Towards the Retrieval of Accurate OD Matrices from Bluetooth Data: Lessons Learned from 2 Years of Data

Gabriel Michau, Alfredo Nantes, Edward Chung

Smart Transport Research Centre
Queensland University of Technology,
QLD 4000 Brisbane
Australia

Gabriel.Michau@gmail.com

Submitted for the 36th Australasian Transport Research Forum (ATRF) annual conference
Queensland University of Technology, Brisbane
Wednesday 2 to Friday 4 October 2013

Abstract:

The Bluetooth technology is being increasingly used to track vehicles throughout their trips, within urban networks and across freeway stretches. One important opportunity offered by this type of data is the measurement of Origin-Destination patterns, emerging from the aggregation and clustering of individual trips. In order to obtain accurate estimations, however, a number of issues need to be addressed, through data filtering and correction techniques. These issues mainly stem from the use of the Bluetooth technology amongst drivers, and the physical properties of the Bluetooth sensors themselves. First, not all cars are equipped with discoverable Bluetooth devices and the Bluetooth-enabled vehicles may belong to some small socio-economic groups of users. Second, the Bluetooth datasets include data from various transport modes; such as pedestrian, bicycles, cars, taxi driver, buses and trains. Third, the Bluetooth sensors may fail to detect all of the nearby Bluetooth-enabled vehicles. As a consequence, the exact journey for some vehicles may become a latent pattern that will need to be extracted from the data. Finally, sensors that are in close proximity to each other may have overlapping detection areas, thus making the task of retrieving the correct travelled path even more challenging.

The aim of this paper is twofold. We first give a comprehensive overview of the aforementioned issues. Further, we propose a methodology that can be followed, in order to cleanse, correct and aggregate Bluetooth data. We postulate that the methods introduced by this paper are the first crucial steps that need to be followed in order to compute accurate Origin-Destination matrices in urban road networks.

Introduction

The complete knowledge of travel demand is the cornerstones for many applications, from transport demand modelling, to design of traffic management schemes (Willumsen 1978).
Knowing the actual demand is important, in order to establish the effectiveness of the network in handling the need of the road users; and to measure the impact of network changes on the overall traffic flow. For practical reasons, this knowledge is very hard to forecast; often the demand is determined through a comparison between the current traffic situation and individual's *stated* preferences(Bates 1982, Louviere 1988, Hensher 1994, Fujii and Gärling 2003); or through forecasting models that rely on assumptions about the evolution of the traffic state. In any case, a good estimate of the present state of the network is a key, preliminary point to any mobility analysis, and therefore a problem of great interest. The state of the network can be described by several indicators, such as the *travel time*, which helps to quantify the level of congestion of the network; and the *Origin/Destination* (OD) *matrix,* often used to track traffic volumes, over space and time.
To obtain the OD matrices, the area covered by the network is usually partitioned into smaller geographic zones, which are in turn represented by their centroids. In general, associated with these centroids are the *power of attraction* (or a potential of being a destination) and *power of production* (or a potential of being an origin). The OD matrices are double-entry tables, *M*. Each element M_{ij} of the matrix contains a census of the volume of journeys, from origin *i* to destination *j*.

