SINTEF

Statistical Modelling for Estimation of OD Matrices for Public Transport Using Wi-Fi and APC Data

Jan Erik Håkegård, Tor Andre Myrvoll and Tor Rune Skoglund

Statistical Modelling for Estimation of OD Matrices for Public Transport Using Wi-Fi and APC Data

Jan Erik Håkegård, Tor Andre Myrvoll
SINTEF Digital
Trondheim, Norway
jan.e.hakegard@sintef.no

Tor Rune Skoglund
FourC
Trondheim, Norway
Email: trs@fourc.eu

Abstract—In this paper, statistical models are proposed to estimate trip level origin-destination (OD) matrices for public transport based on Wi-Fi data traffic. Wi-Fi monitoring equipment installed in 32 buses in Stavanger, Norway, collected Wi-Fi data during several months. The median received signal level of frames transmitted by a device and the time interval between the first and last frame are modelled as statistical distributions, conditional on whether the Wi-Fi device is on the bus or not. Based on these models and using passenger load data from Automatic Passenger Counting (APC) systems installed in the buses, the probability for each detected device being on or off the bus is estimated. When tested on large data sets, the proposed statistical method is more accurate than when hard thresholds for median received signal level and time interval of observation are applied.

I. Introduction

Information about where passengers board and alight public transportation vehicles is useful for planning and operation of public transport networks. This information is typically represented in origin-destination (OD) matrices, which must be estimated based on data from different sources. What kind of data that is available varies between types of transportation and between companies [1]. For some networks, boarding information is obtained by the Automatic Fare Collection (AFC) system or SmartCard transactions [2], while alighting information is lacking. Other networks have Automatic Passenger Counting (APC) systems that provide boarding and alighting counts [3]. For some networks, passengers board and alight the vehicles without being registered at all [4]. How well the OD matrix can be estimated depends on the available input data.

Estimation of the OD matrix using different sources of information has been the topic for research during several decades [1], [2], [3]. The topic of this paper is how data from Wi-Fi devices carried by passengers can be used to obtain an estimate of the OD matrix when passenger load data are available from an APC system installed in the bus.

Most smartphones today have an integrated Wi-Fi radio. Recently, the possibility to track mobile devices by sniffing Wi-Fi traffic has attracted significant interest. This is partly because it provides a source of valuable information to obtain statistical data as the percentage of people carrying a Wi-Fi enabled devices is so high [5], [6], and partly because it

represents a privacy threat as it has been made public that private sector companies as well as state sponsored intelligence agencies in several countries are actively attempting to track smartphone users [7], [8]. A priority in this activity has been to assure that all data that are processed and stored are anonymized, making it impossible for anyone with access to the data to link them to any device or to any user.

The contribution of this paper is to apply the method proposed by the authors in [9] to estimate the OD matrix based on monitoring data of Wi-Fi data traffic and APC passenger load data. Wi-Fi monitoring equipment is installed in 32 buses in Stavanger, Norway, and continuously collects Wi-Fi data traffic. The data are then anonymized and transferred to a central server for further processing. The median received signal level of frames associated with a Wi-Fi device and the time interval between the first and the last frame are modelled using statistical distributions, conditional on whether the device is on the bus or not. By optimising the parameters of the statistical distributions, and minimising the difference between the resulting estimated passenger load and the passenger load obtained from the APC system, the probability for each of the detected devices being on the bus is estimated.

The remaining part of the paper is organised as follows. In the next section, relevant parts of the Wi-Fi standard are reviewed. In Sec. III, the data collection and pre-processing procedures are described, before results are reported in Sec. IV. In Sec V, lines of further work are proposed, before the conclusions are drawn in Sec. VI.

II. Background IEEE802.11

The lower layers of the Wi-Fi communication protocol are standardized by IEEE802.11. It can be useful to briefly review the parts of the standard that are of importance for the collection and processing of Wi-Fi data.

