boundary. This code sample includes the basic process for cleaning the text before tokenisation.

```
# TEXT ANALYTICS
# ==============
library(tidyr)  # to melt dataframes for plotting
library(lubridate) # for extracting values from dates
library(quanteda)
library(stringi)
library(reshape2) # to convert matrix to dataframe
library(ggplot2)

# FUNCTIONS
# =========
clean_text <- function(df) {

 df$clean_text <- df$tweet_text

 # remove URLs
 df$clean_text <- gsub("http[^[:space:]]*", "", df$clean_text)

 # remove any digits, replace punctuation with a space
 df$clean_text <- gsub("\\d+", "", df$clean_text)
 df$clean_text <- gsub("[[:punct:]]+", " ", df$clean_text)

 # remove surplus whitespace
 df$clean_text <- gsub("\\s+", " ", df$clean_text)

 return(df)
}

# function to plot term frequency by frequency count in descending order as a bar chart
term_freq <- function(dfm) {

 df <- convert(dfm, to = "data.frame") # convert to dataframe
 df <- df %>% gather(term, freq) # melt for plotting
 df <- df[2:nrow(df), ] # drop first row, it contains old cols
 df$freq <- as.numeric(df$freq)
 df <- df[order(-df$freq), ] # sort by frequency, descending

 labels <- labs(subtitle=paste0('Frequency of term use in geotagged tweets, ', start_date),
     x='term',
     y='frequency',
     legend=NULL,
     caption=""
 )

 plot <- ggplot(data=df, aes(x = reorder(term, -freq), y = freq)) + geom_bar(stat="identity") +
 scale_y_discrete(expand=c(0,0)) +
 labels + t +
 theme(axis.text.x=element_blank(),
   axis.ticks.x=element_blank())

 ggsave(filename=paste0(plots_folder, img_title, 'termfreq_', start_date, '.png'),
plot=last_plot(), device='png',
   scale=0.8, width=20, height=8, units='cm', dpi=300, limitsize=TRUE)

 return(plot)
}

# setup stop words list - use standard English stopwords list + individual letters
myStopWords <- c(stopwords("english"))
myStopWords <- c(myStopWords, "b", "c", "d", "e", "f", "g", "h", "j", "k", "l", "m", "n", "o",
"p", "q", "r", "s", "t", "u", "v", "w", "x", "y", "z")

# MAIN PROG - TEXT ANALYTICS
# ==========================
# clean the tweets
# ----------------
```

```
df <- clean_text(df)
df <- df[order(df$date_time), ]
df_clean <- df

# produce visuals for a date of interest
# ------------------
# specify date range, then run rest of the code as is. Will output a wordcloud file

#1. Specify date/date range
start_date <- '2016-06-06'
end_date <- '2016-06-06'

# 2. run rest ‘as is’
df <- df_clean
df <- subset(df, (df$date >= start_date & df$date <= end_date))

# create corpus
myCorpus <- corpus(df$clean_text) # create corpus from tweets
docvars(myCorpus, "tag") <- df$date # add feature as tag (column from df to group tweets by)

# check how it looks
summary(myCorpus, n=5)
table(docvars(myCorpus, "tag"))

# create a document frequency matrix,
mydfm <- dfm(myCorpus, groups="tag", remove=myStopWords, stem=FALSE)
topfeatures(mydfm, 30) # top words

#plot as wordcloud
textplot_wordcloud(mydfm, max_words = 20) # wordcloud
dev.copy(jpeg, paste0(plots_folder, 'Rplot_wordcloud_', start_date, '.jpg'))
dev.off()

# plot as a bar chart
term_freq(mydfm)
```

Scoring emotion using the Regressive Imagery Dictionary (RID)

The RID is available as a text file to download from Provalis Research:

- https://provalisresearch.com/products/content-analysis-software/wordstat-dictionary/regressive-imagery-dictionary/

It can then be loaded in R as a dictionary to perform text analysis. The following code sample shows setting up the environment, retrieving the tweets from MySQL, preparing the RID, setting up a stopwords list, cleaning up the text and then running the text analysis steps. Each emotion category of interest is scored separately for analysis. The results were then written out to a text file and loaded into Excel for visualisation (this step could have been included in R using ggplot2).

```
# set-up environment
# ------------------
library(quanteda)
library(stringi)
library(RMySQL)
library(reshape2) # to convert matrix to dataframe
library(lubridate) # for extracting values from dates

# retrive twitter texts from DB
db = dbConnect(MySQL(), user='root', password='', dbname='2016', host='localhost')

query = dbSendQuery(db, "SELECT * FROM qeop_twitter")
data = fetch(query, n=-1)

# create time columns for subsetting
data$timestamp <- as.POSIXct(data$timestamp, format='%Y-%m-%d %H:%M:%S')
data$month = month(data$timestamp)
data$day = day(data$timestamp)
data$hour = hour(data$timestamp)
data$wkday = weekdays(data$timestamp)
data$date = date(data$timestamp)


# setup RID dictionary (regressive imagery dictionary)
RIDdict <- dictionary(file = "RID.CAT", format = "wordstat")
```

```
# removing stopwords and left over letters (eg where I'm became I m) + unuseful popular words
myStopWords <- c(stopwords("english"), "b", "c", "d", "e", "f", "g", "h", "j", "k", "l", "m",
"n", "o", "p", "q", "r", "s", "t", "u", "v", "w", "x", "y", "z")

# + removing overly used words that just identify the location plus remaining special characters
myStopWords <- c(myStopWords, "amp", "@", "greater", "park", "london", "stratford", "queen",
"elizabeth", "olympicpark", "olympic", "day", "uk")

# + removing more location and ambiguous words that may skew sentiment analysis
myStopWords <- c(myStopWords, "westfield", "velodrome", "velopark", "hackney", "wick",
"hackneywick", "lee", "valley",
     "leevalleyvp", "east", "ldnoverground", "bow", "bowchurch", "west", "ham", "westham",
"whufc",
     "stadium", "train", "aquatics", "richmond", "centre", "station", "noordinarypark", "newham",
     "clapham", "junction", "today", "copper", "box", "copperbox", "amorbit", "orbit", "tapeast")

# prep data before converting to corpus (pre-parsed in 3rd stage proc)
# -------------------------------------
# replace remaining punctuation with a space (split words)
data$parsed_text <- gsub("[[:punct:]]+", " ", data$parsed_text)

# remove surplus whitespace
data$parsed_text <- gsub("\\s+", " ", data$parsed_text)

# explore use of dictionaries to detect emotions
# -------------------------------------------------
# create corpus
myCorpus <- corpus(data$parsed_text)

# add date (day) as tag
docvars(myCorpus, "tag") <- data$date

# check how it looks
summary(myCorpus, n=5)
table(docvars(myCorpus, "tag"))

# create a document frequency matrix
mydfm <- dfm(myCorpus, groups="tag", ignoredFeatures=myStopWords, stem=TRUE)

topfeatures(mydfm, 30) # top words

plot(mydfm, max.words = 200)

# normalise by document length (means analysis values will be returned between 0 and 1)
dfmRel <- weight(mydfm, "relFreq")

# apply RID dictionary to analyse text
dfmRID <- applyDictionary(dfmRel, RIDdict)

# convert matrix to data frame
df <- as.data.frame(docnames(dfmRID))

# add number of observations the sentiment analysis is based on for each tag
df$num_twts <- table(docvars(myCorpus, "tag"))

# score emotions
affection <- as.vector(dfmRID[, "EMOTIONS.AFFECTION._"])
df$affection <- affection

anger <- as.vector(dfmRID[, "EMOTIONS.AGGRESSION._"])
df$anger <- anger

anxiety <- as.vector(dfmRID[, "EMOTIONS.ANXIETY._"])
df$anxiety <- anxiety

expressive <- as.vector(dfmRID[, "EMOTIONS.EXPRESSIVE_BEH._"])
df$expressive <- expressive

glory <- as.vector(dfmRID[, "EMOTIONS.GLORY._"])
df$glory <- glory

positive <- as.vector(dfmRID[, "EMOTIONS.POSITIVE_AFFECT._"])
df$positive <- positive

sadness <- as.vector(dfmRID[, "EMOTIONS.SADNESS._"])
df$sadness <- sadness

social <- as.vector(dfmRID[, "SECONDARY.SOCIAL_BEHAVIOR._"])
df$social <- social

# write results to file
write.table(df, 'Twitter_Sentiments_ByDate.csv', sep=",", row.names=TRUE, col.names=TRUE)
```

Note: SQL account password has been removed.

Inferring and filtering to location based on text content

The following process was applied to tweets retrieved that matched a keyword query (called ‘search tweets’ in notes). The process is described in chapter six, section 6.3.

Initially saved as weekly comma-delimited files, files containing tweets were then uploaded to a SQL database. A copy of the database was taken to perform the filtering process. The process was completed using a SQL script. The following script was applied to tweets retrieved weekly from May to August 2016

```
/* 1. CREATE TABLE AND COPY TWEETS ACROSS (search tweets only, not geotagged scrapes) */
CREATE TABLE `phd`.`twt_sch_201605` LIKE `phd`.`raw_twt_201605`;

INSERT INTO twt_sch_201605
SELECT * FROM raw_twt_201605 WHERE geotag LIKE 'N';

/* KEEP IF GEOTAGS WITHIN RANGE (CHANGE GEOTAG FROM N TO Y TO AVOID GETTING CAUGHT IN FUTURE
FILTERS HERE */
UPDATE `phd`.`twt_sch_201605` SET geotag = 'Y' WHERE
lat >= 51.5292 AND lat <= 51.5541 AND lng >= -0.0291 AND lng <= 0.0041;

/* DROP THE OBVIOUS STUFF FIRST */
delete from `phd`.`twt_sch_201605`
where
geotag LIKE 'N' AND
tweet_text LIKE 'RT%'; # drop any tweets beginning with RT

/* REMOVE ANYTHING THAT MAY BE REFERRING TO THE PAST OR FUTURE OR OTHER LOCATIONS OR CONTAIN
CONTENT UNLIKELY TO ASSOCIATE INTERACTIONS WITH */
DELETE FROM `phd`.`twt_sch_201605`
WHERE
geotag LIKE 'N' AND (
(tweet_text REGEXP 'jan|feb|mar|apr|may|jun|jul|aug|sep|oct|nov|dec' AND
tweet_text REGEXP
'1|2|3|4|5|6|7|8|9|10|11|12|13|14|15|16|17|18|19|20|21|22|23|24|25|26|27|28|29|30|31') OR
tweet_text REGEXP 'tomorrow|yester|month|week|year|forward|ahead' OR tweet_text LIKE '%other
night%' OR
tweet_text LIKE '%last%' AND (tweet_text LIKE '% week%' OR tweet_text LIKE '%night%' OR
tweet_text LIKE '%month%' OR tweet_text LIKE '%year%') OR
(tweet_text REGEXP 'monday|tuesday|wednesday|thursday|friday|saturday|sunday' AND
(tweet_text LIKE '%this %' OR tweet_text LIKE '%last %' OR tweet_text LIKE '%next %')) OR
(tweet_text REGEXP 'playing|going' AND tweet_text LIKE '% on %') OR tweet_text REGEXP
'ticket|confirm|advance' OR
tweet_text LIKE '%coming to%' OR tweet_text REGEXP 'reminder' OR tweet_text LIKE '%last week%' OR
(tweet_text LIKE '%can\'t%' OR tweet_text LIKE '%cannot%' OR tweet_text LIKE '%can not%') AND
tweet_text LIKE '%wait%' AND tweet_text NOT LIKE '% back%' OR
tweet_text LIKE '%London 2012%' OR tweet_text LIKE '% to host %' OR tweet_text LIKE '% will host
%' OR
tweet_text REGEXP 'real-estate|property|apartment|contract|assistant|available|book' OR
tweet_text LIKE '% rent%' OR tweet_text LIKE '%shepherd%' /* Shepherd's Bush westfield */ OR
tweet_text LIKE '%stratford%' AND tweet_text LIKE '%avon%' OR tweet_text LIKE '%role %' OR
tweet_text LIKE '%job%' OR tweet_text LIKE '% sale%' OR
tweet_text LIKE 'Aspers %' OR tweet_text LIKE '% aspers %' OR
tweet_text LIKE '%reader %' OR tweet_text LIKE '%read %' OR tweet_text LIKE '% read.%' OR
tweet_text LIKE '%Periscope%' OR tweet_text LIKE '% open%' OR
(tweet_text LIKE '% orbit%' AND tweet_text NOT LIKE '%Arcel%') OR # to get rid of tweets
referring to orbits, like the ISS orbit around the planet
tweet_text LIKE '%event %' OR tweet_text LIKE '%HackneyWick%' OR tweet_text LIKE '%Hackney Wick%'
/* Dropping HackneyWick from scraping - too many unrelated tweets */
);

/* NOW USE SCRAPER FIELD TO KEEP TRACK OF TWEETS ESTIMATED TO BE RELEVANT */
UPDATE `phd`.`twt_sch_201605` SET scraper = 'N';

/* keep if geotag value has been set to Y */
UPDATE `phd`.`twt_sch_201605` SET scraper = 'Y' WHERE
geotag LIKE 'Y';

/* update scraper to 'Y' for all tweets with some sign of an interaction
 ------------------------------------------------------------------------------ */
UPDATE `phd`.`twt_sch_201605` SET scraper = 'Y' WHERE
geotag LIKE 'Y' OR /* keeping all geotagged tweets */
tweet_text REGEXP
'today|tonight|morning|afternoon|evening|before|arriv|watching|waiting|exploring|walk|hanging|cau
ght|late|queue|route|spotted|finished|escape|early' OR
tweet_text LIKE '%I am %' OR tweet_text LIKE '%I\'m %' OR tweet_text LIKE '%I was %' OR
tweet_text LIKE '%we are%' OR tweet_text LIKE '%we\'re%' OR
tweet_text LIKE '%back at%' OR tweet_text LIKE '%back in %' OR tweet_text LIKE '% day at %' OR
tweet_text LIKE '% day in %' OR
tweet_text LIKE '%on my way%' OR tweet_text LIKE '%on our way%' OR
```

```
tweet_text LIKE '%ready %' OR tweet_text LIKE '%made it%' OR tweet_text LIKE '%trip to%' OR
tweet_text LIKE '%@ %' OR
tweet_text LIKE '%at %' AND tweet_text LIKE '%meeting%' OR tweet_text LIKE 'at the %' OR
tweet_text LIKE 'and %' OR
(tweet_text REGEXP 'Westfield|Stratford' AND (tweet_text LIKE '% in %' OR tweet_text LIKE '% at
%')) OR
tweet_text LIKE '% view %' OR tweet_text LIKE '% now %' OR
tweet_text REGEXP
'lunch|dinner|shop|eating|drink|coffee|table|seat|sitting|evacuate|alarm|police' OR
tweet_text REGEXP
'best|ugl|hate|lovely|packed|busy|service|selfie|atmosphere|weather|outside|changed' OR
tweet_text LIKE '%fab %' OR tweet_text LIKE '%fabulous%'
;

select * from twt_sch_201605 where scraper LIKE 'Y';
select * from twt_sch_201605 where scraper NOT LIKE 'Y';

/* DELETE RECORDS THAT HAVE NOT BEEN IDENTIFIED AS POSSIBLE INTERACTION */
DELETE FROM `phd`.`twt_sch_201605` WHERE scraper NOT LIKE 'Y';

/* reset geotag to N even if have coordinates, to differentiate keyword scrapes from geotagged
scrapes */
UPDATE `phd`.`twt_sch_201605` SET geotag = 'N';
```