Until now, the Origin Destination matrices themselves have been retrieved through expensive surveys and/or from assignment algorithms, which infer about the OD patterns from the traffic counts. Although effective, these surveys capture *stated* behaviour, as opposed to the *observed* behaviour captured by Automated Vehicle Systems (AVI). As such, these methods may exhibit strong bias, due to the subjective nature of the user perception. On the other hand, Origin Destination Count-Based Estimation relies on strong assumptions, in order to solve the underdetermined systems that may result from the assignment of routes, according to the limited observations (vehicle counts) available.
Recent technological advances have led to the first AVI systems. As the aim of these systems is to track individuals' behaviours, the improvement of computers capacity was a necessary step to enable the processing of the numerous data. Nowadays, the technologies that are largely used for AVI purposes are plate recognition, GPS and Bluetooth track recording, amongst others.
The Bluetooth technology features some major advantages. Firstly, this technology is particularly suitable for urban networks, as it enables the detections of the discoverable Bluetooth devices in the surrounding of the Bluetooth scanners. Secondly, the Bluetooth scanners are easier to install and maintain compared to plate recognition systems. Indeed, the Bluetooth scanners do not require accurate calibration, as the effectiveness of the detection does not depend on the orientation of the scanners or the vehicles. Thirdly, in most cases a single Bluetooth scanner can be used to capture the traffic at the intersections, regardless of the direction of travel of the vehicles. In contrast, many plate recognition systems are usually needed, one for each direction of travel. Finally, the detection is anonymous, in that the electronic identifier (MAC address) of the detected vehicles can be converted into an encrypted (hash) code, at the sensor site. All these advantages make the Bluetooth technology very appealing, as far as concerns the monitoring of traffic.

Related Work

The Bluetooth data has been extensively used as a reliable source for the estimation of travel time along corridors (Malinovskiy, Lee et al. 2011, Araghi, Krishnan et al. 2012, Araghi, Pedersen et al. 2012, Mitsakis, Grau et al. 2013). It has proven to be a reliable and convenient source of data, due to large amounts of samples that can be collected, and the ease to collect them. This kind of data has also been used for analysing the level of congestion at the intersections, based on the detection time, and the duration of transit (Tsubota, Bhaskar et al. 2011). Van Der Zijpp (1997) discussed the potential of AVI systems for the estimation of Origin-Destination matrices. Since then, further research

has been conducted into the Bluetooth-based data collection, for improving the estimation of these matrices. From the Bluetooth-based travel time analysis, Barceló, Montero et al., amongst others, presented a methodology for estimating Origin-Destination Matrices, along corridors (Barceló, Montero et al. 2010) (freeway with 11 entries and 12 exits) and in urban networks (Barceló, Montero et al. 2012), by using a limited number of detectors(48). Analogous work was conducted by Blogg, Semler et al. (2010), who presented two cases studies in the Brisbane metropolitan area: one with two OD pairs and one with 29 detectors. Yucel, Tuydes-Yaman et al. (2012) presented a case study in Ankara for an open system composed of 10 intersections and 4 major roads, equipped with 4 Bluetooth devices. Carpenter, Fowler et al. (2012), discussed a new opportunity offered by Bluetooth sensors concerning the route specific Origin-Destination matrices estimation. Their work was based on a single case study in Jacksonville with 14 detection devices spread along one corridor. Most of these works are based on the data collected by a limited number of Bluetooth sensors, scattered throughout the network. Therefore, the Origin Destination issues have only been considered over a limited geographical area, or it was studied by aggregating several data sources (e.g. traffic counts). The availability of more than 260 Bluetooth scanners, within the Brisbane urban area, may create new opportunities, as far as concerns the retrieval of Origin Destination matrices. This paper aims to present these new challenges and the difficulties that come with them.

First, this dense network of sensors can directly be used for the ‘zoning’ of the studied area. Each sensor is considered as a centroid and a geographical zone is then associated with it (for example based on Voronoi partitions). Through this description of the network, it becomes easy to assign the origin and destination of trips for individual drivers, from the first and last detections observed in the Bluetooth data collected. These first and last detections, however, might not correspond to the actual origin and destination, as the trips might continue outside the Bluetooth covered area. Nevertheless, the missing information about the complete trip is not relevant to our work, as our aim is the analysis of the OD patterns within the urban context.

If the sensors are deployed at the most crucial intersections, graphs can be used to accurately describe the road network covered by the Bluetooth sensors (c.f. Figure 1). Such graphs will have sensor as vertexes and links indicating the road links between sensors.

Figure 1: Brisbane's road networks with Bluetooth sensors (blue circles) and the infered networks (blue links).

