Estimating real-time highstreet footfall from Wi-Fi probe requests

Balamurugan Soundararaj[a], James Cheshire[a] and Paul Longley[a]

[a]Department of Geography, University College London, United Kingdom.

ARTICLE HISTORY
Compiled June 29, 2018

ABSTRACT
The accurate measurement human activity with high spatial and temporal granularity is crucial for understanding the structure and function of built environment. With increasing mobile ownership, the Wi-Fi 'probe requests' generated by mobile devices can act as an cheap, scalable and real-time source of data for such establishing such measures. One of the major challanges in using Wi-Fi technology for estimating human activity is eliminating biases caused by an uncertain field of measurement and changing technological landscape without compromising the privacy of the users of the mobile devices. In this paper we demonstrate that, with the application of class intervals and a novel graph based technique, we can overcome the challanges and reliably measure real-time pedestrian footfall at retail highstreets.

KEYWORDS
Highstreet footfall; Wi-Fi Probe requests; Sensors; MAC Randomisation

1. Introduction

New and developing technologies today provide the infrastructure over which movements and interactions can be measured and monitored in the 'sentient city' (Amin and Thrift 2017). These include mobile phone networks, which can triangulate user locations relative to networks of masts, use of GPS to locate users of social media services, and WiFi beacon connectivity to access the Internet. These technologies offer differing levels of spatial precision, where mobile telephony and WiFi generally being less reliable and offer lower precision than GPS to the end users, while being more advantageous for broader mobility studies (Pinelli *et al.* 2015).

There has been considerable research into the usefulness of these technologies to understand patterns of movement in cities in near real time (Candia *et al.* 2008, Gonzalez *et al.* 2008, Calabrese *et al.* 2013). Most of this research has focused upon technical specification of accuracy or precision (Song *et al.* 2010, Lane *et al.* 2010), with somewhat less attention devoted to the ways that the characteristics of the technologies and of their human users conspire to create possible bias in the inclusiveness of usage patterns. This is perhaps most manifest in the plethora of papers that analyse the distribution and content of publicly available Twitter social media feeds, which likely constitute an extremely biased and self-selecting subset of users that geo-enable their posts, drawn from a user base that itself bears no identified correspondence to any known background population (Lansley and Longley 2016). Even analysis of mobile

CONTACT Balamurugan Soundararaj. Email: s.bala@ucl.ac.uk

phone data, usually derived from industry players that have significant market share and user bases that may be representative of local populations, may exclude groups such as tourists from distand origins or subscribers to third party services that share distinctive characteristics (Di Luzio *et al.* 2016) These examples illustrate the issues that underpin the assembly and analysis of consumer data, which can be considered as a distinctive class of Big Data that arise from the interactions between humans and customer-facing organisations such as retailers, domestic energy suppliers, transport providers and suppliers of social media and communications (Longley *et al.* 2017). Consumer data account for an ever-increasing real share of all of the data that are collected of citizens, but a fundamental characteristic of consumer-led markets is that no single provider has a monopoly in market provision, and therefore issues of market share and segment generate bias in analysis. The source and operation of this bias is unknown in the absence of extensive and context sensitive attempts to triangulate consumer data with data of known provenance relating to clearly defined populations (Lansley and Longley 2016). In similar ways to other classes of Big Data, consumer data are best throught of as digital 'exhaust', or a by-product created by or harvested from consumer transactions.

In this paper, we evaluate the value of data collected from a Great Britain wide network of 800 devices, installed in order to characterize the footfall patterns of a scientifically balanced sample of retail centres. These devices are located in shop windows and record the probes emitted by mobile phones and other WiFi enabled devices. These are deemed to be consumer data because devices carried by consumers routinely probe for a consumer service, specifically a WiFi connection. Monitoring the probes from such devices provides an indication of the presence of their users, regardless of whether or not Internet connectivity is established. Our core motivation is to appraise the usefulnessof WiFi probe records harvested from our network of sensors in order to indicate levels of pedestrian activity. More broadly still, in our future research we intend to classify the nationwide network of footfall profiles as part of a programme of research to understand the form and functioning of retail areas at a time of far-reaching structural change for the retail industry.

To this end, it is important to first undertake a thorough conceptual and technical appraisal of our consumer data source. In technical terms, screening of information present in the 'probes' and classifing them based on their characteristics is essential in order to remove the ones emitted by devices that do not indicate pedestrian activity, such as network enabled printers and other fixed devices. Related to this, a method to figerprint Wi-Fi probes is necessary to remove probes from individuals' devices that in conceptual terms should not be considered part of footfall – as when, for example, an employees is seated in an office within range of the sensor device. A calibration of sensor measurement is also essential on two grounds: first, individuals may carry multiple devices, or no device at all; and second, the positioning and orientation of the sensor in the retail unit may lead to systematic over- or under-enumeration. These sources of bias in measurement must be accommodated by manual recording of footfall at each location and generalization of these sample survey results to all locations and time periods. As we describe in detail below, manual validation of the data needs to be undertaken in parallel with technical profiling of the mix of consumer devices that probe our sensors, since the effectiveness of data cleaning procedures discussed in this paper differ between individual locations and configurations.

2. Background and Method

In the past decade Wi-Fi has emerged as the most commonly used technology in providing high speed internet access to mobile devices such as smartphones, tablets and laptops in public and private spaces. This has resulted in multiple Wi-Fi networks being available at almost every location in dense urban environments. Traversing through this overlapping mesh of Wi-Fi networks, modern mobile devices with Wi-Fi antennae regularly broadcast a special type of signal known as 'Probe Requests', in order to discover Wi-Fi networks available to them. This helps these devices to connect and switch between the WiFi networks seamlessly.

