(This is a sample cover image for this issue. The actual cover is not yet available at this time.)

Transportation Research Part B 46 (2012) 1097–1122

Contents lists available at SciVerse ScienceDirect

Transportation Research Part B

journal homepage: www.elsevier.com/locate/trb

Arterial travel time forecast with streaming data: A hybrid approach of flow modeling and machine learning

Aude Hofleitner [a,*], Ryan Herring [b], Alexandre Bayen [a,c]

[a] UC Berkeley, Department of Electrical Engineering and Computer Science, Berkeley, CA 94720-1758, United States
[b] UC Berkeley, Department of Industrial Engineering and Operations Research, Apple Inc., Cupertino, CA, United States[1]
[c] UC Berkeley, Department of Civil and Environmental Engineering, Berkeley, CA 94720-1758, United States

ARTICLE INFO

Article history:
Received 19 March 2011
Received in revised form 22 March 2012
Accepted 23 March 2012

Keywords:
Arterial traffic
Estimation
Forecast
Streaming data
Machine learning
GPS probe data

ABSTRACT

This article presents a hybrid modeling framework for estimating and predicting arterial traffic conditions using streaming GPS probe data. The model is based on a well-established theory of traffic flow through signalized intersections and is combined with a machine learning framework to both learn static parameters of the roadways (such as free flow velocity or traffic signal parameters) as well as to estimate and predict travel times through the arterial network. The machine learning component of the approach uses the significant amount of historical data collected by the *Mobile Millennium* system since March 2009 with over 500 probe vehicles reporting their position once per minute in San Francisco, CA.

The hybrid model provides a distinct advantage over pure statistical or pure traffic theory models in that it is robust to noisy data (due to the large volumes of historical data) and it produces forecasts using traffic flow theory principles consistent with the physics of traffic. Validation of the model is performed in two different ways. First, a large scale test of the model is performed by splitting the data source into two sets, using the first to produce the estimates and the second to validate them. Second, an alternate validation approach is presented. It consists of a 3-day experiment in which GPS data was collected once per second from 20 drivers on four routes through San Francisco, allowing for precise calculation of actual travel times. The model is run by down-sampling the data and validated using the travel times from these 20 drivers. The results indicate that this approach is a significant step forward in estimating traffic states throughout the arterial network using a relatively small amount of real-time data. The estimates from our model are compared to those given by a data-driven baseline algorithm, for which we achieve a 16% improvement in terms of the root mean squared error of travel time estimates. The primary reason for success is the reliance on a flow model of traffic, which ensures that estimates are consistent with the physics of traffic.

1. Introduction and background

In numerous parts of the world, traffic congestion has a significant impact on economic activity. An essential step towards active congestion control is the creation of accurate, reliable traffic monitoring systems. Historically, these systems have been mostly limited to highways and have relied on public or private data feeds from dedicated sensing infrastructure, which often includes loop detectors, radars and video cameras.

[1] Afffiliation during redaction of the article.

* Corresponding author.
E-mail address: aude.hofleitner@polytechnique.edu (A. Hofleitner).
URL: http://www.eecs.berkeley.edu/~aude (A. Hofleitner).

0191-2615/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.trb.2012.03.006

For highway networks, it has become common practice to perform both system identification of highway parameters (free flow speed, traffic jam density and flow capacity) and estimation of traffic state (flow, density, length of queues, bulk speed and shockwave location) at a fine spatio-temporal scale (Work et al., 2010; Bickel et al., 2007) using dedicated infrastructure. These approaches heavily rely upon both the availability of data and highway traffic flow models developed over the last half century (Lighthill and Whitham, 1955; Richards, 1956; Daganzo, 1994). These models and data assimilation algorithms have been used to transform this data into usable traffic information (see Work et al. (2010), Thiagarajan et al. (2009), Horvitz et al. (2005), and Krause et al. (2008) for a discussion specific to highways). These highway traffic monitoring systems leverage Kalman filtering (Sun et al., 2004) (or other analogous techniques) and system identification methods to estimate both the macroscopic state of the highways (flow, density, length of queues, bulk speed and shockwave) as well as highway parameters (free flow speed, traffic jam density and flow capacity).

