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 December 3, 2018

ABSTRACT
The accurate measurement of human activity with high spatial and temporal granularity is crucial for understanding the structure and function of the built environment. With increasing mobile ownership, the Wi-Fi 'probe requests' generated by mobile devices can act as a cheap, scalable and real-time source of data for establishing such measures. The two major challenges we face in using these probe requests for estimating human activity are: filtering the noise generated by the uncertain field of measurement, and clustering anonymised probe requests generated by the same devices together without compromising the privacy of the users. In this paper we demonstrate that we can overcome these challenges by using class intervals and a novel graph based technique for filtering and clustering the probe requests which in turn, enables us to reliably measure real-time pedestrian footfall at retail high streets.

KEYWORDS
Pedestrian footfall; Urban sensing; Wi-Fi probe requests; MAC Randomisation

1. Introduction

New and developing 'smart' technologies today provide the infrastructure over which movements and interactions of people can be measured and monitored in the 'sentient city' (Amin and Thrift 2017). This is making it possible to reinvigorate conceptualisations of city as the locus of human activities supplementing night time geographies of residence (Martin *et al.* 2015) with geographies of shopping behaviour (Lloyd and Cheshire 2018), workzone geographies (Singleton and Longley 2018) and studies of movement trajectories (Campbell *et al.* 2008). This is rendering activity-based conceptions of human behaviour central to the analysis of hardship and opportunity in, and around, the smart city (Venerandi *et al.* 2015).

Sentient technologies include mobile phone networks, which can triangulate user locations relative to networks of masts, the use of GPS to locate users of social media services, and Wi-Fi access points providing internet access. These technologies offer differing levels of spatial precision, where mobile telephony and Wi-Fi are generally less reliable and offer lower precision than GPS to the end users, whilst simultaneously being more advantageous for broader mobility studies (Pinelli *et al.* 2015). There has been considerable research into the utility of these technologies to understand patterns

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

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 representing usage patterns across the entire smart city. Analysis of mobile phone data, usually derived from industry players that have significant market share and user bases representative of local populations, may also exclude groups such as tourists from distant 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. Consumer data can be considered as a distinctive class of Big Data which 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 about citizens, but a fundamental characteristic of consumer-led markets is that no single provider has a monopoly in market provision; therefore issues of market share and segment generate bias in the 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 thought of as digital 'exhaust', or a by-product created by, or harvested from, consumer transactions.

In this paper, through a set of experiments, we evaluate the value of data collected a network of 800 devices (CDRC 2016) installed across Great Britain 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 Wi-Fi enabled devices. The data collected from these devices are deemed to be consumer data because devices carried by consumers routinely probe for Wi-Fi connection which is a consumer service. 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 usefulness of Wi-Fi probe requests harvested from our network of sensors as a method of indicating 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 function of retail areas at a time of far-reaching structural change for the retail industry.

To this end, it was important to first undertake a thorough conceptual and technical appraisal of our consumer data source. In technical terms, screening the information present in the 'probe requests' and clustering them based on their characteristics was essential in order to remove those emitted by devices which do not indicate pedestrian activity, such as network enabled printers and other fixed devices. Related to this, a method to fingerprint Wi-Fi probes was necessary in order to remove probes from individuals' devices that in conceptual terms should not be considered part of footfall; for instance when , an employee is seated in an office within range of the sensor device. A calibration of sensor measurement was 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 the generalization of these sample survey results to all locations and time periods. As we describe in detail below, manual validation of the data needed to be undertaken in parallel with the technical profiling of the mix of consumer mobile devices that probed our sensors, since the effectiveness of data cleaning procedures discussed in this paper differ between individual locations and configurations.

2. Background

In the past decade Wi-Fi has emerged as one of 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 network interface 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 Wi-Fi networks seamlessly.

Probe requests are low level signals standardised by IEEE 802.11 specification (iee 2016) for service discovery and is implemented in any Wi-Fi capable device irrespective of the manufacturer or the model. This ubiquity and standardisation makes 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, unlike GPS, the location of the Wi-Fi enabled mobile device cannot be directly inferred from Wi-Fi, there are reliable methods to triangulate the location of mobile devices from the locations of known access points(AP) and the signal strength reported by them (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 (Sørensen and Berglund 2006, Musa and Eriksson 2012) which can be used similar to the GPS trajectories to understand individual travel patterns (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 (Xu *et al.* 2013, 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) along with information present in the probe requests one even model interactions between the users (Cheng *et al.* 2012, Barbera *et al.* 2013, Cunche 2014, Cunche *et al.* 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 their users' privacy by anonymising the globally identifiable portion of the probe requests, (Greenstein *et al.* 2008). Various methods have been devised to overcome this anonymisation process such as estimating the device model information from a known dataset of manufacturers and device behaviours (Martin *et al.* 2016); Scrambler attack using a small part of the physical layer specification for Wi-Fi (Vo-Huu *et al.* 2016, Bloessl *et al.* 2015); and timing attack where the packet sequence information along with information elements 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 breach of privacy of the users of the mobile devices by revealing their identifiable personal information.