In a nutshell, our task involves the 'retrieval' of the OD matrices, rather than their 'estimation'. The major differences between this work and previous research are:

- *A more comprehensive knowledge of each journey*. Through the Bluetooth sensors, these journeys are directly available, and do not need to be estimated, for example, through route assignment algorithms.
- *The opportunity to deal with observed trips and travel times, instead of counts*. From these new types of data it is easier to retrieve Origins and Destinations, and enable the retrievals of route specific O/D matrices (Carpenter, Fowler et al. 2012). Route specific matrices are more detailed than ordinary OD matrices, in that they only concern the user of a particular link or path giving information about their origins and destinations.

In the following sections, we will present the challenges that come with the retrieval of the OD matrices. Through a case study conducted in the Brisbane urban area, we will show how the data is affected.

Challenges

Missed detections analysis and recovery

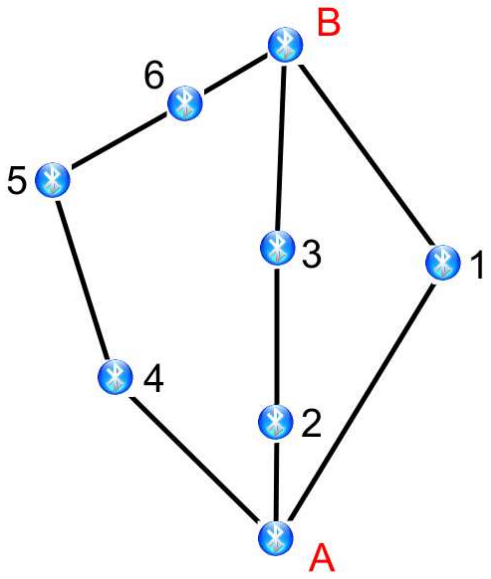

Figure 2: If a user was detected at sensor A and B it was detected twice whereas it should have been detected at least 3 times (in fact 3, 4 or 5 times). Therefore we know that at least one third of the detections are missing.

From the analysis of each pair of successive detections, for each scanned device, we have observed that for more than half of the pairs, at least one detection was missing (c.f. Figure 2). To estimate the minimum number of miss-detections (lower bound) between pairs of scanners, we developed the following heuristic. We first look at the shortest path between scanners, using the Dijkstra algorithm (Schrijver 2003). In our modified version of Dijkstra, the number of detections is used as the cost of a path. This choice is motivated by the observation that the high density of sensors in Brisbane leaves very few possibilities for a driver not to follow the shortest path between two successive detections. Our algorithm takes a list *L* of sequences of detections, as an input. Each sequence contains all the detections D_n for one specific MAC address (i.e. driver) over some chosen period of time (1day in our case study)..

$$L\{i\} = [D_n]_{n\in[1,N]}$$

The list is sorted by increasing time. The output of the algorithm is a list *Tr* of sequences where each sequence corresponds to a journey. The output list contains the index of the Bluetooth device, and the detections D_n, belonging to the same journey. If the same device did several trips during the studied period of time, it would have as many entries in *Tr* as the number of its trips. A detection D_n is composed of the tuple (Id_n, T_n); where Id_n is the index of the detector, and T_n is the time at which the detection occurred.

To create this output list *Tr*, for each sequence *L{i}*, each pair of two consecutive detections *(D_n, D_{n+1})* was analysed. The speed $S_{n,n+1}$ of the device between these two consecutive detections was computed as follow:

$$S_{n,n+1} = \frac{dist(Id_{n+1}, Id_n)}{T_{n+1} - T_n}$$

Where *dist(A,B)* is some metric distance (e.g. Euclidean) between the detectors of indexes *A* and *B*. (Figure 2)

If the speed between two detections is lower than 1km/h, then these detections are unlikely to be part of the same journey and the sequence of detection is cut into two potential different journeys. As for very close sensors, the speed might take any value (caused by noise on the recorded time). We found reasonable to separate both detections if the interval exceeded one hour, that is

If $\mathrm{S_{n,n+1}} < 1\ km/h$ **and** $\mathrm{T_{n+1}} - \mathrm{T_n} > 1\ \mathrm{hour}$ (criteria 1)
$Tr \leftarrow \{i, [D_{1\dots n}]\}$

In this case, D_{n+1} is considered as the beginning of a new journey, for the user *i*.
If the speed is higher than 20km/h, or the inter-time lower than 10 minutes, it is assumed that the detections belong to the same journey. Based on the shortest path, missing detections are then computed and added to the detection sequence.