Scoring similarity between documents

Similarity between documents was calculated using the cosine similarity measure available within the Quanteda package in R. The data is first organised as a corpus and grouped in documents. The similarity between pairs of documents is then scored, producing a matrix. In the example below, the corpus is a single date and the data is grouped by hour into documents.

```
# Another look at similarity (March 2017)
# note: Quanteda package updated. Some older code may have issues. This script has been updated

# Plucking a date and then examining hourly word use

# set-up environment
# ------------------
library(quanteda)
library(stringi)
library(RMySQL)
library(reshape2) # to convert matrix to dataframe
library(lubridate) # for extracting values from dates
library(dplyr)

# removing stopwords and left over letters (eg where I'm became I m)
myStopWords <- c(stopwords("english"), "b", "c", "d", "e", "f", "g", "h", "j", "k", "l", "m",
"n", "o", "p", "q",
     "r", "s", "t", "u", "v", "w", "x", "y", "z")

# + removing overly used words that may be misleading due to ambiguity plus remaining special
characters
myStopWords <- c(myStopWords, "amp", "@", "greater", "park", "london", "stratford", "queen",
"elizabeth", "olympicpark", "olympic",
     "queenelizabetholympicpark", "qeop", "uk")

# retrive twitter texts from DB
db = dbConnect(MySQL(), user='root', password='', dbname='2016', host='localhost')

query = dbSendQuery(db, "SELECT * FROM qeop_twitter")
data = fetch(query, n=-1)

# create time columns for subsetting
data$timestamp <- as.POSIXct(data$timestamp, format='%Y-%m-%d %H:%M:%S')
data$month = month(data$timestamp)
data$day = day(data$timestamp)
data$hour = hour(data$timestamp)
data$wkday = weekdays(data$timestamp)
data$date = date(data$timestamp)

# prep data before converting to corpus (pre-parsed in 3rd stage proc)
# ------------------------------------
# replace digits and remaining punctuation with a space (split words)
data$parsed_text <- gsub("\\d+", " ", data$parsed_text)
data$parsed_text <- gsub("[[:punct:]]+", " ", data$parsed_text)

# remove surplus whitespace
data$parsed_text <- gsub("\\s+", " ", data$parsed_text)

# fix known word issues
```

```
data$parsed_text <- gsub("ac dc", "acdc", data$parsed_text)
data$parsed_text <- gsub("west ham", "westham", data$parsed_text)
data$parsed_text <- gsub("whufc", "westham", data$parsed_text)

# *************************
# MAIN PROGRAM
# *************************

# take day as subset
# ==================
day_date <- '2016-07-23'

# simiarity by hour
# =================
mydata <- subset(data, date == day_date)

# create corpus
myCorpus <- corpus(mydata$parsed_text)

# add feature to group into documents as tag
docvars(myCorpus, "tag") <- mydata$hour

# create a document frequency matrix
mydfm <- dfm(myCorpus, groups="tag", remove=myStopWords, stem=FALSE)

topfeatures(mydfm, 30) # top words

# word similiarity between documents
# ----------------------------------
compare_words <- similarity(mydfm, docnames(mydfm), method="cosine", sorted="FALSE",
margin="document")

# write results to file
output <- paste0('similarity_', day_date, '.csv')
write.table(compare_words, output , sep=",", row.names=TRUE, col.names=TRUE)
```

Note: SQL password has been removed.

Contextual vocabulary analysis

The development of a contextual vocabulary occurred in Python using NLTK and network analysis packages. This was to enable the model to be implemented in real-time. The real-time model has not been included in the thesis.

The code sample below encompasses all steps outlined in chapter six: n-gram analysis, topic modelling with LDA and topic modelling with network analysis and community detection.

Set-up environment and load/prep data

```
# SET-UP
# ======
# plot charts in notebook (comment out when saving to file)
%matplotlib inline

import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import os, re, string, csv
import random
from datetime import datetime
import pymysql.cursors
import collections

#import nltk
from nltk.corpus.reader.plaintext import PlaintextCorpusReader
from nltk.collocations import *

import itertools as it
import networkx as nx
from gensim import corpora, models

from pyf_regexp import *
from pyf_nlp1 import *

pd.options.display.float_format = '{:40,.4f}'.format # specifies default number format to 4
decimal places

# RETRIEVE AND PREP DATA
```

```
# ======================
# Connect to MySQL and get data
# ----------------------------
connection = pymysql.connect(host='localhost',
        user='root',
        port=3306,
        password='',
        db='phd',
        charset='utf8mb4',
        cursorclass=pymysql.cursors.DictCursor)

print(connection)

query = 'select * from raw_twt_201609 where date_time >= "2016-10-22 04:00:00" AND date_time <=
"2016-10-22 23:59:59"'

df = pd.read_sql(query, connection)

# parse the tweets
# ----------------
# slice out tweet text as list
origtext = df['tweet_text']

# extract metadata from list
htags = extract_meta(origtext,'htag') # all hashtags
users = extract_meta(origtext, 'user') # all user mentions

# clean up text for tokenisation
parsed = text_parse(origtext)
parsed = punc_parse(parsed)

htags = punc_parse(htags)
users = punc_parse(users)

# extract user mentions as metadata

# append back to dataframe
df['clean_text'] = parsed
df['hashtags'] = htags
df['mentions'] = users

# drop duplicates
# ---------------
# brute force removal of potential bot content
# will remove records that are identical in the clean_text field (i.e. ignore URLs)
# will keep the earliest one (lowest status_id)
keep_one = df.groupby(['clean_text']).status_id.transform(min)
df = df[df.status_id == keep_one]
```

Note: SQL password removed.

Content Analysis

```
texts = df['clean_text'].astype('str') # slice out parsed texts as a list of strings

# Tokenise and remove noise
# -------------------------
# define default noise list
stopDef = 'nltk'

# define domain-specific additions
stopTerms = ['nan']

tokens = text_to_words(texts)  # word break the tweets (will be list of tweets, each tweet is
list of words)
tokens = remove_noise(tokens, stopDef) # then remove noise words
tokens = custom_noise(tokens, stopTerms) # then remove noise words

# convert tokens to a NLTK text
text = words_to_text(tokens) # will be single document of words, won't define individual tweets

# TERM FREQUENCY
# --------------
fdist = nltk.FreqDist(text)
print(fdist)
print("")
n = 20
print(str(n) + " most common terms:\n")
print(fdist.most_common(n)) # most frequently occurring
fdist.plot(n, cumulative=True)

# most frequently occurring longer words
min_length = 5
min_freq = 10
```

```
vocab = set(text)
long_words = [w for w in vocab if len(w) >= min_length and fdist[w] >= min_freq]
print("words with at least " + str(min_length) + " chars occurring at least " + str(min_freq) + "
times:")
sorted(long_words)

print("least common terms:")
uncommon = [w for w in vocab if len(w) < 4 and fdist[w] < 2]
sorted(uncommon)
# consideration: remove high freq outliers + drop low freq?
```

Collocation and n-gram analysis

```
# Collocation
# -----------
n = 20
print("top " + str(n) + " word pairs:\n")
text.collocations(num=n)

# bi-grams
# --------
bi_freq = 10 # drop n-grams that occurr fewer times than this in dataset
n = 20   # max number to display on page

bigram_measures = nltk.collocations.BigramAssocMeasures()
finder = BigramCollocationFinder.from_words(text)
finder.apply_freq_filter(bi_freq) # ignore n-grams that occur less than bi_freq
pmi_bigrams = finder.score_ngrams(bigram_measures.pmi) # score n-grams using PMI
if n > len(pmi_bigrams):
 n = len(pmi_bigrams)
i = 0
print("highest scoring bi-grams: (PMI)\n")
while i < n:
 print(pmi_bigrams[i])
 i +=1

print("\nmost frequently occurring bi-grams:")
sorted(finder.ngram_fd.items(), key=lambda t: (-t[1], t[0]))[:n] # display frequencies

# note: score is Pointwise Mutual Information - balances frequency with likelihood of word pair

# tri-grams
# ---------
trigram_measures = nltk.collocations.TrigramAssocMeasures()
tri_finder = TrigramCollocationFinder.from_words(text)

tri_freq = 5 # drop n-grams that occurr fewer times than this in dataset

tri_finder.apply_freq_filter(tri_freq) # ignore n-grams that occure less than n times
pmi_trigrams = tri_finder.score_ngrams(trigram_measures.pmi) # score n-grams using PMI

n = 20 # max number to display on page

if n > len(pmi_trigrams):
 n = len(pmi_trigrams)
i = 0
print("highest scoring tri-grams: (PMI)\n")
while i < n:
 print(pmi_trigrams[i])
 i +=1

print("\nmost frequently occurring tri-grams:")
sorted(tri_finder.ngram_fd.items(), key=lambda t: (-t[1], t[0]))[:n] # display frequencies
```

Topic modelling using LDA

```
# TOPIC MODELLING
# ---------------
# create a dictionary using tokens extracted from document/dataset of interest
dictionary = corpora.Dictionary(tokens)

# use dicionary to construct a document-term matrix
# (converts dictionary into a bag-of-words, tuple of term id and term frequency)
dtm = [dictionary.doc2bow(word) for word in tokens]

# create LDA model
numTopics = 3
numPasses = 20

ldamodel = models.ldamodel.LdaModel(dtm, num_topics=numTopics, id2word=dictionary,
passes=numPasses)
```

```
numWords = 10 # number of terms most probable to appear in topic
print(ldamodel.print_topics(num_topics=numTopics, num_words=numWords))

# REFINED TOPIC MODELLING
# -----------------------
# model based on nouns and adjectives only, ignore all other words

# create a dictionary using tokens extracted from document/dataset of interest
dictionary = corpora.Dictionary(tokens)

# 1. tag the words based on their type (part-of-speech - noun, adjective etc.)
tagged_tokens = [nltk.pos_tag(token) for token in tokens]

# 2. create dataset that consists only of the nouns and adjectives
noun_phrases = [[token for token, tag in tokens if re.match(r'NN*|JJ*', tag)]
    for tokens in tagged_tokens]

# use dicionary to construct a document-term matrix
# (converts dictionary into a bag-of-words, tuple of term id and term frequency)
dtm = [dictionary.doc2bow(word) for word in noun_phrases]
print("ready")

# create LDA model
numTopics = 3
numPasses = 20

ldamodel = models.ldamodel.LdaModel(dtm, num_topics=numTopics, id2word=dictionary,
passes=numPasses)

numWords = 10 # number of terms most probable to appear in topic
print(ldamodel.print_topics(num_topics=numTopics, num_words=numWords))

# expanded range of topics
numTopics = 5
numPasses = 20

ldamodel = models.ldamodel.LdaModel(dtm, num_topics=numTopics, id2word=dictionary,
passes=numPasses)

numWords = 10 # number of terms most probable to appear in topic
print(ldamodel.print_topics(num_topics=numTopics, num_words=numWords))
```