There is a clear gap in research for exploring methodologies for estimating the number of unique mobile devices from a set of anonymised probe requests, without the need to reveal their original device information. Such technique has various applications 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 even aim to generate a real-time census of the city (Kontokosta and Johnson 2017). 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 risking the breach of privacy of the users involved.

3. Methodology

The primary aim of this research is to enable us to collect a series of probe requests and process them into a usable pedestrian footfall count. We do this by using a Wi-Fi receiver to collect probe requests broadcast by mobile devices, filtering out the background noise and aggregating them based on the device that generated them. In this section, 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 challenges. 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 APs, quick geolocation 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 so that device can immediately connet to networks when the functionality is switched back on. Since some devices even use the probe requests as a less accurate form of localisation they keep sending probe requests as well. 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 information we can infer from these requests are,

(1) **Media Access Control (MAC) address** which is an name identifying 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) **Time stamp** 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** received by the mobile device.

The MAC address is the primary unique identifier for the mobile device. It has two parts, first part is the Organisationally Unique Identifier (OUI) which gives information about the manufacturer of the device and the second is unique to the device. In modern devices, to protect the users privacy, the MAC address can also be randomised (hence non unique) and is marked as such. Though sequence number of the packet is strictly unique to a mobile device, we hypothesize that we can use them to estimate the number unique devices as demonstrated by (Vanhoef *et al.* 2016), where optional information present in the probe requests - Information Elements (IE) along with the probe requests have been used to fingerprint the devices. This approach has become increasinly difficult as mobile phone manufacturers have severely limited the IEs present in the probe requests thus leading us to explore methods which uses only the sequence numbers. This also affects the established commercial solutions using Wi-Fi probe requests such as Blix, Walkbase, Euclid Analytics, RetailNext etc. There have been another solution proposed by (Hong *et al.* 2018) where the authors tried to solve the similar problem using a hidden markov models based trajectory inference alogorithm but the scope is limited to enclosed, exit controlled public spaces such as shopping malls, railway stations, etc.

Data collection was done with the help of custom sensors built from modifying the hardware used in Smart street sensors (CDRC 2016) and updating them with custom software. The sensor is essentially a Raspberry-Pi device with Wi-Fi and 3G modules. It keeps the Wi-Fi module in 'Monitor' mode and uses the open source software - Wireshark (Combs and Contributors 2018) to passively collect all packets sent to 'broadcast', marked with type as management' and subtype as 'probe requests'. The MAC address in these probe requests is obfuscated at the device level 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. The random salt used in the hashing algorithm was rotated regularly to further mitigate the risk of de-anonymisation of the hash. Though hashing cannot completely ensure anonymisation as discussed in (Demir *et al.* 2014) it can sufficiently obfuscate the data which along with a secure process of data handling gives us reasonable security. A overall schematic of the data collection and storage process is shown in Figure 1. The ground truth on number of pedestrian footfall was recorded using a custom Android application - Clicker (Soundararaj 2018). This app logs accurate timestamps for each time the surveryor records a pedestrian crossing the designated cordon line at the location. In addition to counting the pedestrians manually this also makes the mobile device used constantly probe for networks which also gives us a know device to calibrate our methodology against.

The next step after collecting data was to estimate the footfall or pedestrian activity from them. We identified the following potential uncertainties which arise from our

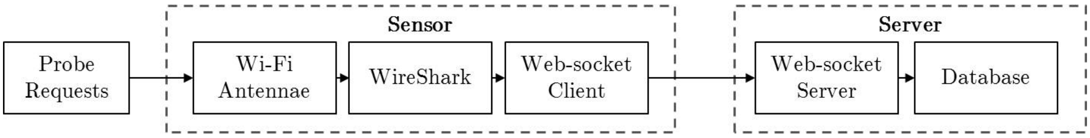


Figure 1. Schematic diagram showing the process of collecting and storing probe requests using the sensor

collection methodology.