A. Frame types and sub-types

When mobile devices with the Wi-Fi radio activated are not connected to a Wi-Fi access point, they transmit probe request frames as part of the active scan procedure. Access points receiving probe requests answer with a probe response frame, providing the mobile device with information about networks within hearing distance. Probe requests are of interest, as

they represent the only frame sub-type transmitted by mobile devices not connected to any network. Hence, to be able to detect devices not connected to an access point, probe request frames must be collected for further processing.

Access points on their side transmit beacon frames at regular time intervals (typical approximately 100 ms). Beacons are of particular interest, as generally only fixed access points transmit beacons. An exception is mobile devices with the Wi-Fi radio in tethering mode, i.e. taking the role of an access point. These are however assumed to represent only a small part of the mobile devices. Devices transmitting beacons are therefore discarded, as we are only interested in mobile devices.

Most of the other frame sub-types may be transmitted by both mobile devices and access points, and the best choice will be to include them in the processing.

B. Frame format

The frames contain a header and a payload.

1) Frame headers: The header contains several fields, and some of these are of importance for the algorithm. Most important is the source MAC address, as it allows the algorithm to link frames transmitted by the same device.

Another field that is of importance for the algorithm is the sequence number, which is incremented for each transmitted frame. The sequence number range is from 0 to 4095 for most devices. For some devices however, it is between 0 and 15. This is probably done as a measure to make tracking of devices not connected to an access point more difficult.

2) Frame payload: The payload contains several Information Elements (IEs). Which IEs that are permitted in a frame depend on the frame type and sub-type. Some IEs contain information about the capabilities of the Wi-Fi radio, and are assumed to be similar for the same vendor, type of device and perhaps the version of operative system. In [7], the IEs of probe requests are analysed, and the entropy of the elements assessed based on data sets.

C. Local and global MAC addresses

The MAC address is a 48-bit number. The seventh bit is called the local bit. If it is 0, the MAC address is global, which means that it is fixed and unique for the Wi-Fi radio of the device. When a mobile device is connected to a network, it always includes its global MAC address in the header of the frames it transmits. If the local bit is 1, the device generates another MAC address that is included in the frame headers. This MAC address is not unique. With a few exceptions, only probe requests are transmitted using local MAC addresses.

1) Randomization of local MAC addresses: Devices transmitting probe requests with the local bit set do in addition randomize the MAC address. This means that the MAC address changes about every 5-15 seconds. The reason for this feature is privacy concerns, and the goal is to make tracking of the device more difficult. It is currently mainly Apple products that randomize the MAC address. This may however change in the future, as also recent Android and Windows versions allow MAC randomization.

2) Hashing of global MAC addresses: For frames transmitted with global MAC addresses in the header, privacy regulations will be violated if the data are stored without any further processing to assure anonymity. Therefore, global MAC addresses are hashed by the monitoring device before being transmitted to the central server for processing and storage. How this is done as described in Sec. III-B2.

III. SIGNAL PROCESSING

A. Data sources

Data from five different sources are required to estimate the OD matrix, each with its own format.

The format of the monitor data is compressed csv files, where information from each sensed Wi-Fi frame representing an event forms one line. The monitoring devices use the radiotap standard for reception of frames. It provides information about radio parameters such as received signal level and frequency channel, in addition to a timestamp. Position data are obtained by an external GNSS receiver.

Real-time data are provided by a Service Interface for Real-time Information Vehicle Monitoring (SIRI VM) web service. The SIRI data contain among others timestamps, locations, bus line and trip ID.

Information about the trips is collected from a NeTEx web service. The NeTEx data contain information about the bus stops, including name, location, and scheduled arrival and departure times.

Finally, APC data are received from the bus operator. The data contain information about scheduled and real arrival times at each stop, the boarding and alighting counts for each stop, the passenger load between each of the stops, and finally the distance between stops in meters. It must be noted that APC data alone do not give the OD matrix, as there is no link between where each passenger boards and alights the bus.