Network and LDA analysis of hashtags

```
hashtags = df['hashtags'].astype('str') # slice out hashtags as strings

# create word list of hashtags for each tweet
tokens = text_to_words(hashtags)  # word break the tweets (will be list of tweets, each tweet is
list of words)

# kill noise
stopWords = ['nan', ';']
tokens = custom_noise(tokens, stopWords)

# convert tokens to a NLTK text
text = words_to_text(tokens) # will be a single doc of words, won't define individual tweets

# TERM FREQUENCY
# --------------
fdist = nltk.FreqDist(text)
print(fdist)
print("")
n = 20
print(str(n) + " most common hashtags:")
print(fdist.most_common(n)) # most frequently occurring
fdist.plot(n, cumulative=True)

# NETWORK ANALYSIS
# ----------------
# select type of centrality ('degree' or 'betweenness' (not doing 'closeness')
#measure = 'degree'
#measure = 'betweenness'
#measure = 'closeness'
measure = 'eigenvector'

print("measurements for " + measure + " centrality\n----\n")
# construct network (links based on co-occurrence of terms within tweets)
edgelist = [edge for token in tokens for edge in it.combinations(token, 2)]
print("number of nodes " + str(len(tokens)))
print("number of edges " + str(len(edgelist)))
print("")

# construct graph using tuples from edgelist
```

```
G = nx.Graph(edgelist)

# calcuate centrality measure
if measure == 'degree':
 index = nx.degree_centrality(G)
if measure == 'betweenness':
 index = nx.betweenness_centrality(G)
if measure == 'closeness':
 index = nx.closeness_centrality(G)
if measure == 'eigenvector':
 index = nx.eigenvector_centrality(G)

# sort highest first
sorted_index = sorted(index.items(), key=lambda x:x[1], reverse=True)

# Top noun phrases by degree centrality (role as hub):
for word, centr in sorted_index[:40]:
 print(word, centr)

# viewing/extracting the network
outputfile = "Hashtags_22ndOct_FullNet" + measure

# export results as an edge list to analyse in R
networkfile = outputfile + ".txt"
nx.write_edgelist(G, networkfile)

# export network results to view in Gephi
networkfile = outputfile + ".gexf"
nx.write_gexf(G, networkfile)

# network model output
%pylab inline
%config InlineBackend.figure_format = 'png'
plt.rc('figure', figsize=(12, 7))
G.remove_nodes_from([n for n in index if index[n] == .0])
node_size = [index[n]*10000 for n in G]
pos = nx.spring_layout(G)
nx.draw_networkx(G, pos, node_size=node_size, edge_color='y', alpha=.4, linewidths=0)

# Testing LDA on the full set of hashtags, to compare approaches
# --------------------------------------
# use dicionary to construct a document-term matrix
# (converts dictionary into a bag-of-words, tuple of term id and term frequency)
dtm = [dictionary.doc2bow(word) for word in tokens]

# create LDA model
numTopics = 3
numPasses = 20

ldamodel = models.ldamodel.LdaModel(dtm, num_topics=numTopics, id2word=dictionary,
passes=numPasses)

numWords = 10 # number of terms most probable to appear in topic
print(ldamodel.print_topics(num_topics=numTopics, num_words=numWords))
```

Note: network output for viewing and visualizing in Gephi.

Community detection within hashtag network

Note: this stage was originally developed and tested in R (testing different community detection algorithms). Ported to Python to develop a single end-to-end model.

```
# COMMUNITY DETECTION (using Louvain method)
# -------------------
import community

# remove nodes with low centrality score
min_centrality = 0
G.remove_nodes_from([n for n in index if index[n] <= min_centrality])

# partition network into communities (community = Louvine method)
partition = community.best_partition(G)

topics = []
count = 0
h_min = 4 # minimum number of hashtag terms (vocab) required to form a topic

for com in set(partition.values()) :
 count = count + 1
 list_nodes = [nodes for nodes in partition.keys() if partition[nodes] == com]

 # create topic if has vocabulary of h_min to h_max hashtag terms
```

```
 if (len(list_nodes) >= h_min):
  topics.append(list_nodes)


# for each topic, find the node with the highest degree to use as label
# -----------------------------------------------------------------------
degree_index = nx.degree_centrality(G) # create index of nodes using degree as centrality measure
sorted_index = sorted(degree_index.items(), key=lambda x:x[1], reverse=True) # sort by degree
measure (descending)

sorted_topics = []
i = 0
while i < len(topics):
 ordered_topic = []
 ordered_topic = [x[0] for x in sorted_index if x[0] in set(topics[i])] # loop over sorted index
and match to set
 sorted_topics.append(ordered_topic)
 i += 1

# note: using x[0] because just want the label from the index, not the degree measure (in
position x[1])
# print sorted_index[1][0]

print("Number of nodes: " + str(len(tokens)))
print("Number of edges: " + str(len(edgelist)))
print("Number of communities partitioned: " + str(len(set(partition.values()))))
print("Topics detected: (vocabulary of " + str(h_min) + "+ hashtags) " + str(len(topics)) + "\n")


# create labels and strings of hashtags
labels = []
strings = []
i = 0
while (i < len(sorted_topics)):
 label = sorted_topics[i][0]
 string = ', '.join(sorted_topics[i])
 labels.append(label)
 strings.append(string)
 print(label + ": " + string + "\n")
 i += 1
```

B.7 Analyses for chapter seven

The analyses and visualisations produced in chapter seven were the culmination of techniques developed over the previous three chapters. A set of functions were developed to perform each stage of the contextual analysis: first, create a grid to profile the landscape and establish its ambient context, then to analyse the incident impact.

The code has the settings for the data source and each landscape pre-configured. However, it could be developed further to enable these parameters to be input at the start of the function, enabling it to be applied to any landscape with an available source of real-time data to analyse.