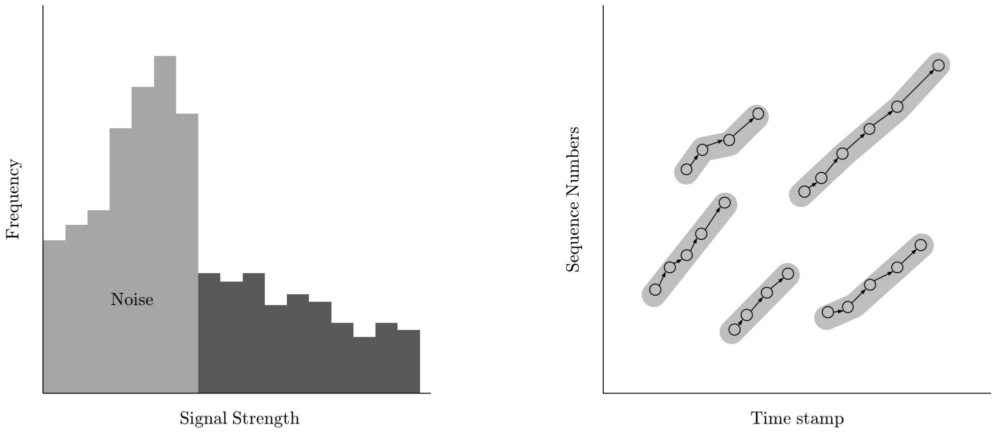
(1) **Background noise** - since the extent to which Wi-Fi signals travel differs subject to various factors such as interference and humidity, it is close to impossible to restrict our data collection to a finite area of interest. This can lead to a significant background noise at certain locations. 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. It is important to note that this method may not work effectively on study locations with complex configurations such as the source of noise and the area of study being located at the same distance from the sensor. This aspect is explored in detail in the broader case study in the following sections.
(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 translates 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 internal and external validation methods to tackle each of these uncertainties.

3.1. *Filtering with Signal Strength*

One of the clues that we can use to estimate the distance between the mobile 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 was found not to be feasible 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. This severely limits our ability in 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 various locations.

We assume that in configurations where a specific source of background noise is at a constant distance, there must be a distinct pattern 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 signal strength



(a) Distribution of signal strengths (dBm) 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

corresponding to the distance between the sensor and the phone shop irrespective of the local conditions as shown in Figure 2. We could identify these breaks in the data using traditional one dimensional clustering 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 address, is being anonymised by new devices, we need to find other information present in the probe request for use as 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 approach to be 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 lead us to look for a more generalisable approach which is independent of the device model.

From our initial look at the data we found that OUI and the sequence number of the packet is the most promising information to achieve this. First we divide our dataset into sets of probe requests with randomised and non-randomised MAC addresses by looking at the second character of the vendor part of the MAC address and if it is E, A, 2 or 6, then these addresses are identified to be randomised. We keep the MAC address as the unique identifier for the non-randomised requests and further divide the randomised ones in to sub categories based on their OUI. We then identify unique mobile devices from within these sets and assign a unique identifier to each device.

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.

Table 1. Comparison of clustering algorithms with a sample of 40000 probe requests

Algorithm	Time (s)	MAPE (%)
Quantile	0.002	27 %
K-Means	0.007	-23 %
Hierarchical Clustering	172.520	-9 %
Bagged Clustering	0.135	-30 %
Fisher	3.034	-30 %
Jenks Natural Break	556.279	-30 %

- 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 in terms of both time and sequence number.

The nodes were then assigned a unique id based on the unique connected component they belong to as shown in Figure 2. This unique identifier is used in the place of MAC address for aggregation for the anonymised probe requests. Though the recycling of sequence number after 4096 leads to multiple unique ids being reported from a single device from a sample experiment done on randomised probe requests sent by "Google" devices we found that the sequence number is reset for around 0.05% of them. We hence assume this to be inconsequential.

3.3. *Calibrating with Ground Truth*

Since mobile device 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 a sample count done manually at each location. We can then calculate an adjustment factor or an offset for each location by comparing the sensor based counts and ground truth and in turn 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 out periodically at these locations to improve the quality of the estimation.

