
Article:
Nantes, A, Ngoduy, D, Miska, M et al. (2015) Probabilistic travel time progression and its application to automatic vehicle identification data. Transportation Research Part B: Methodological, 81 (1). pp. 131-145. ISSN: 0191-2615

https://doi.org/10.1016/j.trb.2015.09.001

Probabilistic Travel Time Progression and its Application to Automatic Vehicle Identification Data

Article *in* Transportation Research Part B Methodological · September 2015

Impact Factor: 2.95 · DOI: 10.1016/j.trb.2015.09.001

4 authors:

Alfredo Nantes

Queensland University of Technology

25 PUBLICATIONS **67** CITATIONS

SEE PROFILE

Dong Ngoduy

University of Leeds

56 PUBLICATIONS **490** CITATIONS

SEE PROFILE

Marc Miska

Queensland University of Technology

40 PUBLICATIONS **131** CITATIONS

SEE PROFILE

Edward Chung

Queensland University of Technology

146 PUBLICATIONS **651** CITATIONS

SEE PROFILE

Probabilistic Travel Time Progression and its Application to Automatic Vehicle Identification Data

Alfredo Nantes[a,c], Dong Ngoduy[b,c], Marc Miska[a], Edward Chung[a]

[a] *Smart Transport Research Centre, Queensland University of Technology, QLD-4001, Brisbane, Australia*
[b] *Institute for Transport Studies, The University of Leeds, Leeds LS2 9JT, United Kingdom*
[c] *Corresponding authors. email: a.nantes@qut.edu.au or d.ngoduy@leeds.ac.uk*

Abstract

Travel time has been identified as an important variable to evaluate the performance of a transportation system. Based on the travel time prediction, road users can make their optimal decision in choosing route and departure time. In order to utilize adequately the advanced data collection methods that provide real-time different types of information, this paper is aimed at a novel approach to the estimation of long roadway travel times, using Automatic Vehicle Identification (AVI) technology. Since the long roads contain a large number of scanners, the AVI sample size tends to reduce and, as such, computing the distribution for the total road travel time becomes difficult. In this work, we introduce a probabilistic framework that extends the deterministic travel time progression method to dependent random variables and enables the off-line estimation of road travel time distributions. In the proposed method, the accuracy of the estimation does not depend on the size of the sample over the entire corridor, but only on the amount of historical data that is available for each link. In practice, the system is also robust to small link samples and can be used to detect outliers within the AVI data.

Keywords: off-line travel time estimation, long corridor, Automatic Vehicle Identification (AVI), data impoverishment problem, probabilistic travel time progression

1. Introduction

Travel time is a key component for the performance evaluation of transport systems. Moreover, travel time is easily perceived by the road users so the provision of travel time information may have a great impact on the route choice of travellers (Mahmassani and Liu, 1999). As is turns out, the provision of travel time information may reduce traffic congestion significantly and improve the performance of the whole network system (Ben-Akiva et al., 1991). Because of its critical role in traffic monitoring, extensive research has been conducted on the estimation and prediction of travel times on both freeways and urban roadways (Bhaskar et al., 2011, 2014, Coifman, 2002, Coifman and Krishnamurthy, 2007, Du et al., 2012, Ndoye et al., 2011, Sun et al., 2008, van Lint et al., 2005, van Lint and van der Zijpp, 2003). The widely used sensors to collect travel time data are loop detectors, which have been found to suffer from high maintenance costs and poor

reliability (Rajagopal and Varaiya, 2007). In recent years, rapid advances in information technology have led to cheaper data collection systems that are enriching the sources of empirical data for use in transport systems including travel time estimation and real-time traffic prediction problems (Deng et al., 2013, Herrera and Bayen, 2010, Herrera et al., 2010, Hofleitner et al., 2012a,b, Jenelius and Koutsopoulos, 2013, Kwong et al., 2009, Liu and Ma, 2009). Data collection systems are classified as fixed sensors and mobile sensors. Fixed sensors are inductive loop detectors, Automatic Number Plate Recognition (ANPR) and Automatic Vehicle Identification (AVI) systems. AVI systems provide traffic information at the location where the sensors are installed. On the other end, mobile sensors, such as GPS equipped vehicles and other Automatic Vehicle Location (AVL) systems, provide data for the entire journey of the vehicle equipped with such sensors (e.g. trajectory data). This paper mainly focuses on the travel time estimation problems using AVI data, such as Bluetooth.

The application of Bluetooth data for travel time estimation became available about a decade ago, when Murphy et al. (2002), Pasolini and Verdone (2002), Sawant et al. (2004) attempted to investigate the use of BT for Intelligent Transport System (ITS) services, and found that the Bluetooth equipped devices in moving vehicles could be discovered. The Bluetooth scanners use the Bluetooth discovery protocol to 'sniff' the unique electronic identifier, the Machine Access Control (MAC) address, of the transiting devices. Vehicles carrying discoverable Bluetooth devices (e.g. car kits, mobile phones and headsets) can be detected by Bluetooth scanners installed at multiple locations along the road network. Both the MAC address and the detection time of the traveller are recorded by the sensors and can be used for measuring the experienced travel times (Haghani et al., 2010). The increasing number of Bluetooth-enabled devices among road users, anonymity of MAC addresses, flexibility of deployment and maintenance of Bluetooth scanners have resulted in increasing scientific investigation into the cost-effectiveness of Bluetooth for travel time estimation (Aliari and Haghani, 2012, Bhaskar and Chung, 2013).

It has been widely reported that point-to-point data sources like the Bluetooth are self-sufficient, as far as the prediction of travel times is concerned (Khoei et al., 2013, Khosravi et al., 2011, Qiao et al., 2013). However, the assumption that is often made is that the data samples are large enough for computing the statistics of interest and that any variation from the expectation can safely be labelled as zero-mean, additive Gaussian noise. In fact, the travel time statistics are meaningful only when the sample is large; and the Bluetooth sample is usually rather small. Moreover, as we will show, the travel time distribution may exhibit multi-modal structures that cannot be completely defined through statistics like mean, median and standard deviation. Some cluster-based approaches have been proposed to address these issues by treating travel time as a random variable and, simultaneously, dealing with the problem of small samples (Jenelius and Koutsopoulos, 2013, Ramezani and Geroliminis, 2012). However, these approaches are computationally expensive and their accuracy depends on the way the AVI data is clustered (Ramezani and Geroliminis, 2012). Furthermore, it is not clear how the length of the road and the number of scanned links within it

affects the estimation, when real data is used (Jenelius and Koutsopoulos, 2013).

In an attempt to address all of these issues, we propose a novel probabilistic framework which extends the deterministic travel time progression rule to the case of random variables. We will show that such a framework can be implemented using non-parametric, polynomial-time algorithms whose accuracy only depends on the amount of link-based data available. Besides being effective and efficient at estimating travel time, in a off-line fashion, our probabilistic approach leads naturally to a simple Bayesian outlier detector of travel times. The rest of this paper is organized as follows. In Section 3, we will introduce the AVI Data Impoverishment problem, which may prevent researchers and practitioners from estimating travel times over long corridors. In Section 4, we will discuss the deterministic approach to travel time progression and explain why such an approach may fail to properly describe experienced travel times, in the case of long arterial roadways. We will then present our probabilistic travel time progression framework and its non-parametric implementation, in Section 5. Section 6 introduces a discrete version of the travel time distribution derived in Section 5, which is used for empirical validation. The accuracy and robustness of the system and the probabilistic outlier detection method are discussed in Section 7, using real data. Conclusions will be drawn in the last section.

2. Notation

For convenience, the notation in Table 1 below defines the symbols used in this paper.

Table 1: Table of symbols.

Symbol	Definition
N	number of scanned nodes of the target road
β	probability of successful detection by a AVI scanner
L	number of time and travel time bins used for the distribution histograms
$K_{n-1,n}$	number of AVI samples available for link $(n-1,n)$
$\tau^{(k)}_{n-1,n}$	travel time expirenced by traveller k over link $(n-1,n)$
$\mathcal{T}_{n-1,n}$	overall travel time over link $(n-1,n)$
$\varepsilon^{(k)}_n$	time at which traveller k has departed from node n
$\mathcal{E}_n$	departure time from node n
$[t_a, t_b]$	observation time interval

3. The AVI Data Impoverishment Problem

Travel times are extracted from AVI data, such as Bluetooth, by simply computing the difference between the time, $\varepsilon^{(k)}_n$, at which a traveller k is detected at some destination node n (e.g. downstream intersection of

a link), minus the time, $\varepsilon_{n-1}^{(k)}$, at which the same traveller is detected at some origin node n (e.g. upstream intersection of a link)

$$\tau_{n-1,n}^{(k)} = \varepsilon_{n}^{(k)} - \varepsilon_{n-1}^{(k)} \tag{1}$$

where $\tau_{n-1,n}^{(k)}$ denotes the travel time experienced by traveller k from origin $n-1$ to destination n. Computing travel times over specific corridors typically involves collecting data from individual travellers, in order to extract meaningful travel time statistics – e.g. average, median and variance – about the target link or road. Unfortunately, as the number of scanned links in the route increases, the AVI data sample tends to reduce in size and the sampling variation or *error* grows. This is a consequence of the miss-detection rate of the AVI scanners, as well as the patronage of the target road. Let us consider an example of a route of N scanned nodes and let β be the probability of successful detection proper to all scanners. Assuming that the detections are all independent of each other, e.g. the scanners do not interfere with one another, the probability for a traveller to be detected by all N scanners along the road is

$$\Pr(N) = \beta^N \tag{2}$$

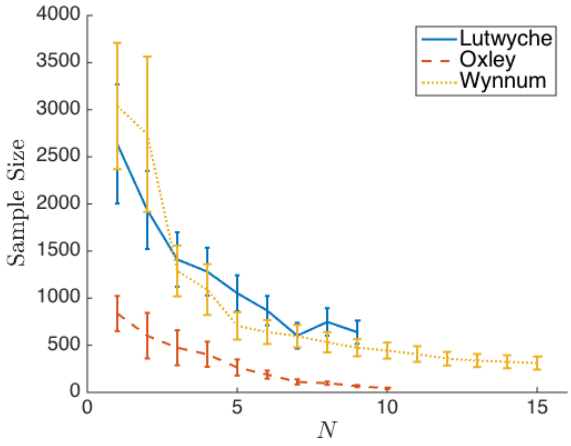

Figure 1: Visualization of the AVI-DI issue in the Bluetooth data, for 3 major arterial roads in Brisbane, namely Lutwyche Rd., Oxley Rd. and Wynnum Rd. The graph shows the trend of the average Bluetooth sample size, as the number of nodes increases. Sample size here signifies the number of travellers that have been detected at each intersection of the road, in the correct order, i.e. $1, 2, \ldots, N$. Vertical bars indicate the standard deviation around the average. The data used for these statistics covered a period of 30 days.