Modelling the socio-spatial dynamic of a landscape

The following code sets up a uniform grid aligned to LandScan coordinates to produce an active population estimate for ambient conditions (no abnormal events) and study socio-spatial dynamics.

```
# Using OpenSignal June 2017 dataset to establish a base ambient context for all locations
#
# All outputs to be saved to 'landscape_profile', prefixed with location (loc_label)
#
###################

# ======
# DEVICE
# ======
# settings that are device-specific
root_folder <- '<FOLDER>'
lab_title <- 4
lab_caption <- 5

# ========
# LOCATION
# ========
# specify location being analysed, will be used to prep specific datasets and parameters
# and will ensure all outputs are saved with this as the start on filenames
# note: also an 'all-london' label for london-wide plots (single optional prog after Setup 1)

loc_label <- 'londonbridge' # 'all-london', 'westminster' or 'londonbridge' or 'oxfordcircus'

# ========================================================
#
# SETUP PART 1: LIBRARIES, DEFAULTS AND GENERAL FUNCTIONS
#
# Run all of this before selecting location to analyse and visualise
# Only need to do this once. Is applicable to all locations
#
# ==============================

library(dplyr)
library(tidyr)
library(ggplot2)   # for plots (and theme)
library(lubridate)  # date/time calcs
library(rgdal)   # for readOGR and SpatialPointsDataFrame
library(OpenStreetMap) # for map background

# set point at which plot will switch from standard form
options(scipen=7)

# set coordinate referencing system (for changing CRS but not reprojecting)
crs_wgs84 <- "+init=epsg:4326" # lat/lng
crs_osm <- "+init=epsg:3857" # OSM projection

# Set coordinate systems for reprojecting
proj_wgs84 <- '+proj=longlat +datum=WGS84'
proj_osm <- '+proj=merc +a=6378137 +b=6378137 +lat_ts=0.0 +lon_0=0.0 +x_0=0.0 +y_0=0.0 +k=1.0
+units=m +nadgrids=@null +wktext +no_defs'

# ------------------
# GENERAL PARAMETERS
# ------------------
rhythm_start <- '2017-06-08'
rhythm_end <- '2017-06-28'
```

```
rhythm_range <- '8 to 28 June 2017'

# increments for generating grids of uniform cells (LandScan scale and pixel (1/16th Landscan
cell) scale)
lscan_increment <- 0.00833333 # extra decimal place
pixel_increment <- 0.00208333

map_size <- 16 # dimensions for saving maps as plots (is a square so same width and height)

# map labels for plots
labels_map <- labs(title=NULL,
     subtitle = 'OpenStreetMap',
     x="Background: map tiles Stamen Design, under CC BY 3.0;\nmap data by OpenStreetMap, under
ODbL. Projection: EPSG:3857",
     y=NULL,
     caption = NULL)

# map theme adjustments
t_map <- theme(axis.text = element_text(colour="white"),
    title = element_text(size=11),
    axis.ticks = element_line(colour="white"), # hiding axis data because is OSM number system,
not meaningful
    axis.title.x = element_text(size=10),
    legend.text = element_text(size=10)
)

chart_height <- 8 # dimensions for saving charts as plots (landscape rectangle)
chart_width <- 12

# chart theme adjustments
t <- theme_bw() +
 theme(panel.border=element_blank(), # removes border around chart area
  axis.text = element_text(size = 10, colour='#444444'),
  axis.ticks = element_line(colour='gray'))

# ------------------
# MAP PREP FUNCTIONS
# ------------------
# RETRIEVE OSM BACKGROUND MAP TILES (for map_type 'osm' or 'stamen')
map_background <- function(map_type='osm', map_design='osm', map_zoom=15) { # z = map zoom

 if (map_type == 'osm') {
 ul <- c(LatMax,LngMin)
 lr <- c(LatMin,LngMax)
 basemap <- openmap(ul,lr, zoom=map_zoom, type=map_design, mergeTiles=FALSE)
 }
 else {
 tile_server <- paste0("http://tile.stamen.com/", map_design, "/{z}/{x}/{y}.png") # Stamen Design
 basemap <- openmap(c(LatMax,LngMin),c(LatMin,LngMax), zoom=map_zoom, type=tile_server,
mergeTiles=FALSE)
 }

 return(basemap)
}

# SNIP DATA TO A BOUNDING BOX
# - requires a 'lat' and 'lng' column in data being snipped
snip_to_box <- function(df, LatMin, LatMax, LngMin, LngMax) {

 df <- subset(df, df$lat > LatMin & df$lat < LatMax)
 df <- subset(df, df$lng > LngMin & df$lng < LngMax)

 return (df)
}

# REPROJECT FROM LAT/LNG TO OSM COORDINATES
# - requires column names for lat and lng to be specified as 'y' and 'x'
# - a and b will be the names for the OSM coordinate columns (default 'osm_x' and 'osm_y')
# - appends the OSM coordinates as new columns to original dataframe
reproject <- function(df, x='lng', y='lat', a='osm_x', b='osm_y') {
 dataset_map_coords <- df[c(x, y)]
 dataset_map_data <- as.data.frame(df[ ,c(1)]) # doesn't matter which column, is dropped at end
 dataset_map <- SpatialPointsDataFrame(coords=dataset_map_coords, data=dataset_map_data)

 # set CRS and reproject to OSM for OSM number system
 dataset_map@proj4string # check first, should be NA - not yet been set
 proj4string(dataset_map) <- CRS(crs_wgs84) # set the current coordinates system
 dataset_map <- spTransform(dataset_map, CRS(proj_osm)) # reproject to OSM

 # convert back to dataframe with OSM cooordinates and then amend to original dataset
 newdf <- as.data.frame(dataset_map)
 names(newdf)[names(newdf) == x] <- a; names(newdf)[names(newdf) == y] <- b
 newdf <- newdf[, 2:3] # just want to keep the OSM coordinates to append back to dataset
(dropping data column)
 df <- cbind(df, newdf)
 return(df)
}
```

```
# ADD CELL IDS TO READINGS (based on which cell the reading falls within)
# - x and y vals start from SW corner of grid and increment up as integers (x1, x2 etc.)
add_cell_ids <- function(df) {

 # using find interval - will determine which interval the reading falls within for grid
 binxy <- data.frame(x=findInterval(df[,2], x_list), # lng <- note: LNG for X
      y=findInterval(df[,1], y_list)) # lat <-  LAT for Y
 df <- cbind(df, binxy)
 df <- mutate(df, cell=paste0('x', x, 'y', y)) # identifies south west corner of pixel

 return(df)
}

# CREATE GRID COORDINATES FOR MAP PLOTS
# - assign coordinates to x, y points of the grid so can plot on OSM
add_grid_coords <- function(df) {

 d <- df %>% group_by(cell, x, y) %>% summarise(dummy=n()) %>% ungroup()
 d <- d[, 1:3] # dropping dummy data, just wanted to provision the list of cells and x, y vals

 d$x <- as.numeric(d$x)
 d$y <- as.numeric(d$y)
 d <- mutate(d, xMin = x, xMax = x+1, yMin = y, yMax = y+1) # gets values for coords in grid

 # assign coordinates for grid cells (could probably do this more efficiently with lapply)
 rows <- length(x_list) - 1
 cols <- length(y_list) - 1
 for (i in 1:rows) {
 d$xMin[d$x == i] <- x_list[i]
 d$xMax[d$x == i] <- x_list[i+1]
 }
 for (i in 1:cols) {
 d$yMin[d$y == i] <- y_list[i]
 d$yMax[d$y == i] <- y_list[i+1]
 }

 # only want data that falls within the grid created.
 # potential for last x and y val to be slightly under/over max val of bounding box used to
create grid
 # due to creating cells by incrementing from SW corner (increment is a fraction, minor rounding
error)
 # drop cell IDs where x or y = 0 (below min val) or x or y > than row-1 and cols-1 (above max
val)
 d <- subset(d, d$x != 0 & d$y != 0 & d$x <= rows & d$y <= cols)

 # get OSM values for grid cells so can plot spatially (need centre and width, height)
 d <- mutate(d, ctrX = xMax-(xMax-xMin)/2, ctrY = yMax-(yMax-yMin)/2)
 d <- reproject(d, x='ctrX', y='ctrY', a='osm_x', b='osm_y')
 d <- reproject(d, x='xMin', y='yMin', a='min_x', b='min_y')
 d <- reproject(d, x='xMax', y='yMax', a='max_x', b='max_y')
 d <- mutate(d, width = max_x - min_x, height = max_y - min_y) # calculate width and height of
cells

 return(d)
}

# ------------------------
# DATA ANALYSIS FUNCTIONS
# ------------------------
# min-max normalisation of values
normalise <- function(x) {
 return ((x - min(x)) / (max(x) - min(x)))
}

# sort out date & time in datasets
setup_datetime <- function (df) {

 # - set data to local timezone - BST (will include adjusting for DST in March and October, means
9am is always 9am)
 # expects dataset to have a column 'date_time' that is a full date and timestamp
 df$date_time <- as.POSIXct(df$date_time, format = '%Y-%m-%d %H:%M:%S')
 df$date_time <- force_tz(df$date_time, tzone="GMT")
 df$date_bst <- with_tz(df$date_time, tzone="Europe/London")
 df$date <- date(df$date_bst)
 df$mth <- month(df$date_bst)
 df$day <- day(df$date_bst)
 df$hr <- hour(df$date_bst)
 df$wday <- ifelse(wday(df$date)==1,7,wday(df$date)-1) # numeric values

 return(df)
}

# ========================
# VISUALISATION FUNCTIONS
# -------------
# MAPS
```

```
# ---------

# plot data on OpenStreetMap map tile
map_plot <- function(df, map_type, map_bg) {

 # OSM map tile (background)
 map_layer <- autoplot(basemap)

 # data layers
 if (map_type == 'data_single') { # plot map layer + data points
 plot <- map_layer +
  geom_point(data=df, aes(x=osm_x, y=osm_y), alpha=0.4, colour='#666666')
 }

 else if (map_type == 'data_double') { # plot two sets of data points
 df1 <- subset(df, df$date_bst <= time_incident)
 df2 <- subset(df, df$date_bst > time_incident)

 plot <- map_layer +
  geom_point(data=df1, aes(x=osm_x, y=osm_y), alpha=0.6, shape=17, colour='#5e3c99') + # purple
triangle (RColorBrewer)
  geom_point(data=df2, aes(x=osm_x, y=osm_y), alpha=0.6, shape=16, colour='#e66101')  # orange
circle (ditto)
 }

 else if (map_type == 'data_grid') { # plotting aggregate data in grid, orange shading
(ColorBrewer)
 plot <- map_layer +
  geom_tile(data=df, aes(x=osm_x, y=osm_y, width=width, height=height, fill=count), alpha=0.6) +
  scale_fill_continuous(limits=c(0, 1), low="#7f2704", high="#fee6ce") # ColorBrewer orange hues
 }

 else if(map_type == 'tweets') { # plotting tweets, reducing alpha, increasing size (few data
points)
 plot <- map_layer +
  geom_point(data=df, aes(x=osm_x, y=osm_y), alpha=0.8, size=1.6, colour='#2166ac') # dark blue
 }

 else { # just plot map layer, no data (background)

 plot <- map_layer
 }

 # add grid layer if specified (want it on top of any data plots)
 if (map_bg == 'osm_lscan') { # LandScan only
 plot <- plot +
  geom_tile(data=grid_lscan, aes(x=osm_x, y=osm_y, width=width, height=height, fill = NULL),
alpha=0, colour="#dd0000", size=0.8)
 }

 if (map_bg == 'osm_pixel') { # Pixel grid with Landscan outline on the top
 plot <- plot +
  geom_tile(data=grid_pixel, aes(x=osm_x, y=osm_y, width=width, height=height, fill = NULL),
alpha=0, colour="#555555", size=0.8) +
  geom_tile(data=grid_lscan, aes(x=osm_x, y=osm_y, width=width, height=height, fill = NULL),
alpha=0, colour="#dd0000", size=0.8)
 }

 if (map_bg == 'osm_pixel_lscan') { # Pixel grid with Landscan outline on the top snipped to
LandScan grid
 plot <- plot +
  geom_tile(data=grid_pixel_lscan, aes(x=osm_x, y=osm_y, width=width, height=height, fill =
NULL), alpha=0, colour="#555555", size=0.8) +
  geom_tile(data=grid_lscan, aes(x=osm_x, y=osm_y, width=width, height=height, fill = NULL),
alpha=0, colour="#dd0000", size=0.8)
 }

 # add labels and theme, save plot (and return for visual in console)
 plot <- plot + labels_map + t_map
 plot_save <- paste0(loc_label, '_', plot_base, '_', map_data, '_', map_bg)

 ggsave(filename=paste0(plots_folder, plot_save,'.png'), plot=last_plot(), device='png',
   scale=1.0, width=map_size, height=map_size, units='cm', dpi=300, limitsize=TRUE)

 return(plot)
}

# ----------
# BAR CHARTS
# ----------
# plot Month of data (counts per date)
plot_dates <- function(df, chart_title, chart_name, chart_type, faceting='no') {

 labels <- labs(subtitle=paste0(chart_title, ""),
     x="date",
     y=NULL,
     caption='weekends highlighted'
 )
```

```
 if (chart_type == 'single') { # 'single' to highlight weekends
 weekends <- subset(df, df$wday >= 6)

 plot <- ggplot() +
  geom_bar(aes(y=count, x=date), data=df, stat='identity') +
  geom_bar(aes(y=count, x=date), data=weekends, fill='#999999', stat='identity')
 }
 else { # 'multi' if plotting by LandScan cell
 labels[lab_caption] <- ''
 plot <- ggplot() +
  geom_bar(aes(y=count, x=date, fill=lscan), data=df, stat='identity', position='dodge')
 }

 if (faceting == 'yes') { # if faceting by cell
 plot <- plot + facet_wrap(~cell)
 }

 plot <- plot +
 scale_x_discrete(expand = c(0,0)) +
 scale_fill_grey(start = 0.25, end = 0.75) +
 t + labels

 plot_save <- paste0(loc_label, '_', plot_base, '_', chart_name)

 ggsave(filename=paste0(plots_folder, plot_save,'.png'), plot=last_plot(), device='png',
   scale=1.0, width=chart_width, height=chart_height, units='cm', dpi=300, limitsize=TRUE)

 return(plot)
}

# plot Day of data (caounts per hour or per minute)
plot_time <- function(df, time_type, chart_type, chart_title, chart_name) { # can plot hours or
mins aggregate, faceting by cell ID

 # counts must be in column 'count', aggregation method (mins or hours) must be in column 'cat'
 # if grid-based, faceting by cell id

 labels <- labs(subtitle=paste0(chart_title, ""),
     x=time_type,
     y=NULL,
     caption=NULL
 )

 plot <- ggplot() +
 geom_bar(aes(y=count, x=cat), data=df, stat='identity', position='dodge')

 if (time_type == 'mins') {

 # add in vertical line for time of incident

 labels[lab_caption] <- 'dotted vertical line indicates time of incident'

 plot <- plot +
  geom_vline(xintercept = incident_time, linetype='dotted', colour='#b2182b', size=1.0)

 }

 if (chart_type == 'grid') { # facet plots if grid (plot per cell id)

 plot <- plot +
  facet_wrap(~cell) # will facet by cell ID
 }

 plot <- plot + t + labels

 plot_save <- paste0(plot_base, '_', chart_name)

 ggsave(filename=paste0(plots_folder, plot_save,'.png'), plot=last_plot(), device='png',
   scale=1.0, width=chart_width, height=chart_height, units='cm', dpi=300, limitsize=TRUE)

 return(plot)
}

# ----------
# BOX PLOTS
# ---------
boxplot_single <- function(df) {
 labels <- labs(subtitle=paste0(chart_title, ""),
     x=x_label,
     y=NULL,
     caption=NULL
 )

 plot <- ggplot(df, aes(x=cat, y=count)) +
 geom_boxplot(fill = '#cccccc') +
 labels + t
```

```
 plot_save <- paste0(loc_label, '_', plot_base, '_', chart_name)

 ggsave(filename=paste0(plots_folder, plot_save,'.png'), plot=last_plot(), device='png',
   scale=1.0, width=chart_width, height=chart_height, units='cm', dpi=300, limitsize=TRUE)

