Understanding sources of measurement error in the Wi-Fi sensor data in the Smart City

Karlo Lugomer[1], Balamurugan Soundararaj[2], Roberto Murcio[3], James Cheshire[4] and Paul Longley[5]

Department of Geography, University College London

January 13, 2017

Summary

Data quality audits are a necessary precursor to quantitative analysis of human activity patterns using primary data collected using automated sensors. This paper reports detailed exploration of the sources of measurement errors that potentially impact upon the quality of the footfall data collected as part of the Consumer Data Research Centre SmartStreetSensor project. Depiction and analysis of activity patterns is integral to numerous applications in urban management, retail and transport planning, and emergency management, yet most analysis to date has remained focused upon data pertaining to night-time residence as from the Census of Population and daytime estimates through sample surveys or traffic counts. Here we investigate how Wi-Fi signals from mobile devices can be used to estimate levels of human activity at different times and locations and argue about the opportunities and issues arising when using them for estimating footfall.

KEYWORDS: Wi-Fi sensors, footfall, pedestrian flows, measurement error, calibration

1. Introduction

The accurate measurement and estimation of levels of human activity and their spatial and temporal distributions make a fundamental first step towards their understanding and management (Louail, 2014). Such distributions are highly granular and dynamic in both dimensions (Steenbruggen, 2014) and an accurate estimation is crucial for decision-making processes in numerous applications (urban management, retail, transport planning, emergency management). Traditional sources of demographic data, however, have been coarse in terms of both spatial and temporal scales and therefore limited to night-time residence (e.g. Census) and sporadic daytime estimates (e.g. sample surveys and traffic counts). Census data, while being comprehensive, has a low frequency of update (Harris and Longley, 2002). Although sample surveys and traffic counts get updated more frequently with a high resolution, a coherent collection of such data has practical issues, the most prominent one being a requirement of the significant input of the manual labour. Wi-Fi sensors bridge the aforementioned gaps, enabling for continuous footfall data collection, with most of the process done automatically. However, resulting estimates are not error free and require further validation. This paper identifies some of the problems encountered when using Wi-Fi sensor data as a proxy for human activity patterns and suggests possible ways to resolve them.

[1] karlo.lugomer.14@ucl.ac.uk
[2] s.bala@ucl.ac.uk
[3] r.murcio@ucl.ac.uk
[4] p.longley@ucl.ac.uk
[5] james.cheshire@ucl.ac.uk

2. Methodology

Out of the various signals emitted by wireless mobile devices, we focus on capturing the Wi-Fi probe request frames sent to search for new Wi-Fi networks (IEEE Standards Association, 2013). Since Wi-Fi capability and hence the probe request mechanism to connect to access points is almost ubiquitous on mobile devices, this approach is one of the most advantageous ones. It is not only non-intrusive and passive, thus improving the participation rate, but also has MAC address of the device which, when hashed, can act as a unique identifier without compromising participants' privacy. As it has been said, this approach is not free of errors. In addition to the difficulty of accurately determining and maintaining an exact range for each sensor, we noted a major source of measurement error in the high prevalence of the Wi-Fi capabilities in electronic devices, which enables sensors to detect Wi-Fi signals from a wide range of devices, such as printers. In this paper, we primarily focused on identifying and quantifying these errors.

2.1 SmartStreetSensor

The SmartStreetSensor project was set up as a collaboration between University College London (UCL) and Local Data Company (LDC) for installing, collecting and analysing Wi-Fi probe requests across the UK using proprietary hardware and software developed by LDC. The project aims to use these probe requests captured at retail establishments across the UK and accurately estimate the corresponding footfall. In January 2017, there have been around 500 sensors installed (Figure 1), covering a large number or retail areas throughout England, Scotland and Wales.

