2
Review of Literature

Understanding the scale, nature and dynamics of distribution of population in space and time has been a central premise of academic research in various fields of study such as human geography, sociology, urban planning and architecture. This granular knowledge of where people are and how they move is critical in practical decision making in various industries such as real estate for valuing places, retail for business planning and emergencies for risk management and evacuation. The first challenge in any of these research is collecting precise and accurate data. This started as research into methods estimating and interpolating highly granular data from existing regional level aggregate datasets and as the technology improved through the later half of twentieth century, the research methodologies adopted the new technologies to utilise these more granular sources of data. Though new technologies provide immense opportunity in collecting large amounts of data which were previously impossible they also introduce their share of uncertainties. Hence it becomes imperative that we understand the evolution of these techniques and methodologies along with the research that used them to build our rationale behind any further research.

Moreover with the proliferation of mobile devices and wireless internet connectivity, even day to day activities are being digitised leading to the creation of large amount of easily accessible data which are generated passively in an unstructured manner. The users' acceptance to the collection and analysis of such data has also been improving until recently [1]. There has also been rising concerns regarding user privacy along with the development of more accurate methods to track them. In this context, the critical task in all these research is to solve the problem of balancing these two by choosing the right technologies and devising the appropriate methods.

In this chapter we conduct a systematic survey of literature in this broad area of 'distribution and dynamics of human activity'. The aim of this survey is to evaluate where we are at the research and understand how we got here. First we conduct a comprehensive survey of over 300 publications which discuss this area of research and we then look at the major themes and trends in the last 70 years. We discuss these themes in detail looking at the aims and achievement of these research while highlighting the opportunities and gaps they leave. We also look at the timeline of their publication and discuss the evolution of the research along with the changes in the technology landscape. We then classify the studies by looking at the technologies employed by them and note the trends in the adoption and phasing out of relevant technologies. We also

[1] Alfred Kobsa. User acceptance of footfall analytics with aggregated and anonymized mobile phone data. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 8647 LNCS, pages 168–179, 2014. ISBN 9783319097695. DOI: 10.1007/978-3-319-09770-1_15

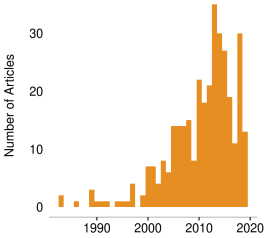

Figure 2.1: Growth of research in the areas of understanding distribution and dynamics of human activity since 1980.

Measured in the number of papers published

try and understand the advantages and disadvantages of these techniques and develop a theoretical framework to understand when and how to use them effectively answer research questions. Finally we summarise the literature survey focussing on the major gaps still left in our understanding and the new, emerging areas where more research is necessary. We also critically evaluate these areas of research in terms of priority and feasibility to develop our research questions and devise a plan for our research.

We begin our literature review with a systematic search for academic publications studying the distribution and dynamics of human activity. We start this search from a set of reviews pertaining to the topic such as Ratti et al. [2006], Jiang et al. [2013], Steenbruggen et al. [2013], Arribas-Bel [2014] and Li et al. [2016]. From these reviews we further expand by navigating the citation networks and finding research which are relevant to our search. Though this does not give us a perfectly comprehensive set of research literature, it does provide us with a representative sample of all the different directions of the research conducted in the area. Through this process, we identified around 325 relevant research publications which deal with the collection, measurement, analysis, visualisation and discussion of population at a granular level. The research started around 1950s looking at estimating day-time urban population at a granular level using existing broader data employing various estimation methods [2]. Though this served as a starting point, the pursuit of such granular data and their applications in corresponding fields didn't pick up until the start of the 21st century fuelled by the digital revolution that followed growth of internet. Figure 2.4 shows the yearly volume of research published since 1980. We can observe that though there were some research conducted through 80s and 90s the real push forward came around beginning of the millennium when mobile phones adoption skyrocketed. In addition to the early 2000s we also see a substantial increase in interest in the beginning of the next decade fuelled by the smartphone revolution which completely changed the research avenues in-terms of volume and types of data available and methodologies available to tackle them. The area of research is multi-disciplinary encompassing academic interest and commercial applications in Geography, Urban Studies, Urban Planning and Management, Emergency planning and Management, Economics, Computer Science and Engineering etc.

[2] Donald L Foley. Urban daytime population: a field for demographic-ecological analysis. *Social Forces*, pages 323–330, 1954. ISSN 0037-7732; and Robert C Schmitt. Estimating daytime populations. *Journal of the American Institute of Planners*, 22(2): 83–85, 1956. ISSN 0002-8991

2.1 *Research Themes*

In this section we look at the major themes and questions tackled by this knowledge base. We start by classifying the research into the major and minor themes explored in them as shown in Figure 2.2. The tree-map shows the volume of research in corresponding themes measured in terms of number of publications. We can observe that the research is conducted in five major areas - population studies focussing on the creating and utilising data on distribution and nature of human activity, mobility and interaction focussing on the changes in these distributions, understanding the nature and function of space from these distribution and change, methods and techniques which can be used to conduct the research and finally issues and solutions related to the privacy of the users while conducting these research. We can also observe

that most of the research apart from developing methods were conducted in the domain of human mobility and social interaction closely followed by the population distribution. In the following sections we discuss these in detail along with their sub themes with the following framework,

1. What are the major lines of questioning?
2. What has been done previously?
3. Where are the opportunities for further research?

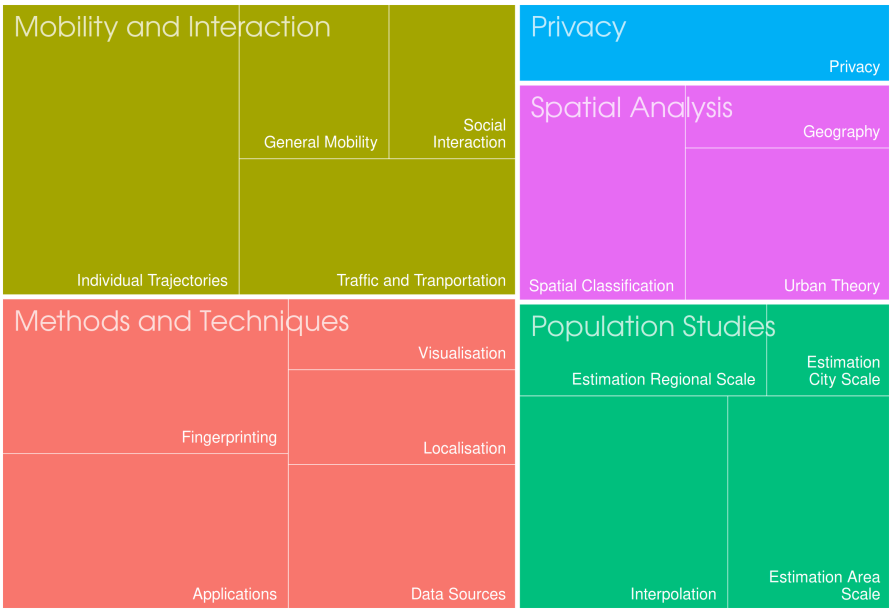

Figure 2.2: Tree-map showing the volume of research conducted under each major themes and their sub-themes.