 return(plot)
}

boxplot_facet <- function(df, grouping='cell') { # facet by cell (landscan or pixel grid)

 labels <- labs(subtitle=paste0(chart_title, ""),
     x=NULL,
     y=NULL,
     caption=NULL
 )

 xvals <- c("Mon", "Tue", "Wed", "Thu", "Fri", "Sat", "Sun")

 layer_facet <- facet_grid(. ~cell)

 plot <- ggplot(df, aes(x=cat, y=count)) +
 geom_boxplot(fill = '#cccccc') +
 scale_x_discrete(breaks=c(1:7), labels=xvals) +
 layer_facet +
 labels + t


 plot_save <- paste0(loc_label, '_', plot_base, '_', chart_name)

 ggsave(filename=paste0(plots_folder, plot_save,'.png'), plot=last_plot(), device='png',
   scale=1.0, width=chart_width, height=chart_height, units='cm', dpi=300, limitsize=TRUE)

 return(plot) # so chart displays on screen as well
}

# -----------------------
# BARS AND LINES COMBINED
# -----------------------
# BARS FOR EACH INDIVIDUAL DATE, LINE FOR MEAN ACROSS DATES

# x-axis is 'cat', y-axis is 'count', will facet_wrap by cell
bars_and_line <- function(df=df, id=df, chart_name, chart_title, chart_height, max_val, plot_type
= 'rhythm') {

 labels <- labs(title=paste0(chart_title, ""), # using title instead of subtitle, need larger
font
     x=x_label,
     y=NULL,
     caption=NULL
 )

 # calculate the mean per hour across the date range
 meandf <- df %>% group_by(cell, hr) %>% summarise(count=mean(count, na.rm = T))
 meandf$cat <- as.factor(meandf$hr)
 meandf$group <- '1' # need to group for each layer in ggplot, so creating dummy value

 # drop readings before 5am (in case there's empty ones in deadzone, start plot from 5am)
 df <- subset(df, df$hr >= 5)
 meandf <- subset(meandf, meandf$hr >= 5)

 plot <- ggplot() +
 geom_bar(data=df, aes(x=cat, y=count, group=date), colour='#333333', size=0.05, fill='#999999',
alpha=0.8, stat='identity', position='dodge') +
 geom_line(data=meandf, aes(x=cat, y=count, group=group), colour='#b2182b', size=1.2)

 if (plot_type == 'incident') { # add a blue line for the day of the incident

 id <- subset(id, id$hr >= 5) # same as for other plots, snip to start from 5am

 id$group <- '1' # need to group for each layer in ggplot, so creating dummy value

 plot <- plot +
  geom_line(data=id, aes(x=cat, y=count, group=group), colour='#2166ac', linetype='longdash',
size=1.2)

 }

 plot <- plot +
 geom_vline(xintercept=5, linetype='dotted', size=1) + # dotted vertical lines for 9am and 5pm
(plotting from 5am)
 geom_vline(xintercept=13, linetype='dotted', size=1) +
 scale_y_continuous(limits=c(0, max_val), expand=c(0,0)) +
 facet_wrap(~cell) +
 labels + t

 plot_save <- paste0(loc_label, '_', plot_base, '_', chart_name)
```

```
 ggsave(filename=paste0(plots_folder, plot_save,'.png'), plot=last_plot(), device='png',
   scale=1.0, width=chart_width, height=chart_height, units='cm', dpi=300, limitsize=TRUE)

 return(plot) # so chart displays on screen as well
}

# ====================
# ANALYSIS PREPARATION (Loading base data file)
# --------------------
# applicable to all locations in study - using same dataset - OpenSignal June data

base_folder <- './' # for saving outputs
plots_folder <- paste0(base_folder, 'landscape_profile/')
data_file <- 'os_london_2017-06_extract.csv' # OpenSignal data to load

# load and prep OpenSignal Data
folder <- paste0(root_folder, 'data/Incidents/')
input_data <- paste0(folder, data_file)
df <- read.csv(input_data)

# - set data to local timezone (means no readings from 01:00 - 02:00 on 26 March due to DST
shift)
df <- setup_datetime(df)

df_source <- df

# SETUP 1 COMPLETE - LIBRARIES, GENERAL PARAMETERS AND FUNCTIONS

# ============================================================
#
# OPTIONAL 1 - plot full outputs from the dataset
#
# ******* ONLY RUN THIS WHEN LOC_LABEL == ALL-LONDON ***********
#
# Will do full outputs of June 2017 OpenSignal dataset for Greater London
# Including map retrieval and plots - will take a long time to run!
# Don't use this section for any other location labels
#
# ============================================================

loc_title <- 'Greater London' # default to start chart title

# CHECK COUNTS
paste0('Number of records in df_source: ', nrow(df_source))
paste0('Number of unique devices in df_source: ', n_distinct(df_source$device))

plot_base <- 'bars_dates_' # start of filename for image saves
chart_height <- 8
chart_width <- 20

# plot count of devices per date
df <- df_source
df <- df %>% group_by(date, wday) %>% summarise(count=n_distinct(device)) %>% ungroup()
df$date <- as.factor(day(df$date))

chart_title <- 'OpenSignal - Unique devices daily across London, June 2017'
chart_height <- 8
chart_width <- 20
chart_name <-'os_Jun17_devices' # to ensure unique filename for plot save
plot_dates(df, chart_title, chart_name, chart_type = 'single')

# SPATIAL SETTINGS FOR OSM MAP TILE AND LANDSCAN GRID (not doing Pixel at this scale)

# Map bounding box (0.0005 beyond min and max values in OpenSignal dataset)
LatMax <- 51.6915 # North
LatMin <- 51.2865 # South
LngMax <- 0.3345  # East
LngMin <- -0.5105 # West

# LandScan grid covering London LandScan dataset (with + 0.00001 on max values to ensure grid
draws within range)
xMin <- -0.5041667 # West
xMax <- 0.312501  # East
yMin <- 51.2875000 # South
yMax <- 51.687501 # North

# get map tile
map_type <- 'stamen' # 'osm' or 'stamen'
map_design <- 'terrain' # pick design for map_type
map_zoom <- 12 # zoom for OSM background (determines map detail)
basemap <- map_background(map_type, map_design, map_zoom)

# snip data to map tile
df <- df_source
df <- snip_to_box(df, LatMin, LatMax, LngMin, LngMax) # data within map tile
df <- reproject(df, x='lng', y='lat')
```

```
df_map <- df

# create LandScan grid data
df_grid <- snip_to_box(df_map, yMin, yMax, xMin, xMax)   # data within LandScan grid

lx <- seq(xMin, xMax, by=lscan_increment)
ly <- seq(yMin, yMax, by=lscan_increment)

# add grid coords to each reading and create results DF for each cell in grid
df <- df_grid
x_list <- lx # list of coordinates for points of the grid on each axis
y_list <- ly #
df_grid <- add_cell_ids(df)    # assign cell IDs to each reading in dataset
grid_lscan <- add_grid_coords(df_grid) # create results DF with each cell ID as a row with its
coordinates

# PLOTTING MAPS
plot_base <- 'map' # start of filename for image saves

map_type <- 'empty'
map_data <- 'nodata' # leave blank, is for filename, not needed unless plotting data

# Plot OSM background (no data)
labels_map[lab_title] <- paste0(loc_title, ' - Map Tile')
map_bg <- 'osm'
map_plot(df, map_type, map_bg)


# plot single date (too many to plot all at once - is just grey goo)
df <- df_map
df <- subset(df, df$date == '2017-06-01')

map_type <- 'data_single' # single layer of data
map_data <- 'osdata_1Jun17' # will form filename along with map_bg
labels_map[lab_title] <- paste0(loc_title, ' - OpenSignal 1 June 2017')
map_bg <- 'osm'
map_size <- 32 # need bigger plot for whole of London
map_plot(df, map_type, map_bg)

map_size <- 16 # set back to default for landscape maps

# ============================================================
#
# SETUP PART 2: SNIP DATA AND LOCATION SPECIFIC SET-UP
#
# Run all of this before proceeding further
# (check specifed correct loc_label at start of script)
#
# ============================================================
#
# OPENSIGNAL DATA SNIP - reduce date period to 1 to 28 June (4 weeks)
# --------------------
df_source <- subset(df_source, df_source$date < '2017-06-29')

# ----------------------
# LOCATION PARAMETERS <- if re-running per location, change loc_label and re-run from here
# ----------------------

if (loc_label == 'westminster') {

 loc_title <- 'Westminster' # default to start chart title

 # Map bounding box
 LatMax <- 51.51022 # North
 LatMin <- 51.49384 # South
 LngMax <- -0.111176  # East
 LngMin <- -0.137676 # West

 # LandScan grid covering the incident
 xMin <- -0.1375 # West
 xMax <- -0.1125  # East
 yMin <- 51.49583333 # South
 yMax <- 51.50416667 # North

 # Extended grid for Pixels (default to 0 if don't know, will keep within LandScan grid)
 exBottom <- 1 # number of rows to add below bottom of LandScan grid
 exTop <- 2  # number of rows to add above top of LandScan grid
 exLeft <- 0  # number of columns to add to left of LandScan grid
 exRight <- 0 # number of columns to add to right of LandScan grid

}

if (loc_label == 'londonbridge') {

 loc_title <- 'London Bridge' # default to start chart title

 # Map bounding box
```

```
 LatMax <- 51.516665 # North
 LatMin <- 51.50027 # South
 LngMax <- -0.0780466  # East
 LngMin <- -0.104266 # West

 # LandScan grid covering the incident
 xMin <- -0.1041667 # West
 xMax <- -0.07916667  # East
 yMin <- 51.50416667 # South
 yMax <- 51.5125 # North

 # Extended grid for Pixels (default to 0 if don't know, will keep within LandScan grid)
 exBottom <- 1 # number of rows to add below bottom of LandScan grid
 exTop <- 0  # number of rows to add above top of LandScan grid
 exLeft <- 0  # number of columns to add to left of LandScan grid
 exRight <- 0 # number of columns to add to right of LandScan grid

}

if (loc_label == 'oxfordcircus') {

 loc_title <- 'Oxford Circus' # default to start chart title

 # Map bounding box
 LatMax <- 51.524567 # North
 LatMin <- 51.50879 # South
 LngMax <- -0.128697  # East
 LngMin <- -0.155197 # West

 # LandScan grid covering the incident
 xMin <- -0.1541667 # West
 xMax <- -0.1291667  # East
 yMin <- 51.5125 # South
 yMax <- 51.5208334 # North

 # Extended grid for Pixels (default to 0 if don't know, will keep within LandScan grid)
 exBottom <- 0 # number of rows to add below bottom of LandScan grid
 exTop <- 0  # number of rows to add above top of LandScan grid
 exLeft <- 0  # number of columns to add to left of LandScan grid
 exRight <- 0 # number of columns to add to right of LandScan grid

}

# --------------
# LOCATION SETUP
# --------------

# 1. GET OSM BACKGROUND FOR LOCATION MAP TILE

map_type <- 'stamen' # 'osm' or 'stamen'
map_design <- 'terrain' # pick design for map_type
map_zoom <- 16 # zoom for OSM background (determines map detail)

basemap <- map_background(map_type, map_design, map_zoom)

# 2. SNIP DATA TO LOCATION MAP, JITTER COORDINATES

df <- df_source
df <- snip_to_box(df, LatMin, LatMax, LngMin, LngMax) # data within map tile

# convert lat and lng into 'old' coords as will be jittering but want to retain originals
df$olat <- df$lat
df$olng <- df$lng
df <- reproject(df, x='olng', y='olat', a='oosm_x', b='oosm_y')

# Jittering the coordinates to 6 decimal places
set.seed(42) # fixing randomisation parameter for reproducibility
df$rnd1 <- runif(nrow(df), -0.000499, 0.000499) # if lat/lng is just snipped to 3 decimal places
df$rnd2 <- runif(nrow(df), -0.000499, 0.000499) # if lat/lng is just snipped to 3 decimal places
df <- mutate(df, jlat=lat+rnd1, jlng=lng+rnd2)
df <- reproject(df, x='jlng', y='jlat', a='josm_x', b='josm_y')

# setting default coords - using jittered coords
df$lat <- df$jlat
df$lng <- df$jlng
df$osm_x <- df$josm_x
df$osm_y <- df$josm_y

df_map <- df
rm(df)


# 3. SETUP GRIDS - ADD CELL IDS TO READINGS AND PROVISION GRID RESULTS DF

# snip and tag OS readings to grid cells, prep results frame with coordinates for spatial plots
# note: cell IDs are the x and y number for the SW corner of the grid (i.e. starts at x1,y1 for
the grid)
```

```
# A. setup grids (lists of x and y values for SW corner of each cell in grid)

# - LandScan grid data
df_grid <- snip_to_box(df_map, yMin, yMax, xMin, xMax)   # data within LandScan grid

lx <- seq(xMin, xMax, by=lscan_increment)
ly <- seq(yMin, yMax, by=lscan_increment)
#grid <- as.data.frame(expand.grid(lx, ly)) # just helps show the pairs of coordinates forming
points in the grid (not used)

# - pixel grid data (may go beyond LandScan grid - need to calculate new boundary first)
exMin <- xMin - (exLeft * pixel_increment)
exMax <- xMax + (exRight * pixel_increment)
eyMin <- yMin - (exBottom * pixel_increment)
eyMax <- yMax + (exTop * pixel_increment)
df_pixel <- snip_to_box(df_map, eyMin, eyMax, exMin, exMax)  # data within Pixel grid (may go
beyond LandScan boundary)

px <- seq(exMin, exMax, by=pixel_increment)
py <- seq(eyMin, eyMax, by=pixel_increment)


# B. find and add cell ID for each reading, then generate a results dataframe listing cells with
coordinates for grid-based plots
# cell ID is xMyM representing position of SW corner in grid (where N = number of rows or cols
and M = N-1)
# if there is data outside the grid, will have a value of 0 or N
# because using SW corner, the last row and last column will be N-1 because the N and/or E value
would be SW for next row/col

# LandScan grid (df_grid)
df <- df_grid
x_list <- lx # list of coordinates for points of the grid on each axis
y_list <- ly #
df_grid <- add_cell_ids(df)    # assign cell IDs to each reading in dataset
grid_lscan <- add_grid_coords(df_grid) # create results DF with each cell ID as a row with its
coordinates

# Pixel grid (df_pixel)
df <- df_pixel
x_list <- px # list of coordinates for points of the grid on each axis
y_list <- py #
df_pixel <- add_cell_ids(df)    # assign cell IDs to each reading in dataset
grid_pixel <- add_grid_coords(df_pixel) # create results DF with each cell ID as a row with its
coordinates

rm(df)

# For pixel grid, append LandScan cell ids (for when want distributions within each landscan
cell)
d <- df_pixel
d$lscan <- '---'

if (loc_label == 'westminster') {
 d$lscan[d$x<5 & d$y >1 & d$y < 6] <- 'ls_x1y1'
 d$lscan[d$x>4 & d$x<9 & d$y >1 & d$y < 6] <- 'ls_x2y1'
 d$lscan[d$x>8 & d$y >1 & d$y < 6] <- 'ls_x3y1'

 grid_pixel_lscan <- subset(grid_pixel, grid_pixel$y > 1 & grid_pixel$y < 6)

}
if (loc_label == 'londonbridge') {
 d$lscan[d$x<5 & d$y >1] <- 'ls_x1y1'
 d$lscan[d$x>4 & d$x<9 & d$y >1] <- 'ls_x2y1'
 d$lscan[d$x>8 & d$y >1] <- 'ls_x3y1'

 grid_pixel_lscan <- subset(grid_pixel, grid_pixel$y > 1)
}
if (loc_label == 'oxfordcircus') {
 d$lscan[d$x<5] <- 'ls_x1y1'
 d$lscan[d$x>4 & d$x<9] <- 'ls_x2y1'
 d$lscan[d$x>8] <- 'ls_x3y1'

 grid_pixel_lscan <- grid_pixel # Oxford Circus pixel grid doesn't extend beyond LandScan grid