If $S_{n,n+1} > 20\ km/h$ **and** $T_{n+1} - T_n < 10min$ (criteria 2)
If *(Id_n, Id_{n+1})* adjacent
$[D_{1\dots n}] \leftarrow [D_{1\dots n}, D_{n+1}]$
Else
$\left[\mathrm{Id_n}, \mathrm{Id_{n'}}, \dots, \mathrm{Id_{n'''}}, \mathrm{Id_{n+1}}\right] = Dijkstra(Id_n, Id_{n+1})$
$T_{n'} = \frac{dist(Id_{n'}, Id_n)}{dist(Id_{n+1}, Id_n)}\ (T_{n+1} - T_n)$
...
$T_{n'''} = \frac{dist(Id_{n'''}, Id_n)}{dist(Id_{n+1}, Id_n)}\ (T_{n+1} - T_n)$
$[D_{1\dots n}] \leftarrow [D_{1\dots n}, D_{n'}, \dots, D_{n'''}, D_{n+1}]$

All sequences whose speed or time intervals did not meet criteria 1 or 2 were left aside, for further studies, as the it was not clear whether they belonged to a single journey or not. For future works, the travel time linked to these detections will be compared to similar ones in the same day, in the same 30 minutes interval; or to other days with similar users' behaviours.

From our experiences, it turned out that 10% of the devices were detected only once a day. Another 9% were detected more than one time, but with isolated detections (every pair of detections satisfies criteria 1). Then, 75% of the remaining detections satisfy the chosen criteria (either criteria 1 or 2 - the speed is below 1km/h or the travel time is above one hour, or, above 20km/h and below 10min). Moreover, only 0.75% of the computed journeys have the same Origin and Destination. We notice that this ratio is highly dependent on both criteria 1 and 2.

From these empirical results, the missed detections can be explained as follows:

- Not all scanners and devices are equally powerful, as some have stronger signals than others. From our dataset we observed that some devices were more likely to be detected, compared to others, as shown on the Figure 3. This assumption is supported by the work of Porter, Kim et al. (2012) highlighting the influence of the antenna on the signal strength and detection.
- The miss-detection rate increases, as the scanning area becomes more crowded with active Bluetooth devices. In fact, it is known that when the number of detectable devices increases, interference may affect the effectiveness of the detection (Franssens 2010, SIG 2013). Finally, the maximum number of devices that can be captured by a scanner is limited (3 devices per second, for the scanners located in the Brisbane area).

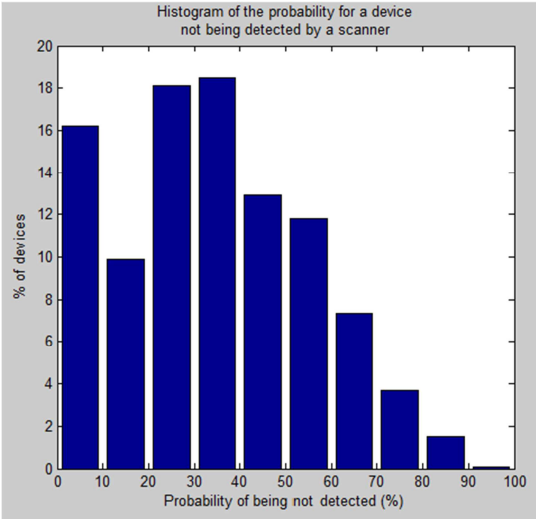

Figure 3: Histogram of the probability for a device not being detected. Two modes are observed. The first mode for a probability of being missed below 10% mainly composed of devices only detected twice by successive detectors and another at 30%.

