BAYESIAN ROUTE CHOICE INFERENCE USING BLUETOOTH TECHNOLOGY

Francisco Garrido-Valenzuela, Pontificia Universidad Católica de Chile (f.garridov@uc.cl)
Sebastián Raveau F., Pontificia Universidad Católica de Chile (sraveau@ing.puc.cl)
Juan Carlos Herrera M., Pontificia Universidad Católica de Chile (jch@ing.puc.cl)

ABSTRACT

Bluetooth technologies can be used to track specific vehicles by their MAC address; this can be done by installing Bluetooth sensors in selected intersections of a study area. However, a vehicle equipped with this technology is not necessarily detected at every intersection as detection probabilities depend on many factors such as weather, nearby infrastructure or vehicles speeds. Given the lack of perfect information, it is necessary to infer the most likely routes chosen by the vehicles.

This study presents a methodology to infer route choices in a network by reconstructing paths between two successive detections of the same vehicle. Probabilities for each route between these detections points are estimated. Once these probabilities are obtained, it is possible to infer the entire route from vehicle's origin to its destination. Being able to predict these travel patterns is essential for transport planning and different policies such as traffic management and route pricing.

The methodology has two steps. In the first step, with the Bluetooth data, a distribution of the time spent by the vehicles at each intersection is calibrated. Additionally, travel time distributions between intersections are obtained. The second step consist in convolving the previous distributions between two successive detections to obtain aggregated distributions for each potential route. Based on the observed travel time and missed detections, Bayesian inference is performed to obtain the probabilities of each route.

Keywords: Bluetooth technology, Route reconstruction, Bayesian inference.

1. INTRODUCCIÓN

The problem of traffic congestion generates a necessity of understanding in detail how travelers move across the city, with the objective of developing tools that help to reduce the resulting externalities. By having a more detailed travel demand estimation it is possible to build more reliable models to design traffic control policies, transport planning, among others. For these purposes, traditional information collection methods on travel patterns (such as origin-destination surveys, loop detectors, license plate recognition, etc.) provide enough data, but the costs of some of them do not allow to be kept data updated and the level of detail of others is not enough to predict the entire trips.

The usage of wireless technologies such as Bluetooth (BT) and Wi-Fi is increasing and researchers are interested in this type of data to replace and/or supply traditional data collection methods (Bhaskar & Chung, 2013). These wireless signals can precede from smartphones, headphones, multimedia control systems in vehicles or others electronic devices, so it is possible to find many of these gadgets in urban trips as travelers carry them. In this way, the signals from these devices can be a good indicator of where people are and how they move along the network.

Uses of these technologies have been reported in the literature to estimate travel times, levels of congestion analysis at intersections, origin-destination matrices estimation, and vehicle routes reconstruction, among others. Regarding route reconstruction, Barceló *et al.* (2012) present a methodology combining BT data and Kalman filters to estimate OD matrices, Blogg *et al.* (2010) and Carpenter *et al.* (2012) develop a route reconstruction methodology on a highway by installing BT detectors, and Musa & Ericksson (2012), Michau *et al.* (2013) and Michau *et al.* (2017) present vehicle route reconstruction in urban context using wireless networks.

This work proposes a new methodology of vehicle route reconstruction in urban context using data from BT detectors. One of the main challenges related to BT data is that antennas may fail to detect vehicles (Michau *et al.*, 2013). On one hand, a model based on Bayesian theory is proposed, which takes into considerations the lack of complete information produced by the missed detection. On the other hand, the model is not limited by the assumption imposed by Musa & Ericksson (2012) that all turning probabilities at an intersection are equal.

The presented methodology has two steps. On the first step, travel time information on links and dwell time at intersections is preloaded on the network. Then, in the second step, an inference of the vehicle route is performed, using the previously loaded information. In this way, it is possible to obtain a table of probabilities per route for each detected vehicle. Finally, a process for evaluating the effectiveness of the methodology is presented, considering different scenarios by modifying the allocation or number of antennas and their detection probabilities.