}

df_pixel <- d

rm(d)


# console outputs to finish this section

paste0('Number of records in df_source after data snip: ', nrow(df_source))
paste0('Number of unique devices in df_source after data snip: ', n_distinct(df_source$device))

paste0('Number of records in ', loc_title, ' landscape: ', nrow(df_map))
```

```
paste0('Number of unique devices: ', n_distinct(df_map$device))


# SETUP 2 COMPLETE - MOBILE SOURCE DATA AND MAP GRIDS ARE READY!

# =====================================================
#
# OPTIONAL 2: COMPARING ORIGINAL AND JITTERED COORDINATES
#
# Can be run for any landscape location.
# Will produce two maps - one with original coordinates, one with jittered
# with LandScan grid to show that coords close to edge may be incorrectly assigned
##
# =====================================================

df <- df_map
df <- subset(df, df$date >= '2017-06-01' & df$date <= '2017-06-02')

plot_base <- 'map' # start of filename for image saves
map_type <- 'data_single' # single layer of data

# PLOTTING MAP WITH JITTERED DATA POINTS (will be used for analysis)

# just the OSM map
map_data <- 'osjit_1Jun17' # will form filename along with map_bg
labels_map[lab_title] <- paste0(loc_title, ' - Jittered Coordinates')
map_bg <- 'osm'
map_plot(df, map_type, map_bg)

map_bg <- 'osm_lscan'
map_plot(df, map_type, map_bg)


# PLOTTING MAP WITH RAW DATA POINTS (rounded to three decimal places)
df$osm_x <- df$oosm_x
df$osm_y <- df$oosm_y

map_data <- 'osorig_1Jun17' # will form filename along with map_bg

# just the OSM map
labels_map[lab_title] <- paste0(loc_title, ' - Rounded Coordinates')
map_bg <- 'osm'
map_plot(df, map_type, map_bg)

# with LandScan grid
labels_map[lab_title] <- paste0(loc_title, ' - Rounded Coordinates')
map_bg <- 'osm_lscan'
map_plot(df, map_type, map_bg)

# =====================================================
#
# MAIN PROG 1: GENERAL ANALYSIS AND VISUALISATION
#
# Can choose and run any plots, starting from dotted underline
# Applies to all locations (using parameters for day before, during and after incident)
# no need to edit anything, just choose which plots to run to produce outputs
#
# =====================================================
# ---------------
# LANDSCAPE MAPS
# ---------------
# map_type = 'empty' or 'data_single' (1 data layer) or 'data_double' (2 data layers)
# for 2 data layers, have used RColorBrewer colour-blind friendly choices + shape change
# - first layer is purple triangles, second is orange circles
# map_bg = 'osm', 'lscan', 'pixel' or 'pixelext'

plot_base <- 'map' # start of filename for image saves

# PLOTTING MAP BACKGROUNDS (empty, no data)
map_type <- 'empty'
map_data <- 'nodata' # leave blank, is for filename, not needed unless plotting data

# Plot OSM background (no grids or data)
labels_map[lab_title] <- paste0(loc_title, ' - Map Tile')
map_bg <- 'osm'
map_plot(df, map_type, map_bg)

# Plot OSM with LandScan grid
labels_map[lab_title] <- paste0(loc_title, ' - LandScan Grid')
map_bg <- 'osm_lscan'
map_plot(df, map_type, map_bg)

# Plot OSM with LandScan Pixel grid and Landscan outline
labels_map[lab_title] <- paste0(loc_title, ' - Pixel Grid')
map_bg <- 'osm_pixel'
map_plot(df, map_type, map_bg)
```

```
# PLOTTING MAP WITH RAW DATA POINTS (SPATIALLY JITTERED)

df <- df_map
df <- subset(df, df$date >= '2017-06-14' & df$date <= '2017-06-16')
map_type <- 'data_single' # single layer of data
map_data <- 'osjit_Jun17' # will form filename along with map_bg

# just the OSM map
labels_map[lab_title] <- paste0(loc_title, ' - OpenSignal, 14 - 16 June 2017')
map_bg <- 'osm'
map_plot(df, map_type, map_bg)

# with LandScan grid
labels_map[lab_title] <- paste0(loc_title, ' - OpenSignal, 14 - 16 June 2017, LandScan Grid')
map_bg <- 'osm_lscan'
map_plot(df, map_type, map_bg)

# with LandScan Pixel grid and Landscan outline
labels_map[lab_title] <- paste0(loc_title, ' - OpenSignal, 14 - 16 June 2017, Pixel Grid')
map_bg <- 'osm_pixel'
map_plot(df, map_type, map_bg)

# PLOTTING MAP WITH RAW DATA POINTS (ROUNDED TO THREE DECIMAL PLACES)

df <- df_map
df <- subset(df, df$date >= '2017-06-14' & df$date <= '2017-06-16')
df$osm_x <- df$oosm_x
df$osm_y <- df$oosm_y
map_type <- 'data_single' # single layer of data
map_data <- 'osrnd_Jun17' # will form filename along with map_bg

# just the OSM map
labels_map[lab_title] <- paste0(loc_title, ' - OpenSignal, 14 - 16 June 2017')
map_bg <- 'osm'
map_plot(df, map_type, map_bg)

# with LandScan grid
labels_map[lab_title] <- paste0(loc_title, ' - OpenSignal, 14 - 16 June 2017, LandScan Grid')
map_bg <- 'osm_lscan'
map_plot(df, map_type, map_bg)

# with LandScan Pixel grid and Landscan outline
labels_map[lab_title] <- paste0(loc_title, ' - OpenSignal, 14 - 16 June 2017, Pixel Grid')
map_bg <- 'osm_pixel'
map_plot(df, map_type, map_bg)

# -----------
# DATE CHARTS
# -----------
# use 'chart_name' to create unique filename per chart

# DAILY PLOTS (DATE BASED)

plot_base <- 'bars_dates_' # start of filename for image saves
chart_height <- 6
chart_width <- 20

df <- df_map
df <- df %>% group_by(date, wday) %>% summarise(count=n_distinct(device)) %>% ungroup()
df$date <- as.factor(day(df$date))

# plot Landscape readings
chart_title <- paste0('Number of devices daily within ', loc_title ,' landscape, 1 to 28 June
2017')
chart_name <-'os_dates' # to ensure unique filename for plot save
plot_dates(df, chart_title, chart_name, chart_type = 'single')

# ------------------
# DEVICE FREQUENCIES
# ------------------
# plot numebr of days devices were present within landscape for June dataset
# plot frequency distribution of location accuracy
df <- df_map
df <- df %>% group_by(device) %>% summarise(count = n_distinct(date)) %>% ungroup()

chart_height = 10
chart_width = 20
ggplot(data=df, aes(df$count)) +
 geom_histogram(breaks=seq(0, 30, by=1), colour='grey50') +
 labs(subtitle = paste0('Number of days a device is present, OpenSignal June 2017, ', loc_title))
+
 labs(x='days present', y='devices') +
 t

plot_save <- paste0(loc_label, '_osJun17_presence_freqdistr')
ggsave(filename=paste0(plots_folder, plot_save,'.png'), plot=last_plot(), device='png',
  scale=1.0, width=chart_width, height=chart_height, units='cm', dpi=300, limitsize=TRUE)
```

```
# MAIN PROG 1 COMPLETED

# ======================================================
#
# MAIN PROG 2: AMBIENT CONTEXT - LANDSCAN SCALE
#
# Can choose and run any plots, starting from dotted underline
# Applies to all locations (using parameters for day before, during and after incident)
# no need to edit anything, just choose which plots to run to produce outputs
#
# When creating new plots, make sure has unique plot_base to start all output names
#
# ======================================================

library(reshape)   # for pivoting using cast (upsets expand in tidyr so sitting here after grids
are sorted)

# 1. SUMMARY STATS
# -------------------------
# reality data summary for analyses (values will output to console)
df <- df_map
paste('map readings:', nrow(df))
paste('map devices: ', nrow(df %>% group_by(device) %>% summarise(n())) )

df<- df_grid
df %>% group_by(cell) %>% summarise(devices = n_distinct(device), readings=n())

rm(df)

# -------------------------
# 2. DAILY RHYTHM
# -------------------------
plot_base <- 'box_day'

df <- df_grid
df <- subset(df, df$date >= rhythm_start & df$date <= rhythm_end) # normal rhythm before incident
df <- df %>% group_by(date, wday, cell) %>% summarise(count=n_distinct(device)) %>% ungroup()
df$cat <- as.factor(df$wday)

# Box plot of week days, faceted by LandScan cell
chart_title <- paste0(loc_title, ', ', rhythm_range, ' per LandScan cell')
chart_name <-paste0('os_lscan') # to ensure unique filename for plot save
chart_height <- 7
chart_width <- 20
boxplot_facet(df, grouping='cell')

# Mean counts
subdf <- df %>% group_by(cell, wday) %>% summarise(mean = mean(count)) %>% ungroup() # means per
day of week per cell
subdf <- cast(subdf, wday ~ cell)
summary(subdf) # day average (across all days)

# go to 'Ambient Context' Excel hacks for distribution analysis (copying subdf values across for
quick hack calcs)

rm(df, subdf)

# -------------------------
# 3. HOURLY RHYTHM
# -------------------------
# note: plot functions have dummy df (df named twice) because function needs 2 DF but only one
used here

plot_base <- 'barline_hrs'

df <- df_grid
df <- subset(df, df$date >= rhythm_start & df$date <= rhythm_end) # normal rhythm before incident

df <- df %>% group_by(date, wday, hr, cell) %>% summarise(count=n_distinct(device)) %>% ungroup()
df$cat <- as.factor(df$hr)

# plotting hourly counts for each date as bars, with a line showing mean across dates + incident
date
chart_height <- 7
chart_width <- 30
x_label <- 'hour'
max_val <- max(df$count) # ensures all plots have same y-axis for comparison

# all days
chart_title <- paste0('All days, ', loc_title, ', ', rhythm_range, ' per LandScan cell')
chart_name <-paste0('os_lscan_all') # to ensure unique filename for plot save
bars_and_line(df, df, chart_name, chart_title, chart_height, max_val, plot_type = 'rhythm')

# weekdays
chart_title <- paste0('Weekdays, ', rhythm_range, ' per LandScan cell')
chart_name <-paste0('os_lscan_week') # to ensure unique filename for plot save
```

```
subdf <- subset(df, df$wday < 6)
bars_and_line(subdf, subdf, chart_name, chart_title, chart_height, max_val, plot_type = 'rhythm')

meandf <- subdf %>% group_by(cell, hr) %>% summarise(count=mean(count, na.rm = T)) # copy mean
counts across to Excel hack
meandf <- cast(meandf, hr ~ cell)

# weekends
chart_title <- paste0('Weekends, ', rhythm_range, ' per LandScan cell')
chart_name <-paste0('os_lscan_wkend') # to ensure unique filename for plot save
subdf <- subset(df, df$wday >= 6)
bars_and_line(subdf, subdf, chart_name, chart_title, chart_height, max_val, plot_type = 'rhythm')

meandf <- subdf %>% group_by(cell, hr) %>% summarise(count=mean(count, na.rm = T)) # copy mean
counts across to Excel hack
meandf <- cast(meandf, hr ~ cell)

rm(df, subdf, meandf)

# -------------------------
# 3B. HOUR CURVE SHAPES
# -------------------------
# part calculating here, then copying mean counts across to Excel to hack curve weights across
all three landscapes

df <- df_grid
df <- subset(df, df$date >= rhythm_start & df$date <= rhythm_end) # normal rhythm before incident

df <- df %>% group_by(date, wday, hr, cell) %>% summarise(count=n_distinct(device)) %>% ungroup()
df$cat <- as.factor(df$hr)

# weekdays
subdf <- subset(df, df$wday < 6)
meandf <- subdf %>% group_by(cell, hr) %>% summarise(count=mean(count, na.rm = T)) # copy mean
counts across to Excel hack
meandf <- cast(meandf, hr ~ cell)

# weekends
subdf <- subset(df, df$wday >= 6)
meandf <- subdf %>% group_by(cell, hr) %>% summarise(count=mean(count, na.rm = T)) # copy mean
counts across to Excel hack
meandf <- cast(meandf, hr ~ cell)


# MAIN PROG 2 COMPLETE!

# ================================================================
#
# MAIN PROG 3: AMBIENT CONTEXT - SPATIAL DISTRIBUTION (PIXEL SCALE)
#
# Can choose and run any plots, starting from dotted underline
# Applies to all locations (using parameters for day before, during and after incident)
# no need to edit anything, just choose which plots to run to produce outputs
#
# When creating new plots, make sure has unique plot_base to start all output names
#
# ================================================================

library(reshape)   # for pivoting using cast (upsets expand in tidyr so sitting here after grids
are sorted)

# ------------------------
# 1. PIXEL-BASED PRESENCE
# ------------------------

# PLOT SCALED PIXEL COUNTS AS A GRID (mean daily counts for weekday and weekend) SNIPPED TO
LANDSCAN GRID

plot_base <- 'grid_pixel'

df <- df_pixel
df <- subset(df, df$date >= rhythm_start & df$date <= rhythm_end) # normal rhythm before incident
df <- subset(df, df$lscan != '---')

# calculate mean count per cell per day of week
d <- df %>% group_by(date, wday, lscan, cell, x, y) %>% summarise(count = n_distinct(device)) %>%
ungroup()
d <- d %>% group_by(wday, lscan, cell, x, y) %>% summarise(count = mean(count)) %>% ungroup()

# calculate mean values per cell for weekdays and weekends
d_wkday <- subset(d, d$wday < 6)
d_wkend <- subset(d, d$wday >= 6)
d_wkday <- d_wkday %>% group_by(cell, lscan) %>% summarise (count = mean(count), cat = 'wkday')
%>% ungroup()
d_wkend <- d_wkend %>% group_by(cell, lscan) %>% summarise (count = mean(count), cat = 'wkend')
%>% ungroup()
```

```
# function to plot mean day counts per cell
plot_data <- function(df) {
 df <- merge(grid_pixel_lscan, df, by.x='cell', by.y='cell', all.x=T)
 df$count[is.na(df$count)] <- 0 # NA means if no devices present in cell, set to 0
 df <- mutate(df, count = count/max(count)) # scale to max val of 1 for plotting
 map_plot(df, map_type = 'data_grid', map_bg = 'osm_pixel_lscan')
}

map_data <- 'wkday' # plot means per cell for weekdays
df <- d_wkday
labels_map[lab_title] <- paste0(loc_title, ' - Mean weekday counts per pixel')
plot_data(df)

map_data <- 'wkend' # plot means per cell for weekends
labels_map[lab_title] <- paste0(loc_title, ' - Mean weekend counts per pixel')
df <- d_wkend
plot_data(df)


# SAME AS ABOVE BUT USING TRIPS INSTEAD OF UNIQUE DEVICES AS THE COUNT (USING SAME FUNCTION)
df <- df_pixel
df <- subset(df, df$date >= rhythm_start & df$date <= rhythm_end) # normal rhythm before incident
df <- subset(df, df$lscan != '---')

# calculate mean count per cell per day of week
d <- df %>% group_by(date, wday, lscan, cell, x, y) %>% summarise(count = n_distinct(device)) %>%
ungroup()
d <- d %>% group_by(wday, lscan, cell, x, y) %>% summarise(count = mean(count)) %>% ungroup()

# calculate mean values per cell for weekdays and weekends
d_wkday <- subset(d, d$wday < 6)
d_wkend <- subset(d, d$wday >= 6)
d_wkday <- d_wkday %>% group_by(cell, lscan) %>% summarise (count = mean(count), cat = 'wkday')
%>% ungroup()
d_wkend <- d_wkend %>% group_by(cell, lscan) %>% summarise (count = mean(count), cat = 'wkend')
%>% ungroup()

# ========================================

# ALL DONE!
```

