1
Introduction

Our understanding of built environment and cities in terms of their form and function have evolved significantly since the early 20th century. What started as an field of research focused on the physical form of the spaces and places later moved towards modelling them as a function of the population that lives in them. Rather than viewing the built environment as infrastructure that need to be built, maintained and managed independently, they have been increasingly viewed as the manifestation of the distribution and dynamics of population embedded in them. This paradigm change further broadened in the later part 20th century to include economic and social activity of the population that happens along the fabric of these built environment. Moreover, with the dawn of the information age around the turn of the millennium, the built environment could be viewed as the tangible result of information exchange that happens beneath them where cities can be seen as high density clusters of information exchange rather than a place with concentration of physical infrastructure such as buildings and roads. This information revolution has not only changed researchers' understanding of the underlying forces of the built environment, but also changed how they approach the task of measuring, analysing, modelling and managing them. It provided them with numerous new technologies, methodologies, tools and, most importantly, unprecedented availability of the comprehensive, granular data generated from the fundamental functions of the built environments. Availability of these data and tools has turned numerous disciplines upside down where the research, rather than taking a top-down 'systems' approach, now tackles the problems using bottom-up, 'data first' approach.

We are currently in the age of 'data deluge' where the amount of data generated in the world far exceeds the capacity to analyse and derive insights from them. This deluge of data has been accelerated to an extent that 90% of all the data ever generated in the world has been created in the last 2 years. With the popularisation of wearable technologies and internet of things, this trend is not expected to change any time soon. Moreover, almost all of the day to day activities of people such as banking, bill payments, ticketing in public transport, taxi hire, social communications, fitness tracking have been getting digitised and generate large amount of unstructured data in the process. Collecting data for research has changed from a highly structured, designed endeavour to

low cost, scraping activity where data about the population are collected and stored across the world regularly but never to be used. Most of the data collection activity has also become 'passive' i.e collected without any effort from the side of participants vastly increasing the capacity of the data collection process leading to the trend for 'big data' where everyone with troves of data looking to use advanced and automated data-mining techniques to extract values from them. The above two phenomena - attempt to model the physical environment as a function of information exchange and the unprecedented availability of data has lead to quite significant volume of research where a variety of data sources have been used to understand variety of aspects of the built environment. For example, Cycling in cities is studied from the data generated by data from the cycle-share docks, functional regions of a country can be derived from call detail record data and population and demography studied through data from social media such as twitter.

This frenzy of data generation and its use is not without its pitfalls as well. One of the major disadvantages in these attempts is the risk they pose to the privacy of the users whose data is being collected and analysed. With personal mobile devices becoming mainstream, almost every data point generated has a person behind it. The rush into the information age and the social media platforms happened at a much faster rate before their ramifications to privacy of the study subjects could be understood properly. Even when the data collected does not contain personal information directly, most of them can reveal personal and potentially sensitive information when linked with other sources of data. For example, anonymised cycle ride trajectories itself might not be interesting information, but combined with other movement datasets such as taxi trips and payment information can disclose the identity along with residences of the people who use them. This has prompted major concern and backlash from the users and subsequently regulators in the past decade. These concerns are addressed in industry and research using both technology and regulation. From the technology perspective, all the stakeholders who generate, collect or use the data try to use cryptography to either anonymise, obscure or encrypt any personal information as much as possible. In terms of regulation, legislation efforts such as General Data Protection Regulations (GDPR) have been introduced to influence the behaviour of these stakeholders by introducing comprehensive rules and punitive measures for non compliance. Though both these approaches ultimately try to protect user's privacy and personal information, they also pose one of the greatest challenge to research which use passively collected user data. In the next 5 years it can be expected that every freely available data source would be protected from the unfretted use which we see today. Wherever this protection is not possible, it can be expected that the data would obscured or anonymised to remove any risk to the privacy of users making it imperative that research adapts to these changes by looking for ways to overcome the challenges posed by them.