B. Pre-processing of data

The parameters that are used by the algorithm are extracted from the monitoring data, and all events with the same source MAC address are then associated with the same device. Frames with a local source MAC address in the header are processed in one way, while frames with a global source MAC address in the header are process in another way.

1) Pre-processing of probe request frames with local MAC: Probe requests transmitted by a device that randomizes the local MAC address will be associated with different devices as the MAC address changes. This will introduce errors in the OD matrix estimation algorithm. One option is to disregard all probe requests with randomized MAC. If the percentage of devices that do not randomize the MAC address is sufficiently large, a good estimate of the probabilistic OD matrix may still be obtained. Combined with APC data, the probabilistic OD matrix may provide a good estimate of the actual OD matrix.

An alternative approach is to try to counteract the randomization applied by the manufacturer. Such algorithms are proposed in the literature [10]. In [7], a two-step approach is proposed. First, IE fingerprinting is used to divide the

devices into groups. Hence, devices with the same IEs in the payload of the probe requests are associated with the same group. Then, the sequence numbers are used to connect devices belonging to the same group. Profiting from the fact that sequence numbers are not reset when the MAC address changes, and defining thresholds for time separation T_{th} and sequence number distance S_{th}, devices that have implemented randomized MAC addresses can be tracked. Both thresholds were set relatively high, to 1000 seconds and 450, respectively.

2) Pre-processing of frames with global MAC: Frames with a source MAC address that is not anonymized by the transmitting device, i.e. frames having a global source MAC address, need to be anonymized due to privacy regulations. Simple hashing of the MAC address is however not sufficient according to legal advisers, since known MAC addresses then can be checked towards the hash value at a later stage. This problem is solved by attaching randomly generated and distributed salt values to all Wi-Fi monitors. These salts are used together with the source MAC address to create one-way hash values. The computations are performed in RAM and in real time. The salts are automatically discarded from all systems when their valid time intervals expires. This implies that the same MAC address, if seen during a later journey, will not generate the same hashes. Due to this, it is also impossible to check if a known MAC address is present in the data set. This also implies that the travel pattern for a single device over longer time than a typical journey is not possible to generate.

C. Estimation of OD matrix using Wi-Fi and APC data

The algorithm estimating the OD matrix must basically contain two steps. First, the most likely boarding and alighting stops for each detected device (if it is on the bus or not) must be estimated. Then, a subset of the detected devices must be selected as being on the bus.

The most straightforward way to estimate the boarding and alighting stops of a device is to set the boarding stop to the stop closest in time or distance to the first event associated with the device, and the alighting stop to the stop closest in time or distance to the last event associated with the device. An alternative approach is proposed by [11], which is based on the assumption that the time interval between consecutive events follows an exponential distribution. It calculates the probability for each boarding-alighting pair for a device using the timestamps of the events, the number of events, and the bus stop arrival times. This approach proves however to be significantly less accurate than the straightforward approach selecting the closest stops to the first and last event.

Selection of the subset of devices being on the bus can be done using hard thresholds or by taking a probabilistic approach.

1) Using hard thresholds: The most straightforward approach to determine which Wi-Fi devices that are on the bus is to use hard thresholds. The parameters used in this work are the time interval between the first and last event associated with a device and the median received signal level of all events associated with the device. We define the indicator I_k to be 1

if device k is on the bus and 0 when device k is not on the bus:

$$I_k = \begin{cases} 1, & \Delta t_k \geq \Delta t_{th} \text{ and } m_k \geq m_{th} \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

where Δt_k is the time interval between the first and last event associated with device k, Δt_{th} is the threshold value for this time interval, m_k is the median received signal level for all events associated with device k, and m_{th} is the corresponding threshold value.