Appendix C: Visual Samples

The following images are a series of screenshots of frames for animations used to visually inspect the spatial and temporal distribution of readings before analysis for studies presented in chapters four and five. The animations were designed for viewing on a computer monitor rather than as a print-out. Visuals are included here for reference.

Wi-Fi readings were aggregated to hourly counts of devices connecting to each access point. Devices were tagged with a behaviour attribute to indicate spatial familiarity: regulars (present on at least three days, 'habitual') versus tourists ('explorers'). OpenSignal readings are plotted as individual points with devices also tagged to indicate spatial familiarity. The readings are animated using the software application Processing. The map background is provided by OpenStreetMap.

The first two images compare Wi-Fi readings at different hours of the same day, 5am versus 11am and 5pm versus 11pm respectively on Tuesday 22 March 2016.

The next two images show readings for the Wi-Fi and webcams on two different event days: The Sports Relief charity event that took place at venues across the park on Sunday 20th March 2016 and a charity football match that took place on the South Lawn on Saturday 27th March 2016.

The final two images show readings for OpenSignal, comparing two different times on the same day – Tuesday 20th June 2017, and comparing two Saturdays, one when no event occurred in the park (24th June 2017) and one when a concert took place at the London Stadium (17th June 2017).

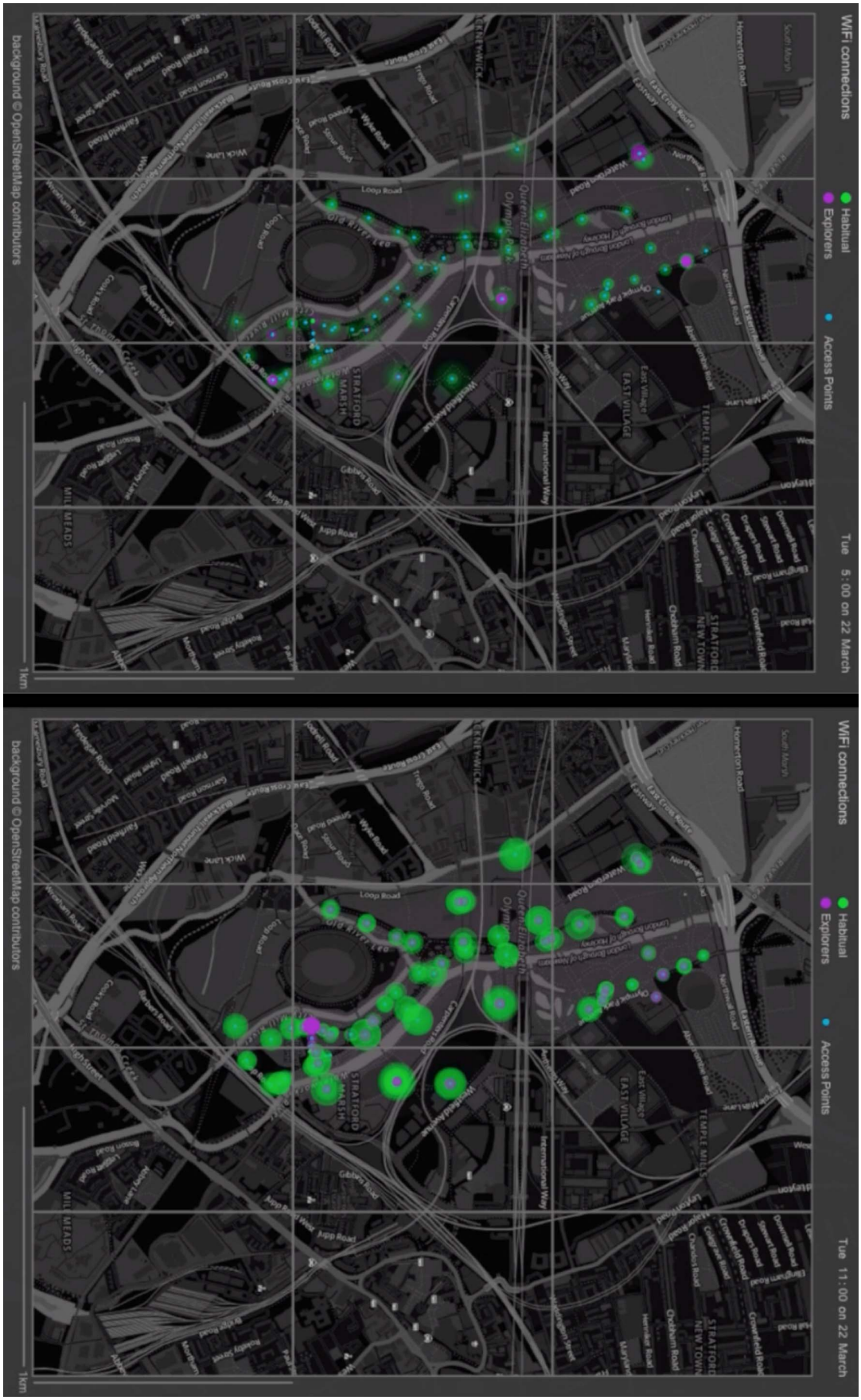


Figure 125. Animation frame comparing 5am and 11am on Tuesday 22 March 2016

Wi-Fi hourly device counts classified by spatial familiarity (habitual - green or explorer - purple).

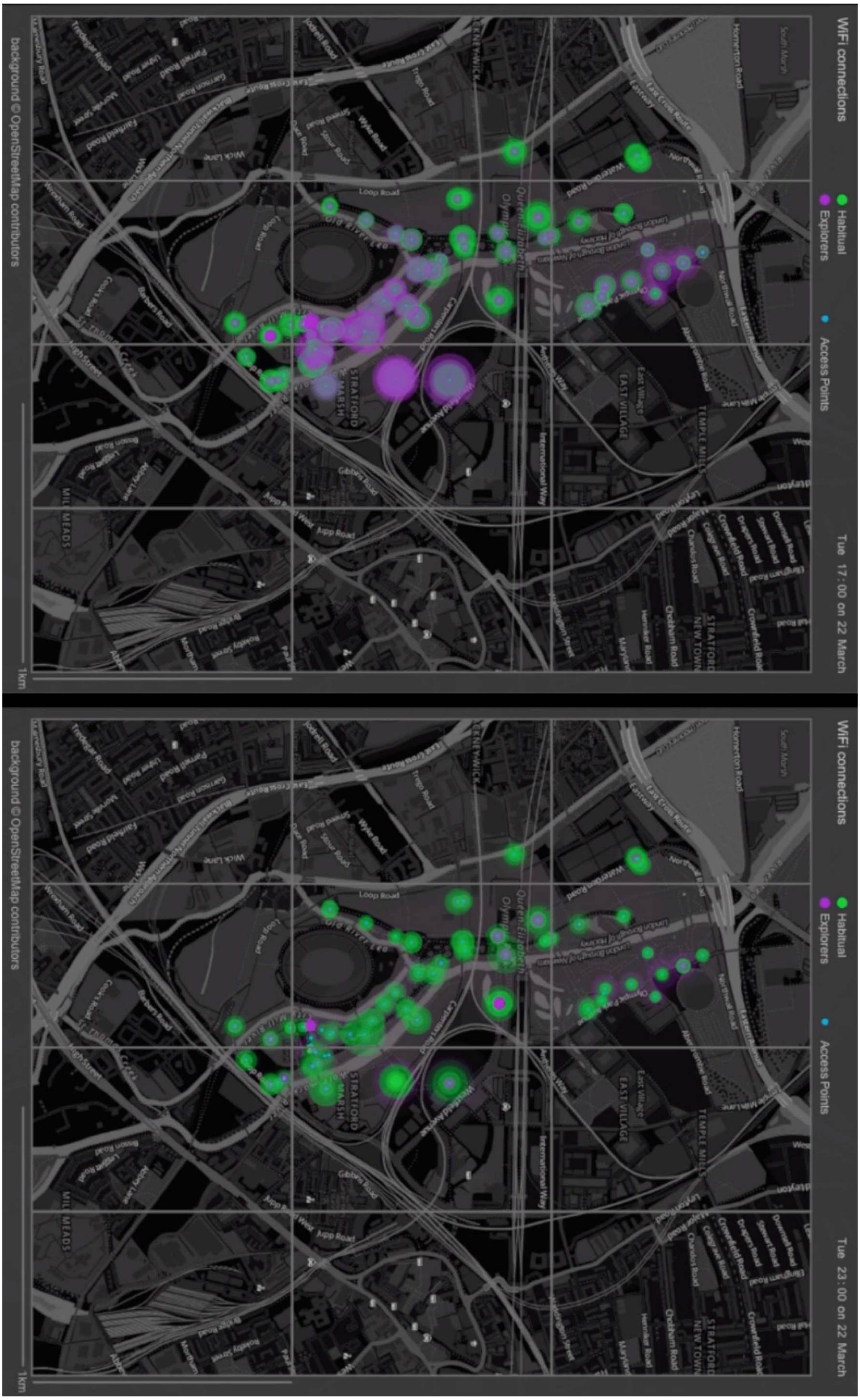


Figure 126. Animation frame comparing 5pm and 11pm on Tuesday 22 March 2016

Wi-Fi hourly device counts classified by spatial familiarity (habitual - green or explorer - purple).

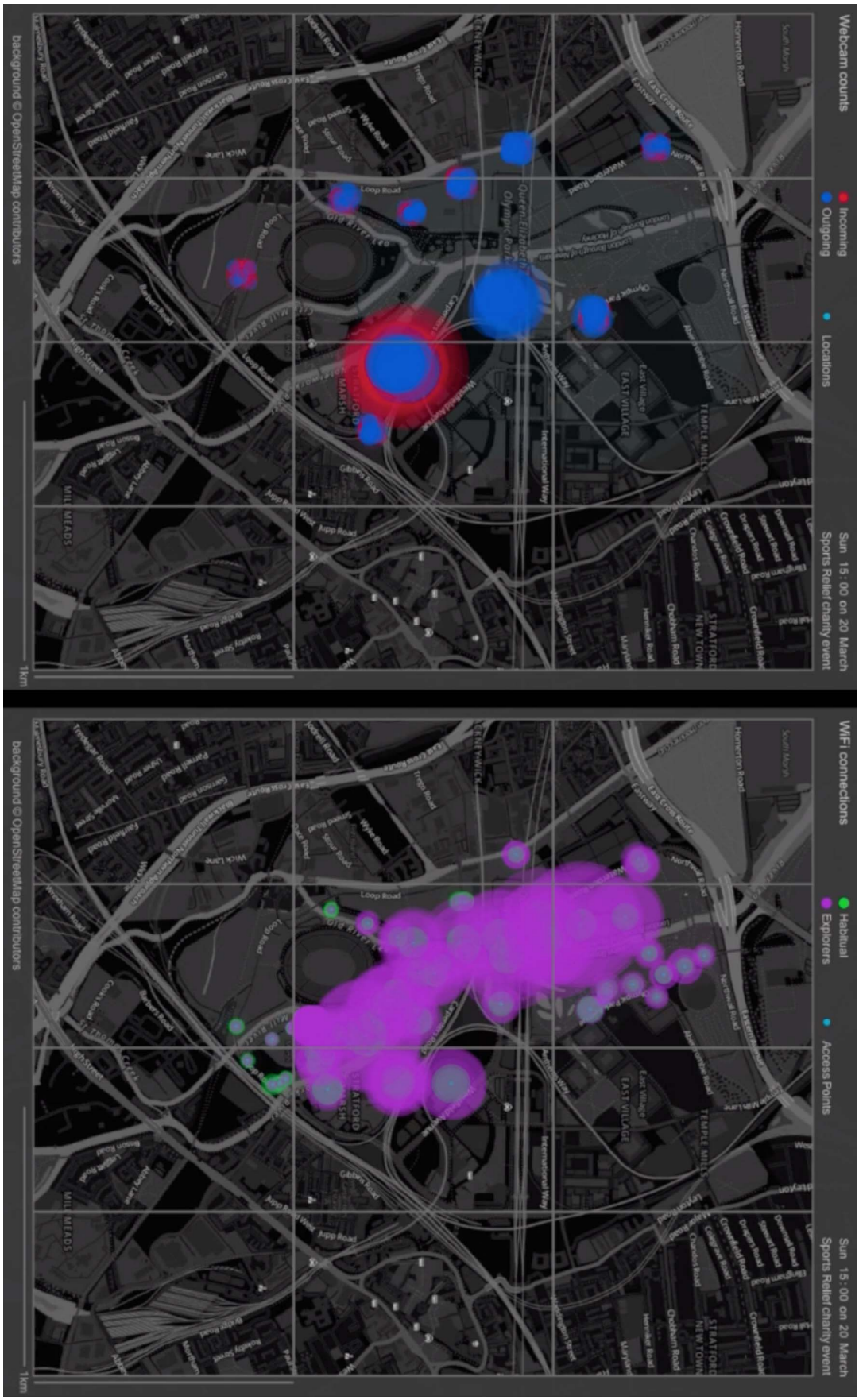


Figure 127. Animation frame at 15:00 on 20 March 2016 (Sports Relief event)

Camera headcounts on left (incoming – blue, outgoing – red); Wi-Fi hourly device counts on right (habitual – green, explorer – purple); Events held at various venues across the park.