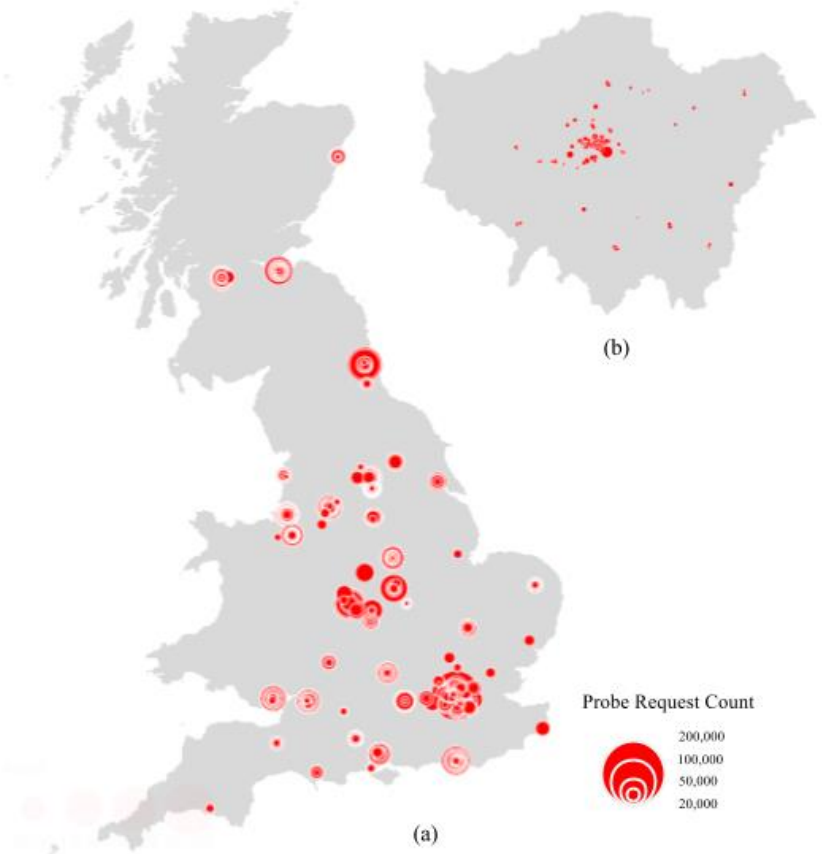


Figure 1: Spatial distribution of Wi-Fi sensors in: (a) Great Britain, (b) Greater London

Figure 2 shows the distribution of sensors ranked based on the total number of probe counts captured. The lack of a normal probability distribution suggests that the probe request counts are not random and may be a function of processes operating within these urban areas.

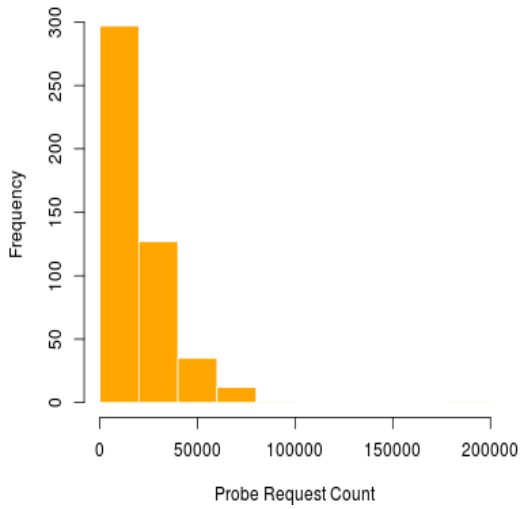


Figure 2: Distribution of number of sensors ranked based on the number of probe requests they collected on 20-12-2016

3. The sources of measurement error in Wi-Fi sensor data

Footfall may be defined as a number of pedestrians passing by in front of the shop window on the whole sidewalk width in a given period. Sensor count is defined as a number of raw, unprocessed MAC addresses of all the devices captured by the sensor in a given period. Number of the devices detected by the sensor and actual footfall detected on site do not normally match, as there are two main groups of sources of uncertainty governing this discrepancy: overcounting and undercounting factors (Figure 3).