2.1.1 *Population Studies*

Though Foley [1954] and Schmitt [1956] started this line of research in 1950's with the discussion on estimating daytime population using broader datasets it was not until the 80s significant volume of research kicked off in this area of study. From 80s until mid 2000's numerous studies were conducted on measuring and studying the population at a granular level both spatially and temporally. The focus of the research around this time was primarily on interpolation from the larger datasets created using censuses, regional or national level sample surveys and other centrally collected sources of data. There have been numerous fairly successful attempts with methodologies where a broad dataset such as regional level population summaries and modelling or interpolating more granular data from them by augmenting with other sources of data such as street networks [Reibel and Bufalino, 2005], remote sensing [Sutton, 1997, Yuan et al., 1997, Chen, 2002] etc. Dobson et al. [2000, 2003], Bhaduri et al. [2002, 2007] and [Mennis, 2003, Mennis and Hultgren, 2006] are examples of such research methodology. These studies were almost done on a city scale or above with mostly modelling or interpolation methods since the data sources were few and were centrally collected.

Around 2005, there was a sharp shift in research where the interpolation methods were replaced by highly available granular data collected over cellular network. Studies were conducted on estimating population densities, presence of tourists, general activity pattens using data from cellular networks. Most of these research were conducted at a far larger geographic scale looking at things at an area level [Pulselli et al., 2008, Girardin et al., 2009, Phithakkitnukoon et al., 2010, Yuan and Raubal, 2016]. There were efforts in using device level sensors such as global positioning system(GPS), Wi-Fi and Bluetooth to detect population distribution and socio-geographic routines [Calabrese et al., 2010, Rose and Welsh, 2010, Farrahi and Gatica-Perez, 2010]. There have been studies on looking at people distribution as granular as queue lengths as discussed by [Wang et al., 2013] to city level dynamic population mapping where the limitations of traditional datasets generated through censuses and surveys [3].

[3] Pierre Deville, Catherine Linard, Samuel Martin, Marius Gilbert, Forrest R Stevens, Andrea E Gaughan, Vincent D Blondel, and Andrew J Tatem. Dynamic population mapping using mobile phone data. *Proceedings of the National Academy of Sciences*, 111(45):15888–15893, 2014. ISSN 0027-8424. DOI: 10.1073/pnas.1408439111

Around the 2015, along with the data collected directly from the mobile devices,the data that are generated by the users activity on these devices are became more important. Social media data such as twitter [Lansley and Longley, 2016b] and other consumer data such as loyalty cards [Lloyd and Cheshire, 2018], smart cards [Ordonez and Erath, 2012] etc. have also become a significant sources of data for such research. Recently, with increased concerns and legislation on privacy, there have been studies which go back to the effort of interpolating granular data from broader datasets but using more data and processor intensive technologies such as agent based modelling, deep learning, small area estimation [Crols and Malleson, 2019, Shibata and Yamamoto, 2019, Rao and Molina, 2015] etc.. Though there have been a lot of work done in most of the directions in this research area, the clear gap arises due to the absence of a continuous, granular and sufficiently longitudinal data-sets to complement the methodologies that have been developed.

2.1.2 *Human Mobility and Interaction*

Study of movement of people is one of the major areas of research which have significantly benefited from the decentralised collection of data at a granular level [4]. In addition to being useful in their own right, these data were in turn used to augment traditional models of travel behaviour, traffic and transport to provide a better understanding of human movement over time and space [Janssens, 2013]. The major themes of research within this area are, Movement of people in space and time with emphasis on understanding the built environment, social interaction between these people with a sociology perspective and traffic and transportation studies with a infrastructure perspective. There is significant volume of research which dealt with recording and analysing the trajectories of the users to understand their movement patterns enabled by the unprecedented availability of detailed data from mobile devices and this is discussed in detail along with the discussion of the technologies used in Section 2.3.

[4] M Castells. Grassrooting the space of flows. i wheeler, aoyama and warf [eds.] cities in the telecommunications age, 2000

2.1.3 *Methodology and Techniques*

Research in this are focused around 5 major topics,

1. Localisation - Research into using the mobile devices and data generated from them as a cheaper alternative to Global Positioning Systems and remote sensing.
2. Data Sources - Identifying and formalising new data-sources as the technology develops
3. Application - Applying these identified data sources to answer questions and solve problems in different disciplines.
4. Visualisation - simplifying, visualising and interpreting these high volume of unstructured, noisy datasets.
5. Device fingerprinting - overcoming the difficulties posed by the anonymisation process and extract useful information.

Localisation of mobile devices without the use of expensive additional infrastructure such as GPS is one of the earliest ideas pursued in this aspect [Bulusu et al., 2000, He et al., 2003, Moore et al., 2004, LaMarca et al., 2005]. This research, when reversed, could also lead to the tracking of these devices in space without the aforementioned infrastructure thus providing a inexpensive, easy way to collect mobility data. The sensors which are already present in the phones such as Bluetooth [Bandara et al., 2004], Wi-Fi [Zarimpas et al., 2006], cellular radio [Dil and Havinga, 2011, Ahas and Mark, 2005] etc. have been considered to be used for localisation of the devices. This has been particularly important in the field of indoor localisation where GPS doesn't usually work [5]. When seen from the other perspective the same technologies and methods can enable us to collect presence and movement data on people indoors [Roy and Chowdhury, 2018a,b, Jia et al., 2019, Nikitin et al., 2019, Kulshrestha et al., 2019, Deng et al., 2018].

[5] Nobuo Kawaguchi. Wifi location information system for both indoors and outdoors. In *International Work-Conference on Artificial Neural Networks*, pages 638–645. Springer, 2009

The identification of data sources started with looking at the 'real time' city examining the digital landscape created by the citizens their electronic devices [Townsend, 2000]. This was furthered by the notion of 'instrumenting' the city and developing methods and techniques under the umbrella of smart cities and internet of things [O'Neill et al., 2006, Sruthi, 2019]. Since there have been research looking at the wireless data collected from positioning technologies [Bensky, 2007] and cellular network [Kiukkonen et al., 2010, Steenbruggen et al., 2015] and even crowdsourcing as method of collection [Shin et al., 2013] leading towards a framework for computational urban planning [Kontokosta, 2015]. With the effort to formalise them as valid sources of data, there have also been research looking at the biases in them such as mobile phone ownership [Wesolowski et al., 2013, Kobus et al., 2013].