When the subset of devices being on the bus are selected, and as the boarding and alighting stops for each device are estimated, the passenger load between each of the stops can be estimated as:

$$\hat{l}(i) = \hat{l}(i-1) + \hat{b}(i) - \hat{c}(i), \ i = 1, \ldots N_s, \quad (2)$$

where $\hat{b}(i)$ is the estimated number of passengers boarding at stop i, $\hat{c}(i)$ is the estimated number of passengers alighting at stop i, $\hat{l}(i)$ is the estimated number of passengers between stop i and stop $i+1$, and N_s is the number of stops. In most cases, $l(0)$ will be zero. In some cases with circle lines, there will however be passengers on the bus when it arrives at the first stop so that $l(0) > 0$. Similarly, the bus is not necessarily empty after the last stop, so that $l(N_s) > 0$. The system must therefore have a memory that extend one trip for each bus.

There is no ground truth OD data available. It is therefore not possible to quantify the accuracy of the OD matrix estimation using metrics like the Hellinger Distance (HD) or relative performance (RP) as proposed in e.g. reference [3]. Instead, a metric measuring the difference between the resulting estimated passenger load and the passenger load obtained from the APC data is applied:

$$\varepsilon_{\Delta t_{th}, m_{th}}(\hat{L}, L) = \frac{\sum_{i=1}^{N_s-1} |\hat{l}(i) - l(i)|}{\sum_{i=1}^{N_s-1} l(i)}, \quad (3)$$

where $\hat{L}$ and L are the vectors of estimated and APC passenger loads along the route, respectively. The vector $\hat{L}$ depends on the devices selected to be on the bus, which in its turn is a function of the thresholds. The thresholds are then selected to minimize the distance between the estimated passenger loads and the passenger loads from the APC system:

$$(\Delta t_{th}, m_{th}) = \arg \min_{\Delta t_{th}, m_{th}} \left\{ \varepsilon_{\Delta t_{th}, m_{th}}(\hat{L}, L) \right\} \quad (4)$$

2) Using probability density functions: In [9], a statistical approach is proposed as alternative to hard thresholds. Rather than using hard thresholds, m and Δt are assumed to follow statistical distributions, conditional on whether the device is on the bus or not. The median received signal level is assumed to be normally distributed with parameters (μ_0, σ_0^2) when the device is off the bus, and with parameters (μ_1, σ_1^2) when the device is on the bus. Two different distributions have been implemented for the time interval between first and last event associated with a device, the exponential distribution and the log-normal distribution. The exponential distributions are characterised by the means λ_0 and λ_1 for devices off and on the bus, respectively. The log-normal distributions
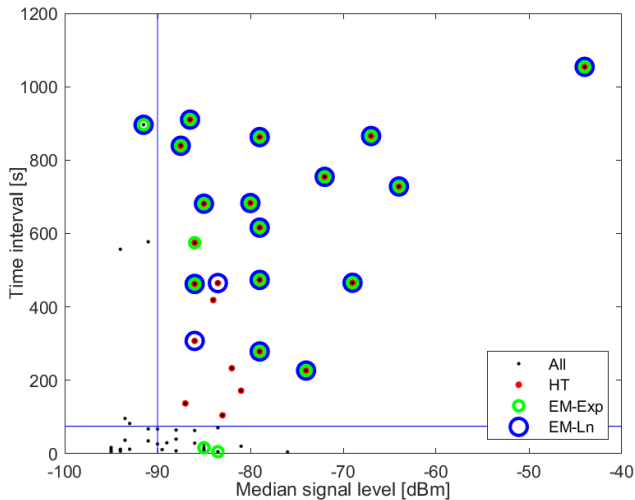
Fig. 1. Median received signal level-time interval diagram using hard thresholds (HT) and probabilistic approach (EM-Exp and EM-Ln)

Fig. 2. Histogram and pdf for the median received signal level for devices off the bus (upper plot) and devices on the bus (lower plot)

are similarly characterised by the parameters $(\mu_{ln,0}, \sigma^2_{ln,0})$ and $(\mu_{ln,1}, \sigma^2_{ln,1})$. The λ value of the exponential distribution corresponds to the inverse of the mean time interval between the first and last event.