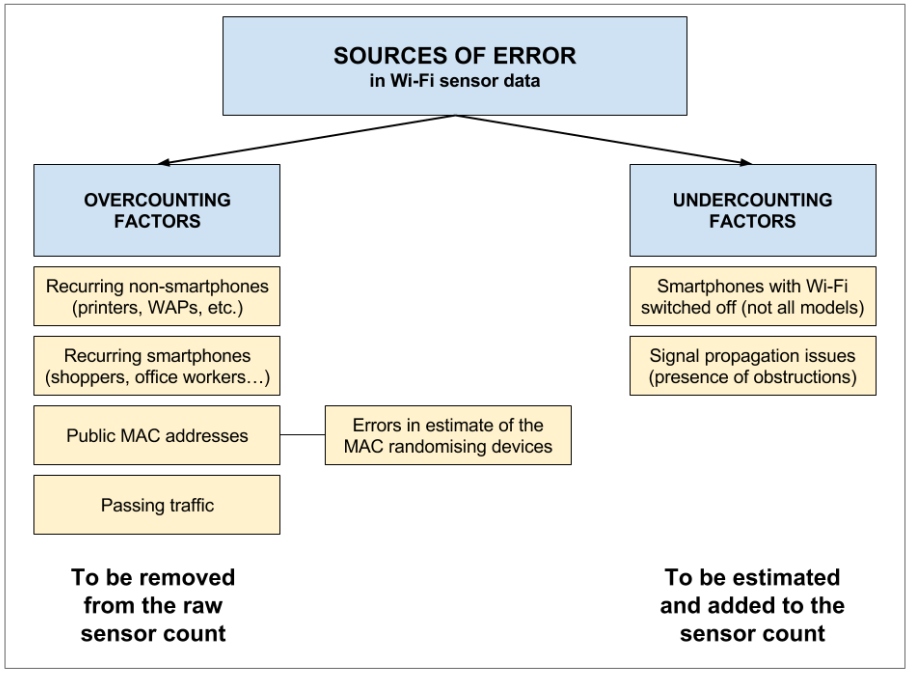


Figure 3: The sources of error in the Wi-Fi sensor data

Overcounting factors cause sensors to count more MAC addresses then there are pedestrians in front of the store. They comprise devices inside the store, as well as devices in the residential properties and offices in the immediate surroundings of the sensor. Those devices are not limited to smartphones, but also include printers, scanners, computers, wireless access points, etc. The key advantage to managing these sources of uncertainty is the possibility to programmatically remove them from total footfall estimate by detecting vendor parts of MAC addresses belonging to the manufacturers of devices other than smartphones and by detecting and removing MAC addresses that remain near the sensor throughout longer periods of time. However, problems still persist due to the fact that some vendors have, in recent years, introduced randomisation of MAC addresses of their customers' devices (Vanhoef et. al., 2016). This means that pedestrians owning devices with newer OS versions installed cannot be detected directly and their number has to be modelled separately.

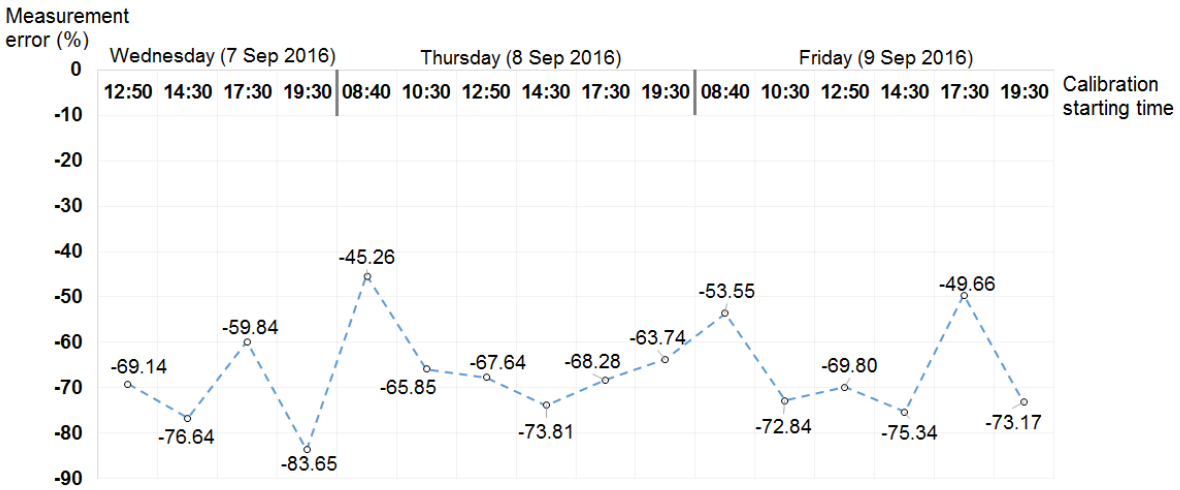
On the other hand, undercounting factors are much more difficult to manage. They include the proportion of pedestrians that either do not carry a smartphone, or have a Wi-Fi turned off, as well as any pedestrian of interest that has not been detected due to decay of signal strength with distance or due to physical obstacles in the store interior, such as walls, blocking the effective data transmission (Jiang et. al., 2015). Wi-Fi signal also propagates differently in different directions and its strength varies between different manufacturers more than it does between the sensors of the same type (Dimitrova et. al., 2012).