Probe requests are low level signals standardised by IEEE 802.1b/g specification (IEEE 2013) as the first step in establishing a Wi-Fi based connection between two devices and is implemented in any Wi-Fi capable device irrespective of the manufacturer or the model. This ubiquity and standardisation make them an excellent source of open, passive, continuous, and wireless data generated by Wi-Fi capable devices present at any given time and location. Considering the unprecendented levels of mobile device ownership in recent years, we can in turn use this data to understand the population distribution in highly dynamic urban environments with high spatial and temporal granularity (Freudiger 2015, Kontokosta and Johnson 2017).

While a Wi-Fi based method to collect data offers us various advantages such as, easy scalability and efficiency in terms of cost and time, It also introduces few systematic biases, uncertainities in the collected data along with the serious risk of infringing on the privacy of the mobile users. In this paper, using a set of probe requests and manual counts collected at various high street locations across London, we demonstrate that pedestrian footfall at these locations can be estimated with considerable precision and accuracy while protecting the privacy of the pedestrians.

Though WiFi is a 'location-less' technology, there are reliable methods to triangulate the location of Wi-Fi enabled mobile devices by the known locations of APs and the signal strength reported by them from the mobile devices (He *et al.* 2003, Moore *et al.* 2004, LaMarca *et al.* 2005). 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 WiFi communication the device has with multiple known APs (Sørensen and Berglund 2006) which can be used similar to the GPS trajectories to understand individual travel patterns (Kim, 2006;(Rekimoto *et al.* 2007, Sapiezynski *et al.* 2015), crowd behaviour (Abedi *et al.* 2013, Mowafi *et al.* 2013), vehicular (Lu *et al.* 2010) and pedestrian movement (**?**Fukuzaki *et al.* 2014, Wang *et al.* 2016). Such data can also be used in transportation planning and management to estimate travel time (Musa and Eriksson 2011) and real time traffic monitoring (Abbott-Jard *et al.* 2013).

Using techniques demonstrated by Franklin *et al.* (2006) and Pang *et al.* (2007) and using the information present in the probe requests one can also successfully track people across the sensors or access points collecting them (Cunche *et al.* 2014), infer their trajectories (Musa and Eriksson 2012), and even model interactions between them (Cheng *et al.* 2012, Barbera *et al.* 2013, Cunche 2014) such as predicting which of them are most likely to meet again (Cunche *et al.* 2012). Using the semantic information present in these probe requests it even is possible to understand the nature of population at a large scale (Di Luzio *et al.* 2016).

Though extensive research has been carried out on this subject with feasible and favorable results, in recent years, one of the major challenges faced in such attempts has been the increasing attempt by mobile phone manufacturers to protect the users' privacy by anonymising the globally identifiable portion of the probe requests, (Greenstein *et al.* 2008). There are various methods which have been devised to overcome this anonymisation process such as decomposition of OUIs where detailed device model information is estimated by analysing an already known dataset of OUIs (Martin *et al.* 2016); Scrambler attack using a small part of the physical layer specification for WiFi (Vo-Huu *et al.* 2016, 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 globally unique device information (Vanhoef *et al.* 2016, Martin *et al.* 2017). These approaches usually result in serious risk of infringement of the privacy of the users of the mobile devices by revealing their identifiable personal information.

There is a clear gap in the research for exploring methodologies which enable us to estimate the number of unique mobile devices from a set of anonymised probe requests, without the need to reveal their original MAC addresses. Such technique has various applications in numerous fields such as uncovering the urban wireless landscape (Rose and Welsh 2010), revealing human activity at large scales (Qin *et al.* 2013), estimating pedestrian numbers in crowds (Schauer *et al.* 2014, Fukuzaki *et al.* 2015) and even counting people in hyper local scales such as queues (Wang *et al.* 2013) With enough infrastructure to collect such information we can aim to generate a real-time census of the city (Kontokosta and Johnson 2017). It has also been demonstrated by (Pinelli *et al.* 2015) through series of experiments on a telecom operator dataset that using such network-driven approach is more advantageous compared to the widely used event-driven approaches. With this background we set out to device and implement a methodology to reliably estimate human activity such as pedestrian footfall from Wi-Fi probe requests without infringing the privacy of the users involve.

3. Methodology

The primary aim of this research is to enable us to collect a series of probe requests and process them into an usable pedestrian footfall count. We do this by using a Wi-Fi receiver to collect probe requests broadcasted by mobile devices, filtering out the background noise and aggregating them based on the device that generated them. We begin by looking at the characteristics of probe requests in detail, device a methodology to collect these probe requests in public areas, examine the systemic biases and uncertainties in the data collection method and device data processing methods to overcome these challanges. Finally we compare the processed footfall counts to the ground truth recorded by primary surveys.

Probe requests are a special type of management packets broadcast by Wi-Fi enabled devices as part of the various functions such as scanning for available access points (AP), quick geo-location by triangulation based known APs, etc. These are broadcast by all Wi-Fi enabled devices regardless of the manufacturer, type or model of the devices though there is some variation on the frequency and the information transmitted through them. In some cases, such as Android devices, these are broadcast even when the Wi-Fi functionality has been turned off by the user. Thus these signals can be used to reliably identify the presence of Wi-Fi enabled mobile devices.