4. Pilot Study

To start we designed a small pilot study to validate the filtering and clustering methodology against the scale and complexity of data collected on an open public area such as a retail high street. We also aim to find the algorithm which is best suited for classification of signal strengths as 'low' and 'high' to filter out the background noise. The data was collected at Oxford Street, London on 20 December 2017 from 12:30 to 13:00 hrs, where Wi-Fi probe requests were collected using the sensor described in Chapter 3 and pedestrian footfall was manually recorded using the Android app (Soundararaj 2018). Being located at one of the busiest retail locations in the United Kingdom, the Wi-Fi sensor captured approximately 60,000 probe requests during the half hour period, and 3,722 people were recorded manually walking on the sidewalk during that time. The count was done at the location in front of the store where the surveyor carried the sensor on a backpack for the automated count and used a phone

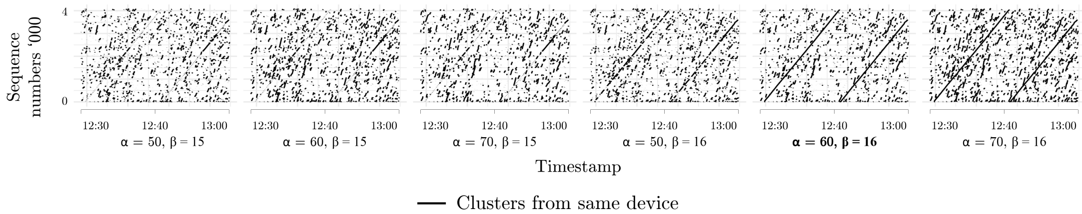


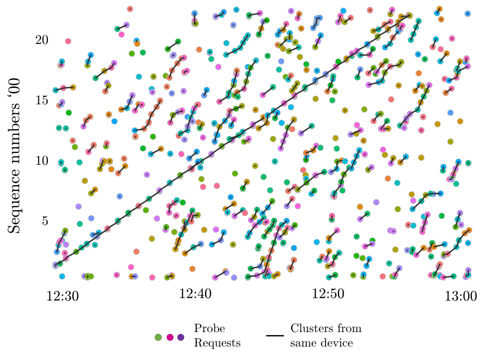
Figure 3. The clustering process was repeated with both increasing sequence number threshold (α) and time threshold (β), until we arrived at the lowest parameters where the know device (black line) is clustered as a single device.

for doing the manual count. The sensor is kept as close to the store window as possible and the manual count is done as a cordon count in fron of the store.

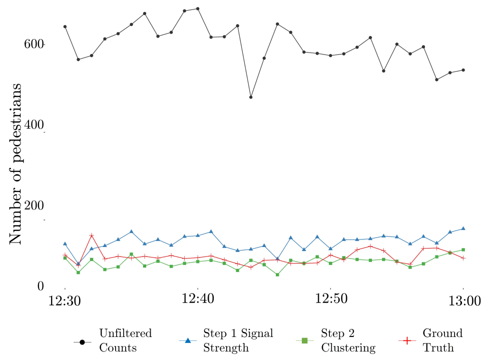
As a first step we just aggregated the probe requests by their MAC address for every minute to generate a minute by minute count of the number of people around the sensor assuming each MAC address corresponds to a mobile device and hence a pedestrian. We then compare this preliminary 'footfall' count to the actual number of pedestrians recorded manually to check for the robustness. We use Mean Absolute Percentage Error (MAPE) measure of robustness of the count since it provides a simple and quick measurement while the street conditions ensure that there are no intervals without any footfall. We find that the MAPE in the raw counts compared to the ground truth is around 425%. This suggests the presence of large amount of noise in the data which might be generated due to sources of uncertainties discussed in Chapter 3 thus demonstrating the need for filtering the data.

We then classified the probe requests as ones with "high signal strength" and "low signal strength" using various one dimensional clustering algorithms such as k-means, quantile, hierarchical clustering, bagged clustering, fisher and jenks natural breaks. The results are shown in Table 1. We found that while hierarchical clustering and jenks gave us fairly low errors, they were too resource intensive for practical use with a larger dataset. We also found that k-means gave quickest results with the lowest MAPE closely followed by quantile algorithm. The cut-off point or threshold for the collected data with which we could classify them as high and low was -71 dBm. We then removed all the probe requests which reported 'low signal strength' and repeated the same aggregation process as before to produce footfall count. This process resulted in a footfall count with a net MAPE of 30%. Though the results are encouraging we are still not completely confident that our filtering process is indeed removing noise or has any correlation the configuration of sensor or position of the mobile devices. These need to be addressed with a larger survey with multiple location of varying orientations.