Because real scanners do fail to detect sometimes, we have that $0 \leq \beta < 1$. This entails that the probability above decreases exponentially as N increases. If the travel times are estimated from the travellers that are detected at each scanned node, the travel time sample too is doomed to decrease exponentially with N. In such cases, the upper bound of the sample size is established by the number of detections over the least traversed link, which in turn, greatly depends on the amount of travellers that actually traverse that link.

We will use the term *AVI Data Impoverishment* (AVI-DI) to denote the combined effect of miss-detections and route patronage on data reduction. Figure 1 shows an example of AVI-DI in the case of Bluetooth data, for 3 arterial stretches of the Brisbane metropolitan area, in Australia. As expected, the number of travellers that are detected at each scanned node (e.g. intersection) of the road decreases exponentially with the number of nodes. The AVI-DI issue may prevent road authorities from obtaining accurate travel times measurements for long corridors, due to insufficient AVI data.

4. Deterministic Travel Time Progression

Logically, if travel time is estimated from the travellers that are detected at each scanned node, then the travel time data about the individual sub-links is at least as abundant as the data for the entire road stretch. Thus, as far as this type of data (i.e. AVI) is concerned, the road travel time is best estimated from the sum of link travel times. Recovering the road travel time from the individual links is typically carried out through a method known as *travel time progression* (Friesz et al., 1993, Nanthawichit et al., 2013, Shim et al., 2011). Using the formalisation of Friesz et al. (1993), given a road defined as a sequence of N connected nodes, we will let $\tau_{n,n-1}$ be the travel time for link $(n-1,\ n)$; and $\varepsilon_n(t)$ the departure time from node $n \in \{1, \ldots, N\}$, given that the departure from node 1 occurs at some given time t. The expression for the total road travel time $\tau_{1:N}$, from node 1 through N, is

$$\tau_{1:N}(t) = \sum_{n=2}^{N} [\varepsilon_n(t) - \varepsilon_{n-1}(t)] = \varepsilon_N(t) - t \tag{3}$$

where we have defined

$$\varepsilon_n(t) = \varepsilon_{n-1}(t) + \tau_{n-1,n}(\varepsilon_{n-1}(t)) \qquad \forall \quad 1 < n \leq N \tag{4}$$

$$\varepsilon_1(t) = t \tag{5}$$

If the link-travel time functions $\tau_{n-1,n}$ are known or can be determined for any time t, for example from the statistics extracted form the AVI data, one can easily determine the road travel time $\tau_{1:N}(t)$. Despite its simplicity, this approach fails to address two important issues related to the nature of travel time. On the one hand, two or more vehicles exiting the origin at the exact same time and travelling along the same road may experience different travel times. This difference mainly depends on the possibility of vehicles to overtake each other and on the decision to break or accelerate at the traffic lights, for instance, when the lights turn amber. Put it differently, the travel time is partially influenced by the randomness component of the human behaviour and, as such, it should be treated as a random variable. The random nature of travel time can be observed in Figure 2-a , which shows how real Bluetooth travel times are distributed over a period of time, represented by the grey area on the left diagrams. Notice that the related distribution (upper right diagram) has two prominent peaks (modes), supposedly due to the presence of a signalised

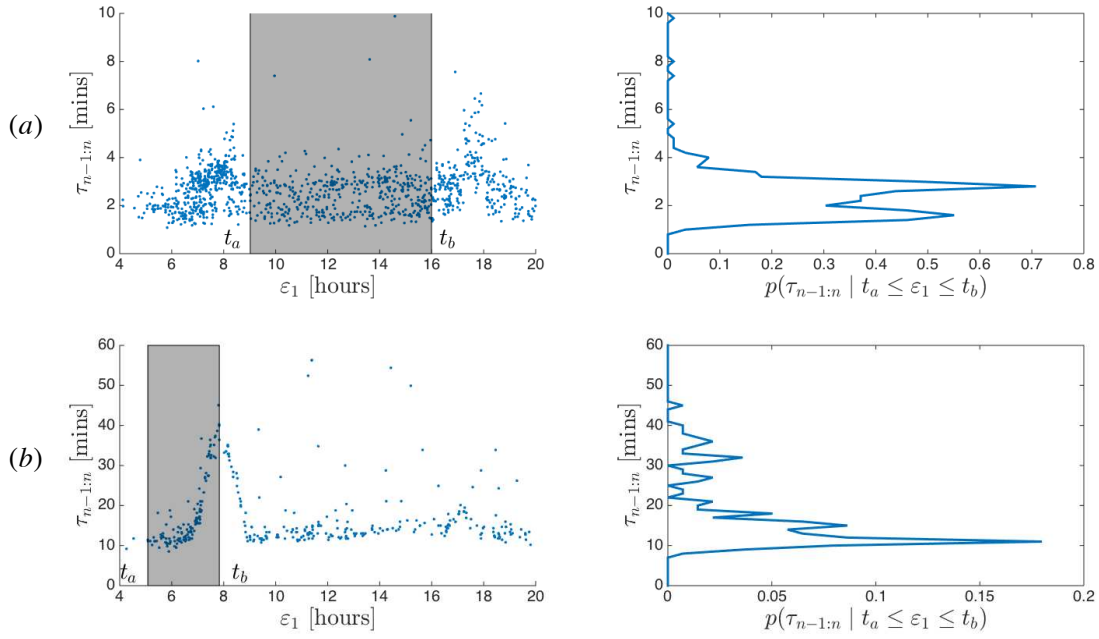

Figure 2: Travel time distributions. (a) Travel times for a link with a signalised intersection between upstream and downstream ends. The travel time distributions on the left concerns the observation time intervals (grey rectangles). (b) Travel times over a long arterial roadway. The observation interval encompasses regions of both free-flow and congested traffic.

intersection between the node $n-1$ and n. Another issue with the deterministic progression method is that it only allows dealing with instantaneous departure times, such as t. From a practical perspective, however, it may be more interesting to estimate travel times over arbitrary observation time windows, e.g. between t_a and t_b, like in Figure 2. It is expected that large observation time windows will yield multi-modal travel time distributions, as a consequence of incorporating data from both free-flow and congested traffic (see Figure 2-b).

5. Probabilistic Travel Time Progression

This section presents the main contribution of our work, which is, accounting for the random nature of travel time when progressively summing up the link travel times. In fact, the problem of estimating road travel time distributions from link data has previously been tackled by Ramezani and Geroliminis (2012) through the ordinary (non-progressive) sum of independent random variables. In order to account for the dependency between data from across adjacent links, the authors identify regions of homogeneous density within such data. Conditioned on these regions, link travel times become independent and can be summed up using convolution. This method produces a set of road travel time distributions, for the target road, which are then mixed together into a single target distribution. The disadvantage of this approach is, however, that the quality of the estimation depends on the definition of the conditioning regions, as acknowledged by

the authors. Moreover, the complexity of the method is exponential on the number of regions and links.

In the approach we present here the relationship between variables from adjacent links is defined through the travel time progression rule introduced earlier, rather than being learned from the data. As we will see in Section 6, this allows solving the problem in polynomial time. For consistency of notation, we will use the symbol $\mathcal{E}_n$ to denote the random variable 'departure time from node n' and $\mathcal{T}_{n-1,n}$ to indicate the random variable 'travel time over link $(n-1, n)$'. The recursive relationship between these variables is similar to that of (4), and is depicted in Figure 3-b, through a Bayesian network. In this network, directed edges represent conditional distributions, whereas vertices represent random variables. For example, for the conditional distribution $\Pr(\mathcal{E}_3 \mid \mathcal{E}_2, \mathcal{T}_{2,3})$ there are links from node $\mathcal{E}_2$ and $\mathcal{T}_{2,3}$ to $\mathcal{E}_3$, whereas for $\Pr(\mathcal{E}_1)$ there are no incoming links. A vertex in the graph is shaded (e.g. $\mathcal{E}_1$ in Figure 3-b) to indicate that the corresponding random variable is set to some observed value. We note that such a Bayesian network is static, in that, the values of random variables do not change overtime. Instead, the probability distributions of all variables is fixed and shaped by the behaviour of travellers, once they have completed their trips from origin to destination. As we shall see, such distributions can be elicited from historical AVI data.