Being a first step of connection initiated by the mobile device, these packets have information regarding the characteristics of the mobile device itself. Some of the key

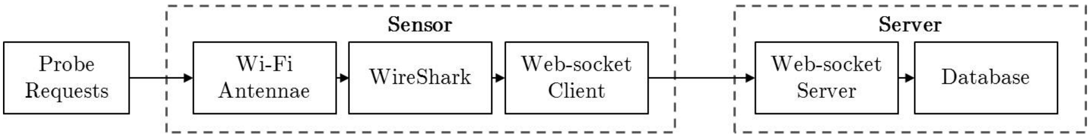


Figure 1. Schematic diagram showing the process of collecting probe requests using the sensor

information we can infer from these requests are,

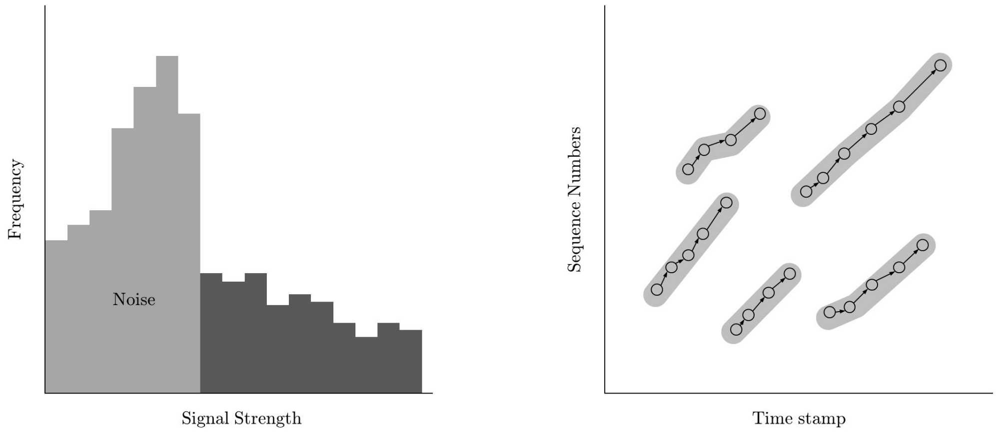
(1) **Media Access Control (MAC) address** which is an unique identifier for the wireless hardware of the mobile device,
(2) **Sequence number** of the request for the mobile device to keep track of the responses,
(3) **Timestamp** at which the request was received by the AP,
(4) Total **length** of the request in number of bits, and
(5) The **strength of the signal** which transmitted the request.

The MAC address is the primary unique identifier for the mobile device. It has two parts, first part is called an Organisation Unique Identifier (OUI) which gives information about the manufacturer of the device and the second is unique to the device. The MAC address can be randomised (hence non unique) and is marked as such. Though sequence number and length of the packet are not strictly unique, we hypothesize that we can use them to estimate unique devices.

Data collection was done with the help of custom sensors built from modifing the Smart street sensor (CDRC 2016) hardware and updating them with custom software. The sensor is essentially a Raspberry Pi connected with Wi-Fi and 3G antennae. It keeps the Wi-Fi module in 'Monitor' mode and uses the open source software - wireshark cite to passively collect all packets sent to 'broadcast', marked with type - 'management' and subtype - 'probe requests'. The MAC address in these probe requests is anonymised using a cryptographic hashing algorithm and transmitted through 3G connection to a central database via web-sockets protocol, where it is stored in a PostgreSQL database for further analysis. A overall schematic of the data collection process is shown in Figure 1. The ground truth on number of pedestrian footfall was recorded using a purpose built Android application cite.

The next step after collecting data was to estimate the footfall or pedestrian activity from them. We identified the following major challanges which arise from our collection methodology.

(1) **Background noise** - since the extent to which Wi-Fi signals travel differs subject to various factors such as interference and humidity, it is close to impossible to restrict our data collection to a finite area of interest. This can lead to a signicant background noise at certain locations. E.g. a phone shop or a bus stop located next to the study area can increase the number of probe requests received by the sensor.
(2) **MAC randomisation** - The mobile devices in recent years have been using randomised 'local' MAC addresses for probe requests to protect the users from being tracked. This makes it impossible to tell if the probe requests are being sent by the same mobile device which is being stationed next to the sensor. This along with the previous problem can further increase the magnitude of error by several fold.
(3) **Mobile ownership** - Since the rate of mobile ownership can vary widely across geography and demography, we cannot assume that every mobile device trans-



(a) Distribution of signal strengths showing the filtering of background noise

(b) Clustering probe requests as nodes in a graph using increasing sequence numbers

Figure 2. Schematic diagrams explaining the methods for filtering by signal strength and clustering using sequence numbers

lates to one pedestrian footfall. In addition to this, there is a long term overall increase in mobile ownership which may lead to the number of probe requests collected overtime.

We propose the following methods to tackle each of these challages.

3.1. *Filtering with Signal Strength*

One of the clues that we can use to estimate the distance between the mobie device and the sensor is the strength of the signal received by the sensor. The obvious approach here is to try and establish a relationship between the signal strength and distance first and use this to filter out the unwanted probe requests. This approach has numerous pitfalls and uncertainities since the decay of signal strength with distance is not always constant. It varies with atmospheric conditions, presence of obstructions between the source and target, the nature of these obstructions and the strength (power level) of the source transponder. This severely limits our ability to establishing a simple conversion between reported signal strength and distance. There is a need for a method which takes in to account these variables across varous locations.

We hypothesise that in configurations where a specific source of background noise is at a constant distance, there must be a distinct break in the number of probe requests reporting signal strength corresponding to that distance. For example, if there is a phone shop next to our sensor where hundreds of phones regularly send probe requests there should be a sharp rise of number of probe requests with reported singal strength corresponding to the distance between the sensor and the phone shop at any given set of conditions as shown in Figure 5. We could identify these breaks in the data using traditional one diamensional classification algorithms such as Jenks natural breaks, k-means, quantile and hierarchical clustering, etc. Since we are only looking for the break in the data and not for absolute values, the methodology should apply for all the variations due micro site conditions thus reducing the overall noise in the collected data.