The next challenge was to identify probe requests which are generated by the same device irrespective of MAC randomisation process. We use the algorithm defined in Chapter 3 and assign a unique identifier or signature to each probe request independent of the MAC address. Since there is neither prior research nor documentation on a universal behaviour of phones in randomising their MAC addresses, we use the stationary device - the one used for manual counting as a reference and find out the suitable time threshold, α and threshold for sequence numbers, β to be 16 seconds and 60 respectively via trial and error. This process is shown in Figure 3 This process is done on top the filtering done based on signal strength and only for the probe requests with randomised MAC addresses. Figure 4 shows the results of this clustering process



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



(b) Comparison of sensor based counts with the ground truth.

Figure 4. The results of the pilot study demonstrating the validity of the methodology.

Table 2. 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

on a small set of randomised probe requests. The probe request with different randomised MAC address is shown by the colored points and the line joining them shows the ones belonging to the same cluster hence expected to be generated by the same device. We finally aggregate the probe requests as before but with the device signature rather than just MAC addresses this results in a footfall count with a MAPE of -18%. A comparison of minute by minute counts resulting from different filtering processes along with the ground truth is shown in Figure 4 showing the promising effectiveness of the methods.

To conclude, from the pilot study we found that both filtering and clustering methods we devices work on complex real world data and results in a final pedestrian counts within a MAPE of 20%. We also found 'k-means' and 'quantile' are best algorithms for clustering signal strengths and the threshold for time and sequence numbers for the clustering algorithm is around 16 and 60 respectively.

5. Case Study Implementation

The methodology set out above was implemented in five different Central London locations at different times. Sensors were installed and data collected for extended periods of time. We also carried out manual counting at these locations across different times of the day. We then applied the methodologies discussed earlier to arrive at estimated pedestrian footfall and compared them with the corresponding manual counts. We finally evaluated the effectiveness of the processes with the Mean Absolute Percentage Error (MAPE) at the locations and report our findings below.

The locations at which the data were collected are shown in Table 2. The locations were chosen for their diverse site conditions and unique sources of noise around the potential location of the sensors. The position of the sensor at these locations with

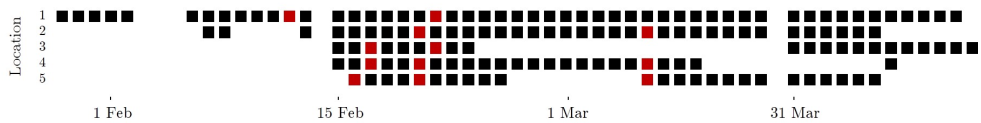
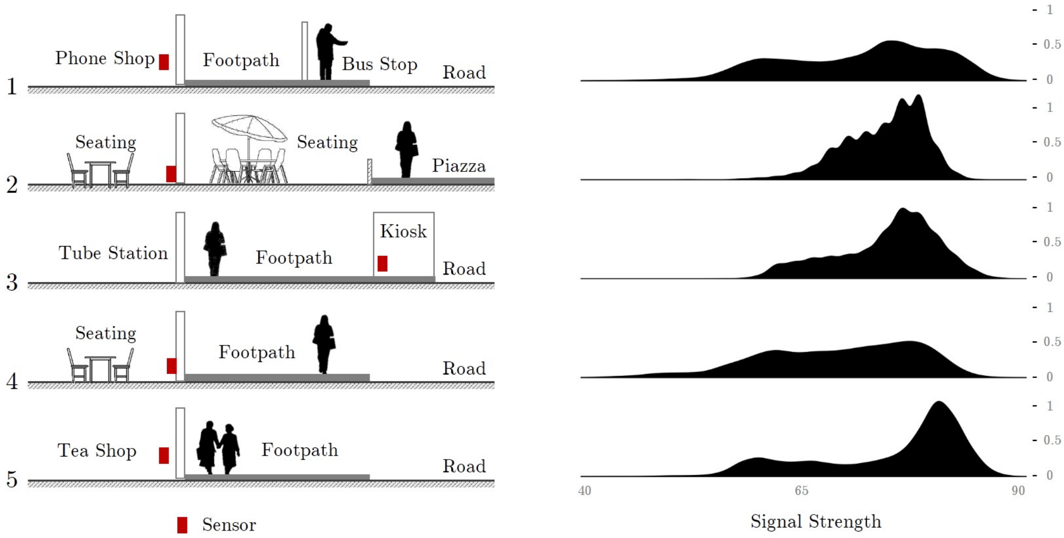


Figure 5. Data collection schedule showing the days when sensors were active at their corresponding locations. The red squares show that manual counting of pedestrians was also done on that day.