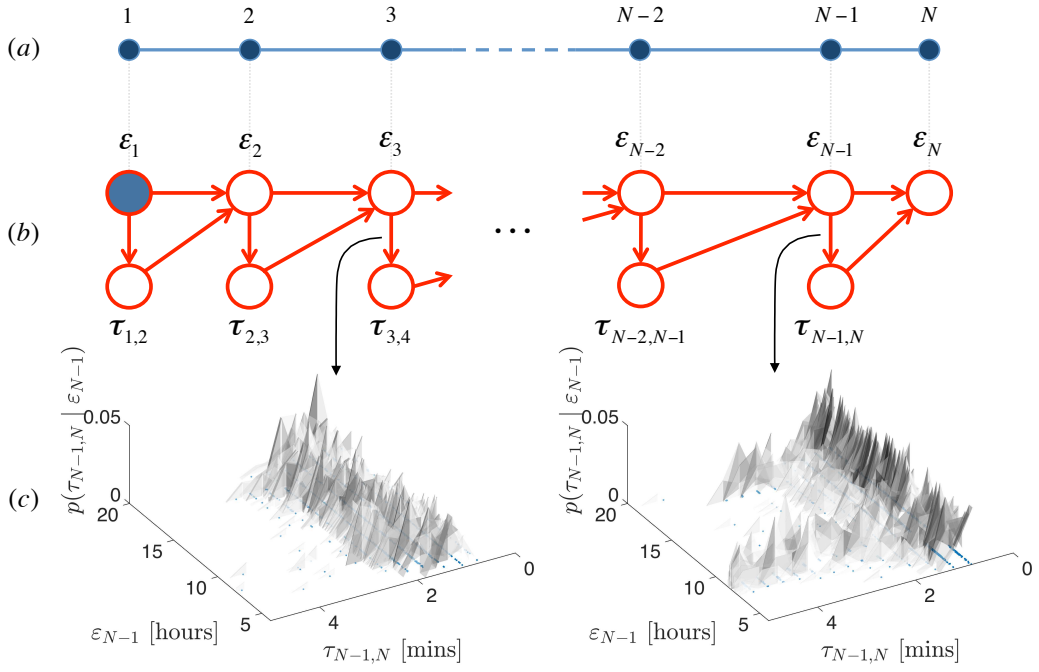

Figure 3: Probabilistic Travel Time Progression. (a) Representation of a road stretch with N nodes. (b) Probabilistic graphical representation of travel time progression for the target road. (c) Examples of conditional distributions over link travel times extracted from the Bluetooth data.

Similarly to (3), we are interested in the posterior over road travel time $\mathcal{T}_{1:N}$, for a route from node 1 through N, conditional to the fact that the departure time $\mathcal{E}_1$ from the origin node occurred between time

t_a and time t_b. More specifically, we wish to compute the posterior probability distribution

$$\Pr(\mathcal{T}_{1:N} = \tau_{1:N} \mid t_a \leq \mathcal{E}_1 \leq t_b) \qquad \forall \quad \tau_{1,N} \quad \text{and} \quad t_a < t_b \tag{6}$$

The problem, posed in these terms, is more general than that the deterministic travel time progression in (3). The road travel time $\mathcal{T}_{1:N}$ is now allowed to take on a set of possible different values, each with an associated probability. This probability will depend on the width of the observation time window $[t_a, t_b]$, the probability distribution of $\mathcal{E}_1$ and the probability distributions of the link travel times $\mathcal{T}_{n-1,n}$.

Lemma 1 (Posterior over departure time)**.** *Let $p_{\mathcal{T}_{n-1,n}}(\tau_{n-1,n} \mid \varepsilon_{n-1})$ be the conditional probability density function of $\mathcal{T}_{n-1,n}$ evaluated at $\mathcal{T}_{n-1,n} = \tau_{n-1,n}$, given $\mathcal{E}_{n-1} = \varepsilon_{n-1}$; and let $p_{\mathcal{E}_{n-1}}(\varepsilon_{n-1} \mid \varepsilon_1)$ be the conditional density function of $\mathcal{E}_{n-1}$ evaluated at $\mathcal{E}_{n-1} = \varepsilon_{n-1}$, given $\mathcal{E}_1 = \varepsilon_1$. Then*

$$\Pr(\mathcal{E}_n = \varepsilon_n \mid \mathcal{E}_1 = \varepsilon_1) = \int p_{\mathcal{T}_{n-1,n}}(\varepsilon_n - \varepsilon_{n-1} \mid \varepsilon_{n-1})\, p_{\mathcal{E}_{n-1}}(\varepsilon_{n-1} \mid \varepsilon_1)\, d\varepsilon_{n-1} \tag{7}$$

Proof. From (4) we know that $\mathcal{E}_2$ is a deterministic node, for its value is specified exactly by the values of its parents $\mathcal{E}_2 = \mathcal{E}_1 + \mathcal{T}_{1,2}$. The posterior over the departure time from node 2 is therefore

$$\Pr(\mathcal{E}_2 = \varepsilon_2 \mid \mathcal{E}_1 = \varepsilon_1) = \Pr(\mathcal{E}_1 + \mathcal{T}_{1,2} = \varepsilon_2 \mid \mathcal{E}_1 = \varepsilon_1) \tag{8}$$

that is, the posterior over $\mathcal{E}_2$ is a copy of the posterior over $\mathcal{T}_{1,2}$, shifted to the right by ε_1 units

$$\Pr(\mathcal{E}_2 = \varepsilon_2 \mid \mathcal{E}_1 = \varepsilon_1) = \Pr(\mathcal{T}_{1,2} = \varepsilon_2 - \varepsilon_1 \mid \mathcal{E}_1 = \varepsilon_1) \tag{9}$$

Let us now consider the posterior over the departure time $\mathcal{E}_3$ from node 3. Similarly to the previous case, $\mathcal{E}_3 = \mathcal{E}_2 + \mathcal{T}_{2,3}$ and we can write

$$\begin{aligned} \Pr(\mathcal{E}_3 = \varepsilon_3,\, \mathcal{E}_2 = \varepsilon_2 \mid \mathcal{E}_1 = \varepsilon_1) &= \Pr(\mathcal{T}_{2,3} = \varepsilon_3 - \varepsilon_2,\, \mathcal{E}_2 = \varepsilon_2 \mid \mathcal{E}_1 = \varepsilon_1) \\ &= \Pr(\mathcal{T}_{2,3} = \varepsilon_3 - \varepsilon_2 \mid \mathcal{E}_2 = \varepsilon_2,\, \mathcal{E}_1 = \varepsilon_1) \\ &\quad \Pr(\mathcal{E}_2 = \varepsilon_2 \mid \mathcal{E}_1 = \varepsilon_1) \end{aligned} \tag{10}$$

Since we are interested in $\Pr(\mathcal{E}_3 = \varepsilon_3 \mid \mathcal{E}_1 = \varepsilon_1)$, we will marginalize (10) over $\mathcal{E}_2$

$$\Pr(\mathcal{E}_3 = \varepsilon_3 \mid \mathcal{E}_1 = \varepsilon_1) = \int p_{\mathcal{T}_{2,3}}(\varepsilon_3 - \varepsilon_2 \mid \varepsilon_2,\, \varepsilon_1)\, p_{\mathcal{E}_2}(\varepsilon_2 \mid \varepsilon_1)\, d\varepsilon_2 \tag{11}$$

$$= \int p_{\mathcal{T}_{2,3}}(\varepsilon_3 - \varepsilon_2 \mid \varepsilon_2)\, p_{\mathcal{E}_2}(\varepsilon_2 \mid \varepsilon_1)\, d\varepsilon_2 \tag{12}$$

The transformation from (11) to (12) is a consequence of the conditional independence property of the Bayesian network in Figure 3-b , that is

$$\Pr(\mathcal{T}_{n-1,n} \mid \mathcal{E}_{n-1}, \mathcal{E}_{n-2}, \ldots, \mathcal{E}_1) = \Pr(\mathcal{T}_{n-1,n} \mid \mathcal{E}_{n-1}) \tag{13}$$

Using the induction method, we arrive at Eq. (7). □

Lemma 1 enables expressing the posterior over $\mathcal{E}_n$ recursively. Indeed, such posterior is computed from the conditional posterior over $\mathcal{E}_{n-1}$ which, in turn, is computed from the posterior over $\mathcal{E}_{n-2}$, and so forth up to $\mathcal{E}_2$. It is worth noticing that the right hand side of (7) is the convolution integral of the joint density function $p_{\mathcal{T}_{n-1,n},\mathcal{E}_{n-1}}(\varepsilon_n - \varepsilon_{n-1}, \varepsilon_{n-1} \mid \varepsilon_1)$. This integral defines the distribution of the sum of the two dependent random variables $\mathcal{T}_{n-1,n}$ and $\mathcal{E}_{n-1}$, in agreement with equation (10).

Our initial goal was to find the posterior over the road travel time $\mathcal{T}_{1:N}$, given that $\mathcal{E}_1$ was defined within a time interval $[t_a, t_b]$.

Lemma 2 (Posterior over road travel time)**.** *Let $p_{\mathcal{E}_N}(\varepsilon_N \mid \varepsilon_1)$ be the conditional probability density function of $\mathcal{E}_N$ evaluated at $\mathcal{E}_N = \varepsilon_N$, given $\mathcal{E}_1 = \varepsilon_1$; and let $p_{\mathcal{E}_1}(\varepsilon_1)$ be prior probability density function of $\mathcal{E}_1$ evaluated at $\mathcal{E}_1 = \varepsilon_1$. Further, let $[t_a, t_b]$ be an arbitrary support for $p_{\mathcal{E}_1}$. Then*

$$\Pr(\mathcal{T}_{1:N} = \tau_{1:N} \mid t_a \leq \mathcal{E}_1 \leq t_b) = \frac{\int_{t_a}^{t_b} p_{\mathcal{E}_N}(\tau_{1:N} + \varepsilon_1 \mid \varepsilon_1)\, p_{\mathcal{E}_1}(\varepsilon_1)\, d\varepsilon_1}{\int_{t_a}^{t_b} p_{\mathcal{E}_1}(\varepsilon_1)\, d\varepsilon_1} \tag{14}$$

