Estimating travel times using Wi-Fi sensor data

Hoogendoorn, Serge; Daamen, Winnie; Duives, Dorine; Yuan, Yufei

Publication date
2016

Document Version
Submitted manuscript

Published in
TRISTAN 2016

Citation (APA)
Hoogendoorn, S., Daamen, W., Duives, D., & Yuan, Y. (2016). Estimating travel times using Wi-Fi sensor data. In *TRISTAN 2016: The Triennial Symposium on Transportation Analysis* (pp. 1-4)

Estimating travel times using Wi-Fi sensor data

Serge P. Hoogendoorn
Faculty of Civil Engineering and Geosciences
Delft University of Technology, The Netherlands
Email: s.p.hoogendoorn@tudelft.nl

Winnie Daamen, Dorine C. Duives, and Yufei Yuan
Faculty of Civil Engineering and Geosciences
Delft University of Technology, The Netherlands

1 Introduction

Determining travel time information from Wi-Fi (or Bluetooth) sensors is not trivial due to various (often technical) reasons. In this contribution, we focus on the problem of distinguishing *travel time* from *the time people spend* performing activities (e.g. fuelling the car, standing still to watch the scenery, buying a train ticket). More specifically, we will consider pedestrian data collected during a large-scale event in the city of Amsterdam called SAIL, where visitors walk along a route while watching and visiting tallships, eating and drinking (https://www.sail.nl/EN-2015). In this specific type of application, travel time information is required to provide information about the delays due to crowding, while the time spent on performing activities does not reflect such crowding effects.

In this contribution, we will present a novel statistical approach to estimate travel time distributions from data collected by Wi-Fi sensors. As we will see, this results in a non-trivial identification approach requiring advanced statistical modelling techniques.

2 Modelling

Let us consider a pair of Wi-Fi sensors that are collecting both information about the unique ID of a Wi-Fi device i (the MAC address) and a timestap t_i. By matching the IDs of the devices at distinct sensors, the time it takes the device to move from the one sensor to the other can be easily determined. These times are referred to as area presence times of device i and are denoted by h_i (in minutes). In the remainder, we will assume that devices are carried by visitors, and that a visitor carries at most one device. An observation of a Wi-Fi device then corresponds to the observation of a pedestrian. Although this may not always be the case, for the sake of the method presented in this contribution, the error that is made due to this assumption does not influence the main message of this contribution.

The area presence times h_i consist of the travel (or walking) time x_i and the activity duration time u_i. Here, the activity duration time is the time spent on everything besides walking (watching an event, spending time buying a drink, eating, etc.). Clearly, for the area presence time we have $h_i = x_i + u_i$.

2.1 Walking time distribution

We will assume that the walking times x_i follow some random distribution X. The motivation for this is that not all people will walk at the same speed. The walking speed is influenced by many (personal) different factors, such as age, gender, trip purpose, as well as by the prevailing traffic conditions. Let $g(x)$ denote the probability density function (pdf) of X. Due to the many factors determining the shape of the distribution, and given the fact that these influences are (at this stage) hard to quantify, we will not further specify the shape of the distribution. However, we will assume that there is a value T^* for which $g(x) = 0$ for all $x > T^*$. In other words, T^* denotes the maximum walking time we expect to observe.

2.2 Activity duration distribution

Similar to the walking times x_i, we assume that the activity duration times u_i follow some random distribution. Since the number of factors that could influence the (total) activity duration is large, and their influence is hard to predict beforehand, a commonly used assumption (e.g. in business process simulation) is that of an exponential distribution. In the remainder, we thus assume that the total duration of the activities is completely random; the plausibility of this assumption will be checked during the data analysis presented in the full paper. We will denote the pdf of the distribution of the activity duration by $r(h)$.

Now, let ϕ denote the probability that a visitor walking through an area without performing an activity. The pdf of the distribution H of all area presence times h_i can be written as follows:

$$f(h) = \phi\ g(h) + (1 - \phi)\ r(h) = g_1(h) + r_1(h) \tag{1}$$

Without going into detail, in our work on *headway distribution modelling*, we have shown that r satisfies:

$$r(h) = \frac{1}{A} \lambda \exp(-\lambda h) \int_0^h g(\tau) d\tau \tag{2}$$

where A is the normalisation constant satisfying $A = \int_0^\infty \lambda \exp(-\lambda\tau) g(\tau) d\tau$ and where $\lambda \geq 0$ denotes the average activity performance duration. λ is the main parameter describing the form of the distribution of the activity duration time.

3 Estimation approach

The main contribution of the estimation approach is presented in the ensuing, which is based on the approach first proposed by the authors in (Hoogendoorn, 2005).