2. BLUETOOTH TECHNOLOGY AND DATA

BT is a wireless network that belongs to Wireless Personal Area Networks (WPAN) used to transmit information at close range between a group of devices. The connection exists only if the devices are active with BT mode enable to provide their Medium Access Control (MAC), a code of numbers and letters that identifies each device as unique on the network. The linking is made thanks to the MAC as identifier, which is picked up by the receiver to make the communication.

A Bluetooth sensor (BS) is a device that can be used to observe and registry the communications made by devices. As each gadget can represent a vehicle, it is possible to record the MACs and detection times of cars entering in influence area of antenna. Thus, by locating several BS in specific locations, it is possible to build a database with the position and time of each detection for each entity.

Table 1 shows an example of the data generated by a detection, where BS-1 is the georeferenced BT sensor, MAC is the unique code that identifies a vehicle and then detection time. However, BS may fail to detect all the devices within the sensor area, so there is a detection probability. This detection probability depends on several factors such as: vehicles speeds when passing through the antenna, nearby infrastructure, surrounding networks, weather conditions, among others.

Table 1. Example of detection data

Bluetooth Sensor	MAC Address	Time Stamp
BS-1	0A:B6:8B:24:E6:50	21-02-16 15:24:13

3. METHODOLOGY

This study presents a methodology to infer route choices of BT-equipped vehicles that move along an urban network. To do this, exclusively BT data is used and the lack of perfect information due to missed detections is considered. The methodology has two steps: (i) "Preload of information in the network" where a distribution of the time spent by the vehicles at each intersection is calibrated. Additionally, travel time distributions between intersections are obtained, both distribution are calibrated using BT data. The second step (ii) "Inference of the route" consist in convolving the previous distributions between two successive detections to obtain aggregated distributions for each potential route. Based on the observed travel time and missed detections, Bayesian inference is performed to obtain the probabilities of each route.

To apply the methodology, it is essential to have a defined BS equipped urban area. As shown in Figure 1a, a specific area of the city is selected and this must have BS (blue dots in Figure 1b) scattered and installed in a set of intersections of the network. Once the study area has been chosen, a graph (G) is defined from the roads and BS configuration, each sensor will be a node in the graph and each street will be a directional link based on direction of the traffic flow. G is a (N, L) graph type, where N is the set of nodes (or BS) of the network and L the set of directional links. Figure 1c show an example of how set up the graph from the real network (Figure 1b).

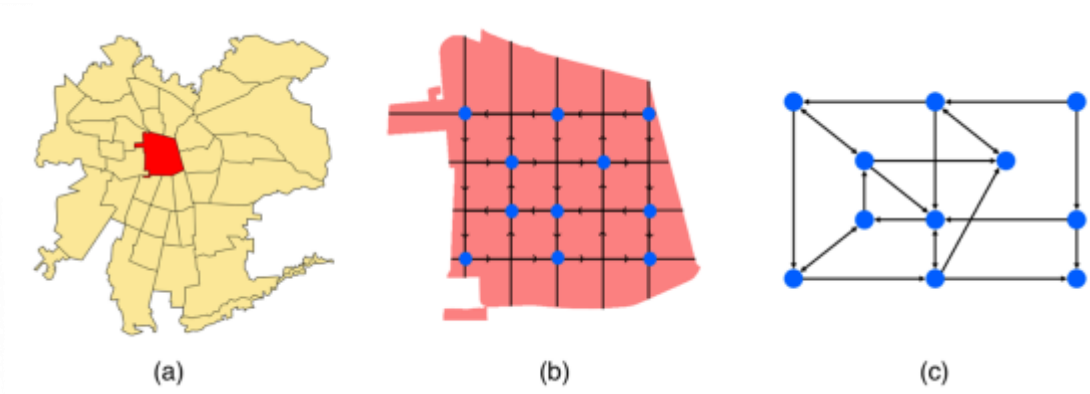

Figure 1. (a) Selected area from a city.; (b) Details of selected area.; (c) Network graph.