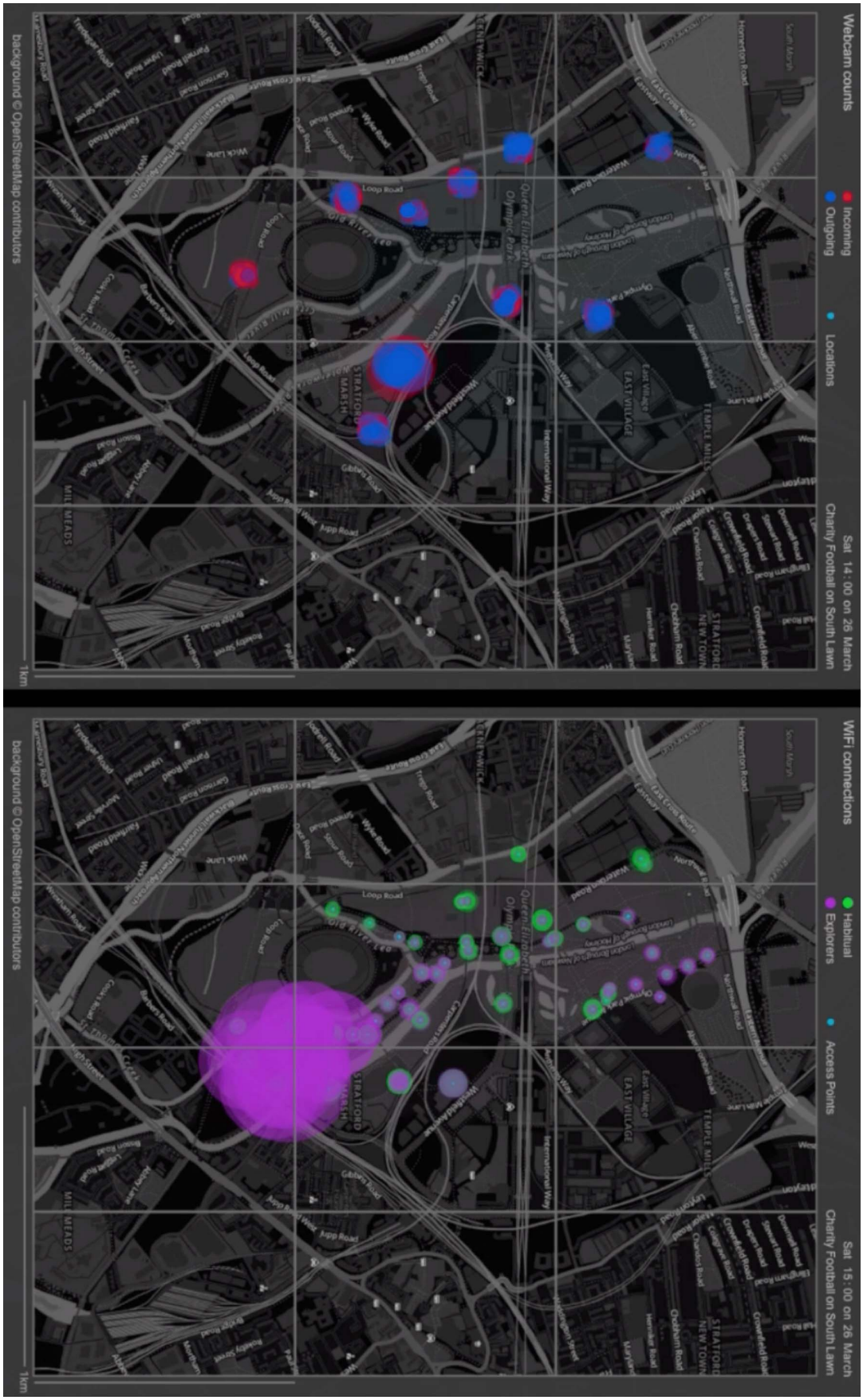


Figure 128. Animation frame at 15:00 on 26 March 2016 (Charity football)

Camera headcounts on left (incoming – blue, outgoing – red); Wi-Fi hourly device counts on right (habitual – green, explorer – purple); Event held on the South Lawn.

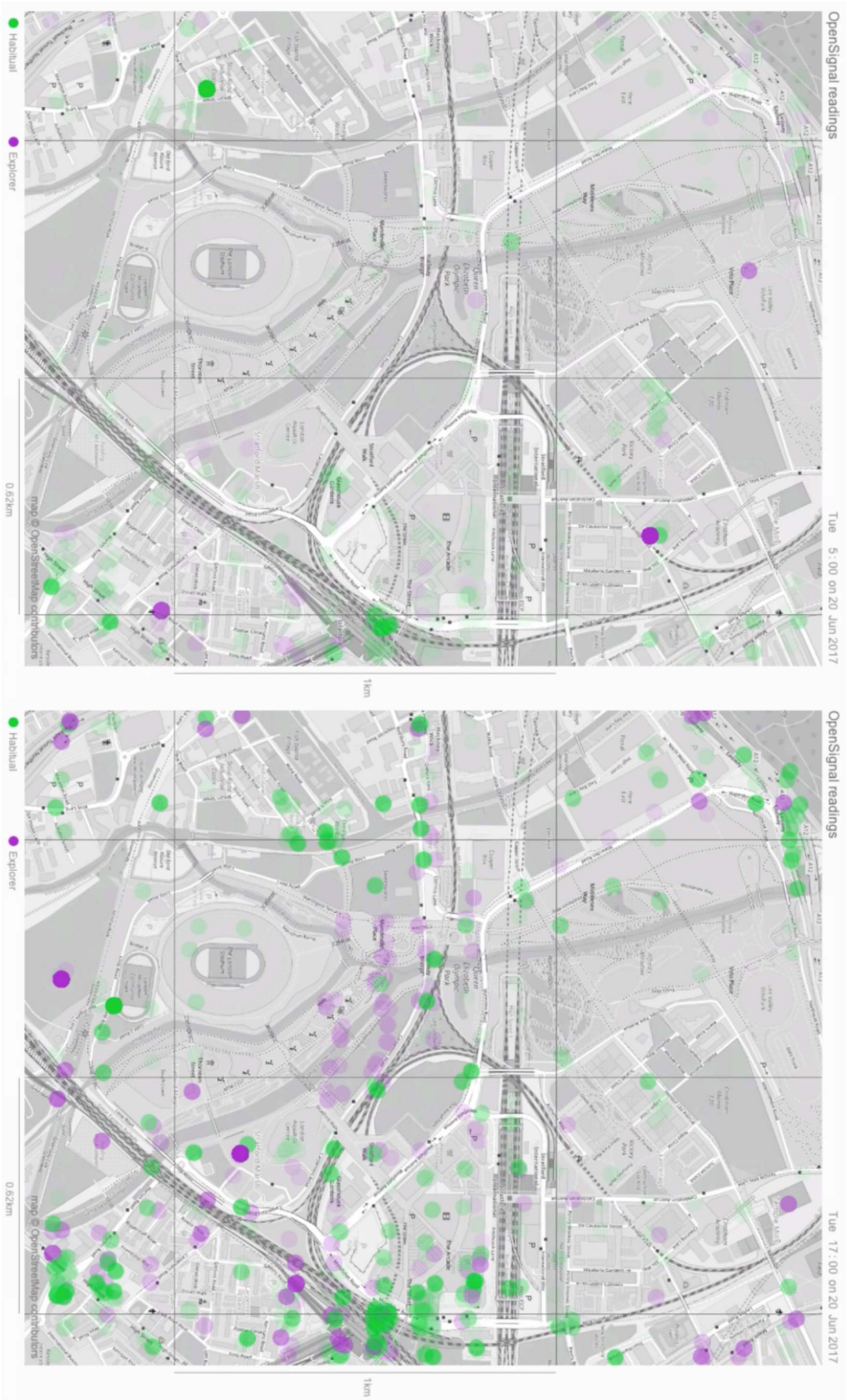


Figure 129. Animation frame comparing 5am and 5pm on Tuesday 20 June 2017

OpenSignal devices classified by spatial familiarity (habitual - green or explorer - purple).

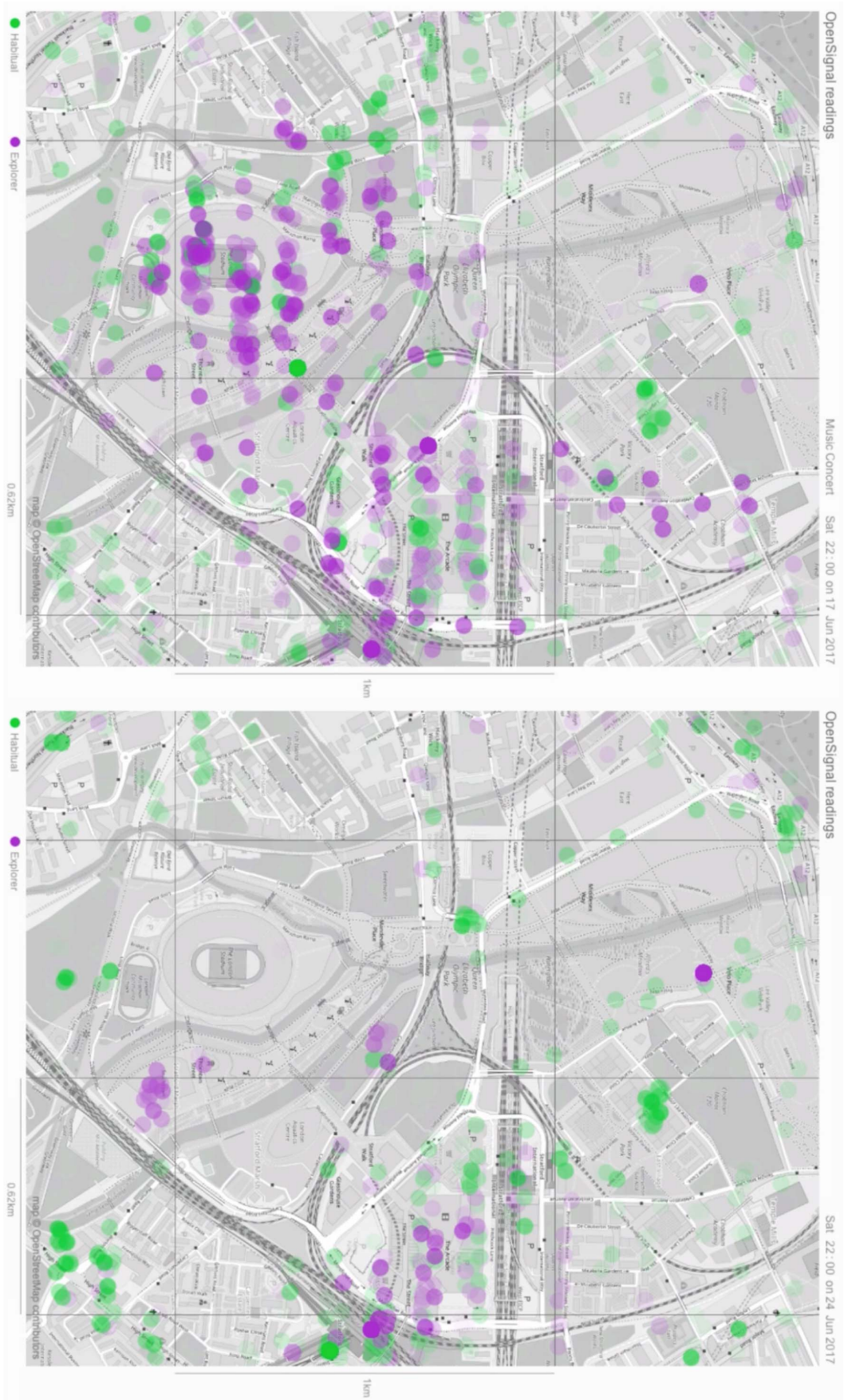


Figure 130. Animation frame comparing Saturdays at 22:00 in June 2017

OpenSignal devices classified by spatial familiarity (habitual - green or explorer - purple); The image on the left is an event Saturday (music concert at the London Stadium); The image on the right is a non-event Saturday.