For arterials (the secondary network), traffic monitoring is more challenging: probe vehicle data is the only significant data source with the prospect of global coverage in the future. It comes from various sources with specific challenges:

- *Fleet data* (FedEx, UPS, taxis, etc.) provides information from one minute sampled GPS data (the current standard in the United States) but with specific spatio-temporal travel patterns (fleets avoid congestion).
- *Participatory sensing* (GPS enabled smartphone or aftermarket device data or 2-way navigation device), for example Garmin, INRIX, Google, Nokia or Waze. This data is unpredictable, sparse, and no single company has ubiquitous coverage.
- *Vehicle re-identification* (*e.g.* RFID, magnetic signature (Kwong et al., 2009), Bluetooth readers, Automated Plate Recognition Cameras) is also used for traffic monitoring, with deployment of readers along some small portion of the transportation network.

The aforementioned features of probe vehicle data, including the lack of ubiquity and reliability, the variety of data types and randomness of the corresponding spatio-temporal coverage, make it challenging for fully characterizing macroscopic traffic model parameters and doing state estimation with these models for large arterial networks. The accuracy of GSM positioning makes it challenging to use this data source for arterial traffic estimation, even though it provides accurate travel time estimates on highways (Liu et al., 2008).

Microscopic models have mainly focused on single intersections (or a small number of intersections) using important data availability assumptions (including signal timing, vehicle counts or high penetration rate travel time measurements (Ban et al., 2009)). Wireless technology provides travel time measurements of a high proportion of the flow of vehicles (Kwong et al., 2009) through vehicle magnetic signature re-identification. This information remains limited to the equipped road which represents, today, a marginal fraction of the arterial network. Geroliminis and Daganzo (2008) developed macroscopic flow models for the secondary network, but the parameters require site-specific calibration experiments. The physics of arterial flows is governed by the presence of traffic lights, often with unknown cycles, intersections, stop signs, and parallel queues. Collecting these detailed parameters is tedious and hence only documented for some sections of few cities.

In light of these challenges, a statistical approach for characterizing the macroscopic state of traffic is well-suited toward designing a robust, scalable arterial traffic monitoring system. Such an approach makes it possible to account for the high variability of arterial traffic while learning the distinct patterns from past data. Real-time data is then fused with the learned patterns to identify the current state of traffic. Following this approach, neural networks and state-space neural networks (Van Lint et al., 2005; Liu et al., 2006), graphical networks (Bayesian networks and Markov Random Fields) (Herring et al., 2010; Park and Lee, 2004; Sun et al., 2006; Furtlehner et al., 2007), regression techniques and time series analysis (Geroliminis and Skabardonis, 2006; Herring et al., 2010) have been introduced to produce short-term traffic predictions for both freeway and arterial traffic with promising results. These articles model the spatio-temporal dependencies of the links of the network which provides more robustness when little or no data is available on some parts of the network. However, none of these articles present a comprehensive modeling approach of arterial traffic flow.

Zhang and Taylor (2007) successfully applied Bayesian networks to automated incident detection. Our approach is based on the similar idea that traffic theory can be formulated in a statistical framework to improve estimation capabilities while leveraging prior information on the model. The fundamental flow conservation laws governing the physics of traffic can be used as a basis for designing a statistical inference framework for learning key traffic parameters. To our knowledge, integration of traffic flow theoretic models into machine learning algorithms is still an emerging field, for which few contributions exist. The efficient use of such models in a statistical inference framework is precisely the contribution of this article.

In this article, we use well-established traffic flow modeling approaches relying on hydrodynamic theory (Lighthill and Whitham, 1955; Richards, 1956; Daganzo, 1994) as the basis for a Bayesian network formulation. First, we recall how *analytical* probability distribution of travel times between *arbitrary* locations can be derived from kinematic wave theory, following the derivations performed by Hofleitner et al. (2012) and Zheng and Van Zuylen (2010). The probability distribution of travel times are parameterized by a minimal set of *link parameters* (signal timing, link capacity and characteristics of the free flow speed) and conditioned on the *state variable* of that link (queue length, *i.e.* location of the last vehicle stopping in the queue on the link). The dynamic evolution of the queue length of each link is parameterized by *intersection parameters* (turn movements and arrival of vehicles in the network). The algorithm presented in this article enables the learning of the link and intersection parameters characterizing the traffic dynamics, even under low penetration rates of probe vehicles characteristics of today's technology in the United States. Estimating these static model parameters is particularly important

Fig. 1. *Mobile Millennium* system, *Left*: cumulated raw probe vehicle data collected on a typical day, for one of the feeds of *Mobile Millennium*. Each dot corresponds to a point where a probe vehicle emitted its position data. *Center*: traffic monitoring system output displayed on a phone. *Right*: web interface of real time monitoring system.

as they are often difficult or impractical to measure directly for an entire network (consisting of many thousands of links). This statistical description of the dynamics of the network enables the estimation of traffic conditions with missing data (no or too little data available on a set of links of the network) in real time, as well as the short term prediction of traffic evolution. Distributions over travel times between points in the network are computed from these estimated parameters. We focus on travel times because they represent one of the most important metrics for drivers.