Proof. Using the product and chain rule we have

$$\Pr(\mathcal{T}_{1:N} = \tau_{1:N} \mid t_a \leq \mathcal{E}_1 \leq t_b) = \frac{\int_{t_a}^{t_b} p_{\mathcal{T}_{1:N}}(\tau_{1:N} \mid \varepsilon_1)\, p_{\mathcal{E}_1}(\varepsilon_1)\, d\varepsilon_1}{\int_{t_a}^{t_b} p_{\mathcal{E}_1}(\varepsilon_1)\, d\varepsilon_1} \tag{15}$$

where $p_{\mathcal{T}_{1:N}}(\tau_{1:N} \mid \varepsilon_1)$ is the conditional probability distribution of $\mathcal{T}_{1:N}$ evaluated at $\mathcal{T}_{1:N} = \tau_{1:N}$, given that $\mathcal{E}_1 = \varepsilon_1$. From (3), we have that $\mathcal{T}_{1:N} = \mathcal{E}_N - \mathcal{E}_1$; that is, the travel time is equal to the difference between the departure time from the destination node minus the departure time from the origin node. Thus, given $\mathcal{E}_1 = \varepsilon_1$, the conditional distribution over road travel time is obtained by shifting the distribution over $\mathcal{E}_N$ to the left by $\tau_{1:N}$ units, that is

$$\Pr(\mathcal{T}_{1:N} = \tau_{1:N} \mid \mathcal{E}_1 = \varepsilon_1) = \Pr(\mathcal{E}_N = \tau_{1:N} + \varepsilon_1 \mid \mathcal{E}_1 = \varepsilon_1) \tag{16}$$

Applying this result to (15) we obtain (14). □

Note that $\Pr(\mathcal{E}_N \mid \mathcal{E}_1 = \varepsilon_1)$ is computed recursively via (7) from the posteriors over $\mathcal{E}_{N-1}, \mathcal{E}_{N-2}, \ldots, \mathcal{E}_2$. The shape of the probability density functions introduced here depends on the classes of users for which the travel time is to be determined, such as cars, taxies, bikes or pedestrians; on the hindrances along the road, such as traffic signals. Eq. (14) describes formally how the target road distribution is linked to the travel time density function $p_{\mathcal{T}_{n-1,n}}(\tau_{n-1,n} \mid \varepsilon_{n-1})$ for each link $(n-1, n)$, the prior over departure time from origin node $\mathcal{E}_1$ and the observation time window $[t_a, t_b]$. The method presented here is fully Bayesian; it does not require the integration of traffic models and can be used for off-line travel time estimations, whenever enough AVI data is available over the target observation window. Other methods have been recently proposed to enable the on-line prediction of traffic state, including travel time, in urban networks using heterogeneous data (Nantes et al., 2015). Such methods rely on traffic flow models, as a means of predicting the next state

of traffic, which are embedded in the Bayesian estimation fitlers such as Kalman filters or particle filters (Ngoduy, 2008, 2011, Wang and Papageorgiou, 2005, Wang et al., 2007).

6. Implementation

In this section we will show how a discrete version of the target travel time distribution in (14) can be constructed from historical AVI data, such as Bluetooth, by using a numerical approach. Since we wish to keep the number of parameters of the system to a minimum, we propose to use histograms to represent the probability distributions. In particular, we will use 2-D histograms to represent all conditional probability density functions. Such histograms are maintained through 2-D arrays; with the 2 dimensions referring to the 2 random variables involved. Accordingly, the value of $p_{\varepsilon_1}(\varepsilon_1 = i \mid \varepsilon_1 = j)$ will be stored in some 2-D array, at row i and column j. Likewise, prior distributions will be represented though 1-D histograms maintained through 1-D arrays; thus, the value of $p(\mathcal{E}_1 = i)$ will be stored in some array, at position i. The 2-D arrays have have fixed size $L \times L$; whereas the 1-D arrays will have fixed size L. Here, L is the number of bins in which both time and travel time are partitioned. Finally, we will assume that these bins have uniform unit width, in order to simplify the mechanism of travel time estimation. In order to fulfil our goal, we first need to estimate the distribution over departure times $\Pr(\mathcal{E}_n = \varepsilon_n \mid \mathcal{E}_1 = \varepsilon_1)$, for all nodes $n = \{1, 2, \ldots, N\}$. We propose Algorithm 1 to carry out this task. From these distributions, the distribution over road travel time $\Pr(\mathcal{T}_{1:N} = \tau_{1:N} \mid t_a \leq \mathcal{E}_1 \leq t_b)$ can finally be computed, as in Eq. (14). This can be achieved by using Algorithm 2.

Algorithm 1 assumes that the arrays representing the functions $p_{\mathcal{T}_{n-1,n}}(\tau_{n-1,n} \mid \varepsilon_{n-1})$ are available from the input set $\mathcal{P}$, for instance, through the function GET_DISTRIBUTION, in line 6. The algorithm also requires the value n indicating the node up to which $\Pr(\mathcal{E}_n = \varepsilon_n \mid \mathcal{E}_1 = \varepsilon_1)$ is to be computed; as well as the number of bins L. In line 8, the convolution function computes the integral in Eq. (7). Before this operation can be carried out, however, the posterior over departure times for the previous node needs to be computed. Hence, the recursive call of the function DEPARTURE_TIME_CDF, in line 7, using the previous node, $n-1$, as an input. The recursion calls will terminate in lines 3, in which the probability $p_{\mathcal{E}_1}(\varepsilon_1 \mid \varepsilon_1)$ is computed. Such a conditional distribution is represented though the identity matrix. Although recursion is not necessary, it is used here for illustrative purposes, as a reflection of the method proposed in the previous section.

Algorithm 2 computes the road travel time distribution through Algorithm 1. It takes a data set $\mathcal{D}$ of point-to-point measurements such as Bluetooth, along with the boundaries t_a and t_b of the observation time period. The algorithm starts by initialising the set of link-based travel time distributions $\mathcal{P}$ (line 1) and setting the array $\boldsymbol{c}$ (line 2) to zero. This array represents the road travel time posterior and has size $(L-1)$, where L is a parameter indicating a sufficiently large observation time, that is at least as big as t_b. Lines 2 through 8 compute $p_{\mathcal{T}_{n-1,n}}(\tau_{n-1,n} \mid \varepsilon_{n-1})$ for each link, by extracting the link-specific travel times

Algorithm 1 Distribution over departure time from node n, given $\mathcal{E}_1$

1: **function** DEPARTURE_TIME_CDF($\mathcal{P}$, n, L)
2: **if** n=1 **then**
3: $\mathbf{E} \leftarrow$ IDENTITY_MATRIX(L) ▷ $\mathbf{E}_{i,j} \leftarrow p_{\varepsilon_1}(\varepsilon_1 = i \mid \varepsilon_1 = j)$
4: **return E**
5: **end if**
6: $\mathbf{T} \leftarrow$ GET_DISTRIBUTION($\mathcal{P}$, n) ▷ $\mathbf{T}_{i,j} \leftarrow p_{\mathcal{T}_{n-1,n}}(\mathcal{T}_{n-1,n} = i \mid \mathcal{E}_{n-1} = j)$
7: $\bar{\mathbf{E}} \leftarrow$ DEPARTURE_TIME_CDF($\mathcal{P}$, $n-1$, L) ▷ $\bar{\mathbf{E}}_{i,j} \leftarrow p_{\varepsilon_{n-1}}(\varepsilon_{n-1} = i \mid \varepsilon_1 = j)$
8: $\mathbf{E} \leftarrow$ CONVOLUTION($\mathbf{T}$, $\bar{\mathbf{E}}$, L) ▷ $\mathbf{E}_{i,j} \leftarrow p_{\varepsilon_n}(\varepsilon_n = i \mid \varepsilon_1 = j)$
9: **return E**
10: **end function**

11: **function** CONVOLUTION($\mathbf{T}$, $\mathbf{E}$, L)
12: $\mathbf{C} \leftarrow$ ZEROS(L, L) ▷ zero matrix of size $L \times L$
13: **for** $i \leftarrow 0, L-1$ **do**
14: **for** $k \leftarrow i, L-1$ **do**
15: $\mathbf{C}_{k,i} \leftarrow \sum_{j=0}^{k} \mathbf{T}_{k-j,j}\, \mathbf{E}_{j,i}$
16: **end for**
17: **end for**
18: **return C**
19: **end function**

and departure times from the dataset $\mathcal{D}$ (line 5) and generating the non-parametric distribution over travel time $\tau_{n-1,n}$, for each departure time ε_{n-1} (line 6), though the function HISTOGRAM_CDF. Here, the travel times $\tau_{n-1,n}^{(k)}$ are computed as in Eq. (1) (e.g. from the Bluetooth data) and $K_{n-1,n}$ denotes the number of travellers that have been detected at both nodes $n-1$ and n. Algorithm 1 is then called in line 9, in order to compute the posterior over departure time, for destination node N. Line 10 extracts the departure times of all travellers that have traversed the entire corridor (denoted by $K_{1:N}$), through the function GET_EPSILON. The prior $p(\mathcal{E}_1)$ is represented by a 1-D histogram, computed in line 11 through the function HISTOGRAM. Finally, lines 12 through 15 compute the target distribution $\Pr(\mathcal{T}_{1:N} \mid t_a \leq \mathcal{E}_1 \leq t_b)$, using a discrete version of Eq. (14).