(a) Installation configuration of sensors at the survey locations (not to scale)

(b) Density distribution of signal strength reported in collected probe requests (lower values show higher signal strength)

Figure 6. Distribution of signal strengths across locations

respect to the context is shown the Figure 6. We can see that Location 5 is the ‘cleanest’ with one clear stationary source of noise (phone shop) while location 2 is the most complex due to the proximity of seating areas to the sensor. The sensors were operational through out February and March, while manual counts were conducted in these locations in half hour sessions on at least two different days. For the purposes of comparing with ground truth, we considered the data from sensors which correspond to the 12 sets of available manual counts. The schedule of data collection is shown in Figure 5.

We start by looking at the distribution of the signal strength reported by the probe requests across the locations. From the density plot shown in Figure 6 we can see that there is clear relation between the distribution of the signal strength and the distance and complexity of the source of noise. We can see that while location 5 shows clean difference between low and high signal strengths, location 2 is almost normally distributed. Intuitively we expect that location 2 and 4 must be harder to classify than location 1 3 and 5. We run the k-means clustering algorithm and filter out the probe requests which are randomised and have signal strengths less than the second break (threshold). It is important to note that we are dealing with relative thresholds of signal strengths which can vary with location and time of the analysis. We then aggregate then probe requests for every minute by counting the number of Unique MAC addresses present in every minute. We also remove devices that dwell around the sensor by removing the MAC addresses which reappear from within the past hour.

Table 3. Results of footfall estimation at each location as Mean Absolute Percentage Error (MAPE) after each step of the 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

The results of the first stage of filtering process along with the thresholds are shown in Table 3. Confirming our intuition, we see that the location 2 has the most MAPE followed by location 4 while rest of them have highly reduced MAPE. It is significant that this method alone reduces our margin of error by a 50 - 100% from the raw counts without any cleaning. This makes the signal strength filtering quick and ideal method for practical applications which doesn't require absolute numbers such as creating large aggregated indexes to show long term trends. We also see that the success of the signal strength filtering can be improved significantly by installing sensors such that the pedestrians and noise are at different distances from the sensor and are the field of measurement is distinct from the surroundings however noisy it might be.

We then run the sequence numbers based clustering process on the rest of the probe requests to reduce the MAPE by almost 50 - 100% on all the sensors except for location 3. Location 3 is an outlier among all the other sensors since it is the only one with large amount of pedestrians very close to the sensor. This may be causing the over filtering caused by the previous process. We finally run the calibration process where we calculate the adjustment factors from the ratio between the manual counts to the counts calculated for the sample period as shown in Table 3. We the use them to adjust the counts to achieve a MAPE ranging from 10 - 50%. We can see that the sensors with people moving right next to them tend to under-count with our methodology while sensors with seating next to them tend to over-count significantly. However, using the filtering process, we can reduce the error to almost 10% closer to that of the ground truth.

6. Conclusion

Sentient technologies make measurement of the human activities that are the life blood of the smart city possible. Yet the data that they harvest are frequently relevant only to the sub-groups within society that avail themselves of particular goods and services – such as social media applications, transport modes or retail offers. In each of these cases, it is necessary to remember that the resulting data are by-products of consumer transactions, and will as a consequence, only pertain to users of the relevant goods or services. If the smart city is to be socially inclusive, it therefore follows that sentient data must represent entire populations, whether by design or by triangulation with external, population wide, sources. This is a non-trivial task, since the ebbs and flows of smart device-enabled citizens rarely pertain to any clearly defined population in either administrative or functional terms (Massam 1975).

Our objective here has been to collect, rather than re-use, data on smart city functioning, by recording Wi-Fi probes and ultimately reconciling them with manual counts in order to infer ambient populations. The internal validation methodology set out in