Our experiments show that even with limited amounts of data (representative of today's data availability), we can obtain accurate travel time estimates. We use data collected by one of the feeds of *Mobile Millennium* Bayen et al. (2011), a fleet of 500 vehicles sampled every minute. The system receives an average of 500,000 data points per day (see Fig. 1) for the San Francisco vicinity. As a basis for comparison, we also develop a simple model and algorithm for processing probe data, denoted *baseline* model and described in Section 5. The improvement of our statistical model over this baseline model is substantial.

The remainder of the article is organized as follows. In Section 2, we present the traffic model and the underlying assumptions. We summarize how probability distributions of travel time between any two locations are derived from this model and model the spatio-temporal statistical dependencies between the links of the network (Section 3). In Section 4, we describe the algorithm developed to learn the parameters of the network and then infer and predict traffic conditions and distributions of travel time across the network (EM Algorithm using particle filtering). In Section 5, we present the results of the model on a subset of the data collected to date.

2. Traffic modeling

2.1. Traffic model and assumptions

We make the following standard assumptions on the dynamics of traffic flow, commonly made in the transportation engineering literature:

1. *Hydrodynamic fluid assumption:* Following classical traffic flow theory, we model vehicular flow as a continuum and represent it with macroscopic variables of *flow* $q(x,t)$ (veh/s), *density* $\rho(x,t)$ (veh/m) and *velocity* $v(x,t)$ (m/s). The definition of flow gives the following relation between these three variables: $q(x,t) = \rho(x,t)v(x,t)$. We make the assumption

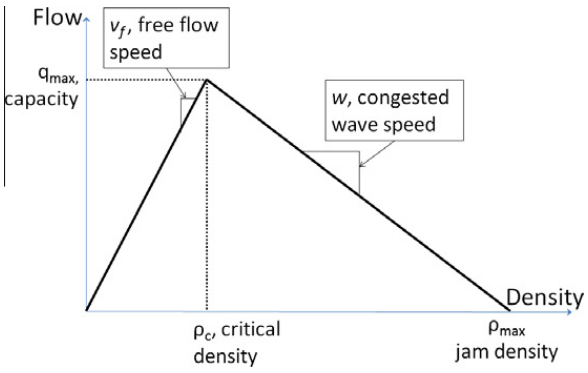

Fig. 2. The fundamental diagram: empirically constructed relation between flow and density of vehicles.

of a triangular fundamental diagram (Fig. 2), as used for arterial traffic estimation and control in different contributions (Geroliminis and Skabardonis, 2010; Zhang and Kim, 2005), some of which are validated with experimental data. Moreover, the differences in fundamental diagram do not change significantly the dynamics of traffic on a large scale (Blandin et al., 2011). The triangular fundamental diagram is parameterized by v_f, the free flow speed (m/s), $\rho_{\max}$, the jam (or maximum) density (veh/m) and $q_{\max}$, the capacity (veh/m). From these parameters, we can derive the critical density $\rho_c = q_{\max}/v_f$ and the congested wave speed $w = q_{\max}/(\rho_{\max} - \rho_c)$. For a given road segment of interest, the vehicles arrive into the link with a specific spatial spacing: this incoming arrival spatial spacing corresponds to the arrival density ρ_a. These quantities will appear with indices later in the text when required.