3.1 Sheffield and London case studies

In order to check whether and how the portion of measurement error attributable to the undercounting sources of uncertainty varies temporally, two sensor locations have been visited: one in Sheffield over sixteen 20-minute periods throughout two and a half days and one in London over fourteen 15-minute periods throughout two days. Sensor counts were devoid of all the unwanted devices, devices with randomised MACs have been estimated and resulting processed counts were compared against the groundtruth derived by manually counting the passers-by directly in front of the sensors on the whole width of the sidewalk.

Since removal of the overcounting sources of uncertainty has taken place prior to further calibration, residual error is mostly negative and its variation is shown on Figures 4a and 4b.

(a)



(b)

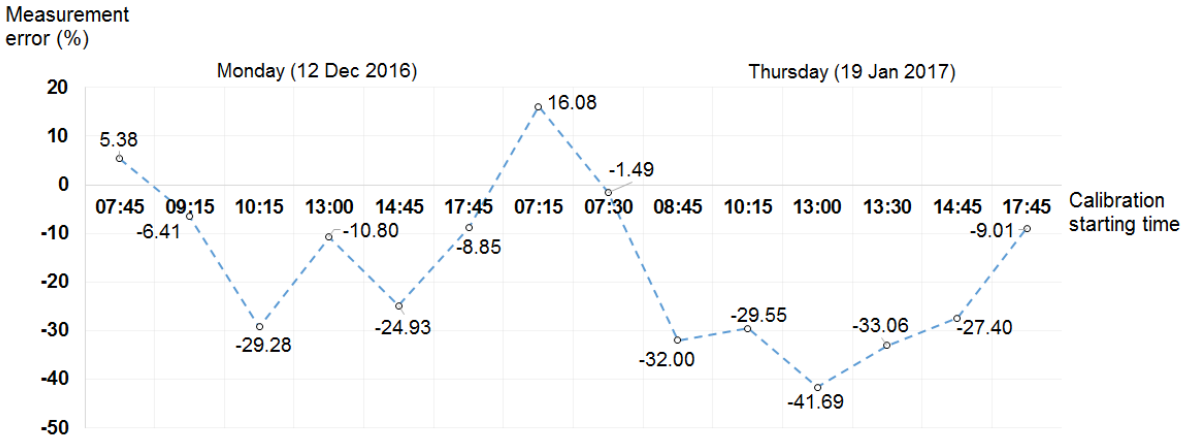


Figure 4: Temporal variation of residual error after initial data cleaning recorded by the sample sensor in: (a) Sheffield (b) London

Magnitude of the remaining error appears to be high and changes significantly throughout the day. Also, the variation does not follow the same pattern at both sensor locations under consideration. This means that more effective calibration requires taking manual counts on multiple occasions on every sensor location and applying different adjustment factors, or their average, to minimise the overall measurement error. Adjustment factor α can be defined as a ratio between the actual footfall and the footfall observed by the sensor (Equation 1):

$$\alpha = \frac{M}{\psi} \tag{1}$$

where M is the manual count of pedestrians at a specified location and time period and ψ is the processed sensor count at the same location and time.

Adjustment factor is then multiplied with processed sensor counts at all other times to derive the actual footfall.

If we randomly pick a period during the day and calculate the adjustment factor based only on one 15 or 20-minute period for each sensor, there is a danger of picking up a period where adjustment factor is far from average, thus significantly increasing the overall error. During the validation process, researchers should, whenever possible, take multiple counts throughout a day and then calculate the average value to minimise the measurement error. It can be seen that this kind of validation reduced the average measurement error on both locations substantially, bringing the absolute value of the measurement error from 68.46% to 0.83% in case of the studied sensor in Sheffield and from 25.03% to 4.59% in case of sensor located in London (Table 1).