3.1 Preload information on network

The objective of this step is to represent the different network congestion states, considering observed travel times. For doing this, dwell time distributions at each node (DT_i) and travel time distributions at each link (TT_j) are calibrated, where $i \in N$ and $j \in L$. Traffic congestion and other possible events could change these distributions, and therefore it is important to update them regularly (e.g. every 10 minutes). So, each distribution is denoted by DT_i^k and TT_j^k, where k is the period of the day.

3.1.1 Dwell time distribution at nodes (DT_i)

The dwell time in nodes is the time vehicles spend at the intersections of the network. This time incorporates the effects of traffic lights, traffic jams or another phenomenon that involves stopping or slowing down. Three cases can occur when a vehicle passes through the influence area of a BS: (i) the vehicle is not detected, (ii) the vehicle is detected only once, and (iii) the vehicle is detected more than once.

Cases (i) and (ii) cannot be used to analyze the time a vehicle spent at the node, so to calibrate the dwell time distributions only case (iii) vehicles are used. This could produce a bias as the data collected will come mostly from low-speed vehicles. Therefore, faster vehicles spend less time in the influence area of a BS than slower cars, thus less detections can be made (vehicle of cases (i) and (ii)). Figure 2 shows how dwell time is obtained for a specific vehicle.

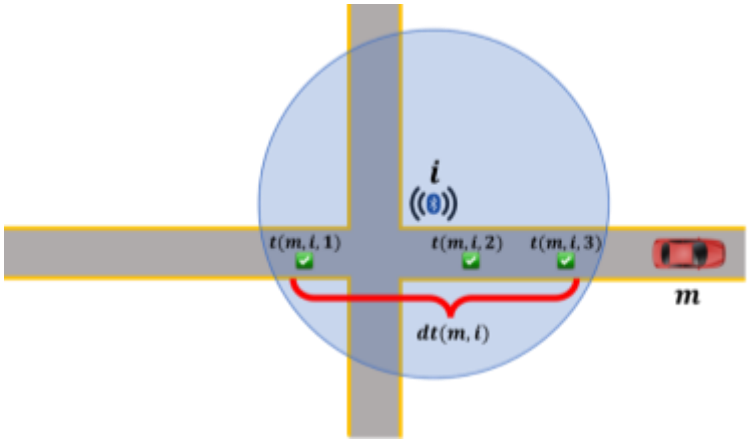

Figure 2. A vehicle detected many times at the same intersection.

For vehicles of case (iii) an approximation of the dwell time is calculated (Equation 1) considering D detections are made around BS i.

$$dt(m,i) = t(m,i,D) - t(m,i,1) \quad (1)$$

Where $dt(m,i)$ is the time that the vehicle MAC m stay around BS at node i, $t(m,i,1)$ is the moment of the first detection of vehicle m at node i and $t(m,i,D)$ is the moment of the last detection of vehicle m at node i. Equation 1 can lead to wrong results when a vehicle passes more than once through the node because the first and last detection would be from different trips (e.g. the first from its morning trip and the last from its evening trip). To address this issue, τ_a is defined as the maximum time that a vehicle can be in the influence area of BS. Then an algorithm is applied to calculate the set of dwell times $(dt(m,i)_n)$ for the n trips of vehicle m at node i and each value is smaller than τ_a.

Once all $dt(m,i)_n$ of each node (for all vehicles) are calculated, these data will be classified according to the time that occur each detection. DT_i^k is the distribution of dwell times, calibrated with the set of detections that occurs at time period k at node i.

3.1.2 Travel time distribution at links (TT_i)

The travel time in links is the time while the vehicles are traveling between two adjacent BS. This time incorporates the effects of pedestrian crossings, road accidents, vehicle queues, etc. Three cases can be defined when a vehicle travel between BS: (i) the vehicle is not detected in any BS, (ii) the vehicle is detected only by one BS and (iii) the vehicle is detected in both BS.

As with the classification of dwell time, cases (i) and (ii) cannot be used to know the vehicle's travel time, so to calibrate the travel time distributions once again are only employ only vehicles from case (iii) are considered. Another bias can be generated due to tending to consider slower vehicles. Figure 3 shows how travel time is obtained for a specific vehicle.