- The position of the detectors is of great importance, as Bluetooth signals are weakened by physical obstacle (walls, billboard, …). In addition, Brennan Jr, Ernst et al. (2010) have shown that the vertical position of the Bluetooth scanner has an influence of the effectiveness of the sensor.
- The weather as a strong influence on the signal strength.
- Not all Bluetooth devices are always in discoverable mode. (e.g. some devices may become undiscoverable after a few minutes of non-use)
- The scanners detection process can be described as an *inquiry cycle* during which the detector will send inquiry messages on a broad range of frequencies and waiting for devices to answer (Peterson, Baldwin et al. 2004). However, this inquiry cycle needs some time to complete. It is advised (Peterson, Baldwin et al. 2004, SIG 2013) that a Bluetooth device should remain in a discoverable mode (or *inquiry substate*) for 10.24 seconds, within the detection zone of a scanner. Therefore, a device moving at a speed of above 72km/h have a small probability of not being detected by a scanner with a scanning radius of 100m (200m in 10 seconds).

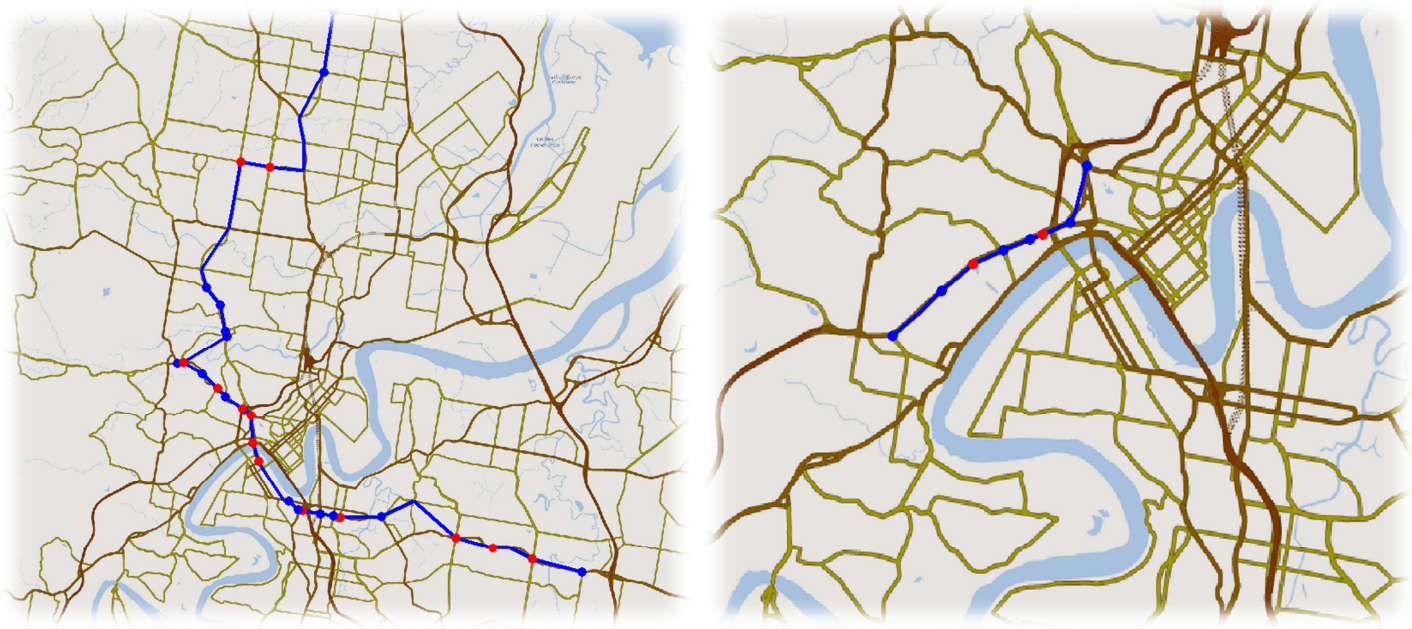

Figure 4: Example of Trips with missing detections (red)