Algorithm 1 and 2 are polynomial-time. In particular, Algorithm 1 is $\mathcal{O}\left((N-1)L^3\right)$ due to the convolution of 2-D arrays repeated for each one of the $N-1$ links. As for Algorithm 2, the 2-D histogram produced in line 6, for an unsorted array in $\mathcal{D}$ is $\mathcal{O}((N-1)KL^2)$, assuming there were at most K data samples for each link. Line 11 is linear in K and lines 13 through 15 are $\mathcal{O}(L(t_b - t_a))$. Since Algorithm 1 is called outside the for-loops of Algorithm 2 (line 9), the complexity of the system is $\mathcal{O}\left((N-1)ML^2\right)$, where

Algorithm 2 Distribution over road travel time, given $t_a \leq \mathcal{E}_1 \leq t_b$

1: **function** ROAD_TRAVEL_TIME_CDF($\mathcal{D}$, t_a, t_b)
2: $\quad\mathcal{P} \leftarrow \varnothing$ ▷ initialize set of travel time posteriors
3: $\quad\boldsymbol{c} \leftarrow \text{ZEROS}(L)$ ▷ 1-D zero array of L elements
4: $\quad$**for** $n \leftarrow 2, N$ **do**
5: $\quad\quad\mathbf{D} \leftarrow \text{GET\_TAU\_EPSILON}(\mathcal{D}, n)$ ▷ $\mathbf{D} \leftarrow \begin{pmatrix} \tau_{n-1,n}^{(1)} & \varepsilon_{n-1}^{(1)} \\ \tau_{n-1,n}^{(2)} & \varepsilon_{n-1}^{(2)} \\ \vdots & \vdots \\ \tau_{n-1,n}^{(K_{n-1,n})} & \varepsilon_{n-1}^{(K_{n-1,n})} \end{pmatrix}$
6: $\quad\quad\mathbf{T} \leftarrow \text{HISTOGRAM\_CDF}(\mathbf{D})$ ▷ $\mathbf{T}_{i,j} \leftarrow p_{\mathcal{T}_{n-1,n}}(\mathcal{T}_{n-1,n} = i \mid \mathcal{E}_{n-1} = j)$
7: $\quad\quad$add $\mathbf{T}$ to $\mathcal{P}$
8: $\quad$**end for**
9: $\quad\mathbf{E} \leftarrow \text{DEPARTURE\_TIME\_CDF}(\mathcal{P}, N, L)$ ▷ $\mathbf{E}_{i,j} \leftarrow p_{\mathcal{E}_N}(\mathcal{E}_N = i \mid \mathcal{E}_1 = j)$
10: $\quad\boldsymbol{d} \leftarrow \text{GET\_EPSILON}(\mathcal{D}, 1, N)$ ▷ $\boldsymbol{d} \leftarrow (\varepsilon_1^{(1)}, \varepsilon_1^{(2)}, \ldots, \varepsilon_1^{(K_{1:N})})$
11: $\quad\boldsymbol{e} \leftarrow \text{HISTOGRAM}(\boldsymbol{d})$ ▷ $\boldsymbol{e}_i \leftarrow p(\mathcal{E}_1 = i)$
12: $\quad\eta \leftarrow \left(\sum_{j=t_a}^{t_b} \boldsymbol{e}_j\right)^{-1}$
13: $\quad$**for** $i \leftarrow 0, L - t_b - 1$ **do**
14: $\quad\quad\boldsymbol{c}_i \leftarrow \eta \sum_{j=t_a}^{t_b} \mathbf{E}_{i+j,j}\, \boldsymbol{e}_j$ ▷ $\mathbf{c}_i \leftarrow \Pr(\mathcal{T}_{1:N} = i \mid t_a \leq \mathcal{E}_1 \leq t_b)$
15: $\quad$**end for**
16: $\quad$**return** $\boldsymbol{c}$
17: **end function**

$M = \max(K, L)$. Note that Algorithm 1 needs to be called only once, in order to determine the posterior over departure times. Such posterior can then be utilised to compute the target ravel time distribution for any time interval $[t_a, t_b]$, e.g. through lines 10 though 15 of Algorithm 2. The cost of this latter operation is only $\mathcal{O}(L(t_b - t_a))$.

7. Case study

The framework proposed was tested on the Bluetooth data available for 3 arterial stretches of the Brisbane metropolitan area, in Australia. The experiments concerned a 10,484 m stretch (Wynnum Rd.); a 5,510 m stretch (Lutwyche Rd.); and a 8,405 m stretch (Oxley Rd.). Our system was tested using Bluetooth data, collected over a period of 30 days. This data consisted of tuples of the kind

$$< \text{Scanner\_ID}, \text{Traveller\_ID}, \text{Time\_Of\_First\_Detection} > \tag{17}$$

indicating the identifier of the scanner, the identifier of the detected traveller and the time of first detection as the traveller entered the scanning area. The location of the Bluetooth scanners and the roads targeted for

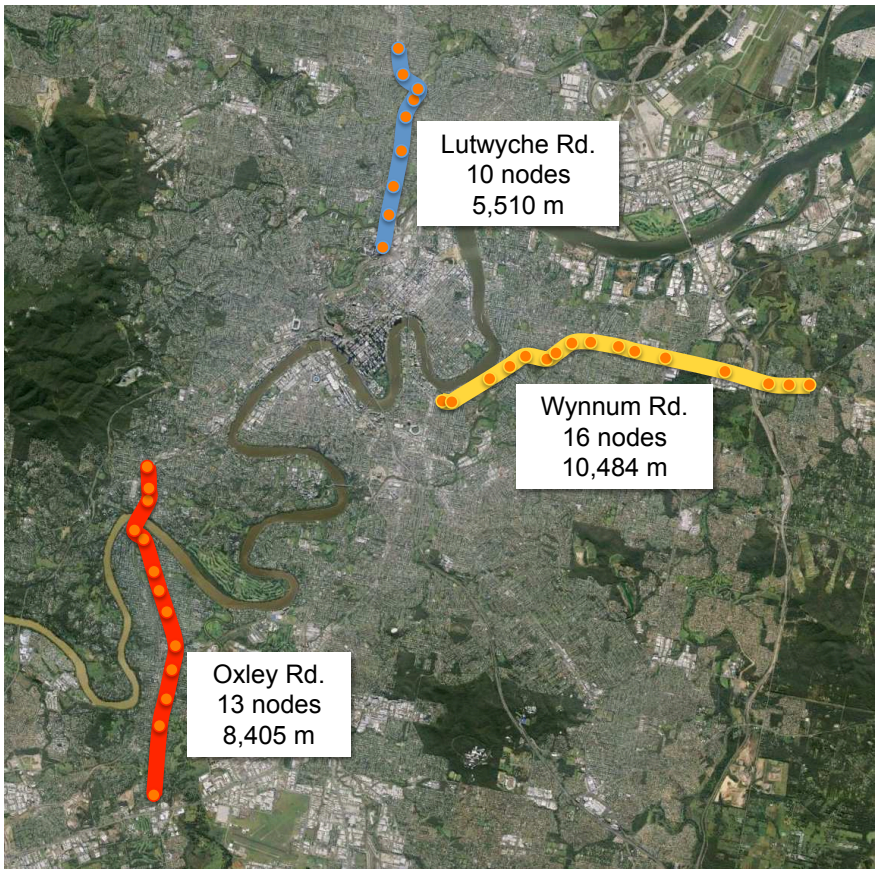

Figure 4: Arterial stretches of the Brisbane area monitored through Bluetooth scanners. The position of the scanners is indicated by filled circles.

our experiments are depicted in Figure 4. Travel times from individual travellers where computed as in (1). Link travel times that exceeded a threshold (1 hour in our experiments) were removed prior to processing, as considered outliers. No other filtering was applied to the data.

7.1. System Accuracy

In order to measure the accuracy of our system, we compared the travel time distribution produced by Algorithm 2 with the actual travel times from the set of Bluetooth travellers that were detected at each node of the target roads, and where thus specific to the experienced travel time over the entire stretch. Let $\mathcal{D}_{\text{links}}$ be the set of Bluetooth travel times specific to the links of the target road; and $\mathcal{D}_{\text{road}}$ the set of Bluetooth travel times specific to the entire target road. We tested the null hypothesis that the travel times in $\mathcal{D}_{\text{road}}$ came from the distribution estimated from $\mathcal{D}_{\text{links}}$, through Algorithm 2. Each experiment consisted of partitioning the domain of $\mathcal{E}_1$ (departure time form origin node) into intervals $[t_a, t_b]$ that were large enough to contain K travel time points from the set $\mathcal{D}_{\text{road}}$; where K was the minimum amount of points chosen for testing ($K = 30$ in our tests). For each of these intervals, we generated a 1-D histogram $\boldsymbol{o}$ from all travel times in $\mathcal{D}_{\text{road}}$ that fell within the interval $[t_a, t_b]$. This same time interval was also used to generate the estimated road travel time distribution $\boldsymbol{c}$, using Algorithm 2, that is, by calling the function ROAD_TRAVEL_TIME_CDF($\mathcal{D}_{\text{links}}$, t_a, t_b). Both $\boldsymbol{o}$ and $\boldsymbol{c}$ were of size L. The comparison between measured and estimated histograms was then carried out through a *chi square* test. Specifically, let $\boldsymbol{o}_i$ be the number of measured road travel times falling within bin i of the measurement histogram; and $\boldsymbol{c}_i$ the value of bin i of the estimated road travel time distribution. The expected number of travel times, $\boldsymbol{e}_i$ for bin i was computed as follows

$$\boldsymbol{e}_i = n\,\boldsymbol{c}_i; \qquad \text{where} \quad n = \sum_{i}^{L} \boldsymbol{o}_i \tag{18}$$

The measure of discrepancy between the observed and estimated histogram computed under the null hypothesis was

$$\bar{\chi}^2 = \sum_{i=1}^{L} \frac{(\boldsymbol{o}_i - \boldsymbol{e}_i)^2}{\boldsymbol{e}_i} \tag{19}$$

The chi square test involves measuring

$$p = \Pr(\chi^2 \geq \bar{\chi}^2) = \int_{\bar{\chi}^2}^{\infty} p_{\chi^2}(x, L-1)\,dx \tag{20}$$

where $p_{\chi^2}(x, L-1)$ is the probability density function of the random variable chi square, χ^2, with $L-1$ degrees of freedom.