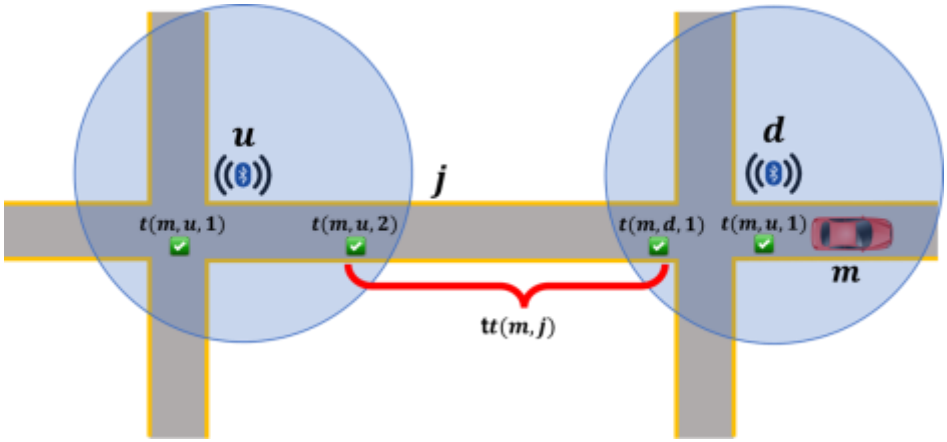

Figure 3. A vehicle detected by an upstream and downstream BS on a road.

For vehicles of case (iii) an approximation of the travel time is calculated (Equation 2) considering the detections that are made at upstream and downstream BS.

$$tt(m,j) = t(m,d,1) - t(m,u,D) \quad (2)$$

Where $tt(m,j)$ is the travel time of the vehicle m in the link j, $t(m,d,1)$ is the moment of the first detection of vehicle m at node d (downstream) and $t(m,u,D)$ is the moment of the last detection (assuming D detections) of vehicle m at node u (upstream). A vehicle can travel more than once through a link in the same trip, so $tt(m,j)_n$ is defined as the travel time of the vehicle m in the link j in the nth link trip. Then an algorithm is applied that calculate the set of travel times $(tt(m,j)_n)$. These times must be lower than a maximum time τ_b, because cases where vehicle trips pass twice in the same link and where the trip start in a link and finish in the same must be not considered on travel time (these cases do not show a correct travel time of the link).

Once all $tt(m,j)_n$ of each link (for all vehicles) are calculated, these data will be classified according to the time that occur each detection. TT_j^k is the distribution of travel times, calibrated with the set of detections that occurs at time period k at link j.

3.2 Inference of the route

With the calibrated dwell time and travel time distributions, it is possible to infer route choice probabilities between two successive detections of the same vehicle in non-adjacent nodes. Considering that $d(m,i,t)$ is a detection of the vehicle MAC m in the node i at time t. The inference is over the route made between $d(m,a,t_a)$ and $d(m,b,t_b)$ is sought, where $t_a < t_b$. It must be considered that $d(m,a,t_a)$ is the last detection in node a (upstream node) and $d(m,b,t_b)$ is the first detection in node b (downstream node). Given this information, notice that there are no detections of vehicle m between t_a and t_b. Figure 4 shows both detections and the route between nodes a and b that it will be inferred.

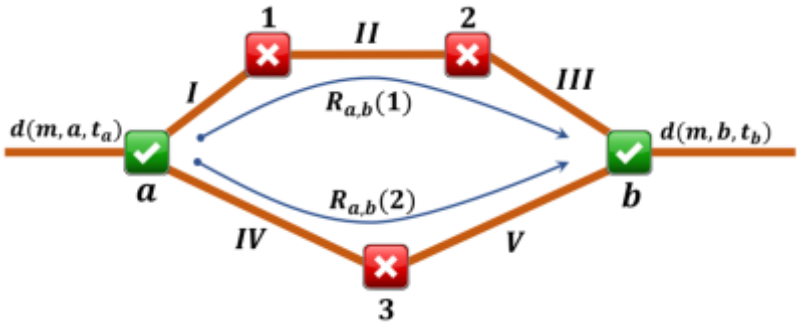

Figure 4. An example of network with two possible routes to go from a to b.