2. *Characterization of the state of traffic assumption*: For each link of the network, traffic conditions are characterized by a traffic state variable. This state variable represents the number of vehicles that stop on the link per light cycle. It is denoted ξ (generically) and will appear with indices later in the text when required.
3. *Time discretization assumption:* We model arterial traffic as a discrete time dynamical system and denote by Δ_t the time discretization (typically Δ_t is in the order of *five to fifteen* minutes). Let t_0 denote the initial time, we assume that the state and flow entering each link is constant on time intervals $[t_0 + t\Delta_t, t_0 + (t+1)\Delta_t]$ for $t \in \{0, \ldots, T\}$. For a specific time of day (*e.g.* Mid-week evening rush hour), the parameters of traffic signals (red time, R and cycle time, C) do not change. According to the time discretization. This work is mainly focused on deriving travel time distributions for cases in which measurements are sparse. Thus, considering the state of the system as piecewise constant does not prevent estimation, as we are interested in *trends* more than in *fluctuations*. Under these assumptions, for intersections signalized by a traffic light, the system exhibits a periodic behavior, in each time interval, dictated by the period of the traffic light. Note that this assumption allows the queue length to change over time, but the fundamental characteristics of the queue (e.g. the maximum length reached during a cycle and thus the number of vehicles stopping in the queue per cycle) remain constant within a time interval. In particular, the number of vehicles stopping per cycle is constant for a link i and a time interval t and is denoted $\xi^{i,t}$.
4. *Transition modeling assumption:* According to the time discretization assumption, the state variables $\xi^{i,t}$ are piecewise constant, with possible discontinuities at the end of each interval. These transitions model the information propagation on the road network by taking into account the spatio-temporal dependencies of the state of the links. Based on the conservation of vehicles, we model these transitions using an approach derived from the *Cell Transmission Model* (Daganzo, 1994). The state of a link during a time interval depends on the state of this link and the adjacent links during the previous time interval, to represent the constraint of supply and demand of downstream and upstream links respectively. The dynamic evolution of the traffic state of each link is probabilistic and parameterized by *turn movement probabilities* from and to neighboring links and *arrival rates* of vehicles in the network. The parameters of the turn movements can be learned historically.
5. *Conditional independence assumption:* We consider a graphical model representing the conditional independence assumptions between the state variables (representing traffic conditions) and the observations. A graphical model is a graph in which the nodes represent random variables. The edges denote the conditional independence structure between the random variables. For more background on graphical models, please refer to Jordan (1999). The random variables represented by the present graph are (i) the *state variables* $\xi^{i,t}$, number of vehicles stopping on a link per light cycle, on each link i at each time interval t and (ii) the set of *travel times* $\mathbf{y}^{i,t}$ measured on each link i at each time interval t. The conditional independence assumptions between the random variables can be formulated as follows:
   (a) Travel time measurements on link i for time interval t are independent and identically distributed given the state $\xi^{i,t}$ (number of vehicles stopping on a link per light cycle) of this link at this time interval. This means that given the state $\xi^{i,t}$, a travel time on link i during time interval t does not depend on the realization of the other travel time measurements on link i during time interval t. Note that the *conditional* independence assumption is much less strong than assuming independence between travel times.
   (b) Travel time measurements on link i for time interval t are independent from all the other random variables given the state $\xi^{i,t}$ of this link at this time interval. This means that given the state $\xi^{i,t}$, a travel time on link i during time interval t does not depend on the realization of the other random variables. It does not depend on the states of the other links at any time intervals nor on the state of link i during time intervals previous or posterior to time interval t nor on the realization of other travel time measurements.
   (c) Conditioned on the state of the adjacent links (including itself) at the previous time interval t, the state $\xi^{i,t+1}$ of link i at time interval $t+1$ is independent from the travel time measurements from anterior time periods and all other anterior state variables. This means that given the states $\xi^{j,t}$ of the adjacent links of link i (including link i), the state of link i during time interval $t+1$ does not depend on the realization of the anterior random variables. It does not depend on the states of the non adjacent links at time interval t nor on the state of any link at time intervals anterior to $t-1$ nor on the realization of travel time measurements during time intervals interior to t. In the following, the set of adjacent links of link i (including link i) is referred to as the *neighbors* of link i.
6. *Data availability assumption:* We receive streaming data in real-time. The data consists of point to point travel time measurements from a small subset of vehicles traveling on the network. Measurements from the past are stored and accessible in real time. The *Mobile Millennium* system, developed by UC Berkeley and Nokia (Bayen et al., 2011) provides such data (see Fig. 1).