Overlapping detections

The location of the sensor is also of great importance regarding the quality of the dataset collected. Firstly, sensors located too close to each other can have overlapping detection zones. Downstream and upstream scanners might therefore detect a device in the reverse order, yielding erroneous patterns of travel as shown in Figure 5. However, this phenomenon can be easily detected, as the previously described algorithm monitors the speed of each device along the trip that has been synthesized. If a trip contains anomalous speeds, and repeated links between two nearby detectors, this trip will be corrected later, with the removal of the repeated pattern.

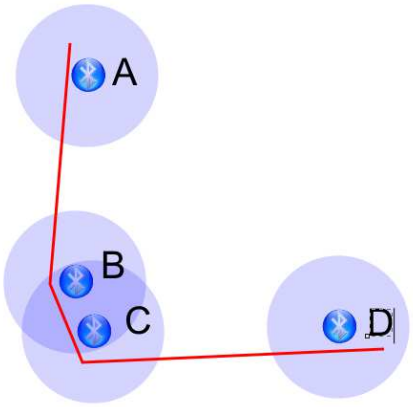

Figure 5: A car following the itinerary ABCD might be detected as ACBD. Therefore the algorithm described previously will compute the itinary ABCBCD. The repetition of the link BC and the anomalous speed resulting makes this effect easily detectable.

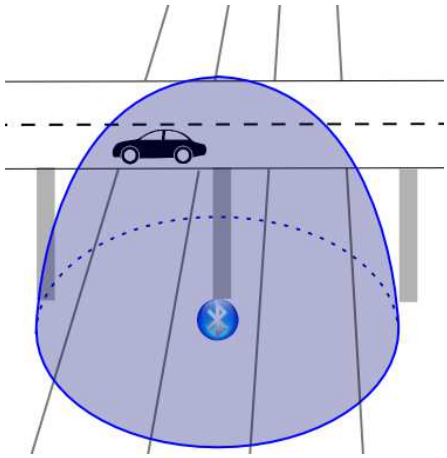

Figure 6: A Bluetooth sensor might detect a car belonging to another corridor than the ones it was installed for. When it happens, the detector seems to be Origin or Destination for the detected device as it will not be detected anymore in the area.

Finally, we observe that the detection area, for some of the sensors, may span across multiple corridors. As a consequence, the traffic that is detected by a sensor may not necessarily belong to the target corridor. Figure 6 shows an example of this phenomenon. In the figure, the detected car is

driving a corridor that is different from the target corridor; that is, the one under the overpass. If no Bluetooth sensors cover this overpass erroneous Origin/Destination patterns may be generated.. Such situations should be detected and properly handled as these scanners will be overestimated Origin or Destination.

Figure 7: The red dot is a sensor located at an intersection below the Pacific Motorway but that detects also cars on it. The red circles are area where sensors overlap.

Uniqueness of MAC address

Although MAC addresses are expected to be unique (SIG 2013) it appeared, from our dataset, that some vehicles are equipped with Bluetooth devices with 'shared' addresses. These artefacts in the data can be easily detected, as they will result in individual vehicles moving at extremely high speed, throughout the network. The algorithm introduced earlier can therefore detect this phenomenon. From our dataset, we observed that around a very small percentage of Bluetooth devices were moving at a speed higher than 120km/h. As such, a solution to this problem could be the removal of the 'suspicious' vehicles from the dataset.