3.2. *Clustering with sequence numbers*

Since our primary unique identifier - MAC adddress, is being anonymised by new devces, we need to find other information present in the probe request for a unique identifier. Obvious approach here is to establish a factor of randomisation and adjust the counts for these probe requests based on this factor. We found this aproach not feasible, since the proportion of devices which randomise the MAC addresses increases over time. There is also a wide variation in the frequency at which the devices randomise the MAC addresses and the method used for the process. This leads us to look for a more generalisable approach which is independent of the device model.

From our initial analysis we found that OUI, length of the packet and sequence number of the packet being the most promising information to acheive this. First we divide our dataset into sets of probe requests with randomised and non-randomised MAC addresses and keep the MAC address as the unique identifier for the latter set. For randomised ones we futher divide them in to sub categories based on their OUIand length of the packet. Since the length tends to stay unique to specific models of devices we are left with the task of identifying the unique mobile devices from within these distinct models.

The proposed algorithm creates a graph where the probe requests represented the nodes, and links are created between them based on the following rules:

- A link could go only forward in time.
- A link could go from low to high sequence numbers.
- A link could exist between nodes with a maximum time difference of α - time threshold.
- A link could exist between nodes with a maximum sequence number difference of β - sequence threshold.
- A node could have only one incoming link and one outgoing link, which is the shortest of all such possible links.

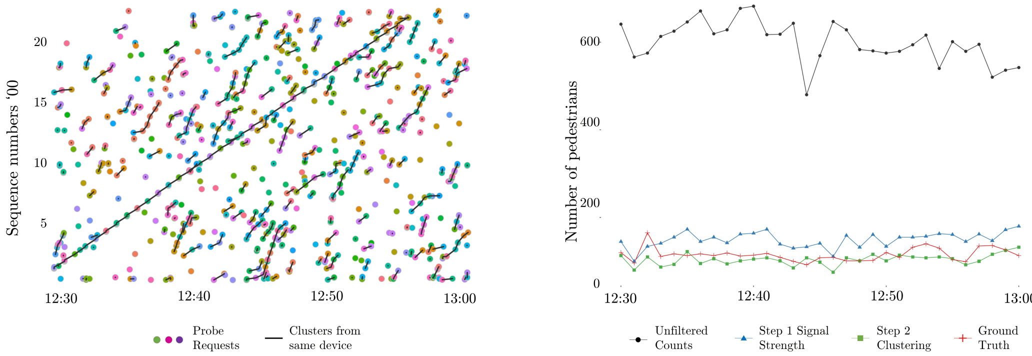
The nodes were then classified based on the unique connected component they belong to as shown in Figure 5. This classification was assigned as the unique identifier for the anonymised probe requests in the place of MAC address. Though the recycling of sequence number after 4000 leads to multiple classifications reported on single device, the maginitude of error is greatly reduced.

3.3. *Calibrating with ground truth*

Since mobile phone ownership is an external uncertainty to our study and could arise from variety of spatio, temporal and demographic factors, we propose to solve this using external source of information. We hypothesize that an adjustment factor could be arrived at for each location of data collection, comparing the sensor based counts and ground truth and it can be used to adjust the data reliably to reflect the ground truth in absolute numbers for the future. This calibration can be carried over periodically and the frequecy of which will improve the quality of the estimation.

4. Pilot Study

To start we designed a small pilot study with the aim to validate the classification and clustering methodology against the scale and complexity of data collected on a open



(a) Clustering probe requests based on increasing sequence numbers present in them.

(b) Comparision of counts after filtering with manual counts

Figure 3. The results of the clustering process and the comparison with ground truth

public area such as a retail highstreet. We also want to find the algorithm which is best suited for classification of signal strengths. The data was collected at Oxford Street in London on 20 December 2017 from 12:30 to 13:00 hrs, where Wi-Fi probe requests sets were collected using the Wi-Fi sensor described earlier and pedestrian footfall was manually recorded using the Android app. Being located at one of the busiest retail locations in the United Kingdom, the WiFi sensor captured approximately 60,000 probe requests during the period, and 3,722 people were recorded manually.

When we just aggregated the probe requests by their MAC address for every minute, the mean error between the sensor counts and the manual counts was observed to be on average 425%. This suggested that there was a large amount of noise in the data which might have included signals from devices outside the area where the manual count was conducted and anonymised probe requests with different MAC addresses from a few devices stationed next to the sensor. We then classified the probe requests as "high signal strength" and "low signal strength" using 'k-means' classification algorithm which resulted in the lowest mean error percentage closely followed by 'quantile'. The cut-off point or threshold for the collected data was -71 dBm. We eliminated the noise from devices outside the area of interest by removing all the probe requests which reported a "low signal strength" below the threshold. We found this process of filtering highly effective and reduced the mean minute by minute error to 30%.

We then move on to assign an unique field identifying the mobile device generating the probe requests. For the 45% of the probe requests which were not randomised, we kept the MAC addresses as the unique identifier. For the rest of the probe requests we applied the clustering algorithm to assign the unique identifier. With the help of the known stationary device (the mobile device used by the surveyor to record pedestrians manually) and through trail and error, We found the suitable time threshold, α to be 15 seconds and threshold for sequence numbers, β to be 60. Figure 4 shows the clustering of probe requests. The dots are individual probe requests and the red lines connect probe requests within the same cluster which are generated by the same mobile device. We finally combine both normal and anonymised probe requests, aggregate them based on their unique identifier which further reduced the mean error to -18%.