It can be shown that, if the measured travel times that produced $\boldsymbol{o}$ came from the estimated histogram distribution in $\boldsymbol{c}$, then $\bar{\chi}^2$ would be an outcome of the random variable χ^2. Therefore, if p is smaller than a pre-defined significance α (0.01 in our tests) the null hypothesis is rejected, as the measured travel times are likely to come from a different distribution. An example of measured and expected histograms generated

from $\mathcal{D}_{\text{road}}$ and $\mathcal{D}_{\text{links}}$, respectively, is shown in Figure 5. The diagrams on the right show the observation

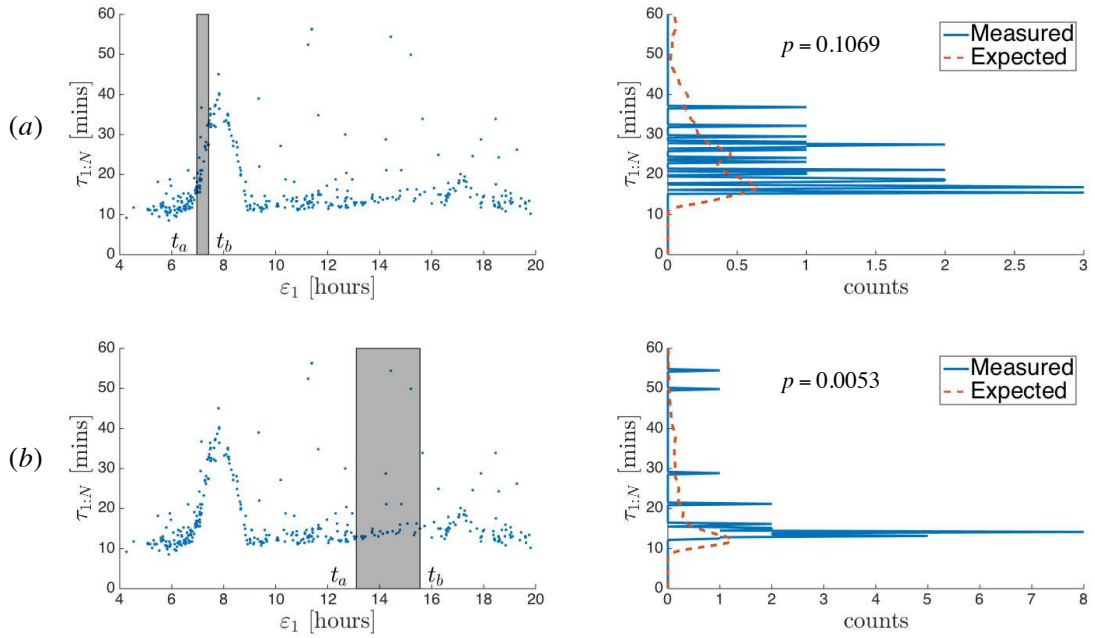

Figure 5: Examples of estimated histograms (right diagrams) from the Bluetooth travel times (dots on the left diagrams) over different observation time intervals, visualised through gray rectangles. The expected histogram (dashed line) was computed from the link travel times in $\mathcal{D}_{\text{links}}$, using Algorithm 2. The measured histogram (continuous line) was computed directly from the road travel times in $\mathcal{D}_{\text{road}}$. The p values of the chi square tests on the right diagrams are computed as in (20).

time intervals $[t_a, t_b]$ that were large enough to contain K road travel time points from $\mathcal{D}_{\text{road}}$. The diagrams on the left visualize the histograms $\boldsymbol{o}$ (continuous line) and $\boldsymbol{e}$ (dashed line), corresponding to the observation time interval chosen. The p values resulting from the chi square test are also reported on the right diagrams.

In order to assess how the accuracy of our method varied with respect to the number of nodes N of the target road, we considered $N-1$ sub-stretches defined as follows: stretch 1 was made of link $\{(1,2)\}$; stretch 2 was made of links $\{(1,2),(2,3)\}$ and so forth up to stretch $N-1$, made of links $\{(1,2),(2,3),\ldots,(N-1,N)\}$. We therefore applied our travel time progression method to each one of these sub-stretches and used the chi squared test to validate the results. Since the Bluetooth data concerned a period of 30 days and given that, on average, the choice of $K = 30$ enabled us to test on several observation time intervals per day, each sub-stretch was tested hundreds of times, yielding hundreds of p values. From the many p values available for each sub-stretch, we could then estimate the probability of the test to succeed, given the number of nodes of the road. This likelihood, which we shall indicate as $\Pr(p \geq \alpha \mid N)$, was simply obtained by counting the instances in which $p \geq \alpha$, versus the total number of tests, for the stretch at hand. The results of this operation are depicted in Figure 6 (left diagram).

As expected, the system is most accurate in the case of one-link stretches, where $\Pr(p \geq \alpha \mid N = 2) \approx 1$.

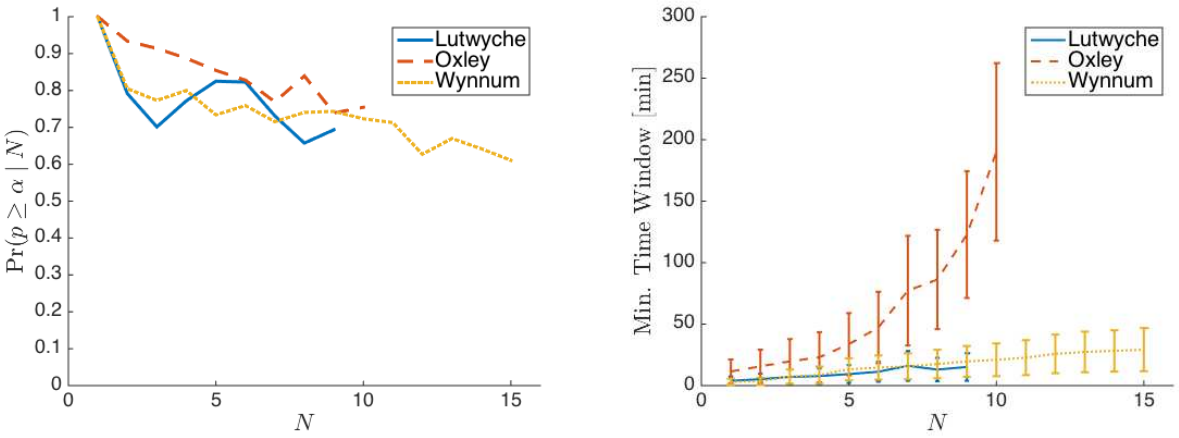

Figure 6: Accuracy tests. (left) Probability of the chi square test to succeed at $\alpha = 0.01$ significance. (right) Width of observation time interval $[t_a, t_b]$ containing at least $K = 30$ travel time points, versus number of nodes of the road. Vertical bars indicate the standard deviation around the mean values.

This likelihood appears to decrease, somewhat linearly, with N. For the longest road (10,484 m) made of 15 links, the chi test was successful about 60% of the time. This seems to contrast with the hypothesis that the road travel time distribution is always equal to the distribution of the progressive sum of link travel times. We propose that the decreasing trend of the likelihood may be linked to the amount and quality of data available. First and foremost, the link travel time distributions are extracted from the AVI sample rather than the entire population of vehicles. if the link data are scarce, the distribution of the sample may not coincide with the distribution of the population. Second, as the number of nodes increases, due to the data impoverishment issue, the sample size reduces or, equivalently, the observation time intervals $[t_a, t_b]$ containing the minimum sample size becomes wider. This phenomenon can be observed in Figure 6 (right diagram), for the 3 roads considered in our experiments. The chi square test requires the number of observations to be large. However, as the observation time interval grows wider, the number of observed travel times in them (30 in our case) may become insufficient for the chi square test to still produce meaningful statistics. A third factor that may contribute to the unexpected trend of the likelihood of p is the latency introduced by the Bluetooth scanners in time-stamping the travellers, as they enter the scanning area. As explained by Nantes et al. (2014), this latency distorts the actual travel time distribution and may yield a carry-over effect on the final estimated distribution.

7.2. System Robustness

In the previous section, we established that the probabilistic travel time progression can effectively be used to estimate the road travel time distribution. The system accuracy depends on the amount of data that is available for the sub-links, and is independent of the amount of road travel time data, for this latter

is not even used for the estimation. In this section, we aim to show that the probabilistic travel time progression can also be more robust, compared to methods that ignore link travel time data. Again, we will use $\mathcal{D}_{\text{links}}$ to denote the set of travel times available for all links of the target road; and $\mathcal{D}_{\text{road}}$ the set of travel times of all travellers that were detected at each node of the target road. In order to measure robustness, each test involved considering a varying fraction $\rho \in \{0.01, 0.1, 0.3, 1\}$ of the total number of road travel times in $\mathcal{D}_{\text{road}}$ and $\mathcal{D}_{\text{links}}$. More precisely, $\rho = 0.01$ indicates that 10% of the data was sub-sampled from $\mathcal{D}_{\text{road}}$ and another 10% was sub-sampled from each link of $\mathcal{D}_{\text{links}}$. These samples were chosen uniformly at random. We repeated this process 30 times for each of the 4 values of ρ, thus generating 120 different data sets for each one of the 3 target roads. For each data set, we used our travel time progression method for estimating the road travel time distribution from the link data sub-sampled from $\mathcal{D}_{\text{links}}$, and fitted two different types of distributions to the road data sub-sampled from $\mathcal{D}_{\text{road}}$. These latter were the histogram and the kernel density distributions. We therefore measured the variation of the shape of the distributions due to random sub-sampling, also known as *variance of the sampler*. In order to quantify the variation due to random sub-sampling between two distribution arrays $\boldsymbol{p}$ and $\boldsymbol{q}$ of size L, we used the symmetric non-negative Kullback-Leibler divergence, defined as

$$\text{Variability}_{\text{KL}} = D_{\text{KL}}(\boldsymbol{p}\|\boldsymbol{q}) + D_{\text{KL}}(\boldsymbol{q}\|\boldsymbol{p}) \tag{21}$$

where

$$D_{\text{KL}}(\boldsymbol{p}\|\boldsymbol{q}) = \sum_{i=1}^{L} \boldsymbol{p}_i \ln \frac{\boldsymbol{p}_i}{\boldsymbol{q}_i} \tag{22}$$

Figure 7 shows examples of the density function estimates using a histogram distribution (a), a kernel density function (b) and our approach (c), for a given road, upon a given observation period $[t_a, t_b]$ and for a given ρ. The smaller the value of the divergence measure, the more similar $\boldsymbol{p}$ and $\boldsymbol{q}$ are to each others. This divergence was computed for each value of ρ, across all distributions of the same type — i.e. histogram, kernel or probabilistic travel time progression (PTTP) – produced from the 30 tests and across various observation time intervals. The outcome of the robustness tests is shown in Figure 7-d.