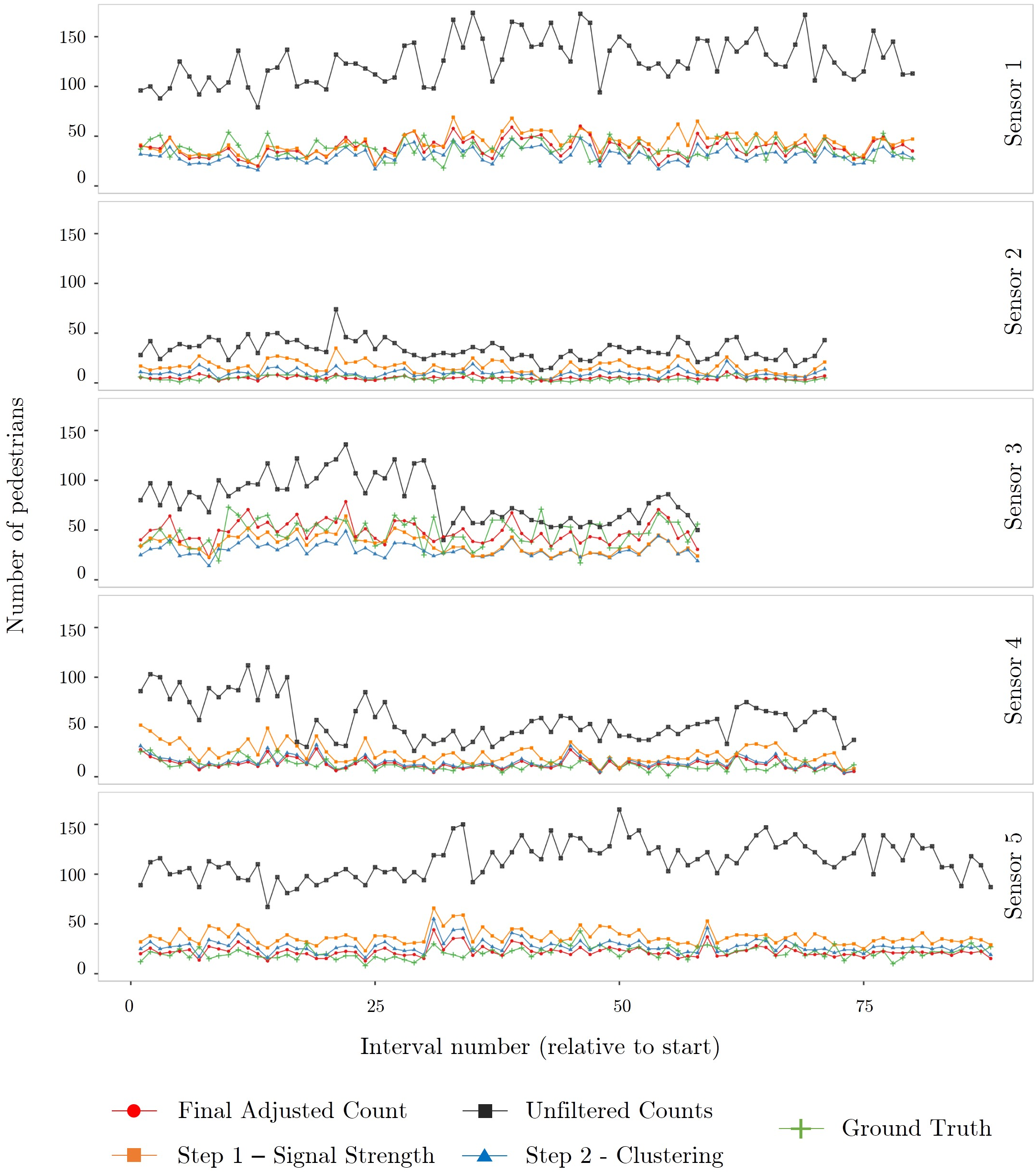


Figure 7. Comparison of the filtering process with the ground truth in all the locations.

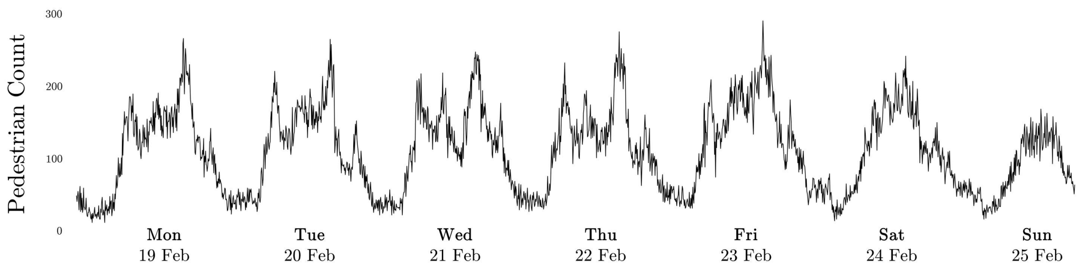


Figure 8. A week of pedestrian footfall at the Strand, London collected by the methodology. The counts are aggregated for 5 minute intervals.

the technical sections of this paper, allied to external validation from pedestrian counts, renders the method inclusive and robust when recording activity levels in retail centres in real time. We have 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 wide network of sensors (Longley *et al.* 2017). 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. It is important to note that the filtering process work based solely 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. External validation then entailed reconciling adjusted counts with the footfall observed at sample locations. This 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.