The minute by minute comparision of counts from the filtering processes along with the ground truth is shown in Figure **??** From the pilot study, we find that both classification and clustering methods work on complex real world data and results in a final pedestrian counts within a error of 20%. We also find 'k-means' and 'quantile'

Table 1. Locations where sensors were installed

ID	Location	Type	Installation notes
1	Camden High Street	Phone Shop	Bus stop in front
2	Central St.Giles Piazza	Restaurant	Seating area on both sides
3	Holborn Underground Station	Information Kiosk	Overlooks station entrance
4	Brunswick Center	Fast Food Restaurant	Has seating area on one side
5	The Strand	Tea Shop	Has phone shop next door

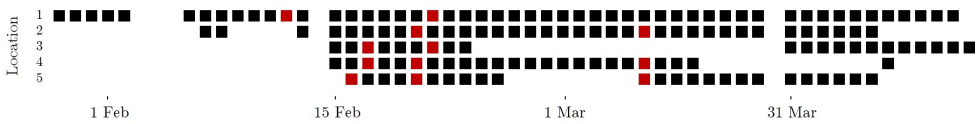


Figure 4. Days when the sensors were active at the corresponding location. The red square shows that manual data collection was also done.

are best algorithms for classifying signal strengths.

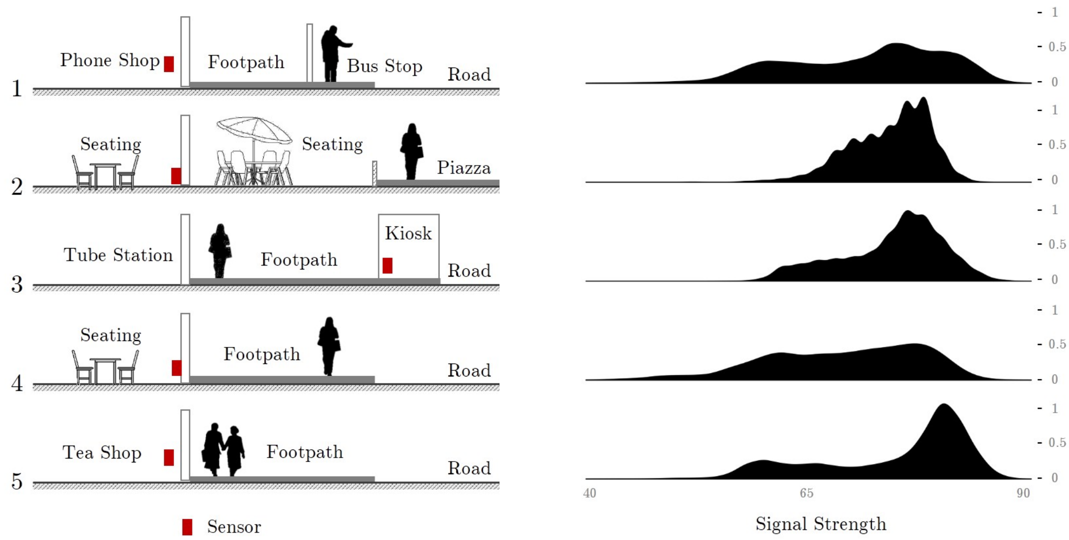
5. Main Study

The main study was designed to validate the results of the pilot study over different locations at different times. We choose five different locations across central London, to install the sensors and collect data for a long period of time. We also carry out manual counting on these locations along with this across different time of the day. We then apply the filtering based on signal strength and sequence numbers and compare with manual counts and evaluate the effectiveness of the process with the mean error per minute on these locations. Finally we calculate the ajustment factor for the first interval of manual counts and check if that works on the consecutive intervals.

The locations where the data were collected are shown in the table. The location are chosen for their variety of configuration and sources of noise. Location 1 is the 'cleanest' of while location 2 is the one with the most complexity. The configuration, installation and data collection schedule is shown in the figure.

Though data was collected for many continous days, for the purposes of comparing with ground truth we just consider the only the data from sensors corresponding to those sessions. We have 12 sets of data over 6 different days. We have atleast two set manual counts for each location for verification of calibration.

First we see how the distribution of the signal strength varies with location and configuration. The density plot for signal strength is shown alogn with configuration in figure. We can see that the signal strength distribution shows distinct patterns of high and low when the installation is that there is a clear distant source of noise but this distinction get more and more obscure as we move towards difficult installations. For example, location 2 is almost a normally distributed noise as it is too far to pick up any pedestrians but location 5 with a clear view of footpath and a phone shop next door shows clear distinction between the two. Intuitively the classfication algorithm should give us better results in the latter. It is important to note that we are dealing with relative singal strengths, this can vary with location and time of the analysis but we should be still be able to differentiate signal from noise. We run the kmeans classification algorithm and filter out the probe requests which are randomised and have signal strengths less than the second break (or the threshold). We then count the number of Unique MAC addresses present in every minute and remove MAC addreses



(a) Installation configuration of sensors at various locations

(b) Density distribution of signal strength (lower values show higher signal strength)

Figure 5. Distribution of signal strengths across locations

which reappear within 30 minutes of previous appearance. We then compare this with the minute by minute aggreagation for the manual counts and find the average error per minute for the sensor count. The results are shown in the table.We see that the location 2 has the most error 400% confirming our intuition while surprisingly location 3 has the least error -3%. We also see that the error follows the complexity of the installation. It is also very promising that this method alone reduces our margin of error by a 50 - 100% For some practical purposes which do not require absolute numbers, this should be sufficient. e.g. Indexes of activity and change dashboards and short term trends identification.

We see that the success of the signal strength filtering depends on the how we tackle the problems in the installing the devices. The more the device in installed in a way that the field of measurement is distinct from its surroundings (however noisy they might be) the more successful this process is. Closer sources error causes overcounting as we see in 2, 4 but at the same time highly close field of measurement with no big source of error can cause undercounting such as in 3. The complexity of the installation and uniformity of surroundings near the sensors directly coorrelate to the MAPE. The most effective way of measuring pedestrian is to have the sensor next to have the sensor next to clear area of pedestrian activity and in case of absence of any major source of noise, we should not do the signal strength filtering at all. We can even note the major possible sources of noise around the device and use it to fiugre out the right way of partitioning the data e.g. location 2 and 3 should have different pattern of signal strength filtering depending on their configuration.