In addition to the privacy concerns this deluge of data introduces significant technological challenges as well. The have been extensive

research and development from both academia and industry under the umbrella of 'big data' developing theory, methods and tools to tackle the challenges posed by such large assemblages of data to derive meaningful insights from them. This hype for 'big data' promises to solve a lot of the technological and logistical challenges incurred in many disciplines but it needs to be mentioned that it introduces additional overheads in terms of cost and resources. In cases of research projects, blindly jumping on the big data bandwagon without consideration has the potential to cause more problems than advantages. The discipline of Geography, especially geographic information systems and science has a long tradition of dealing with large datasets since the beginning of the field and the recent deluge of data causes issues due to their complexity, latency and lack of structure rather than their sheer volume. Hence it is extremely important to be mindful while adopting the contributions from the big data discourse for research so that the solutions are implemented where the real problems are. There needs to be careful consideration when choosing or designing the methods, tools and frameworks which are used to address the unique requirements of the 'new age' data sources. Moreover there needs to be an inquiry into a framework for how these considerations identified and addressed.

In this context, this research sets out to work on the opportunities presented above by collecting large volumes of passively generated data, solve uncertainties in them arising due to their high variability and their pursuit to protect the privacy of the users and finally produce useful information regarding the distribution and dynamics of footfall in the country. We start with a broad and systematic literature survey on the topic of 'distribution and dynamics of human activity' in Chapter 2 where major themes of research were identified and their evolution in the past 30 years were identified along with the development of technologies that were employed. The literature search resulted in the identification of the best possible data source for further research along with opportunities available for further research. Having identified Wi-Fi as one of the most promising technologies for our research, in chapter 3 we look at Wi-Fi specification in detail especially the 'probe request' mechanism. In addition to studying the standards and specification to identify relevant data available in them, the chapter also discusses the design and implementation of series of small experiments to capture and analyse data in real-world. Three sets of initial experiments were conducted and results from them were used to conduct a longer and broader 'pilot study' collecting data from locations across London. The chapter introduces the 'Smart Street Sensors'(SSS) project - a national level collecting Wi-Fi data at large number of retail locations across UK. The chapter concludes with a detailed evaluation of all the data collected from these experiments and the SSS project in terms of the bias, noise and uncertainties present in them. Chapter 4 deals with processing the Wi-Fi data to remove the identified uncertainties to produce 'clean' and continuous information on volume of footfall at the corresponding locations. The emphasis on not using personal data or methods that can potentially reveal personal information is firmly held through out the

chapter. While in section 4.1 a framework for evaluating the 'bigness' of the data is discussed and subsequently a 'data toolkit' for processing them is devised, in section 4.2 methods to clean the data into realistic estimate of footfall have been discussed. In section 4.3 both the toolkit and methods were put together to architect a 'data pipeline' which digests the continuous stream of data from SSS project and outputs meaningful footfall numbers efficiently. Chapter 5 is an assortment ideas and prototypes applying the output of the research where four major themes are discussed - creating an index for footfall across United Kingdom, detecting events occurring in the context using the changes in the volume of footfall, estimating the flow of pedestrian between locations from the changes in footfall volumes and identifying the nature and relationship between places along with possibilities for further research.

The potential of creating such detailed, long term, national level footfall volumes produced by this research is immense. Such information can be of the major components in building a 'smart city' where availability of detailed, real-time data on the state built environment and its use is vital. It can also help us in our pursuit to accomplish a real-time census of people and their movement in the city. It can not only provide us with snapshots of the state of a retail areas but also help in measuring, modelling and manipulating them in real-time as a dynamic systems which respond to interventions. We can even link these footfall volumes to other sources of data such as commercial consumer datasets and public transport statistics to build a comprehensive picture on the health and efficiency of the city systems. Availability of such dataset can revolutionise academic research in fields such as urban planning, public policy and urban management where the effect of interventions could be objectively measured and analysed. Though this research did not try to explore the applications of this footfall data in detail, it hopes to serve as a solid basis for further studies in a various academic disciplines such as geography, business management, risk management, spatial analysis and computer science which can employ the data to either derive insights about the locations and context or use them as a reference/training dataset source for validating methods and tools. Availability of such national level data on footfall volumes spanning continuously over years can also have massive impact in industries such as retail, transportation, real estate and information technology. As this research has a significant bias towards retail locations, the outputs can be especially of massive value for various stakeholders in the retail industry,

1. *Retailers* can get detailed information on when and where their customers shop which can lead to more efficient business operations.
2. *Customers* can be informed on the popularity of places and when to visit them.
3. *Landlords* can achieve a way to objectively evaluate their properties' values based on their location and also time.
4. *Local Authorities* can be enabled to monitor and manage the health of their retail areas over longer periods of time.