7.3. Probabilistic Outlier Detection

Besides being a robust mechanism of travel time estimation over long roadways, the probabilistic framework introduced here can also be used to determine the boundaries that best separate valid travel time data from outliers. By outliers, we mean travellers whose travel time diverges significantly from the one of the target class (or classes). We have seen earlier that the AVI data concerning the whole road, $\mathcal{D}_{\text{road}}$, consists of departure times from origin ε_1 and related total road travel times $\tau_{1:N}$. Using the travel time progression applied to the link data $\mathcal{D}_{\text{links}}$, we wish to determine whether the measurement $\boldsymbol{x} = (\tau_{1:N}, \varepsilon_1)^T \in \mathcal{D}_{\text{road}}$ is likely to be normal or anomalous, i.e. an outlier. Precisely, we are interested in the probability of $\boldsymbol{x}$ which,

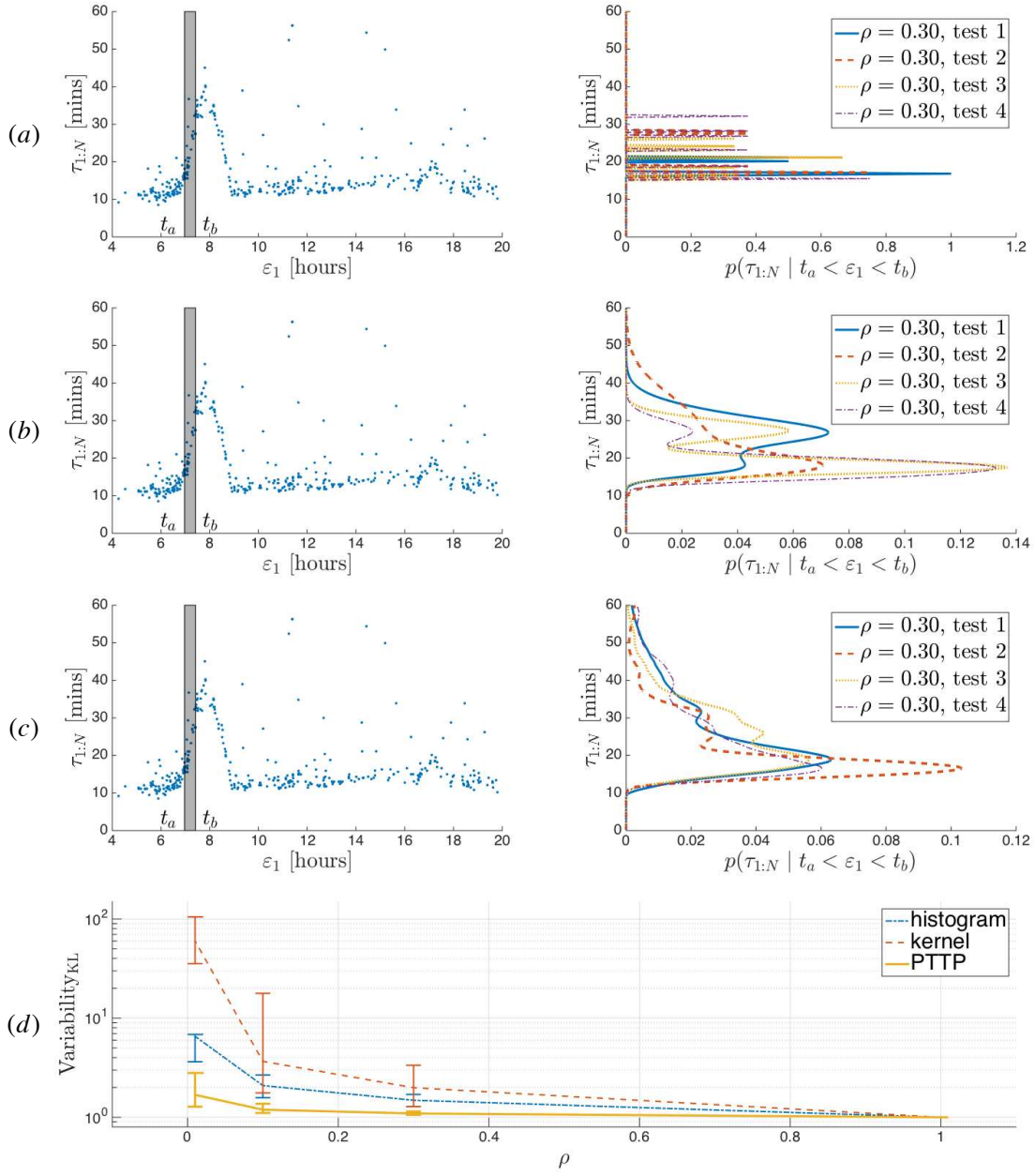

Figure 7: Robustness tests. (a) Histogram distribution fit using only 30% ($\rho = 0.3$) of the data in $\mathcal{D}_{\text{road}}$. (b) Kernel distribution fit using only 30% of the data in $\mathcal{D}_{\text{road}}$. (c) Distribution estimation using our Probabilistic Travel Time Progression (PTTP) and only 30% of the data from each link in $\mathcal{D}_{\text{links}}$. The difference across tests 1 through 4 is due to random sub-sampling. Observation intervals $[t_a, t_b]$ are fixed for all tests. (d) Variability of the travel time distribution across all tests, as a function of the fraction ρ of total sample size. This plot, being in log-scale, shows the values of Variability$_{\mathrm{KL}} + 1$.

by using the chain rule, can be expressed in the form

$$\Pr(\boldsymbol{x}) = \Pr(\mathcal{T}_{1:N} = \tau_{1:N},\, \mathcal{E}_1 = \varepsilon_1) = \Pr(\mathcal{T}_{1:N} = \tau_{1:N} \mid \mathcal{E}_1 = \varepsilon_1)\, \Pr(\mathcal{E}_1 = \varepsilon_1) \tag{23}$$

The posterior over road travel time is given by the probabilistic progression mechanism presented earlier. The prior probabilities $\Pr(\mathcal{E}_1 = \varepsilon_1)$ are estimated from the data in $\mathcal{D}_{\text{road}}$. Clearly, the smaller $\Pr(\boldsymbol{x})$, the higher the chance that the point is an outlier. Thus, the decision on whether to retain or reject $\boldsymbol{x}$ can be based on a threshold. Accordingly, any observation $\boldsymbol{x}$ whose probability is below that threshold is considered anomalous and thus rejected. An example of this type of outlier detection is depicted in Figure 8. In literature, it is common to resort to the Median Absolute Deviation (MAD) and other similar statistics, in

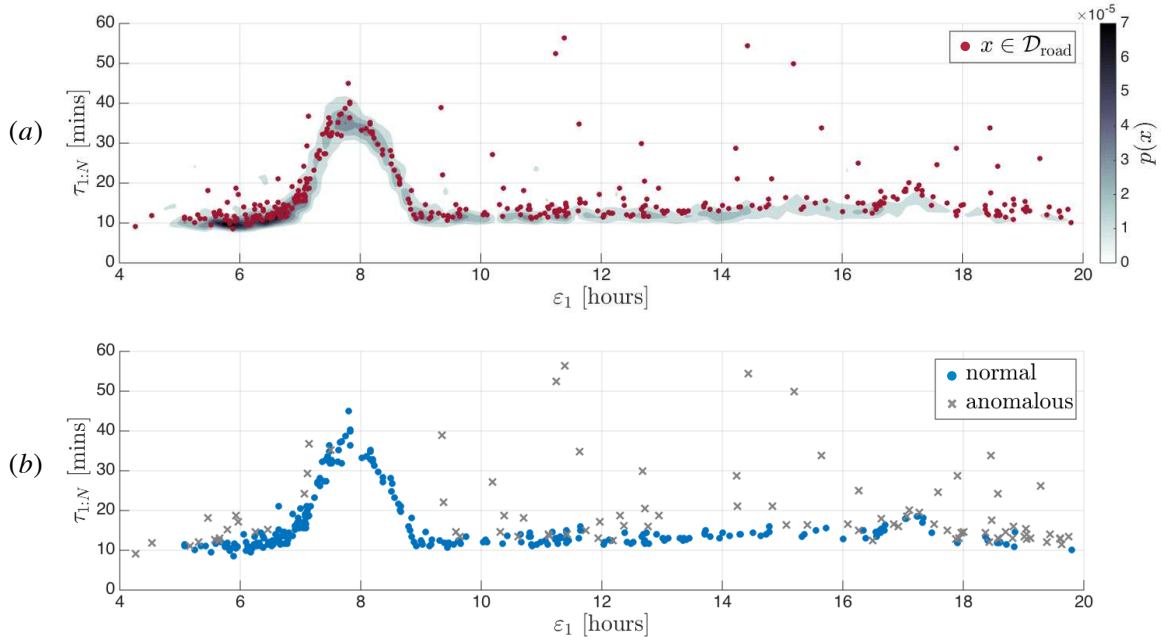

Figure 8: Probabilistic Outlier Detection. (a) AVI data from travellers that have traversed the entire road stretch (dots). The surface of the joint probability $\Pr(\boldsymbol{x})$ is shown in the background; darker regions indicate higher probability values. (b) Outliers (crosses) are the points in $\mathcal{D}_{\text{road}}$ whose probability is below a threshold (9×10^{-10} in this case).