In terms of filtering by sequence numbers we see that the the filtering works in all the locations. It has the similar effect as the signal strength and brings us furthe closer to the manual count. Finally we calculate the adjustment factor from the first 15 minutes and apply that for the rest of the interval. We get the counts down to almost 10% on good locations. The locations which still over count are the ones with seating next to them (2 and 4). We also see that the trends are preserved with the filtering process.

Table 2. Data collected at different locations and the correspondin mean absolute percentage error after each filtering process

Sensor	Signal strength threshold (-dBm)	Adjustment factor	MAPE without any cleaning (%)	MAPE after filtering signal strength (%)	MAPE after filtering sequence numbers (%)	MAPE of final adjusted counts (%)
1	-70	1.25	259	22	-13	9
2	-74	0.51	928	396	206	55
3	-72	1.60	87	-19	-31	10
4	-70	0.88	498	142	52	33
5	-72	0.80	473	84	38	11

Finally to summarize, through a simple raspberry pi based wifi sensors, we collected probe requests sent by user's mobile devices, cleaned them based on the reported signal strength, clustered/ fingerprinted them based on sequence numbers, aggregated them to get pedestrian counts. We then compared it to the read pedestrian count and developed adjustment factors which work for short term future.

6. Conclusion

This paper has made a novel contribution to the measurement of activities in retail centres in real time. To this end, it has described the collection and processing of a novel consumer Big Dataset that enables valid measures of levels of footfall activity which has been scaled across a Great Britain wide network of sensors. In both conceptual and technical terms, it illustrates the ways in which passively collected consumer data can be 'hardened' to render them robust and reliable by using related procedures of internal and external validation. Internal validation addresses the issues of screening out device probes that do not indicate footfall, and the further screening of device probes to 'fingerprint' the effects of MAC randomization. External validation then entails reconciling adjusted counts with the footfall observed at sample locations. The latter procedure makes it possible to generalise from locations at which manual footfall surveys are conducted to all others in the system, and to develop a classification of device locations that are more or less susceptible to noise generation.

It is important to note that the filtering process work based soley on the information present in the probe requests and their temporal distribution. This ensures that although the mobile devices were uniquely identified, there was no further personal data generated by linking the probe requests to the users of the mobile devices. This method essentially gave us a way to estimate the footfall in real-time without identifying or tracking the mobile devices themselves.

This Wi-Fi based footfall counting methodology offers a large number of applications and benefits for real time spatial analysis. Since Wi-Fi based sensors are inexpensive and the data model is scalable, it is possible to use this methodology for a large network of sensors to gather granular data on pedestrian footfall. A snapshot showing week's worth of precise footfall in area around charrig cross, London is shown in the **??** to demonstrate the potential for such dataset. Projects such as SmartStreetSensors (CDRC 2016), may utilise this methodology to overcome the challenges introduced by the implementation of MAC address randomisation. Such precise and granular data also enables us to confidently model the pedestrian flow in urban road networks, and will be an indispensable tool in the smart city framework. It can also be used to understand and classify geographical areas based on the spatio-temporal distribution of the volume of activity in them which we intend to research in future.