2.2. Arterial traffic dynamics

In arterial networks, traffic conditions are driven by the formation and the dissipation of queues at intersections. The dynamics of queues are characterized by shocks, which are formed at the interface of traffic flows with two different densities. In arterial networks, the dynamics of the flow are dependent on the characteristics of the traffic signal. The duration of the red time and the cycle time of a traffic light are respectively noted R and C (see Section 2.1, Assumption 3). The following derivations are based on classical horizontal queuing theory and have been known by the traffic engineering community for many years. We present these standard derivations for completeness as they constitute a basis for the model developed in the present article. For notational simplicity, the reference to the link i and the time interval t are omitted in this section.

We define two discrete traffic regimes: *undersaturated* and *congested*, which represent different dynamics of the arterial link depending on the presence (respectively the absence) of a remaining queue when the light switches from green to red. At the transition between the two regimes the number of vehicles that stop in the link per cycle is the maximum number of vehicles that can exit the link in the duration of a cycle. We call this number of vehicle, the *saturation number of vehicles* ξ_s. As Kimber and Hollis (1979) pointed out, there is a smooth transition between these regimes. The distinct regimes are introduced for the mathematical derivations of the travel time distributions, in particular because of the presence of a remaining queue in the congested regime. Fig. 3 illustrates these two regimes under the assumptions made in Section 2.1. The assumption of a triangular fundamental diagram and the constant arrival density imply the constant speed of formation and dissolution of the queue (respectively denoted v_a and w), computed with the Rankine-Hugoniot jump conditions (Evans, 1998) as

$$v_a = \frac{\rho_a v_f}{\rho_{\max} - \rho_a} \quad \text{and} \quad w = \frac{\rho_c v_f}{\rho_{\max} - \rho_c}. \tag{1}$$

2.2.1. Undersaturated regime

The queue fully dissipates within the green time. This queue is called the *triangular queue* (from its triangular shape on the space–time diagram of trajectories). It is defined as the spatio-temporal region where vehicles are stopped on the link. Its length is called the maximum queue length, denoted $l_{\max}$. The number of vehicles stopping during a light cycle is denoted ξ. From traffic theory, we derive:

$$l_{\max} = R\frac{w v_a}{w - v_a} = R\frac{v_f}{\rho_{\max}}\frac{\rho_c \rho_a}{\rho_c - \rho_a}. \tag{2}$$

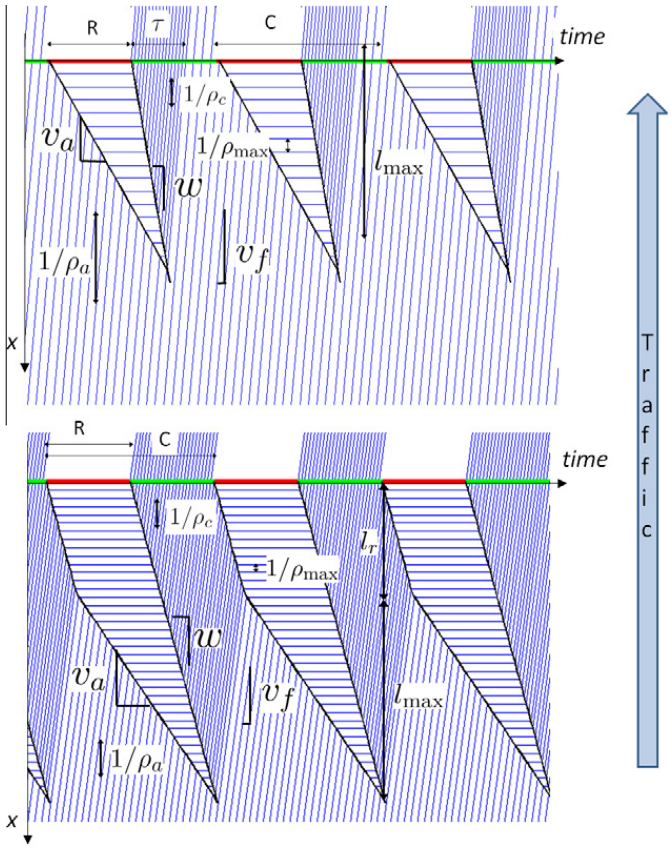

Fig. 3. Space time diagram of vehicle trajectories with uniform arrivals under an undersaturated traffic regime (top) and a congested traffic regime (bottom).