The set of all reasonable possible routes (i.e. without loops and long detours) between nodes a and b is constructed and called $R_{a,b}$. The expression $R_{a,b}(i)$ is used to denote the ith route of the set $R_{a,b}$. Each route has a sequence of nodes and links. For example, in Figure 4, the route $R_{a,b}(1)$ has 3 links and 2 nodes, while the route$R_{a,b}(2)$ has 2 links and 1 node. With the listed routes a Bayesian inferences is performed based on "probabilities of total travel time" and "probabilities of missed detections". Then a table probabilities per routes is calculated. This process is explained in the next points.

3.2.1 Bayesian inference

Dwell time distribution, travel time distribution and number of potentially missed detection are the information that is known prior to infer the route choice probability. It is known that there are not detections between nodes (a and b). So, it is possible to define that probability as $\Pr\{R = R_{a,b}(i)|Nd = n \wedge t = T_{obs}\}$, this represents the probability of a vehicle have chosen route i ($R = R_{a,b}(i)$), given that this vehicle is not detected by any of n detectors of the route ($Nd = n$) and it spent t_{obs} in the travel from a to b ($t = T_{obs}$). With the purpose of simplify the notation this probability will rewrite as $\Pr\{R_i|Nd \wedge t\}$. Bayes theorem and total probabilities property are used to develop this expression (Equation 3).

$$\Pr\{R_i|Nd \wedge t\} = \frac{\Pr\{Nd|t \wedge R_i\} \cdot \Pr\{t|R_i\} \cdot \Pr\{R_i\}}{\sum_j \Pr\{Nd|t \wedge R_j\} \cdot \Pr\{t|R_j\} \cdot \Pr\{R_j\}} \quad (3)$$

The sum in the denominator considers all elements of set $R_{a,b}$. To evaluate the expression of Equation 3 it is necessary to obtain three types of probabilities: $\Pr\{t|R_i\}$, $\Pr\{Nd|t \wedge R_i\}$ and $\Pr\{R_i\}$. The first one is the probability that a vehicle travel T_{obs} for a specific route, the second is the probability of not be registered when the vehicle travel along the route and the last one is the probability of using route i. The next points show how is possible to calculate these probabilities.

3.2.2 Total travel time probability ($\boldsymbol{Pr\{t|R_i\}}$)

For each route i from $R_{a,b}$, a distribution of total travel time on period k (TTT_i^k) is constructed. To do this, A convolution of dwell time and travel time distribution (that belong to route i) is made (Equation 4). Using the network of Figure 4, TTT_1^k is the result of convolving dwell time distributions (DT_1^k and DT_2^k) and travel time distributions (TT_I^k, TT_{II}^k and TT_{III}^k).

$$TTT_i^k = TT_1^k \circledast TT_2^k \circledast \dots \circledast TT_h^k \circledast DT_1^k \circledast DT_2^k \circledast \dots \circledast DT_n^k \quad (4)$$

Where $\circledast$ is convolving operator, h and n are the number of links and nodes at route i respectively. The variable T_{obs} includes in its value an error generated by the technology. That error can be divided in two components. The first one corresponds to the time difference between the detection at t_a and the real time on which the vehicle left the influence area in a. In a similar manner, the second one is the difference between the real time on which the vehicle entered the influence area in b and the detection at t_b. Its value can be calculated by the expression shown in Equation 5.

$$\varepsilon = \frac{E[DT_a^k]}{det(a)+1} + \frac{E\left[DT_b^k\right]}{det(b)+1} \quad (5)$$

Where each term corresponds to the error generated in each BS, $E[DT_a^k]$ and $E\left[DT_b^k\right]$ are the expected values of each distribution and $det(a)$ and $det(b)$ are the number of detections made in each BS for this specific vehicle. With the total travel time distribution calibrated for each route, it is possible to infer the probability that the observed travel time ($T_{obs} = t_b - t_a$) can be explained by the data TTT_i^k in a specific route (Equation 6). That is the probability that T_{obs} matches with the travel time of the given route. This probability is called Pt_i^k and represents $\Pr\{t|R_i\}$.