The statistical parameters are determined by minimising the difference between the resulting estimated passenger load and the passenger load data obtained from the APC system. The joint optimisation is done using the EM algorithm, as explained in detail in [9]. The algorithm provides a probability value for each device being on the bus. This value can be used as soft information in the OD matrix, or a threshold value can be set to obtain a hard decision for each device.

IV. RESULTS

Monitoring devices are installed in 32 buses in Stavanger, Norway. They have continuously collected data for several months. The results included in this publication are however based on processing of data from three bus lines during three days in August 2017. In order to first illustrate how the algorithms work, results from one typical trip is selected. Then the accuracy of the different approaches are compared for a larger number of trips.

A. Results from one example trip

The trip is selected because it contains a relatively small number of detected devices. The total number of detected devices was 1198. After removing devices assumed to be access points because they transmit beacons, probe responses or data frames with the fromDS bit set, 461 devices are left. Requiring in addition that the device must be on the bus at least over one stop, only 55 devices remain as candidates to being on the bus.

In Fig. 1, the median signal level and time interval for the candidate devices are plotted. The 55 devices are illustrated by a black dot.
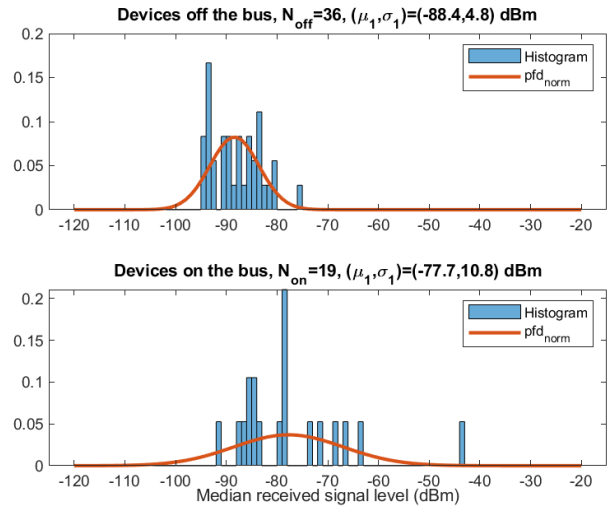
The hard threshold (HT) algorithm calculated the optimal thresholds to be 75 seconds and -90 dBm using eq. (4). The optimal thresholds are illustrated by the blue lines in the figure. All devices observed during a longer time interval than 75 seconds and with median signal level above -90 dBm are assumed to be on the bus, and these are marked by a red dot.

The EM algorithm selects mostly the same devices as the HT algorithm, but there are some differences. When the time interval is modelled as exponential distributions (EM-Exp), 19 devices are selected. 16 of these are also selected by the HT algorithm. When the time interval is modelled as log-normal distributions (EM-Ln), 17 devices are selected. Two of these were not selected when the time interval was modelled as exponential distributions. The figure illustrates well that the algorithms select more or less the same subset of the devices as being on the bus, but that there are a few differences.

Fig. 2 illustrates the resulting pdfs of the median received signal level, and compare them with histograms of the data. The upper plot corresponds to devices that are located outside the bus, and the lower plot to devices that are located on the bus. The difference in mean value for devices off and on the bus is about 10 dB (-88.4 dBm versus -77.7 dBm).

The standard deviation of the pdf of devices on the bus is more than twice as large as for devices off the bus (10.8 dBm versus 4.8 dBm). This illustrates the fact that the received signal levels may vary greatly for frames transmitted by devices on the bus, depending on whether the device is located close to the monitor or not, whether it is in a bag on the floor or in use with line-of-sight to the monitor etc.