The duration between the time when the light turns green and the time when the queue fully dissipates is the *clearing time* denoted τ, sometimes also referred to as *saturation green time*. The maximum number of vehicles that can go through the intersection during a light cycle is the *saturation number of vehicles*, introduced earlier, denoted ξ_s. Recalling that ξ denotes the number of vehicles which stop in the queue per cycle, the relation with the clearing time is given by

$$\tau = (C - R)\frac{\xi}{\xi_s}, \tag{3}$$

and we notice that $\xi = \xi_s$ when the clearing time reaches $C - R$, *i.e.* when the queue fully dissipates as the signal turns red (limit of the undersaturated regime).

2.2.2. Congested regime

In this regime, there exists a part of the queue downstream of the triangular queue called *remaining queue* with length l_r corresponding to vehicles which must stop multiple times before going through the intersection. The number of vehicles which stop in the queue per cycle is denoted ξ. It is the sum of the number of vehicles which stop in the triangular queue (ξ_s vehicles) and in the remaining queue ($l_r/\rho_{\max}$).

All notations introduced up to here are illustrated for both regimes in Fig. 3, except ξ and ξ_s which represent number of vehicles (and are related to the corresponding queue lengths through the maximum density $\rho_{\max}$).

2.2.3. Stationarity of the two regimes

Assumption 2 made earlier implies the periodicity of these queue evolutions for each time interval Δ_t (see Fig. 3). As indicated by the slopes of the trajectories in the figure, vehicles travel at the free flow speed v_f. The distance between two vehicles is the inverse of the arrival density $1/\rho_a$. The time during which vehicles are stopped in the queue is represented by the horizontal line in the queue. The length of this line represents the delay experienced in the corresponding queue. The distance between vehicles stopped in the queue is the inverse of the maximum density $1/\rho_{\max}$. When the queue dissipates, vehicles are released with a speed v_f and a density ρ_c, the distance between two vehicles is $1/\rho_c$.

We next use these two regimes to derive probability distribution functions for the travel time along a link (Hofleitner et al., 2012; Zheng and Van Zuylen, 2010).

2.2.4. Model for differences in driving behavior

We define the free flow pace p_f as the inverse of the free flow speed v_f. To account for differences in driving behavior, we model the free flow pace as a random variable, distributed among the different drivers according to a probability distribution φ. If we choose a family of distributions, it is parameterized by a vector θ_p. In this article, we assume that the free flow pace is distributed according to a Gamma distribution[1] parameterized by $\theta_p = (\alpha, \beta)$. This variability also takes into account small errors in the model such as the ones due to a choice of fundamental diagram, the existence of stochastic overflow queues (Viti and Van Zuylen, 2009) or to non-uniform arrival rates.

2.3. Network model and associated notation

In the derivations, we define a set of independent parameters to characterize the probability distribution of travel times for each link of the network. First, we learn these parameters from historical data. Then, we perform estimation of traffic conditions in real-time from sparse streaming data. The parameters are specific for each link i but we omit the indices i in this section for notational simplicity. We summarize here the variables that are learned historically by the model and that are sufficient to characterize the travel time distribution on each link of the network, conditioned on the number of vehicles in the queue (dynamic state variable).

- Static model parameters (learned historically): cycle time, C, red time, R, saturation number of vehicles, ξ_s, parameters of the free flow pace distribution, θ_p.
- Traffic state (estimated dynamically), ξ (number of vehicles in the queue).

The model only uses two parameters derived from the fundamental diagram (p_f and ξ_s). These two parameters allow for the computation of the critical density and the capacity but not the maximum density $\rho_{\max}$. However, the maximum density (effective length of the vehicles) may be estimated off-line with other means (*e.g.* The *Highway Capacity Manual* (Trans Res Board, 2000)). It may remain constant over time and be the same for links with similar properties. Assuming that the maximum density has been estimated, we can estimate the probability distribution of the other traffic variables, including flow and density of vehicles at any location x and time interval t, where x denotes the distance to the downstream intersection.

The *time evolution* of the state of traffic depends on the probabilistic *assignment* of vehicles to the links of the network. We denote L^k_{in} (resp. L^k_{out}) the set of incoming (resp. outgoing) links of intersection k. We allow for dummy links representing sinks, k_{out} and sources, k_{in} that model vehicles arriving or leaving the network at intersection k (parking, residential roads,

[1] The probability distribution γ of a Gamma random variable $x \in \mathbb{R}^+$ with shape α and inverse scale parameter β is given by $\gamma(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$, where Γ is the Gamma function defined on $\mathbb{R}^+$ and with integral expression $\Gamma(x) = \int_0^{+\infty} t^{x-1} e^{-t}\, dt$.