$$Pt_i^k = \Pr\{T_{obs} - \varepsilon < t \leq T_{obs}\} = \int_{T_{obs}-\varepsilon}^{T_{obs}} TTT_i^k(t)\, dt \quad (6)$$

3.2.3 Missed detections probability ($\boldsymbol{Pr\{Nd|t \wedge R_i\}}$)

Each route from $R_{a,b}$ has a fixed number of BS (number of nodes). The probability of detect a vehicle that travels along the BS area at speed v is called p_v. Thus, if a route has n BS and none of them makes a detection, then the probability of that event (Pd_i^v) is the product of the individual probabilities of missing detections in route i (Equation 7). Using the network of Figure 4, Pd_1^v is $(1-p_v)^2$ and Pd_2^v is $(1-p_v)^1$.

$$Pd_i^v = (1-p_v)^n \quad (7)$$

Where n is the number of BS in route i and v is the mean speed of the vehicle along this route. So, $\Pr\{\mathrm{Nd}|\mathrm{t} \wedge \mathrm{R_i}\} = \mathrm{Pd_i^v}$, where the speed is calculated from the quotient between the distance of nodes (a and b) and $\mathrm{T_{obs}}$ ($v = distance(R_{a,b}(i))/T_{obs}$). Notice that $\mathrm{p_v}$ value is exogenously calculated for different speeds based on field experiments. To do that, it is possible to define different probabilities according speed ranges. Additionally, theoretical functions can be defined for how the detection probability changes based on the vehicles' speeds.

3.2.4 Prior route usage probability ($\boldsymbol{Pr\{R_i\}}$)

This probability is the usage proportion of the different routes. It represents the prior flow assignment f_i on each route i. There are different ways to obtain this flow; for example, it could be assumed that all routes have the same flow, or the vehicle counts in the database could be used to obtain an approximation of the flows. Independently of the method used to estimate the flows, their proportion is calculated using the expression show in Equation 8.

$$\Pr\{R_i\} = \frac{f_i}{\sum_j f_j} \tag{8}$$

As all term in Equation 3 have been defined and can be calculated, it is possible to obtain the route choice probabilities, between successive detections. This process is repeated for each pair of successive detections of a whole trip, obtaining probabilities for each route from origin to destination.

4. METHODOLOGY DEVELOPMENT AND AIMSUN SIMULATION

The methodology was programmed in C# and tested within an Aimsun simulation. In the microsimulation environment, an API was developed to represent the BS detecting behavior. Simulation is chosen over real data to assess the explanatory and forecasting capabilities of the proposed methodology, by comparing the routes followed by each vehicle with the methodology's route choice probabilities (real data does not provide the complete routes taken by the vehicles).

4.1 Aimsun API and BS behavior model

The Aimsun API was programmed in Python and works over any road network. The API is composed by two main components: the first one is the network coding, where nodes and links files are created with the topological information; the second one generates the Bluetooth and real database, the former has the BSs detections and the latter one has the information of the entire vehicles' routes.

The model used to represent the behavior of the BS generates a list of detections with the information of Table 1. To do this, every t seconds the BS tries to detect all nearby vehicles in a d meters radius. Each vehicle has a probability of being detected, which is decreasing with its current speed at time t. Then, a Bernoulli experiment is conducted to know if the BS makes a detection. All detections are recorded in the Bluetooth database.

4.2 Aimsun simulation

The methodology is tested with a 6 by 6 Manhattan network. It has a BS and a traffic light in every intersection (i.e. there are 36 BSs). Each link is 100 meters long (see Figure 5a). The demand level is represented by three periods, as seen in Figure 5b, and it has a Logit route choice behavior with a time-based utility.