Identifying and fingerprinting unique devices and users from noisy, unstructured data is another area of active research under methodologies and techniques [6]. The majority of the work has been done as an extension of localisation where the GPS-less positioning leading to finger printing people and their movement out of the data [Pang et al., 2007a, Pappalardo et al., 2015]. Additionally there are work looking at the tracks collected from Wi-Fi or mobile data and extract unique users out of them [Girardin et al., 2008, Eagle and Pentland, 2009, Jiang et al., 2012]. It is also demonstrated that it is possible to wireless technologies can be used to detect even device free entities [Elgohary, 2013]. These localisation and clustering techniques can also be used for socio-geographical analysis and to understand the patterns of

[6] Bin Jiang and Xiaobai Yao. Location-based services and gis in perspective. *Computers, Environment and Urban Systems*, 30(6):712–725, 2006. ISSN 0198-9715; and Lin Liao. *Location-based activity recognition*. PhD thesis, University of Washington, 2006

activity of people [Licoppe et al., 2008]. There have been quite a few security research on the robustness of the anonymisation techniques while revealing methodologies to overcome limitations imposed by them [Mathieu Cunche, 2016, Chothia and Smirnov, 2010, Krumm, 2007]. Cheng and Wang [2016] was one of the first to look into devising a method to do this in a non-intrusive way which are further extended by Di Luzio et al. [2016], Adamsky et al. [2018] and Dai et al. [2019]. This is currently an active field of research and there is immense opportunity for further research.

Visualising the temporal dynamics of data collected on human activities through decentralised processes poses significant challenges when approached with traditional cartographic concepts [7]. Digital media especially animation has been explored as an option to solve for the temporal dimension [Morrison et al., 2000, Lobben, 2003] but is bound by the cognitive limits of the viewer [Harrower, 2007]. There have been approaches proposed around animations of generated surfaces [Kobayashi et al., 2011] and network-based visualizations [Ferrara et al., 2014] leaving gaps in research for new methods in dynamic geographic visualisation [Fabrikant, 2005] and visualising path and flow of phenomena [Thomas, 2005], particularly of the mobility data collected from cellphones [Sbodio et al., 2014]. This provides us with a promising opportunity for research in methods for visualising high frequency, hyper-local pedestrian data within the limits of cognition of the viewer.

[7] Alan M MacEachren and Menno-Jan Kraak. Research challenges in geovisualization. *Cartography and Geographic Information Science*, 28(1):3–12, 2001. ISSN 1523-0406; and Elaine J Hallisey. Cartographic visualization: an assessment and epistemological review*. *The Professional Geographer*, 57(3):350–364, 2005. ISSN 0033-0124

2.1.4 *Spatial Analysis - Theory and Modelling*

Traditional and modern geography was dominated by the study of centrally collected data acquired through extensive field surveys and remote sensing. In the last two decades, a significant paradigm change has been introduced by the availability of unprecedented amount of data generated by unconventional sources such as mobile phones, social media posts etc. This move to the post-modern geography has been accompanied by a change in our understanding of the built environment and human geography from a static point of view to a more dynamic definition [8]. This definition is based on the bottom-up mechanisms which make human activity such as information exchange and economy to manifest in the physical built environments as argued by [Batty, 1990, 1997, Batty et al., 2012] and [Batty, 2013a,b].

This transition into the digital age [Graham and Healey, 1999, Tranos and Nijkamp, 2012, Tranos, 2013] has changed the politics of space and time [Massey, 1992] and been more pronounced in the study of urban built environment where technology has redefined the concepts of place and space [Graham and Marvin, 2001, 2002, Sassen, 2001]. With the ability to collect and analyse of data on large complex systems in real-time [Graham, 1997], we are exploring the possibilities of understanding their structure and organisation using concepts of complexity theory [Bettencourt, 2013, Portugali et al., 2012] with more emphasis on their temporal patterns such as the argument towards finding the pulse of the city [Batty, 2010]. With the population getting more and more connected [Castells, 2010], the nature of space/place is being dynamically defined by the population themselves [Giuliano and Small, 1991] and vice versa [Zandvliet and Dijst, 2006]. This flood of hard data [9] was accompanied not only by optimism in its potential [Thomas, 2001] but also

[8] Edward Soja. Postmodern geographies, 1989. URL http://books.google.com/books?id=sNcRAQAAIAAJ

[9] Nature Editorial. A flood of hard data. *Nature*, 435: 698, 2008. ISSN 0028-0836. DOI: 10.1038/453698a

by the questions raised on the challenges in handling the diverse, large scale, non standardised data it produces and the usefulness or representativeness of the resulting analysis [Miller, 2010, Arribas-Bel and Sanz-Gracia, 2014].

However, availability of such data has impressive uses in urban studies [10] especially with advancement of new technologies [Steenbruggen et al., 2013] and possibility of distributed, crowdsourced data collection [Lokanathan and Gunaratne, 2015].

[10] Luis M A Luís M.A. Bettencourt. The uses of big data in cities. *Big Data*, 2(1):12–22, 2014. ISSN 2167-6461. DOI: 10.1089/big.2013.0042. URL http://online.liebertpub.com/doi/abs/10.1089/big.2013.0042

2.1.5 *Privacy*

The ubiquity of personal devices and digitisation of day to day activities through these mobile devices [McMeel, 2018] has provided many opportunities for researchers and industry for collecting, analysing and deriving inputs from them. However at the same this also increased the risk of infringement on privacy of the users whose data is being collected [11]. There is immense value in uniquely identifying and profiling information on people for specialised purposes such as security [Cutter et al., 2006] and law enforcement [Dobson and Fisher, 2003] but also has extreme risks associated when not handled with care [VanWey et al., 2005].

[11] T Scott Saponas, Jonathan Lester, Carl Hartung, Sameer Agarwal, and Tadayoshi Kohno. Devices that tell on you: Privacy trends in consumer ubiquitous computing. In *Usenix Security*, volume 3, page 3, 2007; and John Krumm. A survey of computational location privacy. *Personal and Ubiquitous Computing*, 13(6):391–399, 2009

Strictly protecting personal information while ensuring the information is usable for research by maintaining the uniqueness in the data is the major concern which was addressed by devising frameworks for secure practices in confidentially collecting and using the location data [Duckham and Kulik, 2006, Tang et al., 2006, Lane et al., 2014]. Some efforts sought to accomplish this task through cryptographic hashing algorithms (Pang, 2007) while others aimed to thwart identification and tracking at the device level by techniques such as MAC randomisation [Gruteser and Grunwald, 2005, Greenstein et al., 2008]. Finally though getting consent of users for the collection and use of such information from their mobile devices is challenging, there is a significantly improved acceptance when the process offers value in return such as discounts and monetary benefits [Kobsa, 2014].

There is opportunity in this area for research in applying the cryptographic solutions along with the privacy preserving frameworks to arrive at methods which can extract useful information out of large personal data while obscuring or anonymising them.

