First review comments

Title : Estimating Real-Time Highstreet Footfall from Wi-Fi Probe Requests

Manuscript id : IJGIS-2018-0340

Comments dated : 28-Oct-2018

1 Editor

1.1 Overall Comments

1. While this is an important and timing topic dealing with data uncertainty and privacy in the mobile age, the reviewers have raised concerns on
    - Clarity of the methodology
    - Lack of discussion of wifi-based human activity studies in the big data age
    - Proper use of terminology
    - Performance of the algorithm etc.
2. In lieu of these comments, they have recommended to make substantial revisions to your manuscript
3. Also, since this is a "smart cities" related special issue, I ask you to
    - Add more discussions on smart cities in the introduction and conclusion sections and
    - Link your research more tightly to this theme.

> Thank you for the very helpful comments of the reviewers of our paper/ As requested, we have responded to all comments without adding significant length to the paper. We have also extended our discussion of how effective measurement of footfall is relevant to understanding the functioning of the Smart City. If any of what follows and what has been uploaded is unclear, please do not hesitate to contact me.
>
> Yours sincerely, Balamurugan Soundararaj, (on behalf also of Paul Longley and James Cheshire)

1.2 General Instructions

1. Please try to avoid making the paper longer (keep in mind that the average length of papers in this journal is 18-20 published pages). It is important that you ensure that the spelling and grammar is of a high standard.
2. Please also highlight the changes to your manuscript within the document by using the track changes mode in MS Word or by using bold or coloured text.

3. To revise your manuscript, please select the link below, https://mc.manuscriptcentral.com/ijgis?URL_MASK=a32caf3bbd49412cac8d1cfa9d7ebaf8 which will take you your decision letter, and where you need to resubmit your manuscript.
4. If you have used the on-line LaTeX template, upload all those files and a pdf.
5. Please check your English spelling and grammar.
6. There are two hard limits that you need to bear in mind,
    - Image size - Please be aware that the total number of pixels (height X width) in each image must be less than 40,000,000 (40 mp).
    - Total file size - please make sure that the combined, uncompressed, file sizes do not exceed 20Mb. The internal LZW compression, which is an option in TIF, is acceptable however.
    - Please do not save B/W or Greyscale images as RGB.
7. Equations and tables need to be in an editable format.
8. You need to check that the paper title in ScholarOne is the same as the title on the paper and that the Title page contains all authors' names and affiliations as you would want them to appear.
9. When submitting your revised manuscript, you will be able to respond to the comments made by the reviewers in the space provided. You can use this space to document any changes you make to the original manuscript. In order to expedite the processing of the revised manuscript, please be as specific as possible in your response to the reviewers.
10. Please delete any redundant files before completing the submission.

> Thank you for the detailed instructions. All of these have been taken in consideration and checked for creating the revised manuscript for the final submission.

2 Reviewer-1

2.1 Overall Summary

1. This paper is about the estimation of pedestrian footfall using Wi-Fi signal emitted by mobile devices. It deals with the problem of estimating the footfall size based on a noisy and biased data. More specifically it addresses the following issues that may lead to over counting:
    - MAC randomization that may cause a device to generate several MAC addresses;
    - Signals received from outside the interest zone; unequal adoption of Wi-Fi smartphone in the population. To address this issue, a number of techniques are introduced:

    - clustering received signal strength to remove device that are outside the monitored area;
    - clustering probes based on sequence number and timing;
    - finally an adjustment factor is applied to correct the estimated value.

2. The adjustment factor is calibrated for each location using a manual counting. The author conducted a pilot experiment to identify the clustering algorithms and parameters, and then they conducted a second range of experiments at different location over an extended period of time. During each of those experiments, ground-truth count was collected manually. The results show that the presented approach can significantly reduce the counting error compared to a naive solution consisting in counting the distinct number of MAC observed. Nevertheless, there are cases (2) where the error is still large (50%), even if in other cases (1) the error is only of 9%.

> This is a correct surmisal of the remit of the paper. We have also added material to the introduction and conclusion of the paper to better link this remit to the Smart Cities discourse.

2.2 General Comments

1. The idea of clustering frames using sequence number has been introduced in [1]. This work needs to be cited when introducing the proposed clustering algorithm.

   [1] Mathy Vanhoef, Célestin Matte, Mathieu Cunche, Leonardo Cardoso, Frank Piessens. Why MAC Address Randomization is not Enough: An Analysis of Wi-Fi Network Discovery Mechanisms. ACM AsiaCCS, May 2016, Xi’an, China. 2016,

> Thank you. We have included a citation to the source, as well as one other recent addition to this fast-developing literature.