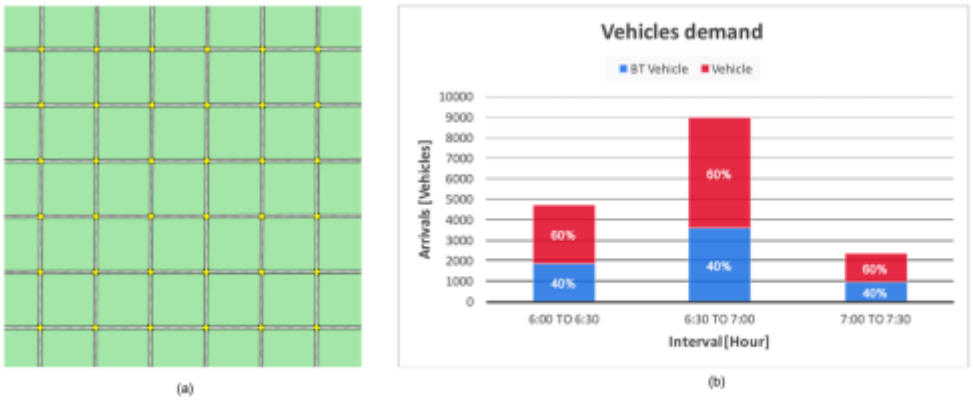

Figure 5. (a) Network of the simulation experiment. (b) Vehicle demand on network

The BS behavior model used in the simulation has four different intervals (t) between detections (1, 2, 5 and 10 seconds). The detection radius is 25 meters and the function used to obtain the detecting probability from the current speed of the vehicles is represented by Equation 9.

$$p_v = 0.8 - 0.01 \cdot v \tag{9}$$

Where v is the speed in km/h and p_v is the detecting probability when a vehicle pass through a BS at speed v.

5. RESULTS

The main indicator used to test the methodology is the First Preference Recovery (FPR) (Ortúzar and Willumsen, 2002). The FPR counts how many times the model assigns the higher probability to the route chosen by the vehicle. Table 2 shows the FPR percentages for two different prior route usage probabilities, as discussed in Section 3.2.4, and four different BS detection intervals (1, 2, 5 and 10 seconds).

Table 2. FPR indicator as a percentage

Model\Interval	1 second	2 seconds	5 seconds	10 seconds
Equals flows	100%	99%	97%	91%
DB counts	100%	99%	97%	92%

Table 2 shows that with a larger detection interval a lower FPR is obtained. There are no significant differences between assuming prior route flows to be equal and using the counts of the database to define the prior route usage probabilities.

Figure 7 shows the absolute flow differences in each link between the real and estimated assignments. Only the cases with 1 and 10 second intervals are shown. In the 10 seconds' scenario, the errors are up to 24 percent, while in the 1 second' scenario they are not greater than 3 percent.