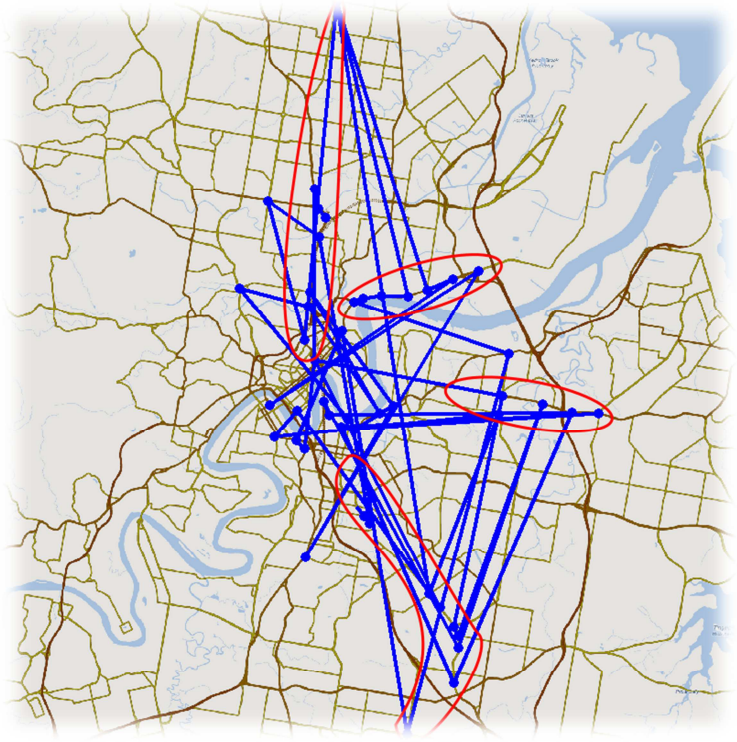

Figure 8: Real detection of a single MAC address between 6:30 and 7:00 am the 3. October 2012 (more than 50 detections). Each link represents two successive detections. The speed computed along the links is often largely over 150 km/h. This sequence reorganised and divided by corridor shows that at least three devices are needed to obtain such sequence with reasonable speed. (red ellipses)

Conclusion and Future Work

The article presented the major issues in the cleaning of Bluetooth data towards the retrieval of OD matrices. As the area covered by Bluetooth networks becomes larger, the data cleansing and correction mechanisms presented here become very important, for each of these issues is likely to affect the accuracy of the results.

As we have discussed earlier, the mode of travel being used is not directly available from the Bluetooth data. Also, the vehicles that are equipped with discoverable Bluetooth devices currently represent a small fraction of the entire traffic. As far as the separation of the modes is concerned, Araghi, Krishnan et al. (2012) have shown that clustering techniques (hierarchical, K-means and two-step) are quite effective to distinguish between motorized and non-motorized users, in uncongested conditions. However, to the best of our knowledge, very little research has been conducted towards distinguishing the various travel modes, within the motorized vehicle class, by only using Bluetooth data. Finally, as for most of the AVI system, Bluetooth sensors cannot give information about the number of travellers per detected vehicle.

In our future research, we will investigate methods for the effective clustering of various transport modes within the Origin-Destination patterns. Then, we will focus on the retrieval of the actual OD Matrices by using the Bluetooth data only. Finally, we will compare such matrices, with other available sources (Household Travel Survey).

References

Araghi, B. N., R. Krishnan and H. Lahrmann (2012). "Application of Bluetooth Technology for Mode-Specific Travel Time Estimation on Arterial Roads: Potentials and Challenges." Trafikdage pa Aalborg Universitet.

Araghi, B. N., K. S. Pedersen, L. T. Christensen, R. Krishnan and H. Lahrmann (2012). Accuracy of Travel Time Estimation Using Bluetooth Technology: Case Study Limfjord Tunnel Aalborg. 19th ITS World Congress. Vienna, Austria.

Barceló, J., L. Montero, M. Bullejos, O. Serch and C. Carmona (2012). Dynamic OD matrix estimation exploiting bluetooth data in Urban networks. Proceedings of the 14th international conference on Automatic Control, Modelling & Simulation, and Proceedings of the 11th international conference on Microelectronics, Nanoelectronics, Optoelectronics, World Scientific and Engineering Academy and Society (WSEAS).