Let $\{h_i\}$ denote the set of area presence times collected during some interval in which the traffic conditions were more or less similar. We have mentioned the existence of some value T^* for which $g(x) = 0$ in case $x \geq T^*$. This means that all observations $h_i \geq T^*$ belong to visitors that are performing activities, and are not just walking. Since we have assumed an exponential distribution, we can estimate the parameter λ as follows (Hoogendoorn, 2005):

$$\hat{\lambda} = \frac{1}{m} \sum_{i=1}^{n} \{h_i | h_i \geq T^*\} \tag{3}$$

where m is the number of observations for which $h_i > T^*$ and n is the total number of observations in the sample. For the estimate of the normalisation constant A, we then get (Hoogendoorn, 2005):

$$\hat{A} = \frac{m}{n} \exp(\hat{\lambda} T^*) \tag{4}$$

Now, let $\hat{f}_n(h)$ denote a distribution function determined from the sample (e.g. a histogram, or a kernel estimate). Then, we can show that by solving the following integral equation:

$$\hat{r}_1(h) = \frac{\hat{A}\hat{\lambda}}{\hat{\phi}} \exp(-\hat{\lambda} T^*) \int_0^h \left(\hat{f}_n(\tau) - \hat{r}_1(\tau) \right) d\tau \tag{5}$$

subject to

$$\hat{\phi} = \int_0^\infty \left(\hat{f}_n(\tau) - \hat{r}_1(\tau) \right) d\tau \tag{6}$$

we can determine the following estimate for the walking time distribution:

$$\hat{g}(h) = \frac{\hat{f}_n(h) - \hat{r}_1(h)}{\hat{\phi}} \tag{7}$$

This leaves us with determining an appropriate value for the threshold T^*. A good way to do this is by looking at the empirical survival function $\hat{S}_n(h)$, and drawing this function on semi-logarithmic paper. From the point where the curve turns into a straight line, we can assume that the distribution follows an exponential shape. Hence, this point is an appropriate value for T^*.

4 Estimation example

In the final part of this extended abstract, we illustrate the results of applying the proposed estimation method on one pair of Wi-Fi sensors that were installed during the SAIL event to monitor the visitor flows. The considered sensors were approximately 650 meters apart. Based on the length of the area, a threshold value of $T^* = 30$ minutes was chosen.

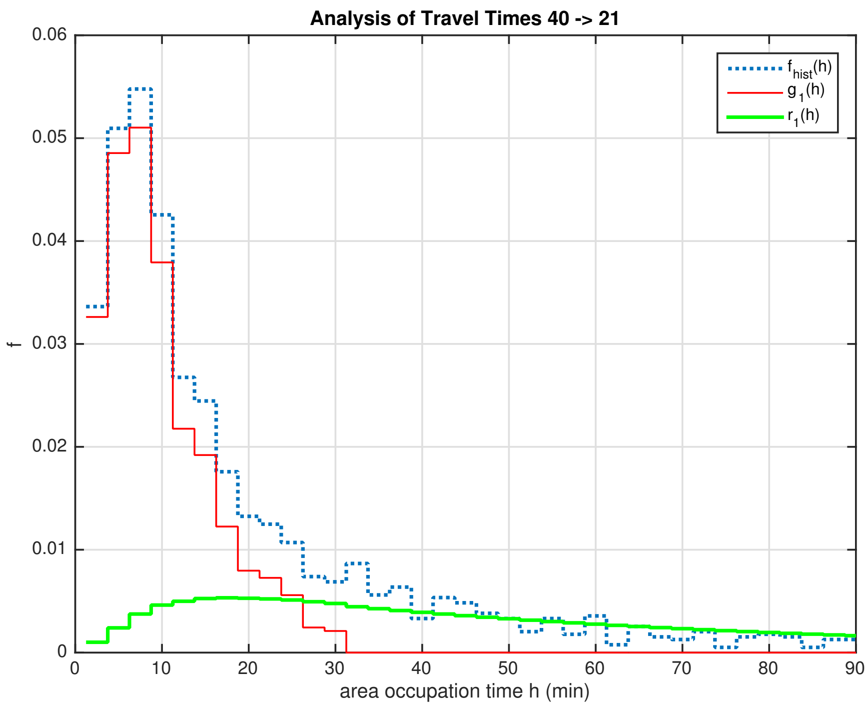

Figure 1: Estimation result for travel time distribution between 2 Wi-Fi sensors.

Fig. 1 depicts the results of the estimation procedure. The green line indicates the distribution of the activity duration period, while the red line specifies the walking time distribution. The sum of the two distributions equals the area presence time distribution (blue line). For this example, we found that $\phi = 0.61$, which implies that 61% of the observed people performed activities during their stay in the area between the two sensors. At the same time, the travel time distribution has a average value of 8.6 minutes and a standard deviation of 6.1 minutes. Both values appear to be very plausible given the average walking time that we can expect at this particular event.

In the full paper, we will provide a comprehensive analysis of the results, including a comparison of this method with an alternative approach and data collected using GPS devices.

References

S.P. Hoogendoorn. Unified approach to estimating free speed distributions. *Transportation Research Part B: Methodological*, 39 (8):709–727, 2005.