2.2 *Research Trends*

Figure 2.3 shows the volume of research done in this topic since 1980 categorised based on their major themes discussed earlier. We can observe that there are distinct trends in the research over time, which evolved around the development of technology in the last two decades. Until 90s the research was mostly centered around population studies on estimating and interpolating granular spatial and temporal information from larger and cross sectional datasets such as census and sample surveys. The period between 2000-2010 there was interest in potential of the new data generated by the digital revolution. We can categorise this as the 'mobile era' where carrying mobile devices become mainstream. This explosion of research coincided with mobile phones becoming more popular and ubiquitous with population in urban

areas and was around development of methods and techniques to utilise the data generated from them. There were also extensive studies in using the datasets to understand human mobility along with a rising concern in the privacy of the users who's data which are being used for these studies.

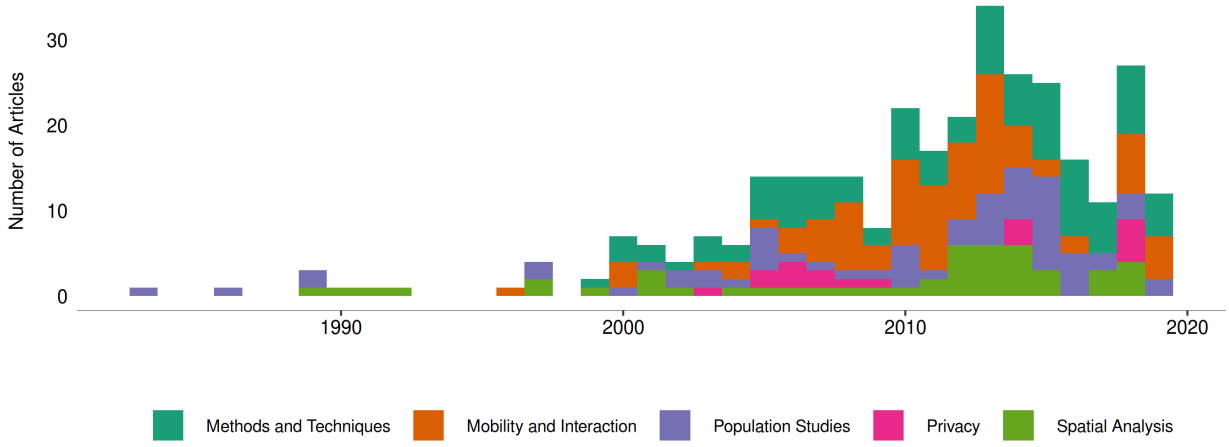

Figure 2.3: Outline of the 'Medium data toolkit' devised to collect, process, visualise and manage the Wi-Fi probe requests data

The release of iPhone in 2008 and the increase in the share of 'smartphones' in the next 10 years sparked the 'smartphone' era. The change made sure that all the mobile devices gaining numerous capabilities such as internet connectivity over Wi-Fi and mobile network, location awareness with global positioning system, movement recognition with accelerometers and connectivity other 'wearable' devices through Bluetooth. This also lead to the digitisation of lifestyle where every aspect of the life being done through these devices over internet while generating huge amount of data on these activities. This sparked the large volume of research on the form and function of space by studying this data and on the dynamics of human population in space and time in the next 5 years. These research were particularly centered around tracking the trajectory of people using the mobile devices they carry with them as the smartphones made it easier to collect the necessary data directly from them rather than depending on a centrally collected datasets from mobile carriers. With the theoretical limit to predictability in human mobility quantified by Song et al. [2010b], the focus on urban mobility has been declining in the past few years which has led to a renewed interest in population studies at a local-local level in real-time. In addition to using the data from the mobile devices, these studies have also been exploring the use of large assemblages of consumer data that are being generated in this connected mobile environment and linking them together to create a fuller picture [12]

Finally, with the increase in use of personal data, there has also been an increase in research regarding the privacy of the users. Along with this, the mobile devices and subsequently the data generated by them are more and more anonymised so that the users cannot be tracked or identified at a personal level. This has given rise to the new trend in research to devise methods to overcome this anonymisation and at the same time research which considers these methods as vulnerabilities and find solutions to make the anonymisation process more robust. There is clear need for methods which

[12] Paul Longley, James Cheshire, and Alex Singleton. *Consumer Data Research*. UCL Press, 2018

anonymise the data sufficiently to protect the identity of the users and at the same time enable us to conduct research in measuring studying population distribution and movement at a granular level.

2.3 *Techniques and technology*

When we look at the literature from the technology perspective, we observe that over the years, the research continuously picks up and applies recent technological developments in the pursuit of understanding the distribution of human activity and population. Figure 2.5 shows the distribution of the research in terms of the main technique/ technology used over the past 50 years. We observe that the earliest attempts started from the exploration of using interpolation and modelling techniques on a broader dataset. As the need for more granular datasets increased there were attempts to devise and utilize bespoke solutions to generate them. When mobile devices became mainstream, the focus shifted to utilize the relevant components of the mobile infrastructure. A significant number of studies were done in utilising data collected from the mobile network, sensors in the mobile devices, especially GPS and Wi-Fi, in addition to the social media content generated from these devices. A detailed account of these studies is given below,

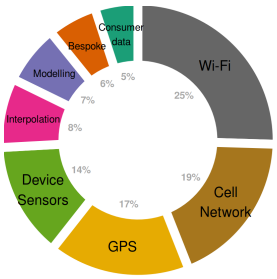

Figure 2.4: Distribution of research across various techniques and technologies

Measured in the number of papers published

2.3.1 *Interpolation and Modelling*

Attempts in using the existing data collected through traditional methods such as census and large scale sample surveys to create spatially and temporally granular and detailed estimates were carried out by applying various interpolation methods such as pycnophylactic, dasymetric interpolation [Tobler, 1979, Mennis, 2003, Mennis and Hultgren, 2006, Hawley and Moellering, 2005, Tapp, 2010, Wismans et al., 2017] along with spatial [Lam, 1983, Martin, 1989, Martin et al., 2015] and temporal interpolation techniques [Glickman, 1986]. These methods along with supplementary data such as remote sensing imagery [Sutton et al., 2001, Chen, 2002] and street networks [Reibel and Bufalino, 2005] were shown to be useful in producing detailed granular population maps at various scales with varying degree of success [Dobson et al., 2000, Bhaduri et al., 2002, Dobson and Fisher, 2003, Bhaduri et al., 2005, 2007]. These approaches have been employed in various applications such as econometric studies [McDonald, 1989], studies on public health [Hay et al., 2005], emergency management [Kwan and Lee, 2005] and flood risk estimations [Smith et al., 2016].