Fig. 3 illustrates the resulting pdfs of the time intervals between first and last detection when it is modelled as an exponential distribution. For frames transmitted by devices off the bus, the mean time interval of observation is 88.3 seconds, while it is 608 seconds for devices on the bus.
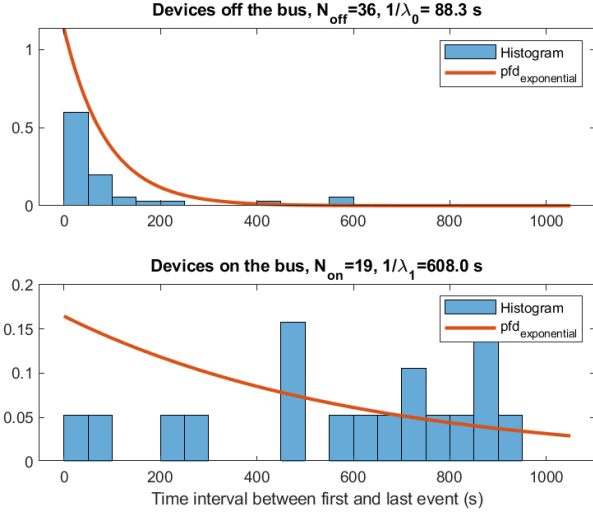
Fig. 3. Histogram and exponential pdf for the time interval of observation for devices off the bus (upper plot) and devices on the bus (lower plot)
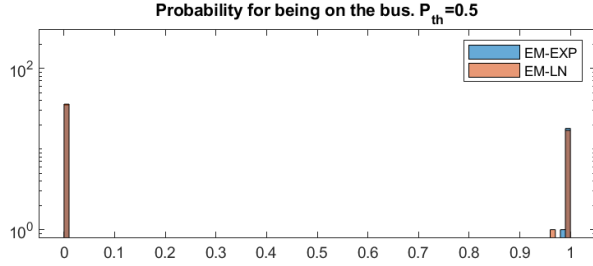

Fig. 4. Histogram of the probabilities of being on the bus using the probabilistic approach

Fig. 5. Comparison between estimated passenger loads, boardings and alightings using hard thresholds and probabilistics approach, and APC data

Fig. 6. CDF of the difference between estimated passenger loads and APC data

The algorithm selects devices not only based on time interval and on median received signal level. A third condition is that the estimated passenger load should be as close to the APC passenger load as possible. A potential source of error is that the APC passenger loads do not correspond to the number of devices on to bus. For one, any APC system is affected by some error. Moreover, there is not a one-to-one relation between passengers and detectable devices. If for instance 90% of the passengers carry a detectable Wi-Fi device, the algorithm may add several detected devices that are not actually on the bus in order to reduce the error between the estimated passenger load and the passenger load from the APC data.

When the statistical parameters are set, the algorithm calculates the probability for each device being on the bus. Fig. 4 shows a histogram over the probabilities for each of the devices being on the bus. The probabilities are in most cases very close to 0 or to 1.

