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

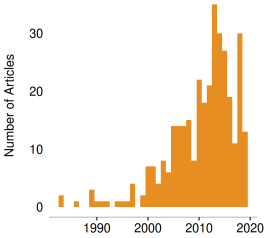

Figure 2.1: Growth of research in the areas of understanding distribution and dynamics of human activity since 1980.

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 [2,3]. 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 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

[3] 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?

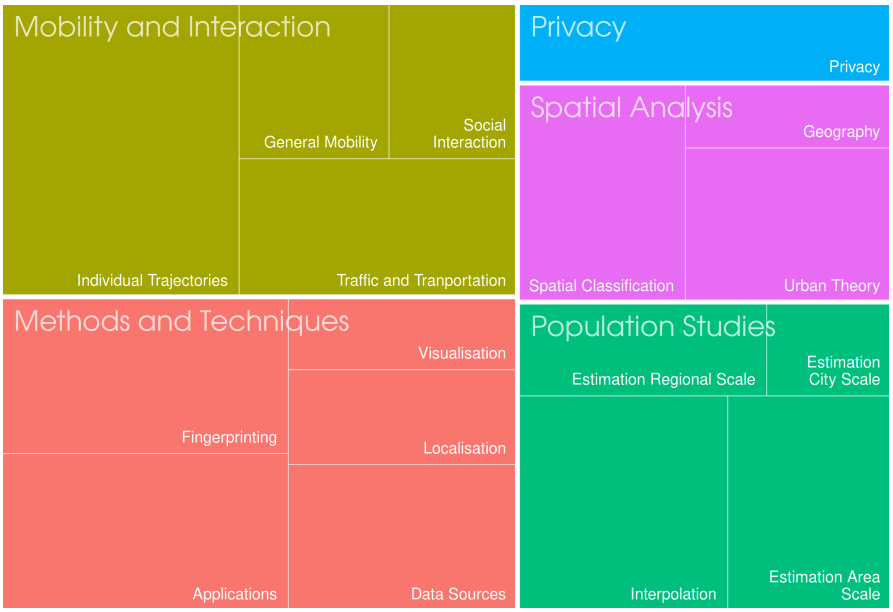

Figure 2.2: Tree-map showing the volume of research conducted under each major themes and their sub-themes.

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 [4].

[4] 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, 2016] 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*

This is one of the major areas of research which have significantly benefited from the decentralised collection of data at a granular level [5]. 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.

[5] 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 gen-

erated 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 [6]. 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, 2018b,a, Jia et al., 2019, Nikitin et al., 2019].

[6] 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]. 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 [Jiang and Yao, 2006, Liao, 2006]. 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 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 [MacEachren and Kraak, 2001, Hallisey, 2005]. 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.

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 [7]. 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].

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

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 [8] was accompanied not only by optimism in its potential [Thomas, 2001] but also by the questions raised on the challenges in handling the diverse, large scale,

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

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 [Bettencourt, 2014] especially with advancement of new technologies [Steenbruggen et al., 2013] and possibility of distributed, crowdsourced data collection [Lokanathan and Gunaratne, 2015].

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 [9]. 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].

[9] 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.

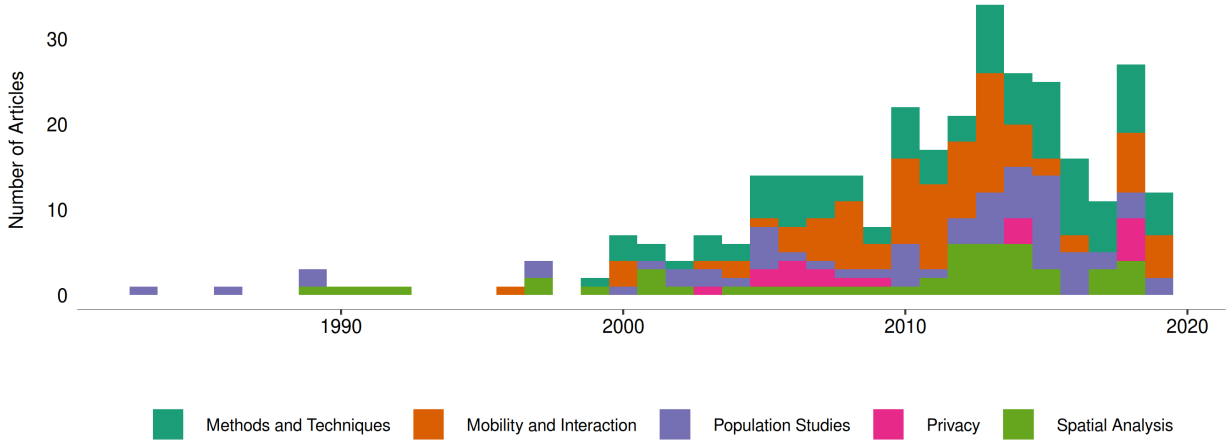

Figure 2.3: Outline of the 'Medium data toolkit' devised to collect, process, visualise and manage the Wi-Fi probe requests data

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 [10], 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 [11]

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 anonymise the data sufficiently to protect the identity of the users and at the

[10] Chaoming Song, Zehui Qu, Nicholas Blumm, and Albert-László Barabási. Limits of predictability in human mobility. *Science*, 327 (5968):1018–1021, 2010. ISSN 0036-8075, 1095-9203. DOI: 10.1126/science.1177170

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

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 the research tend to pick up and apply recent technological developments in the pursuit of understanding the distribution of human activity and population spatially and temporally over the past two decades. 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 the available coarse data and as the need for more granular datasets increased there were attempts to devise and utlize 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 WiFi, in addition to the social media content generated from these devices. A detailed account of these studies is given below,

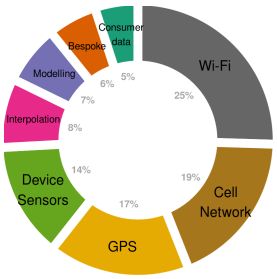

Figure 2.4: Growth of research in the topic

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] along with spatial [Lam, 1983, Martin, 1989] 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, **?**]. 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