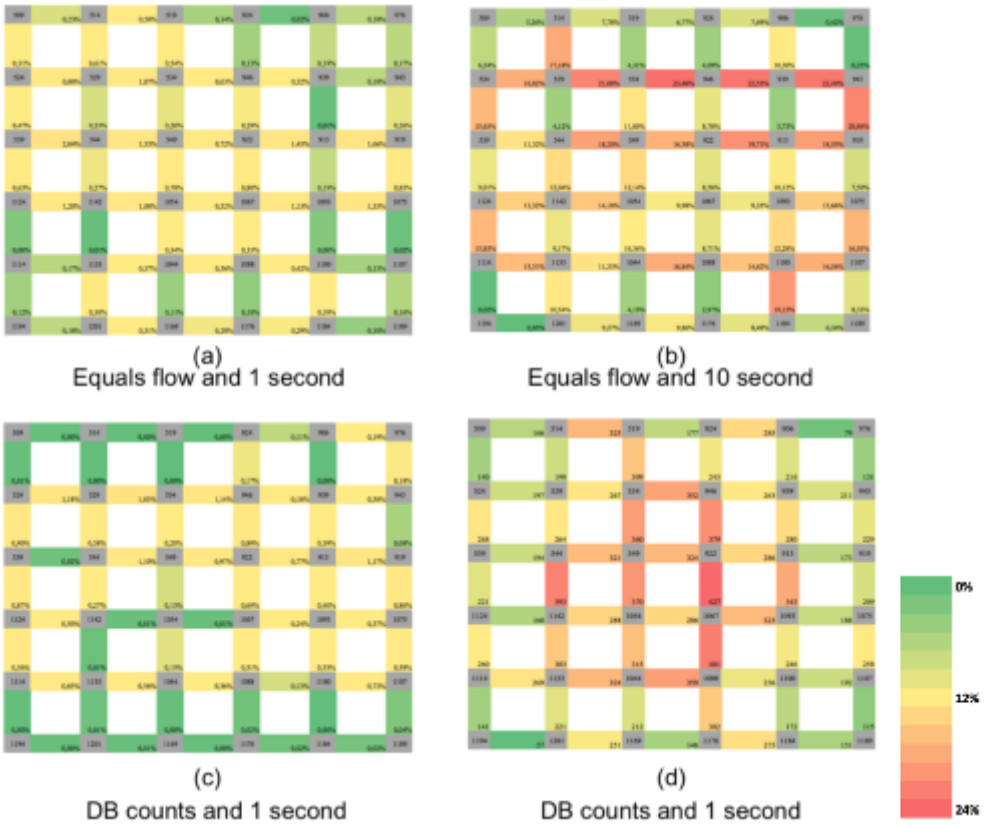

Figure 6. Relatives flows compared with the real flows

6. CONCLUSIONS AND NEXT STEPS

The methodology shows good prediction values, which depend on many factor such as the frequency of BS detections, the number of BSs and the distributions of the BSs. Controlling BSs' detection frequency is an important factor that impacts the quality of the predictions. Higher frequencies show better results.

Prior flows do not seem to affect the quality of the predictions on the simulated network, but it is important to try different BS configurations on the network.

As next steps, other methodologies to obtain different prior flows can be developed. It is important to try different scenarios considering other parameters such as the radius of the BSs, and to know with more details the behavior of the probability of detection trying different function of probability from speed. Finally, compare the methodology with other approaches found in the literature.

This work also offers future lines of research: to develop a methodology that allows to expand the sample of BT vehicles to all with low cost, to develop techniques of data filtering against the large number of different signals that this type of antennas can registry and finally work on the deepening of this methodology.

ACKNOWLEDGMENTS

The authors would like to thank for the support provided by CONICYT thorough FONDECYT project #1160943 and the Center for Sustainable Urban Development (CONICYT/FONDAP 15110020).

REFERENCES

Barcelo, J., Montero, L., Bullejos, M., Serch, O., & Carmona, C. (2012). Dynamic OD matrix estimation exploiting bluetooth data in urban networks. Recent Researches in Automatic Control and Electronics, 116-121.

Bhaskar, A., & Chung, E. (2013). Fundamental understanding on the use of Bluetooth scanner as a complementary transport data. Transportation Research Part C: Emerging Technologies, 37, 42-72.

Blogg, M., Semler, C., Hingorani, M., &Troutbeck, R. (2010, September). Travel time and origin-destination data collection using Bluetooth MAC address readers. In Australasian transport research forum (Vol. 36).

Carpenter, C., Fowler, M., & Adler, T. (2012). Generating route-specific origin-destination tables using Bluetooth technology. Transportation Research Record: Journal of the Transportation Research Board, (2308), 96-102.

Michau, G., Nantes, A., & Chung, E. (2013). Towards the retrieval of accurate OD matrices from Bluetooth data: lessons learned from 2 years of data.

Michau, G., Nantes, A., Bhaskar, A., Chung, E., Abry, P., &Borgnat, P. (2017). Bluetooth Data in an Urban Context: Retrieving Vehicle Trajectories. IEEE Transactions on Intelligent Transportation Systems.

Musa, A. B. M., & Eriksson, J. (2012, November). Tracking unmodified smartphones using wi-fi monitors. In Proceedings of the 10th ACM conference on embedded network sensor systems (pp. 281-294). ACM.

Ortuzar, J. D. D., & Willumsen, L. G. (2002). *Modelling transport* (Vol. 3).