Fig. 5 shows the passenger load estimates resulting from the HT, EM-EXP and EM-LN algorithms together with the numbers from the APC system. The figure also contains the cumulative number of boardings and alightings. According
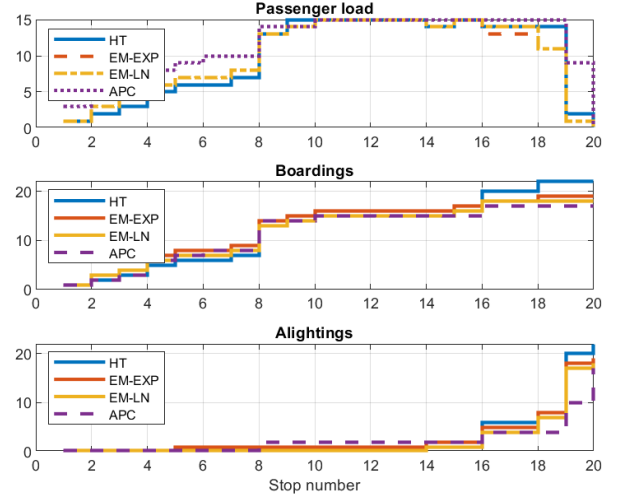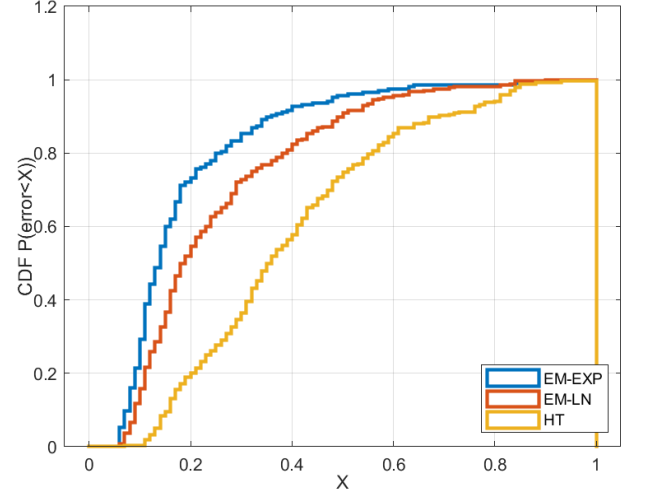
to the APC system, the total number of passengers boarding during the trip was 17. For this trip, the EM-LN algorithm provides boarding and alighting numbers closest to the APC data, while the HT algorithm significantly over-estimates the number of passengers.

B. Passenger load accuracy

In order to assess that accuracy of the algorithms with more confidence, the OD matrix is estimated for a large number of trips, and the corresponding passenger load estimates are calculated and compared to numbers from the APC system.

Fig. 6 shows the cumulative density functions (CDFs) of the error for the three algorithms based on 278 trips. The figure indicates that the EM algorithm outperforms the HT algorithm,

and that it is better to model the time interval between the first and the last frame as an exponential distribution than as a log-normal distribution. For instance, for about 78% of the trips, the EM-EXP algorithm leads to an error lower than 0.2. For the EM-LN algorithm, the error is lower than 0.2 for 55% of the trips, while the error is lower than 0.2 for only about 20% of the trips using the HT algorithm.

V. FURTHER WORK

The work presented in this paper will be continued along several lines. The large amount of data available opens for a more thorough investigation of the statistical characteristics of the key parameters. Analysis of the received signal levels, temporal and spatial characteristics of the events etc. may lead to more accurate statistical distributions for the modelling.

As mentioned earlier, there is in general not an one-to-one relationship between number of detectable Wi-Fi devices on board the bus and the number of passengers. This should be included in the model, for instance by reducing the importance of the error in passenger load in the EM algorithm, or by introducing a constant factor in the HT algorithm accounting for the assumption that only a certain percentage of the passengers can be detected.

In the current implementation of the algorithms, randomized MAC addresses are connected using fixed thresholds for time intervals, sequence number distance, and median signal level. Methods to optimize the thresholds and statistical methods as alternative to hard thresholds should be investigated.

The results presented in this paper rely on the availability of APC data. If APC equipment is installed in only a part of the buses, it could be of interest to train the statistical models on trips with APC data, and then use the trained models on trips without APC data. More work should be done to assess how the statistical parameters of the models vary from day to day, as function of time of day, and between bus lines.

The resulting OD matrix can be as used as a priori information or seed matrix in the Iterative Proportional Fitting (IPF) method [6], [12]. With no a priori information, the IPF procedure starts with a matrix where all downstream stops are equally probable. With a priori information, the base matrix may be closer to reality and the IPF matrix becomes more accurate.