Table 1: Measurement errors for sensor in Sheffield (7 Sep – 9 Sep 2016) and London (12 Dec 2016 and 19 Jan 2017)

Sensor location	Day	Measurement error (%) (without adjustment factor)	Measurement error (%) Adjustment factor taken into account:		
			Minimum	Maximum	Median
Sheffield	Wed	-71.35	-47.67	75.20	-8.41
Sheffield	Thu	-67.04	-39.79	101.58	5.38
Sheffield	Fri	-67.75	-41.10	97.19	3.09
Sheffield	**Average**	**-68.46**	**-42.38**	**92.89**	**-0.83**
London	Mon (12 Dec)	-22.37	-33.12	33.14	8.32
London	Thu (19 Jan)	-26.85	-36.99	25.45	2.06
London	**Average**	**-25.03**	**-35.43**	**28.56**	**4.59**

4. Conclusions and further steps

Even though Wi-Fi sensor data are not so straightforward to use and require additional processing before making them sufficiently accurate for further analysis, they nevertheless present a broad spectrum of opportunities for human activity pattern modelling. This paper has explored some of the possible ways to tackle sources of error in Wi-Fi sensor data, with particular emphasis on the importance of the fieldwork and making sure that processed data correspond to the reality. The measurement errors can therefore be greatly reduced by combining the simple data mining algorithms and groundtruthing. The accuracy of the final derived footfall will depend on the effectiveness of those algorithms, the number of the ground truth samples and the time they were taken.

Our next steps include formulating a more comprehensive and robust calibration method accounting for both overcounting and undercounting sources of measurement error. This will enable us to better understand and model the spatio-temporal variation of the pedestrian flows and use the estimated footfall data to devise the classifications of retail areas based on the temporal activity patterns and other socioeconomic indicators.

5. Acknowledgements

This work is funded by the UK ESRC Consumer Data Research Centre (CDRC). We would like to thank Ivana Mogin and Adrian Cox from the Local Data Company for providing useful advice throughout the research procedure.

6. Biography

Karlo Lugomer is a Geography PhD student at University College London (UCL). His research interests incorporate GIS applications and quantitative methods in human geography, with particular focus on the geodemographics, retail geographies and spatio-temporal activity patterns in urban areas. He previously studied Geography BSc at University of Zagreb and Geospatial Analysis MSc at UCL.

Balamurugan Soundararaj is a PhD student at Department of Geography, UCL. His research interest include the collection, analysis, visualisation and interpretation of large scale, complex of spatial data. He previously studied Advanced spatial analysis and visualisation at UCL and Urban planning from School of Planning and Architecture, Delhi.

Roberto Murcio is a mathematician with a PhD in Complex Systems. His research focus on far-from-equilibrium systems and information theory, applied to urban studies.

James Cheshire is Senior Lecturer in Quantitative Geography at UCL.

Paul Longley is Professor of Geographic Information Science at UCL.

References

Dimitrova, D.C., Alyafawi, I. and Braun, T., 2012, June. Experimental comparison of bluetooth and WiFi signal propagation for indoor localisation. In *International Conference on Wired/Wireless Internet Communications* (pp. 126-137). Springer Berlin Heidelberg.

Harris, R. J. and Longley, P. A. (2002). Creating small area measures of urban deprivation. *Environment and Planning A* 34 (6), pp. 1073–1093.

IEEE Standards Association, 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 GHz*,

Jiang, P., Zhang, Y., Fu, W., Liu, H. Su, X. (2015). Indoor Mobile Localization Based on Wi-Fi Fingerprint's Important Access Point. *International Journal of Distributed Sensor Networks*, 2015, p.45.

Louail, T. et al., 2014. From mobile phone data to the spatial structure of cities. *arXiv:1401.4540v1 [physics.soc-ph]*, 18, pp.1–14.

Reades, J. et al., 2007. Cellular census: Explorations in Urban data collection. *IEEE Pervasive Computing*, 6(3), pp.30–38.

Steenbruggen, J., Tranos, E. & Nijkamp, P., 2015. Data from mobile phone operators: A tool for smarter cities? *Telecommunications Policy*, 39(3–4), pp.335–346.

Vanhoef, M., Matte, C., Cunche, M., Cardoso, L.S. and Piessens, F., 2016, May. 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* (pp. 413-424). ACM.