In addition to these interpolation techniques classic modelling techniques can also be used to estimate daytime populations and demographic structure at hyper-local scales [Jochem et al., 2013, Jia et al., 2014], urban scales [Alahmadi et al., 2013, Abowd et al., 2004] and regional scales [Foley, 1954, Schmitt, 1956, Singleton and Longley, 2015, McCormack, 2017]. The granular data created with such modelling techniques are shown to be useful in urban planning and management [Parrott and Stutz, 1999], emergency management [Alexander, 2002, Cutter et al., 2006] and in modelling traffic and transportation [Lefebvre et al., 2013]. These interpolation and modelling techniques along with granular data produced are also used in classifying spatial areas

and hence understanding the structure of cities in general [McMillen, 2001, 2004, Lee, 2007, Arribas-Bel, 2014]. Though being useful, these techniques are still shown to have limitations and uncertainties [Nagle et al., 2014], which mostly arise from the nature of the input data employed. This leads us to the need for more detailed and frequent collection of data.

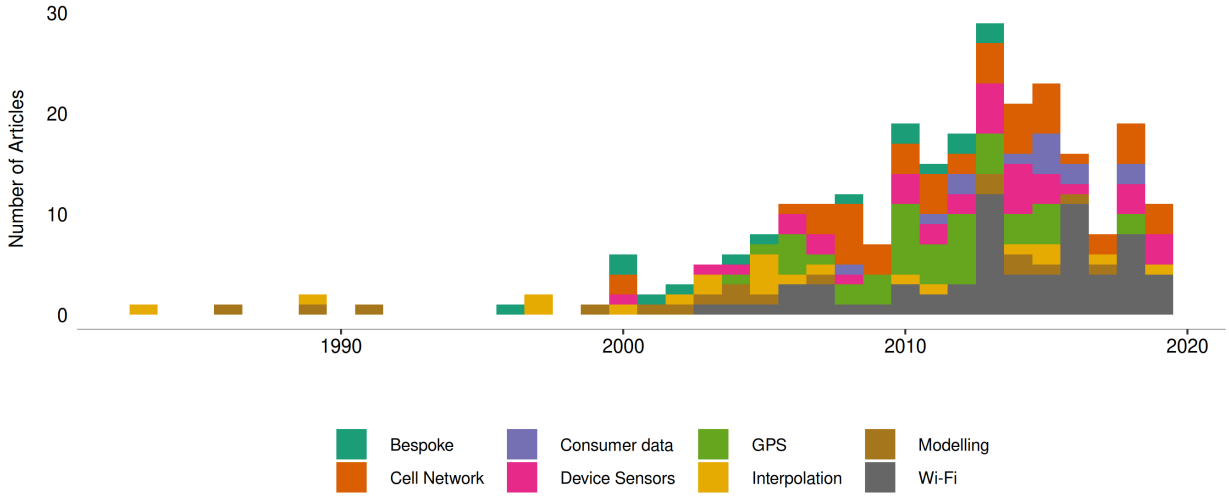

Figure 2.5: Outline of the 'Medium data toolkit' devised to collect, process, visualise and manage the Wi-Fi probe requests data

2.3.2 *Bespoke technologies*

Following this need, there has been efforts to use bespoke or specialised technologies such as cameras [Cai and Aggarwal, 1996, Heikkilä and Silvén, 2004, Kröckel and Bodendorf, 2012], Lasers [Zhao and Shibasaki, 2005, Arras et al., 2008] and radio frequency receivers [Bahl and Padmanabhan, 2000, Yang et al., 2013, Chothia and Smirnov, 2010, Bulusu et al., 2000, Dil and Havinga, 2011] to measure human activity. But the major problem with such solutions is the cost and effort involved in designing and implementing them at urban and regional scales comprehensively. Moreover, being specialised and centralised they tend to be challenging to maintain and update as the technological landscape change. This gives us the need to identify and use techniques which are more general in nature and can be used for longer periods of time which are cheap to install to achieve a more comprehensive coverage.

2.3.3 *Cellular Network*

The rise of mobile phones as ubiquitous personal devices for the broader population has provided us with a viable alternative for collecting data with greater granularity at large scales. Mobile infrastructure consists of both the 'network part', built and managed by the service providers, and the 'user part', which is the phones owned by the users' themselves. The network part, in addition to providing connectivity to the users, also collects information on these devices actively such as communication between the users and passively such as when the phones themselves move from tower to tower. The mobile devices themselves have a variety of sensors such as accelerometer to identify

movement, compass to identify orientation, GPS receiver to deduct geographic position, etc. They also have various communication capabilities such as cellular, Wi-Fi, Bluetooth and Near field communications (NFC) etc. Both of these sensors and communication capabilities can be used as sources of data themselves. With the growth of mobile devices and the infrastructure surrounding it, there has been significant effort in utilising data generated by every component of this complex infrastructure.

The first set of research started to use the cellular network data for urban studies [Jiang et al., 2013, Steenbruggen et al., 2015, Lokanathan and Gunaratne, 2015, Calabrese et al., 2015, Reades et al., 2007]. Even though this approach has been acknowledged to have inherent biases such as ownership bias across particular demographic groups [Wesolowski et al., 2013] the relative advantages such as coverage made them excellent sources of data. Visual exploration of use of such data using interactive interfaces to evaluate quality of service and scenario testing has been tested for the optimisation of public transport [Sbodio et al., 2014]. Such network data with the active and passive information collected from them can be used to create trajectories of people [13], detect their daily routine [Sevtsuk and Ratti, 2010] and classify those routes in terms of function [Becker et al., 2011a]. It was also demonstrated to be useful in understanding overall mobility and flow of people and information [Candia et al., 2008, Krings et al., 2009, Simini et al., 2012, Zhang et al., 2019]. These data can be used to identify asymmetry in flow of people spatially [Phithakkitnukoon and Ratti, 2011], estimate volume and pattern of road usage [Bolla et al., 2000, Wang et al., 2012] and by augmenting the topology to optimise operations [Puzis et al., 2013]. Such datasets have been extensively used in traffic and transportation research to derive origin-destination matrices [Caceres et al., 2007, Mellegard et al., 2011, Iqbal et al., 2014], travel time estimation [Janecek et al., 2012] and traffic status estimation [Demissie et al., 2013, Grauwin et al., 2015].

[13] Johannes Schlaich, Thomas Otterstätter, and Markus Friedrich. Generating trajectories from mobile phone data. In *Proceedings of the 89th annual meeting compendium of papers, transportation research board of the national academies*, 2010

It has been shown that mobile network data can be used to uncover nature of the population such as tourists in specific areas [Girardin et al., 2008] and the interaction between the people in the study area [Campbell et al., 2008]. The structure [Onnela et al., 2007a,b], geography [Lambiotte et al., 2008] and dynamics [Hidalgo and Rodriguez-Sickert, 2008] of such networks have been studied and demonstrated to be useful in predicting their change [Wang et al., 2011, Vajakas et al., 2018] over time. This social networks and their spatio-temporal structure can also be used for classification of land use [Pei et al., 2014, Jia et al., 2018], assessment of spatial patterns [Reades et al., 2009, Steenbruggen et al., 2013] and understanding the broader spatial structure of cities [Louail et al., 2014, Arribas-Bel and Tranos, 2015] and regions [Arhipova et al., 2018]. The data collected from the cellular network when examined at granular levels such as inter-personal communication and economic activity can be used to create estimations of micro area-level population density [Pulselli et al., 2008, Ng et al., 2017] and also the characteristics [Girardin et al., 2009] and the nature of the activity [Phithakkitnukoon et al., 2010]. Aggregated human activity measured from such research in turn can be used to measure and model population dynamics and land use density and mix at broader level [Jacobs-Crisioni et al., 2014, Tranos and Nijkamp, 2015, Tranos et al., 2018]. The spatial patterns thus uncovered can then be applied to

urban planning [Becker et al., 2011b] whilst the temporal patterns uncovered have immense utility for the disciplines such as epidemiology. For example, population influxes measured from changes in mobile network usage can be used to model spread of diseases [Buckee et al., 2015].