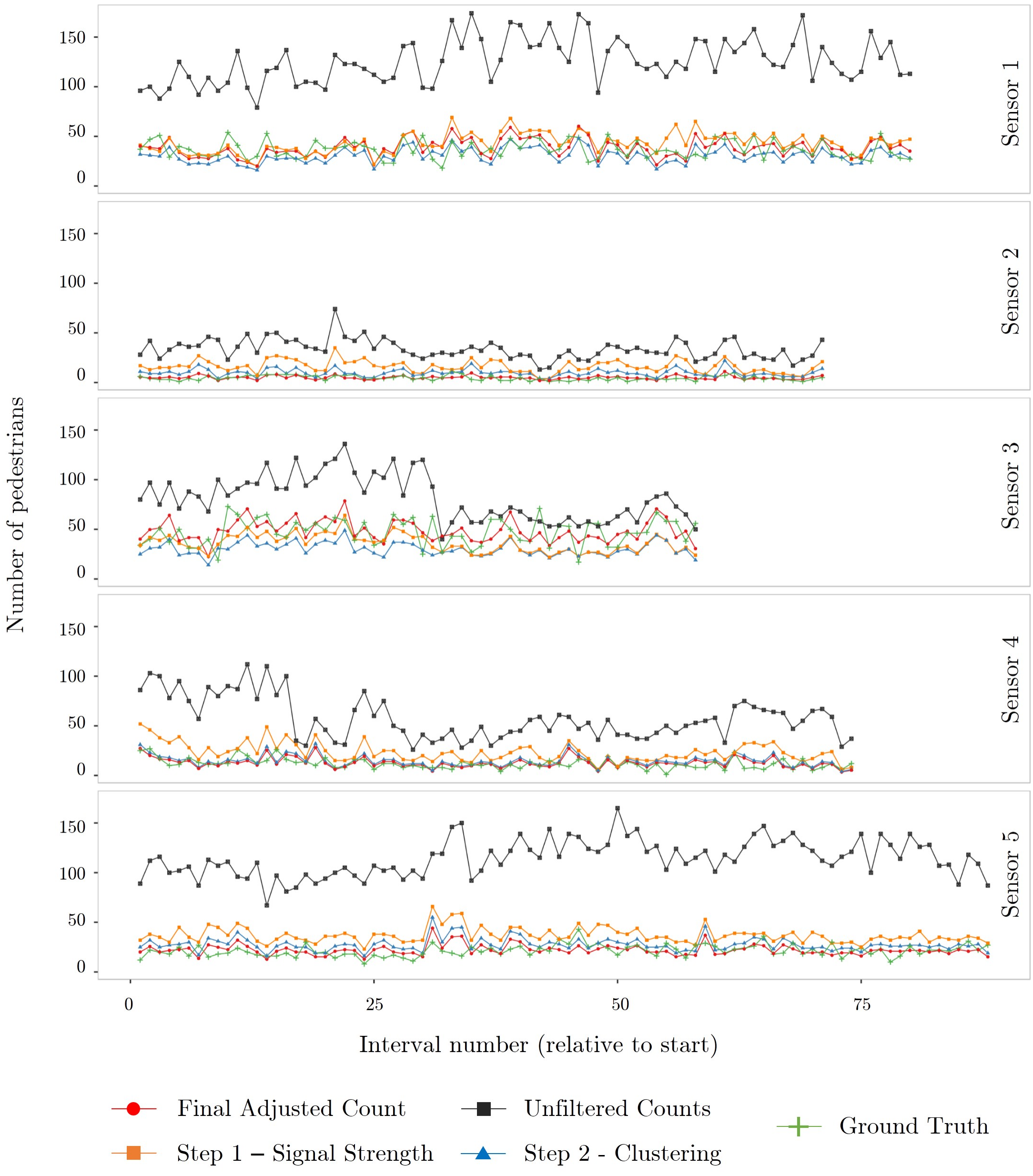


Figure 6. Comparison of all the counts in teh sensors

Acknowledgement

Authors would like to acknowledge the contributions of the Local Data Company in facilitating the data collection process and Dr Roberto Murcio for his inputs.

References

Abbott-Jard, M., Shah, H., and Bhaskar, A., 2013. Empirical evaluation of bluetooth and wifi scanning for road transport. *In*: *Australasian Transport Research Forum (ATRF), 36th, 2013, Brisbane, Queensland, Australia.* 14.

Abedi, N., Bhaskar, A., and Chung, E., 2013. Bluetooth and wi-fi mac address based crowd data collection and monitoring: benefits, challenges and enhancement.

Amin, A. and Thrift, N., 2017. *Seeing like a city.* John Wiley & Sons.

Barbera, M.V., *et al.*, 2013. Signals from the crowd: uncovering social relationships through smartphone probes. *In*: *Proceedings of the 2013 conference on Internet measurement conference.* ACM, 265–276.

Bloessl, B., *et al.*, 2015. The scrambler attack: A robust physical layer attack on location privacy in vehicular networks. *In*: *Computing, Networking and Communications (ICNC), 2015 International Conference on.* IEEE, 395–400.

Calabrese, F., *et al.*, 2013. Understanding individual mobility patterns from urban sensing data: A mobile phone trace example. *Transportation research part C: emerging technologies*, 26, 301–313.

Candia, J., *et al.*, 2008. Uncovering individual and collective human dynamics from mobile phone records. *Journal of physics A: mathematical and theoretical*, 41 (22), 224015.

CDRC, 2016. Smart street sensor project. goo.gl/E4tR8o. [Online; accessed 31-January-2018].

Cheng, L. and Wang, J., 2016. How can i guard my ap?: non-intrusive user identification for mobile devices using wifi signals. *In*: *Proceedings of the 17th ACM International Symposium on Mobile Ad Hoc Networking and Computing.* ACM, 91–100.

Cheng, N., *et al.*, 2012. Inferring user relationship from hidden information in wlans. *In*: *MILITARY COMMUNICATIONS CONFERENCE, 2012-MILCOM 2012.* IEEE, 1–6.

Cunche, M., 2014. I know your mac address: Targeted tracking of individual using wi-fi. *Journal of Computer Virology and Hacking Techniques*, 10 (4), 219–227.

Cunche, M., Kaafar, M.A., and Boreli, R., 2012. I know who you will meet this evening! linking wireless devices using wi-fi probe requests. *In*: *World of Wireless, Mobile and Multimedia Networks (WoWMoM), 2012 IEEE International Symposium on a.* IEEE, 1–9.

Cunche, M., Kaafar, M.A., and Boreli, R., 2014. Linking wireless devices using information contained in wi-fi probe requests. *Pervasive and Mobile Computing*, 11, 56–69.

Di Luzio, A., Mei, A., and Stefa, J., 2016. Mind your probes: De-anonymization of large crowds through smartphone wifi probe requests. *In*: *Computer Communications, IEEE INFOCOM 2016-The 35th Annual IEEE International Conference on.* IEEE, 1–9.

Franklin, J., *et al.*, 2006. Passive data link layer 802.11 wireless device driver fingerprinting. *In*: *USENIX Security Symposium.* vol. 3, 16–89.

Freudiger, J., 2015. How talkative is your mobile device?: an experimental study of wi-fi probe requests. *In*: *Proceedings of the 8th ACM Conference on Security & Privacy in Wireless and Mobile Networks.* ACM, 8.

Fukuzaki, Y., *et al.*, 2014. A pedestrian flow analysis system using wi-fi packet sensors to a real environment. *In*: *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication.* ACM, 721–730.

Fukuzaki, Y., *et al.*, 2015. Statistical analysis of actual number of pedestrians for wi-fi packet-based pedestrian flow sensing. *In*: *Adjunct Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2015 ACM International Symposium on Wearable Computers.* ACM, 1519–1526.

Gonzalez, M.C., Hidalgo, C.A., and Barabasi, A.L., 2008. Understanding individual human mobility patterns. *nature*, 453 (7196), 779.

Greenstein, B., *et al.*, 2008. Improving wireless privacy with an identifier-free link layer protocol. *In*: *Proceedings of the 6th international conference on Mobile systems, applications, and services*. ACM, 40–53.