2. The author considered that the hashing of MAC address is a good way to protect users privacy, as some MAC address collected by the device may not be random. However hashing of MAC address cannot be considered as a proper anonymization technique (see [2] and http://webpolicy.org/2014/03/19/questionable-crypto-in-retail-analytics/); at best it can be considered as an obfuscation. [2] Levent Demir, Mathieu Cunche, Cédric Lauradoux. Analysing the privacy policies of Wi-Fi trackers. Workshop on Physical Analytics, Jun 2014, Bretton Woods, United States. ACM, 2014

> Thank you for pointing this out. In the revised version of the paper we have made reference to this discussion, while also making clear the anonymisation is ensured by the system architecture that processes and curates the data. Specifically we rotated the random salt used to hash the MAC addresses regularly to prevent any kind of tracking of the users across a longer period of time with which the hashing process is carried out.

3. There is a number of commercial solutions doing counting based on Wi-Fi signal. Although the detail of their solution is not public, it would be worthwhile to acknowledge their existence.

> Thank you - this has been done.

4. In section 3.2, it is explained that data is divided into two sets corresponding on random and non-random MAC. Which method is used to perform this classification ?

> The identification of a MAC address as randomised or non-randomised is done looking at the second character of the vendor part of the MAC address. The addresses with E, A, 2, 6 at these locations are identified to be randomised MAC addresses. This has been included in the revised paper.

5. How was the sensor positioned compared to the area monitored used for ground truth ? A more detailed description is required regarding the positioning of the sensor and the size and topology of the surveyed area.

> This is a good point, and is very important for our validation work. We have included a more detailed and explicit discussion of the way that field observation was used to calibrate and cross validate the Wi-Fi results.

2.3 Clarifications & Corrections

1. The topic addressed in this work is interesting and timely. The proposed approach is sound and shows promising results. The fact that the authors conducted field experiments to evaluate their solution is a positive point. Yet before publication, I believe that there are a number of points that needs to be clarified or corrected. Regarding the technical aspects I have a number of minor remarks:

> Thank you very much for these very positive comments. The paper reports on extensive methodological research and field implementation and is, we believe, cutting edge in its finding. It also has important implications for our understanding of the functioning of smart cities, and these implications are discussed in greater detail in the revised version.

2. "Wi-Fi beacon technologies to access the Internet". In Wi-Fi, beacons are management frames used by Access Point for service discovery. They do not provide Internet access. "Wi-Fi Access Point to access ..."

> Thank you. We have corrected and clarified the point.

3. "Wi-Fi antennae regularly broadcast a special type of signal" -> the antenna is just the part ensuring that the signal is correctly emitted, but

it is not the source. In this context it is better to use a more general concept and refer to the "Wi-Fi network interface" (or Wi-Fi interface) as the element broadcasting a signal.

> Thank you. Point clarified in the revised version.

4. The IEEE 802.11 b/g is just a sub part of the 802.11 specification. Here you can refer to the 802.11 specifications as a whole : [“802.11-2012 - 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 802.11-2012 (Revision of IEEE Std 802.11-2007), pp. 1–2793, Mar. 2012.

> Thank you. Noted and added to the revised version.

5. "the first step in establishing a Wi-Fi based connection " is also known as service discovery

> Thank you. Added to the revised version

6. "Wi-Fi functionality has been turned off by the user", regarding this point you may be interested in :

7. "Media Access Control (MAC) address which is an unique identifier for the wireless hardware of the mobile device" The transmitter field of the 802.11 frame can indeed contain a unique MAC address but can also contain a random one as noted by the author. Here it would be better to avoid stating that the content of this field is always a unique MAC.

> Thank you. Again, this has been noted in the revised version of the paper.

8. "The strength of the signal which transmitted the request.". In fact the Received Signal Strength Indicator, a value providing an indication on the strength of the received signal as seen by the receiver, not the strength at the emitter side.

> Thank you. This distinction has been noted in the revised version of the paper.

9. Could you provide more information on the application used for counting (Clicker)

> This simple application was developed specifically for the project for Android platform. More details on the application has been included in the revised manuscript.

10. "Though the recycling of sequence number r after 4000 leads to multiple classifications reported on single device, the magnitude of error is greatly reduced." This sentence is not clear. And the cycle length is not 4000 but 4096 = 2^12.

> Thank you. This point has been noted and accommodated into the revised version.

11. Regarding the issue of cycling sequence number, one could use a distance with a modulo : di,j = di - dj mod 4096.

> Thank you very much for the suggestion. We carried out some tests on a sample data of 1910 probe requests comprising of google phones with randomised MAC addresses captured in the pilot study and found that only around 0.5% of the probe requests encountered the recycling of sequence number. Which, in our opinion, doesn't influence the results in any significant way considering the context. We have also included this in our revised manuscript.