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 a week's worth of precise footfall in area around Charring cross, London is shown in Figure 8 in order to demonstrate the potential for such a dataset. Projects such as SmartStreetSensors (Longley *et al.* 2017), may utilise this methodology to overcome the challenges introduced by the implementation of MAC address randomisation.

The vicissitudes of MAC randomisation, and the provisions of privacy legislation such as EU General Data Protection Regulations mitigate against tracking individuals across the smart city using this approach. This can be modelled using agent-based methods (Heppenstall *et al.* 2011), however. In our own research we have also begun to link store time-lagged till receipts to footfall, and have used such data to better understand the dwell times that characterise such different retail uses as stores with window displays and fast food restaurants. Such analysis not only provides a more nuanced picture of movement through retail areas, but also enables valorisation of micro sites within retail centres. In the UK, for example, this is of immediate practical importance in evaluating business rates on properties, and has still wider implications for the setting of retail unit rental values. There are obvious extensions to understanding the ebbs and flows of activities in the 24-hour smart city.

More broadly still, extensions to this strand of smart city research are likely to seek to differentiate the quality of different elements within footfall according to mission e.g. travel to adjacent workplace zones, leisure, etc., and personal characteristics such as spending power. In this respect, future research may not only simulate linkage of harmonised footfall counts between sensor locations, but also link these in turn to disaggregate origin-destination matrices for bikeshare and other public transport modes. Our own investigations will consider these and other challenges to understanding the functioning of the sentient city.

Acknowledgement

This research was funded by Economic Social Research Council grant – Retail Business Datasafe (ES/L011840/1) through Consumer Data Research Centre. Authors would like to acknowledge the contributions of the Local Data Company in facilitating the data collection process and Dr Roberto Murcio and Dr Tank Green for their inputs.

References

2016. Ieee standard for information technology–telecommunications and information exchange between systems local and metropolitan area networks–specific requirements - part 11: Wireless lan medium access control (mac) and physical layer (phy) specifications. *IEEE Std 802.11-2016 (Revision of IEEE Std 802.11-2012)*, 1–3534.

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.

Campbell, A.T., *et al.*, 2008. Transforming the social networking experience with sensing presence from mobile phones. *In*: *Proceedings of the 6th ACM conference on Embedded network sensor systems.* ACM, 367–368.

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.

Combs, G. and Contributors, 2018. Wireshark - network protocol analyzer. https://www.wireshark.org/about.html.

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.

Demir, L., Cunche, M., and Lauradoux, C., 2014. Analysing the privacy policies of wi-fi trackers. *In*: *Proceedings of the 2014 workshop on physical analytics.* ACM, 39–44.

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.
Heppenstall, A.J., *et al.*, 2011. *Agent-based models of geographical systems*. Springer Science & Business Media.
Hong, H., De Silva, G.D., and Chan, M.C., 2018. Crowdprobe: Non-invasive crowd monitoring with wi-fi probe. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2 (3), 115.
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.
Lloyd, A. and Cheshire, J., 2018. Detecting address uncertainty in loyalty card data. *Applied Spatial Analysis and Policy*, 1–21.
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, D., Cockings, S., and Leung, S., 2015. Developing a flexible framework for spatiotemporal population modeling. *Annals of the Association of American Geographers*, 105 (4), 754–772.
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.
Massam, B.H., 1975. *Location and space in social administration*. E. Arnold.
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.

Singleton, A. and Longley, P., 2018. Developing new and hybrid geodemographic classifications: London's workplace zones. *Applied geography*.

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.

Soundararaj, B., 2018. Clicker. https://play.google.com/store/apps/details?id=com.bala.manualcount.

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.

Venerandi, A., *et al.*, 2015. Measuring urban deprivation from user generated content. *In*: *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*. ACM, 254–264.

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.

Xu, Z., *et al.*, 2013. Pedestrain monitoring system using wi-fi technology and rssi based localization. *International Journal of Wireless & Mobile Networks*, 5 (4), 17.

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.