Though the mobile network provides much more granular and accurate data than interpolation techniques, it is not without its limitations [Yucel, 2017]. The distribution of network infrastructure usually follows the purposes of service coverage and follows commercial decisions. This introduces systematic biases in the data passively collected through them. Moreover, the data actively collected through them has bias based on the volume of usage of services by the customers which can vary widely spatially, temporally and also based on demography. In addition to this because of the coverage, the data collected from mobile service providers pose immense privacy risk when linked to other sources of consumer data. This makes collection of data directly from the devices using the sensors on the device much more robust in certain cases.

2.3.4 *Mobile Sensors*

The most prominent sensors and capabilities present in mobile devices that can be used for distributed urban sensing are Cellular radio, Bluetooth, Wi-Fi, GPS, accelerometer and compass [14]. Since cellular radio is managed by the cellular network and covered in mobile network data, we explore the research done with other sensors. In contrast to planned actively collected data, data passively collected via a distributed network of general purpose devices tends to be larger and more temporally dynamic. For example, an organised survey conducted every month to understand interpersonal communications between people in a team of 50 will result in a 2500 records a month. The same task is done through collecting data on email communication sent by them will result in a same volume records in a day. The challenges and solutions on collecting and analysing such large-scale longitudinal data are discussed by [Laurila et al., 2012, Antonic et al., 2013]. The real time nature of such data also gives us the opportunity to monitor and understand the city in much smaller temporal scales [Townsend, 2000, O'Neill et al., 2006] and the representativeness of such datasets have also been explored [Shin et al., 2013, Kobus et al., 2013]. Data generated from communication networks can be used to understand the structure of urban systems which are becoming increasingly border-less [15]. Similar to the network based data, it can help in understanding human mobility [Asgari et al., 2013, Amini et al., 2014, Zhang et al., 2014] through mining trajectory patterns [Giannotti et al., 2007] and socio geographic routines [Farrahi and Gatica-Perez, 2010]. It is also useful in various traffic and transportation applications for monitoring roads [Mohan et al., 2008] and estimating traffic [Cheng et al., 2006], uncovering regional characteristics [Chi et al., 2014] and extracting land use patterns [Shimosaka et al., 2014]. Apart from GPS and Wi-Fi, there have been efforts in exploring other possibilities such as Bluetooth for location [Bandara et al., 2004, Becker et al., 2019] and aggregate detected Bluetooth activity to monitor freeway status [Haghani et al., 2010]. There have also been successful implementations of frameworks to predict movement of people by combining Wi-Fi and Bluetooth [Vu et al., 2011]. But owing to shorter range and requirement of active engagement from

[14] Nicholas D Lane, Emiliano Miluzzo, Hong Lu, Daniel Peebles, Tanzeem Choudhury, and Andrew T Campbell. A survey of mobile phone sensing. *IEEE Communications magazine*, 48(9), 2010; and Enwei Zhu, Maham Khan, Philipp Kats, Shreya Santosh Bamne, and Stanislav Sobolevsky. Digital urban sensing: A multi-layered approach. *arXiv preprint arXiv:1809.01280*, 2018

[15] Luca Bertolini and Martin Dijst. Mobility environments and network cities. *Journal of urban design*, 8(1):27–43, 2003

the user where they have to actively start the device pairing process, Bluetooth is much less preferable for large-scale data collection than GPS or Wi-Fi. The research on GPS and Wi-Fi based studies are discussed in more detail below.

2.3.5 *Global Positioning System*

In addition to providing a user's location to applications such as maps and navigation, the GPS capability in mobile devices in tandem with Wi-Fi can also maintain a continuous list of locations visited by the device over long periods of time. It works mostly in the background and requires almost no active input from the user to operate. Though very convenient for collecting data, due to the privacy risks associated with it, GPS is often one of the resources in a device that requires explicit user permission to be accessed. The concepts and methodologies for collecting such data were set out by [Asakura and Hato, 2004] and there have been attempts to collect this rich data from volunteers at a large scale along with ancillary data [Kiukkonen et al., 2010] and provide a location based service application for the collection of data [Ratti et al., 2006, Jiang and Yao, 2006, Ahas and Mark, 2005].

The accuracy, convenience and being designed for navigation makes GPS one of the most used technologies for mobility studies [16]. It has been used to analyse and understand individual mobility patterns [Neuhaus, 2009, Shin et al., 2010], which have been shown to have a high order of regularity in spite of the complexity [Brockmann et al., 2006, Song et al., 2010a]. There have been efforts to use this regularity to predict the future location of people [Monreale et al., 2009, Calabrese et al., 2010]. The limitations of predictions have also been quantified [Song et al., 2010b]. There have been successful efforts in extracting behaviours and patterns from such trajectory data [Liu et al., 2010, Cho et al., 2011, Hoteit et al., 2013, Pappalardo et al., 2013] along to understand individual patterns from large assemblages [Giannotti et al., 2011, Calabrese et al., 2013] and vice versa [Wirz et al., 2012]. In traffic and transportation, GPS trajectory from mobile devices is used to estimate [Calabrese et al., 2011] and expand [Jing et al., 2011] OD matrices, detect the mode of travel [Gong et al., 2012, Rossi et al., 2015] and calibrate existing spatial interaction models [Yue et al., 2012] . Since the data is collected at the device level and depends on the activity of the individual, it can be de-anonymised to reveal the nature of the owner of the devices. The possibilities of detecting the activity of the individual from trajectory information is demonstrated by [Liao, 2006, Krumm, 2007]. Patterns [Jiang et al., 2012] and structures in routines [Eagle and Pentland, 2009] can be extracted from these trajectories and can be used for socio geographic analysis of the population [Licoppe et al., 2008, Chen et al., 2018]. It can also utilised in classification of the population at a particular location at a given time [Pappalardo et al., 2015]. Being inherently spatial and activity driven, GPS trajectories have been shown to be useful to identify [Bao et al., 2012], characterise [Wan and Lin, 2013] and automatically label [Do and Gatica-Perez, 2014] significant places of interest. It can also be used for land use detection [Toole et al., 2012, Zhang et al., 2018], classification [Jiang et al., 2015] and the study of urban morphology [Kang et al., 2012]. These GPS trajectories have been shown to be useful in estimating population dynamics at local level and within short durations during social events [Calabrese et al.,

[16] Marta C. González, Cesar A César A. Hidalgo, Albert-László Barabási, Marta C Gonzalez, Cesar A César A. Hidalgo, and Albert-Laszlo Barabasi. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, 2008. ISSN 0028-0836. DOI: 10.1038/nature06958. URL http://www.nature.com/nature/journal/v453/n7196/full/nature06958.html{%}5Cnhttp://www.nature.com/nature/journal/v453/n7196/pdf/nature06958.pdf

2010, Kim and MacEachren, 2014, Deville et al., 2014]. When combined with other data sources can be useful to understand relationship between spatial areas [Long and Thill, 2015].