2.4 Typos and other remarks

1. Challanges
2. "These are deemed to be consumer data because devices carried by consumers routinely probe for a consumer service, specifically a Wi-Fi connection. Monitoring the probes from such devices provides an indication" This sentence is not clear.
3. "Though Wi-Fi F is a ‘location-less’ technology, ", please clarify what you mean by that.
4. "privacy infringement" may not be the best word as infringement relate to a law or a regulation. Privacy breach or leak may be a better choice.
5. Wireshark and not WireShark.
6. "We hypothesize that an adjustment factor could be arrived at for each location of data collection, ""This calibration can be carried over periodically and the frequency to improve the quality of the estimation." something wrong in both sentences.

> Thank you for picking up these minor lapses. All have been corrected or clarified in the revised version of the paper.

3 Reviewer-2

3.1 Overall Summary

1. This paper addressed the data quality issue of Wi-Fi-based probe requests in estimating pedestrian counts. The authors proposed several methods to improve the accuracy of estimating the number of unique mobile devices from a set of anonymized probe requests without revealing their original device information.

> This is correct, although the intended emphasis upon disclosure is a feature rather than a principal motivation. We have also added further discussion of how these principally technical issues relate to smart city data infrastructure issues.

2. This is a crucial topic in mobile data uncertainty, and the manuscript has the potential of becoming a useful reference in the field. However, it was not organized effectively and missing many details in the methodology. Before addressing these problems, it is difficult to assess the reliability and scientific merit of the case study.

> Thank you very much for this endorsement of the motivation for our paper. We hope that following responses and the amendments to the paper will allay your concerns about the methodology and application.

3.2 General Comments

1. First, the literature review section adequately covered Wi-Fi-based human activity studies, but it is in lack of an overview of modeling human activity in the big data era. I would suggest the authors add an overarching paragraph (or a sub-section) in Section 2 discussing the importance and challenges of modeling human activities and urban dynamics based on various types of big geodata, such as mobile phone records, location-based social media, Bluetooth data, etc.

> This is a welcome comment, and we have added material in the introduction and conclusion as to the value of new Big Data sources for developing activity-based representations of behaviour and circumstances in the smart city, This chimes with much of our related research, but we write cognisant of the article length restrictions of IJGIS.

2. The conclusion did not address most of the important results from the case study and should be expanded.

> Yes, we agree, and some wider conclusions about the implications of our research for better understanding the functioning of the smart city have now been added.

3.3 Corrections and Clarifications

Second, the methodology should be better clarified. There are many details missing or in need of a better justification. For example,

1. P4 l50, without providing any explanation, the authors hypothesized that sequence number and length of the packet are sufficient to estimate the number of unique devices. This argument was not elaborated until 2.5 pages later in section 3.2.

   > This material has now been amended and resequenced.

2. P6 l27, please provide more information regarding how you defined the threshold for low/high signal strength when eliminating the background noise. This was also unclear in the case study on P8, where the authors repeatedly mixed up classification and clustering. Neither k-means nor hierarchical clustering is considered a “classification algorithm” as claimed in p8 l26. Very few details were provided regarding the parameter setting of these analyses.

   > Thank you. The reference to clustering has been amended. The specifics of the thresholds used have been included, and the techniques justified in more detail.

3. P6 l30-31, in the phone shop example, the authors mentioned that it is possible to identify background noise based on a sharp rise of the request number; however, this will also eliminate the regular pedestrian flow in that shop and cause additional inaccuracy. Please clarify.

   > Thank you. Clarified in the revised version. It is know with preliminary investigation that the received signal strength indicator decreases with increasing distance between the mobile device and the base station. This change is also found to be exponential as opposed to linear. With this context When we install a sensor to have a clear view of the pedestrian footpath in front of it, we make sure that all the devices beyond a certain distance will report similarly low signal strength. This threshold is identified using the k-means algorithm to find the natural break in the data. We have found that this successfully isolates a farther stationary source of large number of probe requests such as a phone shop or a restaurant next door, Bus stops etc. This doesn't work when the source of error is within the range of the footfall. As shown in our examples 2,4. Here our methodology fails to provide sufficient accuracy.

4. P7 l21, how often does the recycling of sequence number occur, and how will this impact the reliability of your analysis? Please provide more details.

The recycling of the sequence number occurs every 4096 counts. From our experiments we observed that a device staying next to device actively searching recycles the sequence number 1-2 times in a 30 mins interval. On average this is expected to be much less than this. When the device doesn't probe as frequently. We have included more details regarding the cycling of sequence numbers and have mentioned method to overcome this as mentioned
by other reviewer. The change in implementing the method is estimated to be --% which we found to be not significant in our context.

5. P7 Section 3.3, please be more specific about the calibration process and the external source of information to be used here.

Thanks for pointing this out. Following your suggestion, this section has been expanded and clarified to include more details. The calibration is done by adjusting the future counts based on a ratio between sample manual counts and corresponding sensor counts at the location. The external source of information used is the manual counts which is collected outside the sensor

6. P7 l45, the analysis in this research is based on one single sensor for each location; however, in reality, it is very common to estimate travel flows based on multiple sensors (e.g., nearby stores in a shopping mall). How would your method address duplication in multiple sensors?