Barceló, J., L. Montero, L. Marqués and C. Carmona (2010). A Kalman-Filter Approach For Dynamic OD Estimation In Corridors Based On Bluetooth And Wifi Data Collection. Proceedings 12th World Conf. on Transportation Research.

Bates, J. (1982). Stated preference technique for the analysis of transportation behavior. Proceedings of World Conference of Transportation Research.

Blogg, M., C. Semler, M. Hingorani and R. Troutbeck (2010). Travel Time and Origin-Destination Data Collection using Bluetooth MAC Address Readers. Australasian Transport Research Forum (ATRF), 33rd, 2010, Canberra, ACT, Australia.

Brennan Jr, T. M., J. M. Ernst, C. M. Day, D. M. Bullock, J. V. Krogmeier and M. Martchouk (2010). "Influence of vertical sensor placement on data collection efficiency from bluetooth MAC address collection devices." Journal of Transportation Engineering **136**(12): 1104-1109.

Carpenter, C., M. Fowler and T. J. Adler (2012). "Generating Route-Specific Origin-Destination Tables Using Bluetooth Technology." Transportation Research Record: Journal of the Transportation Research Board **2308**(1): 96-102.

Franssens, A. (2010). "Impact of multiple inquires on the bluetooth discovery process: and its application to localization."

Fujii, S. and T. Gärling (2003). "Application of attitude theory for improved predictive accuracy of stated preference methods in travel demand analysis." Transportation Research Part A: Policy and Practice **37**(4): 389-402.

Hensher, D. A. (1994). "Stated preference analysis of travel choices: the state of practice." Transportation **21**(2): 107-133.

Louviere, J. J. (1988). "Conjoint analysis modelling of stated preferences: a review of theory, methods, recent developments and external validity." Journal of transport economics and policy: 93-119.

Malinovskiy, Y., U.-K. Lee, Y.-J. Wu and Y. Wang (2011). Investigation of Bluetooth-based travel time estimation error on a short corridor. Proceedings of the 90th Annual Meeting of the Transportation Research Board, Washington, DC.

Mitsakis, E., J.-M. S. Grau, E. Chrysohoou and G. Aifadopoulou (2013). A Robust Method for Real Time Estimation of Travel Times for Dense Urban Road Networks Using Point-to-Point Detectors. Transportation Research Board 92nd Annual Meeting.

Peterson, B. S., R. O. Baldwin and J. P. Kharoufeh (2004). A specification-compatible Bluetooth inquiry simplification. System Sciences, 2004. Proceedings of the 37th Annual Hawaii International Conference on, IEEE.

Porter, J. D., D. S. Kim, M. E. Magaña, P. Poocharoen and C. A. G. Arriaga (2012). "Antenna Characterization for Bluetooth-based Travel Time Data Collection." Journal of Intelligent Transportation Systems(just-accepted).

Schrijver, A. (2003). Combinatorial Optimization, Polyhedra and Efficiency. Heidelberg, Springer Verlag.

SIG. (2013). "Bluetooth Special Interest Group." from https://www.bluetooth.org/en-us/.

Tsubota, T., A. Bhaskar, E. Chung and R. Billot (2011). Arterial traffic congestion analysis using Bluetooth Duration data. Australasian Transport Research Forum 2011, . P. Tisato, Oxlad, Lindsay, & Taylor, Michael (Eds.). 28 - 30 September 2011, Adelaide Hilton Hotel, Adelaid, SA.

Van Der Zijpp, N. J. (1997). "Dynamic origin-destination matrix estimation from traffic counts and automated vehicle identification data." Transportation Research Record: Journal of the Transportation Research Board **1607**(1): 87-94.

Willumsen, L. G. (1978). "Estimation of an OD Matrix from Traffic Counts–A Review." Institute of Transport Studies, Universities of Leed **Working paper 99**.

Yucel, S., H. Tuydes-Yaman, O. Altintasi and O. Murat (2012). "Determination Of Vehicular Travel Patterns in an Urban Location using Bluetooth Technology."