Detection of transfers and estimation of end-to-end OD matrices would provide additional information to public transport companies. If boarding data from AFC systems or SmartCard transactions can be used to connect multi-hop travels [2], these can be used together with the Wi-Fi data to provide end-to-end OD matrices. Without such data, Wi-Fi data must be used to detect transfers. This is currently challenging due to the anonymization of the MAC addresses. Still, this is a research direction for further research.

The amount of data to transfer to and store at the central server quickly becomes considerable. Work is required to make the complete process more efficient, and to present the result for the bus operator or others in a straightforward way.

VI. CONCLUSIONS

The results presented in this paper illustrate how Wi-Fi data and statistical models can be used to estimate the OD matrix. Passenger load data available from an additional APC system provide information that enables the algorithm to select the detected devices in such a way that the estimated passenger load is close to the APC data. Although ground truth OD data are not available to precisely quantify the accuracy, it seems very likely that a small error in estimated passenger load means small errors in the OD matrix. This method may therefore provide public transportation companies with valuable information not available directly from the APC system when it comes to passenger flows.

REFERENCES

[1] A. Cui, "Bus Passenger Origin-Destination Matrix Estimation Using Automated Data Collection Systems," MS Thesis, Massachusetts Institute of Technology, 2006.

[2] E. Wang, J.P. Attaniccu, N.H.M. Wilson, "Bus Passenger Origin-Destination Estimation and Related Analyses Using Automated Data Collection Systems," Journal of Public Transportation 14, no 4, Dec. 2011.

[3] Y. Ji, R.G. Mishalani, M.R. McCord, "Iteratively Improving the Base Matrix of the IPF Method for Estimating Transit Route-level OD Flows from AOC Data," 13th World Conference on Transport Research, 15-18 July, 2013.

[4] S. Terabe, A. Hidayat, H. Yaginuma, "Wi-Fi sensing technologies for affordable travel data collection from bus passengers," The 11th Conference on Transport Survey Methods, Quebec, Canada, 24-29 September, 2017.

[5] M. Dunlap, Z. Li. K. Henrickson, Y. Wang, "Estimation of Origin and Destination Information from Bluetooth and Wi-Fi Sensing for Transit," Transportation Research Record: Journal of the Transportation Research Board, No 2595, Transportation Research Board, Washington DC, 2016, pp.11-17.

[6] R.G. Mishalani, M.R. McCord, T. Reinhold, "Use of Mobile Device Wireless Signals to Determine Transit Route-Level Passenger Origin-Destination Flows," Transportation Research Record: Journal of the Transportation Research Board, No. 2544, Transportation Research Board, Washington DC, 2016, pp.123-130.

[7] M. Vanhoef et al., "Why MAC Address Randomization is not Enough: An Analysis of Wi-Fi Network Discovery Mechanisms," AsiaCCS, May 2016.

[8] J. Martin, T. Mayberry, C. Dunahoe. L. Foppe, L. Brown, C. Riggins, E.C. Rye, D. Brown, "A Study of MAC Address Randomization in Mobile Devices and When it Fails," Proceedings on Privacy Enhancing Technologies, no. 4, 2017.

[9] T. A. Myrvoll, J.E. Håkegård, T. Matsui, F. Septier, "Counting Public Transport Passenger using Wi-Fi Signatures of Mobile Devices," IEEE ITSC 2017, Yokohama, Japan, Oct. 2017.

[10] C. Matte, M. Cunche, F. Rousseau, M. Vanhoef, "Defeating MAC Address Randomization Through Timing Attacks," 9th ACM Conference on Security and Privacy in Wireless and Mobile Networks (ACM WiSec), July 2016.

[11] Y. Ji, J. Zhao, Z. Zhang, Y. Du, "Estimating Bus Loads and OD Flows Using Location-Stamped Farebox and Wi-Fi Signal Data," Hindawi, Journal of Advanced Transportation, Vol. 2017, Article ID 6374858.

[12] D.S. Navick, P.G. Furth, "Distance-Based Model for Estimating a Bus Route Origin-Destination Matrix," Transport Research Record 1433, 1994.