This is a valid concern. Which we tried to clarify in the revised manuscript. Our motivation is not to attempt to track the movements of individuals, although we do attempt to accommodate measurement variations attributable to micro-site location of the sensors. In related work, we have carried out controlled comparison of sensor measures obtained from adjacent store locations, and examined correspondence with retail unit function. This enables identification of the differences in dwell time between, for example, fast food outlets and clothes shops. This full analysis requires addition of store till data and the analysis is detailed. For this reason we include only a short reference to this extension of our analysis here. We also aim this method to count footfall accurately at locations rather than the flow of footfall between those locations. But as demonstrated in we can infer the relation or flow between the locations using other probabilistic approaches.

Lee J, Nemati S, Silva I, Edwards BA, Butler JP, Malhotra A. Transfer entropy estimation and directional coupling change detection in biomedical time series. Biomed Eng Online. 2012;11:19. Published 2012 Apr 13. doi:10.1186/1475-925X-11-19

Vicente R, Wibral M, Lindner M, Pipa G. Transfer entropy--a model-free measure of effective connectivity for the neurociences. J Comput Neurosci. 2010;30(1):45-67.

7. P10 l48, how was the adjustment factor determined?

> The adjustment factor is calculated as the ratio between the manual counts to the counts calculated from sensor for the sample period. This is used to adjust future counts at the same locations. We have included the clarification on the revised manuscript.

4 Reviewer-3

4.1 Overall Summary

1. This paper presents Wi-Fi based footfall counting methodology, which is timely and interesting topic. The method can be performed to estimate human activity such as pedestrian footfall from Wi-Fi probe requests.

> Yes, Thank you for your encouraging comment.

4.2 General Comments

Some areas where I would like to see more detail:

1. It would have been interesting to understand the temporal characteristics for a week pedestrian footfall. How to consider this situation in your study? Why do you choose time range from 12:30 to 13:00 hrs ?

> Thank you. Our related work (cited below) examines daily, weekly and seasonal variation, and wil be the subject of future papers.
>
> Lugomer, K., & Longley, P. (2018). Towards a Comprehensive Temporal Classification of Footfall Patterns in the Cities of Great Britain (Short Paper). In LIPIcs-Leibniz International Proceedings in Informatics (Vol. 114). Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.
>
> There was no reason for selecting the time in the pilot study. The location was chosen for it it known to have high number of randomised probe requests and has one of the highest footfall in London. For the main study the counting was done such that each location has at least 2/3 counts in the same time interval.

2. Will you give more detailed information about the data and the parameters for various clustering algorithms? Including the number of records, the number of clusters. How about the performance for different clustering methods in your study ?

> Thank you. This part has been expanded and clarified in the revised manuscript. The performance of the algorithms used for one dimensional clustering were evaluated on the basis of the time taken for processing and the resulting mean absolute percentage error. We found that the k-means gives us the best data for the quickest processing time. The graph based clustering was done using two parameters - sequence threshold of 60 and time threshold of 16 seconds. These values were arrived arrived at using trial and error by using a know device as the control mechanism. This has been illustrated further in the revised manuscript.

4.3 Corrections and Clarifications

More specific comments can be found in the following.

1. Minor formatting issues and typos needs to be fixed. Figure 2, no value and units in the x-axis and y-axis. Table 1, and Figure 4 are in wrong location.

> Thank you. This has been corrected in the revised version of the manuscript. Figure 2 is just a schematic showing the idea of cut off and clustering hence there are now values on the axes. We have included the units in the description

2. Related work: How about WiFi-based real-time data analytics?

> We have included reference to such work in our extended discussion of the relevance of the work to the smart cities discourse.

3. Methodology: Will you draw a diagram of the Methodology?

> We have resequenced some of the material in response to Referee 2 and, cognisant of over-all length requirements, think that this (rather than a diagram) addresses your concerns.

4. Performance Analysis: How about the performance for collecting data using PostgreSQL? Did you compare Apache Flink and Spark Streaming for real-time data collecting?

This is an interesting point in the context of big data and tools for tackling big data. Though the data set here is large enough to make desktop computing unfeasible, it is neither big nor fast enough for a large scale big data system. An open source and free toolkit comprising of text processing tools and relational database management system was enough to handle the complexity, size and speed of the data. The toolkit employed here involves PostgreSQL, Bash, Unix tools for text processing (sed, awk, sort, join), Node.js and R. The same toolkit proved to be useful in the larger project (smart street sensor) providing near real-time delivery of data.

5. The reference list needs to be written in a standard format. Some reference papers are lacking specific pages .Please use a standard way of reference list.

Thank you. This has been corrected in the revised version of the paper.