He, T., *et al.*, 2003. Range-free localization schemes for large scale sensor networks. *In*: *Proceedings of the 9th annual international conference on Mobile computing and networking*. ACM, 81–95.

IEEE, 2013. Part 11: Wireless lan medium access control (mac) and physical layer (phy) specifications; amendment 4: Enhancements for very high throughput for operation in bands below 6 ghza.

Kawaguchi, N., 2009. Wifi location information system for both indoors and outdoors. *In*: *International Work-Conference on Artificial Neural Networks*. Springer, 638–645.

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

LaMarca, A., *et al.*, 2005. Place lab: Device positioning using radio beacons in the wild. *In*: *International Conference on Pervasive Computing*. Springer, 116–133.

Lane, N.D., *et al.*, 2010. A survey of mobile phone sensing. *IEEE Communications magazine*, 48 (9).

Lansley, G. and Longley, P., 2016. Deriving age and gender from forenames for consumer analytics. *Journal of Retailing and Consumer Services*, 30, 271–278.

Longley, P., Cheshire, J., and Singleton, A., 2017. *Consumer data research*. John Wiley & Sons.

Lu, H., *et al.*, 2010. Vehicle tracking using particle filter in wi-fi network. *In*: *Vehicular Technology Conference Fall (VTC 2010-Fall), 2010 IEEE 72nd*. IEEE, 1–5.

Martin, J., *et al.*, 2017. A study of mac address randomization in mobile devices and when it fails. *arXiv preprint arXiv:1703.02874*.

Martin, J., Rye, E., and Beverly, R., 2016. Decomposition of mac address structure for granular device inference. *In*: *Proceedings of the 32nd Annual Conference on Computer Security Applications*. ACM, 78–88.

Matte, C., *et al.*, 2016. Defeating mac address randomization through timing attacks. *In*: *Proceedings of the 9th ACM Conference on Security & Privacy in Wireless and Mobile Networks*. ACM, 15–20.

Moore, D., *et al.*, 2004. Robust distributed network localization with noisy range measurements. *In*: *Proceedings of the 2nd international conference on Embedded networked sensor systems*. ACM, 50–61.

Mowafi, Y., *et al.*, 2013. Tracking human mobility at mass gathering events using wisp. *In*: *Future Generation Communication Technology (FGCT), 2013 Second International Conference on*. IEEE, 157–162.

Musa, A. and Eriksson, J., 2011. Wiflow: real time travel time estimation using wi-fi monitors. *In*: *Proceedings of the 9th ACM Conference on Embedded Networked Sensor Systems*. ACM, 429–430.

Musa, A. and Eriksson, J., 2012. Tracking unmodified smartphones using wi-fi monitors. *In*: *Proceedings of the 10th ACM conference on embedded network sensor systems*. ACM, 281–294.

Pang, J., *et al.*, 2007. 802.11 user fingerprinting. *In*: *Proceedings of the 13th annual ACM international conference on Mobile computing and networking*. ACM, 99–110.

Pinelli, F., Di Lorenzo, G., and Calabrese, F., 2015. Comparing urban sensing applications using event and network-driven mobile phone location data. *In*: *Mobile Data Management (MDM), 2015 16th IEEE International Conference on*. IEEE, vol. 1, 219–226.

Qin, W., *et al.*, 2013. Discovering human presence activities with smartphones using nonintrusive wi-fi sniffer sensors: the big data prospective. *International Journal of Distributed Sensor Networks*, 9 (12), 927940.

Rekimoto, J., Miyaki, T., and Ishizawa, T., 2007. Lifetag: Wifi-based continuous location logging for life pattern analysis. *In*: *LoCA*. vol. 2007, 35–49.

Rose, I. and Welsh, M., 2010. Mapping the urban wireless landscape with argos. *In*: *Proceedings of the 8th ACM Conference on Embedded Networked Sensor Systems*. ACM, 323–336.

Sapiezynski, P., *et al.*, 2015. Tracking human mobility using wifi signals. *PloS one*, 10 (7), e0130824.

Schauer, L., Werner, M., and Marcus, P., 2014. Estimating crowd densities and pedestrian flows using wi-fi and bluetooth. *In*: *Proceedings of the 11th International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services*. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), 171–177.

Song, C., *et al.*, 2010. Limits of predictability in human mobility. *Science*, 327 (5968), 1018–1021.

Sørensen, R. and Berglund, T., 2006. Location tracking on smartphone using ieee802. 11b/g based wlan infrastructure at itu of copenhagen.

Vanhoef, M., *et al.*, 2016. Why mac address randomization is not enough: An analysis of wi-fi network discovery mechanisms. *In*: *Proceedings of the 11th ACM on Asia Conference on Computer and Communications Security*. ACM, 413–424.

Vo-Huu, T.D., Vo-Huu, T.D., and Noubir, G., 2016. Fingerprinting wi-fi devices using software defined radios. *In*: *Proceedings of the 9th ACM Conference on Security & Privacy in Wireless and Mobile Networks*. ACM, 3–14.

Wang, W., Liu, A.X., and Shahzad, M., 2016. Gait recognition using wifi signals. *In*: *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 363–373.

Wang, Y., *et al.*, 2013. Measuring human queues using wifi signals. *In*: *Proceedings of the 19th annual international conference on Mobile computing & networking*. ACM, 235–238.

Xi, W., *et al.*, 2010. Locating sensors in the wild: pursuit of ranging quality. *In*: *Proceedings of the 8th ACM conference on Embedded Networked Sensor Systems*. ACM, 295–308.

Zarimpas, V., Honary, B., and Darnell, M., 2006. Indoor 802.1 x based location determination and realtime tracking. *In*: *The IET International Conference on Wireless, Mobile and Multimedia Networks (ICWMMN 2006), Hang Zhou, China*. IET.