From the literature we see that GPS is one of the most precise and accurate user side methods of collecting location of mobile devices. In addition, the data collected is well understood and collection methodologies can be scaled up with minimum resources. That being said, it is also well established that urban sensing methods using GPS of mobile devices has problems of enhanced risk of breach of privacy when executed passively and need explicit user engagement when executed actively.

2.3.6 Wi-Fi

Wi-Fi is a wireless network connection protocol standardised by IEEE [2016]. It is a distributed server-client based system where the client connects to access points (AP). Every device in the network has a unique hardware specific MAC address, which is transmitted between the device and AP before the connection is made. The key feature of Wi-Fi infrastructure is that the network is distributed and the APs can be set up and operated by anyone locally unlike mobile networks. Since they are primarily used for Internet service provision, the protocol has priority for continuity of connectivity so the devices constantly scan for new and better connections. This is done through a probe request, which is detailed in later sections. With this background we can see that Wi-Fi provides a fair middle ground between an entirely network driven approach such as cellular network to an entirely user driven approach such as GPS. Since the network infrastructure is distributed and deployed for Internet it offers near complete coverage, is very resilient, and can encapsulate and reinforce civic space in cities [17].

[17] Paul M. Torrens. Wi-fi geographies. *Annals of the Association of American Geographers*, 98(1):59–84, 2008. DOI: 10.1080/00045600701734133

Though Wi-Fi is a location less technology, there are reliable methods to triangulate the location of the device by the signal strength and the known locations of APs [He et al., 2003, Moore et al., 2004, LaMarca et al., 2005, Dinesh et al., 2017, Lin and Huang, 2018]. This can overcome the usual shortcoming of GPS, which struggles for precision and accuracy in indoor and densely built environments [Zarimpas et al., 2006, Kawaguchi, 2009, Xi et al., 2010]. Utilising this, we can easily and quickly estimate trajectories of the mobile devices just using the Wi-Fi communication the device has with multiple known APs [Xu et al., 2013]. This can be used similar to the GPS trajectories to understand individual travel patterns [Kim et al., 2006, Rekimoto et al., 2007, Sapiezynski et al., 2015], crowd behaviour [Abedi et al., 2013, Mowafi et al., 2013, Shu et al., 2017], vehicular [Lu et al., 2010] and pedestrian movement [Xu et al., 2013, Fukuzaki et al., 2014, Wang et al., 2016, Taylor et al., 2019]. It can also be used in transportation planning and management to estimate travel time [Musa and Eriksson, 2011, Håkegård et al., 2018] and real time traffic monitoring [Abbott-Jard et al., 2013a].

Being a general network protocol designed to be used by mobile devices, Wi-Fi devices relay a range of public signals known as probe request frames on regular intervals throughout its operation, for the purpose of connecting and maintaining a reliable and secure connection for the mobile device [Freudiger, 2015]. These signals can be captured using inexpensive customised hardware,

non-intrusively and in turn to be used for numerous applications. In addition to a uniquely identifiable MAC address, these signals include a range of other information which when combined with the temporal signatures of the signals received can help us understand the nature and identify the devices which are generating these signals. These device/user fingerprinting techniques are demonstrated by [Franklin et al., 2006] and [Pang et al., 2007b] and the unique MAC addresses and associated information can successfully track people across access points [18], their trajectories [Musa and Eriksson, 2012], the relationship between them [Cheng et al., 2012, Barbera et al., 2013, Cunche et al., 2014] and predict which of them will be most likely to meet again [Cunche et al., 2012]. Using the semantic information present in these probe requests it is possible to understand the nature of these users at a large scale [Di Luzio et al., 2016]. Using the received signal strengths from pre placed devices we can monitor the presence and movement of entities that are not even carrying a Wi-Fi enabled device [Elgohary, 2013].

Because of the security and privacy risks posed by the Wi-Fi protocol's use of hardware based MAC address, various methods to strengthen the security have been proposed [Pang et al., 2007b, Greenstein et al., 2008]. The randomisation of MAC addresses has become more mainstream in mobile devices with the introduction of it as a default operating system behaviour in iOS 8 by Apple Inc. Since MAC randomisation is not a perfect solution [Mathieu Cunche, 2016] there have been numerous attempts to fingerprint unique devices from the randomised anonymous information present in the probe request frames for the purposes of trajectory tracking and access point security. The methods used are decomposition of OUIs where detailed device model information is estimated by analysing an already known dataset of OUIs [Martin et al., 2016]; Scrambler attack where a small part of the physical layer specification for Wi-Fi is used [Bloessl et al., 2015]; and finally, the timing attack where the packet sequence information present in the probe request frame is used [Matte et al., 2016, Cheng and Wang, 2016]. A combination of these methodologies has been proven to produce de-anonymised unique device information from randomised MAC addresses [Vanhoef et al., 2016]. In addition to tracking, Wi-Fi probe requests can be aggregated to uncover the urban wireless landscape [Rose and Welsh, 2010] and used to reveal human activity at large scales [Qin et al., 2013], pedestrian numbers in crowds [Schauer et al., 2014, Fukuzaki et al., 2015] and also counting people in hyper local scales such as queues [Wang et al., 2013]. With enough infrastructure we can aim to generate a real-time census of the city [Kontokosta and Johnson, 2016] and also predict the amount of time a device will spend around the sensor as well [Manweiler et al., 2013]. Similar to GPS data this can be used as an additional control layer for interpolation techniques such as map merging [Erinc et al., 2013]. A comparison of various approaches was done by [19] where through experiments on a telecom operator dataset, it was showed that using network-driven mobile phone location data is more advantageous compared to the widely used event-driven ones.