order to effectively remove outliers, prior to any travel time analysis (Ban et al., 2010). The outlier detector proposed here uses the distributions produced by the probabilistic travel time estimator and does no require additional parameters to tweak. By definition, anomalous travel times lie in regions of low probability. Convolving the travel time probability density functions across links has the desired effect of smoothing out the tails of the resulting posterior distributions. As a consequence, using a simple threshold method may suffice to effectively remove most of the "unintended" travel times. Note that by "unintended", we mean travel times other than the target ones. The AVI data may refer to multiple classes of travellers and travel behaviours. These may include, for instance, temporary detours from the target road. If such events are

rare their effect on the road travel time distribution will be negligible, due the convolution operations. By contrast, behaviours that significantly affect the link travel time distributions may have significant impact also on the road travel time distributions. As such, it is important to cleanse the AVI data appropriately, prior to estimating link and corridor travel times

8. Conclusion and Future Work

In this work, we have introduced a novel off-line travel time estimation mechanism that can be used for computing travel time distributions for long roadways, from historical link-travel time data. To achieve this aim, we have proposed a probabilistic framework which generalises the deterministic travel time progression approach to the case of random variables. As the conventional probabilistic travel time progression method, our method does not relies on road travel time data and is therefore expected to produce more reliable estimates, for it effectively overcomes the AVI Data Impoverishment issue. Moreover, treating travel times and departure times as random variables enables handling arbitrarily complex distributions that real AVI data may exhibit. We have tested our system on real Bluetooth data, available in the Brisbane metropolitan area. From our tests, we found that our distribution estimates are consistent with the travel times that are actually observed for the entire road. Like any other data-driven estimator, the accuracy of our method is linked to the amount of data that is available. Nevertheless, we have shown that the travel time distributions estimated through probabilistic travel time progression are generally more robust to sample size changes, in the sense of Kullback-Leibler divergence, compared to those that are learned directly from the road travel time data. Finally, we have shown how the distributions produced through our framework can be used to discern between normal and anomalous travel time behaviours hidden with historical AVI data.

In this paper, we have assumed that the sampling error proper to the AVI scanners can be neglected. In our future research, will relax this assumption by incorporating the effect of the sampling error on the final travel time distribution.

Acknowledgements

This research was financially supported by the UK Engineering and Physical Sciences Research Council (EPSRC) Career Acceleration fellowship grant EP/J002186/1.

References

Aliari, Y., Haghani, A., 2012. Bluetooth sensor data and ground truth testing of reported travel times. Transportation Research Record 2308, 167–172.

Ban, X., J., Li, Y., Skabardonis, A., Margulici, J. D., 2010. Performance evaluation of travel-time estimation methods for real-time traffic applications. Journal of Intelligent Transportation Systems: Technology, Planning, and Operations 14 (2), 54–67.

Ben-Akiva, M., De Palmal, A., Kaysi, I., 1991. Dynamic network models and driver information systems. Transportation Research Part A 25, 251–266.

Bhaskar, A., Chung, E., 2013. Fundamental understanding on the use of bluetooth scanner as a complementary transport data. Transportation Research Part C 37, 42–72.

Bhaskar, A., Chung, E., Dumont, A. G., 2011. Fusing loop detector and probe vehicle data to estimate travel time statistics on signalized urban networks. Computer Aided in Civil and Infrastructure Engineering 26, 433–450.

Bhaskar, A., Qu, M., Chung, E., 2014. A hybrid model for motorway travel time estimation- considering increased detector spacing. Transportation Research Record 2442, 71–84.

Coifman, B., 2002. Estimating travel times and vehicle trajectories on freeways using dual loop detectors. Transportation Research Part A 36, 351–364.

Coifman, B., Krishnamurthy, S., 2007. Vehicle reidentification and travel time measurement across freeway junctions using the existing detector infrastructure. Transportation Research Part C 15, 135–153.

Deng, W., Lei, H., Zhou, X., 2013. Traffic state estimation and uncertainty quantification based on heterogeneous data sources: A three detector approach. Transportation Research Part B 57, 132–157.

Du, L., Peeta, S., Kim, Y. H., 2012. An adaptive information fusion model to predict the short-term link travel time distribution in dynamic traffic networks. Transportation Research Part B 46, 235—252.

Friesz, T. L., Bernstein, D., Smith, T. E., Tobin, R. L., Wie, B. W., 1993. A variational inequality formulation of the dynamic network user equilibrium problem. Operations Research 41 (1), 179–191.

Haghani, A., Hamedi, M., Sadabadi, K. F., Young, S. E., Tarnoff, P. J., 2010. Data collection of freeway travel time ground truth with bluetooth sensors. Transportation Research Record 2160, 60–68.

Herrera, J. C., Bayen, A., 2010. Incorporation of lagrangian measurements in freeway traffic state estimation. Transportation Research Part B 44 (460-481).

Herrera, J. C., Work, D. B., Herring, R., Ban, X., Jacobson, Q., Bayen, A., 2010. Evaluation of traffic data obtained via gps-enabled mobile phones: the mobile century field experiment. Transportation Research Part C 18, 568—583.

Hofleitner, A., Herring, R., Abbeel, P., Bayen, A., 2012a. Learning the dynamics of arterial traffic from probe data using a dynamic bayesian network. IEEE Transactions on Intelligent Transportation Systems 13 (4).

Hofleitner, A., Herring, R., Bayen, A., 2012b. Arterial travel time forecast with streaming data: A hybrid approach of flow modeling and machine learning. Transportation Research Part B 46 (9), 1097–1122.

Jenelius, E., Koutsopoulos, H. N., 2013. Travel time estimation for urban road networks using low frequency probe vehicle data. Transportation Research Part B: Methodological 53.

Khoei, Mohammad, A., Bhaskar, A., Chung, E., 2013. Travel time prediction on signalised urban arterials by applying sarima modelling on bluetooth data. In: Australasian Transport Research Forum (ATRF) 2013.

Khosravi, A., Mazloumi, E., Nahavandi, S., Creighton, Douglas, C., van Lint, J., W. C., 2011. A genetic algorithm-based method for improving quality of travel time prediction intervals. Transportation Research Part C: Emerging Technologies.

Kwong, K., Kavaler, R., Rajagopal, R., Varaiya, P., 2009. Arterial travel time estimation based on vehicle re-identification using wireless magnetic sensors. Transportation Research Part C 17, 586–606.

Liu, H., Ma, W., 2009. A virtual vehicle probe model for time-dependent travel time estimation on signalized arterials. Transportation Research Part C 17, 11–26.

Mahmassani, H. S., Liu, Y. H., 1999. Dynamics of commuting decision behavior under advanced traveller infromation systems. Transportation Research Part C 7, 91–107.

Murphy, P., Welsh, E., Frantz, P., 2002. Using bluetooth for short-term ad-hoc connections between moving vehicles: A feasibility study. In: IEEE Vehicular Technology Conference. Birmingham, AL, pp. 414–418.

Nantes, A., Miska, M. P., Bhaskar, A., Chung, E., 2014. Noisy bluetooth traffic data? ARRB Road & Transport Research

Journal 23 (1), 33–43.

Nantes, A., Ngoduy, D., Bhaskar, A., Miska, M., E., C., 2015. Real-time traffic state estimation in urban corridors from heterogeneous data. Transportation Research Part C doi:10.1016/j.trc.2015.07.005.

Nanthawichit, C., Nakatsuji, T., Suzuki, H., 2013. Application of probe vehicle data for real-time traffic state estimation and short-term prediction on a freeway. In: Annual Meeting of the Transportation Research Board.

Ndoye, M., Totten, V. F., Krogmeier, J. V., Bullock, D. M., 2011. Sensing and signal processing for vehicle reidentification and travel time estimation. IEEE Transactions on Intelligent Transportation Systems 12, 119–131.

Ngoduy, D., 2008. Applicable filtering framework for online multiclass freeway network estimation. Physica A 387, 599–616.

Ngoduy, D., 2011. Kernel Smoothing Method Applicable to the Dynamic Calibration of Traffic Flow Models. Computer-Aided Civil and Infrastructure Engineering 26, 420–432.

Pasolini, G., Verdone, R., 2002. Bluetooth for its? In: The 5th International Symposium on Wireless personal multimedia communications. Vol. 1. pp. 315—319.

Qiao, W., Haghani, A., Hamedi, M., 2013. A nonparametric model for short-term travel time prediction using bluetooth data. Journal of Intelligent Transportation Systems: Technology, Planning, and Operations 17 (2), 165–175.

Rajagopal, R., Varaiya, P., 2007. Health of california's loop detector system: Final report for path to 6300. Tech. rep., University of California, Berkeley.

Ramezani, M., Geroliminis, K., 2012. On the estimation of arterial route travel time distribution with markov chains. Transportation Research Part B 46, 1576—1590.

Sawant, H., Jindong, T., Qingyan, Y., Qizhi, W., 2004. Using bluetooth and sensor networks for intelligent transportation systems. In: Proceedings of the 7th International IEEE Conference on Intelligent Transportation Systems. pp. 767—772.

Shim, S., Shin, K., Choi, K., Namkoong, J., Lee, S., 2011. Estimating path travel time by aggregating link travel times obtained through dedicated short range communication probe data. In: 18th ITS World Congress.

Sun, L., Yang, J., Mahmassani, H. S., 2008. Travel time estimation based on piecewise truncated quadratic speed trajectory. Transportation Research Part A 42, 173–186.

van Lint, J. W. C., Hoogendoorn, S. P., van Zuylen, H. J., 2005. Accurate freeway travel time prediction with state-space neural networks under missing data. Transportation Research Part C 13, 347–369.

van Lint, J. W. C., van der Zijpp, N., 2003. Improving a travel-time estimation algorithm by using dual loop detectors. Transportation Research Record 1855, 41–48.

Wang, Y., Papageorgiou, M., 2005. Real-time freeway traffic state estimation based on extended Kalman filter: a general approach. Transportation Research Part B 39, 141–167.

Wang, Y., Papageorgiou, M., Messmer, A., 2007. Real-Time Freeway Traffic State Estimation Based on Extended Kalman Filter: A Case Study. Transportation Science 41, 167–181.