[18] Mathieu Cunche. I know your mac address: targeted tracking of individual using wi-fi, 2014. ISSN 22638733

[19] Fabio Pinelli, Giusy Di Lorenzo, and Francesco Calabrese. Comparing urban sensing applications using event and network-driven mobile phone location data. In *Mobile Data Management (MDM), 2015 16th IEEE International Conference on*, volume 1, pages 219–226. IEEE, 2015

2.3.7 *Consumer data*

In addition to the direct data from the sensors themselves the content generated from the mobile devices such as social media data or smart-cards [20] can provide a viable proxy for estimating the level and nature of human activity. The use of geo-located tweets on the study of small-area dynamic population estimation [Ordonez and Erath, 2012, Marchetti et al., 2015, McKenzie et al., 2015, Lansley and Longley, 2016b], geo-demographics [Bawa-Cavia, 2011, Longley et al., 2015, Lansley and Longley, 2016a] and global mobility [Hawelka et al., 2014] has been thoroughly explored. These data sources are shown to be useful in social sciences [Crane and Sornette, 2008], abnormal event detection [Chae et al., 2012] and analysing urban environments [Sagl et al., 2012]. It can also be used as a control layer for interpolation techniques we discussed earlier [Lin and Cromley, 2015].

[20] Chen Zhong, Michael Batty, Ed Manley, Jiaqiu Wang, Zijia Wang, Feng Chen, and Gerhard Schmitt. Variability in regularity: Mining temporal mobility patterns in london, singapore and beijing using smart-card data. *PLoS ONE*, 11(2), 2016. ISSN 19326203. DOI: 10.1371/journal.pone.0149222

2.4 *Research Gaps and Opportunities*

In this section we summarise the previous sections to find out the best possible technology for further research and discuss the research gaps and opportunities available to us. Table 2.1 summarises the above discussion to evaluate all the relevant technologies that can be used for the data collection and analysis for the study of human activity at a granular level.

Table 2.1: Evaluation of different technologies or approaches that can be used for data collection. **independence from secondary data collected by a third party.*

Technology	Interpolation	Bespoke	Cellular	GPS	Wi-Fi
Coverage	Local	City	All	Local	All
Certainty	Very Low	High	Medium	High	Medium
Independence*	Low	Very High	Low	Medium	High
Intrusiveness	Low	Medium	High	High	Medium
Granularity	Very Low	Very High	Medium	High	High
Ease of Collection	Medium	Low	Medium	Low	High
Scalability	Medium	Low	High	Medium	High
Privacy Risk	Low	Medium	High	High	Medium

We can observer that Wi-Fi offers the best possible technology in terms of flexibility and scalability for data collection through mobile devices at an individual level while posing some risk to privacy of participants and involves uncertainty regarding the field of measurement.

From the literature search we can summarise that there is a considerable opportunity in the collection and analysis of mobile phone based data for measuring hyper-local, spatio-temporal dynamics of human activity. The potential for research gaps broadly fall under the following areas,

2.4.1 *Granular population analysis*

Opportunity 1: *Design and collection of national/regional, longitudinal, grass root level data set which enables study of population both spatially and temporally.*

Previous research in this area of study has been limited to either national/ regional level studies using centrally collected night time population data such as censuses or to area level studies conducted with mobile devices based technologies. Though there were some efforts in collecting and using mobile phone data at national/ regional level we have never been presented with

such unprecedented level of data available now. The explosion of consumer data both publicly available and privately held presents previously unseen opportunity and also limited by the privacy concerns that arise with them. There is an immense opportunity to collect and standardise a large national level dataset which closely linked to the population distribution and movement in an anonymised way which then can be used to understand the distribution of population and its change. There is a need to extend such effort longitudinally which can give us insights in to the change of such phenomenon in time. This has the potential to enable us to ask broad questions such as,

- What are the trends in the footfall in UK?
- What are the daily rhythm of different cites?
- How much a weather event affect economy of a region?

Such dataset, in conjunction with other consumer data sources, in addition to augmenting each other to improve their quality, can vastly improve our understanding of the structure and dynamics of population.

2.4.2 *Device fingerprinting*

Opportunity 2: *Developing models and methods to identify anomalies in the data and underlying events causing them*

The privacy concerns about the data collected from personal mobile devices has pushed the industry and users to find ways to anonymise the data generate over the last decade. All the mobility studies recording user trajectories across space and time are rendered infeasible with the cryptographic hashing and randomisation techniques employed by the devices. This along with progressive legislation such as General Data Protection Regulation have severely constrained the data available for mobility research. As we see later, even the estimation of ambient population is limited by these developments. Though there are prior research in this area, most are conducted from security perspective evaluating the robustness of the randomisation/obfuscation procedure. These research focus around de-anonymising the obfuscated data to recreate the personal information from them thus demonstrating vulnerabilities and risks for the users. In this context, there is a clear gap for research in to methods to rather carry out fingerprinting of these devices using patterns in the data to create useful information from them without actually de-anonymise the data. This can lead to production of data-sets and methodologies which will enable us to,

- Get accurate estimation of ambient populations.
- Understand the movement of the population in space and time.

2.4.3 *Event Detection*

Opportunity 3: *Developing models and methods to identify anomalies in the data and underlying events causing them*

Having granular spatio-temporal data on population at an area level also enables us to look at the activity of people at this scale. For example, the spike in Wi-Fi activity at a certain area at a certain time can illuminate us with a specific event that is happening in that area. Thought research have been done on this area using social media data, a longitudinal data-set collected using mobile technology can enable us to formalise the models needed to identify anomalies, quantify the causation of such anomalies to real world event. The

near real-time aspect of such research also provides us opportunities in fields of disaster management, smart cities etc. There are opportunities to ask questions such as,

- How did the tube strike affected London?
- What were the hot spots for New years celebration?
- What was effect of a road closure in specific part of the city?

2.4.4 *Pedestrian Flow*

Similar to the device fingerprinting, estimating and understanding pedestrian flow in the street network has immense opportunities since the anonymisation of mobile devices has taken off. Even when the problem of the identifying unique fingerprints of users in the data has not been solved, there is a need to understand the overall performance of the street network in terms of pedestrian flow just from the precise, granular data available, especially when the data source is as unstructured and noisy as the Wi-Fi sensors. This problem can he approached in two ways,

Opportunity 4: *Estimating flow of pedestrians in the street network from Wi-Fi data*

1. Probabilistic approach - Where the relationship between the temporal change in volumes at locations are modelled. For e.g. how much and how often the footfall counts at one location mirrors/ follows other location gives us an idea of how many pedestrians move from one location to the other.
2. Interpolation - Where the relationship between the locations are defined in terms of multiple variables such as how similar they are, how close they are etc. These relationships can in turn used to build a graph of locations and use this graph as a source to interpolate other locations.

2.4.5 *Nature and Change of Places*

Though there are extensive research in using ambient population and people's movement to understand the form and function of the space, the mobile technologies have introduced the opportunity to remove the subjectivity from them. With access to granular and long-term data sets, we can hope to look into the how the places have changed over time and how the external factors such as policy and economy has affected them. There are opportunities to ask questions such as,

Opportunity 5: *Using long term data to detect the nature and change of form and function of a place.*

- How does UK's exit from EU affect its high streets?
- Has a specific area has become more or less vibrant?