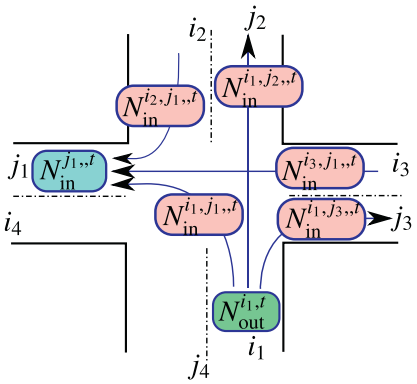

Fig. 4. Schematic representation of an intersection k with incoming links $L_{in}^k = \{i_1, i_2, i_3, i_4\}$ and outgoing links $L_{out}^k = \{j_1, j_2, j_3, j_4\}$. The figure also represents the vehicle assignment during time interval t.

etc.). At time interval t, we define $n_{in}^{i,t}$ (resp. $n_{out}^{i,t}$) the number of vehicles arriving (resp. leaving) link i during a cycle and $N_{in}^{i,t}$ (resp $N_{out}^{i,t}$) the total number of vehicles arriving (resp. leaving) the link during the duration Δ_t of time interval t. In the derivations at time interval t, for two adjacent links i and j (with i upstream of j), we call $n_{in}^{i,j,t}$ (resp. $N_{in}^{i,j,t}$) the number of vehicles arriving to link j from link i during a cycle (resp. during time interval t). These notations are summarized in Fig. 4.

The dynamics of the state of traffic are fully characterized by the turn movements on the network. For an incoming link $i \in L_{in}^k$ and an outgoing link $j \in L_{out}^k \cup k_{out}$ of intersection k, the probability of going from link i to link j is called a *turn probability* and denoted $v^{i,j}$. These variables are non negative and satisfy $\sum_{j \in L_{out}^k \cup k_{out}} v^{i,j} = 1$. The presence of a source at the intersection is modeled for each outgoing link of the intersection $j \in L_{out}^k$ via a Poisson process with intensity λ_j.

In the following, we summarize the derivation of probability distributions $g^i(y_{x_1,x_2}|\xi^{i,t})$ for the travel time y_{x_1,x_2} between two locations x_1 and x_2 on a link i of the network, conditioned on its state $\xi^{i,t}$ at time interval t. Zheng and Van Zuylen (2010) first introduced these derivations for travel time distributions on an arterial link with stochastic queues. Hofleitner et al. (2012) also studied these analytical derivations, focusing on the distribution of travel times between *arbitrary* locations. The set of travel time measurements received for link i during time interval t is denoted $\mathbf{y}^{i,t}$. We also derive transition probabilities for the number of stopped vehicles per cycle on a link i at time $t+1$ given the number of stopped vehicles of the neighboring links at time t. The full set of notation used in this article is available in Appendix A for convenience.

3. Probabilistic model of traffic dynamics

3.1. Modeling the travel time distributions between any two points on a link

Previous work on the characterization of travel time distributions derives the mean average delay and queue length at the end of the green time using analytical expressions and numerical simulation under different arrival processes (Webster, 1958; Van Den Broek et al., 2006; Leeuwaarden, 2006). Other work studies the influence of the stochasticity of overflow queues (Viti and Van Zuylen, 2009) on the probability distribution of travel times. In the present article, we focus our attention on specific aspects of stochasticity which represent significant factors for the variability of travel times among vehicles traveling on an arterial link at the same time: the entrance time with respect to the beginning of the cycle which determines the duration of the delay. This choice is motivated by the desire of identifying *analytically* the effects of the aforementioned *specific* stochastic patterns on the behavior of the system. Compared to previous work related to deriving travel time distributions in arterial networks, Hofleitner et al. (2012) presented analytical derivations between *arbitrary* locations, a feature which is required to incorporate measurements from probe vehicles which send their locations at random places, not necessarily at the beginning and end of links. The details of the derivations are out of the scope of this article and are fully documented in (Hofleitner et al., 2012; Hofleitner and Bayen, 2011) and we summarize the different steps of the derivations:

- Derive the probability of delay δ_{x_1,x_2} experienced between the two locations x_1 and x_2 on the link, parameterized by the network parameters and the traffic state.
- Model the differences in driving behavior, as presented in Section 2.2. Considering a free flow pace p_f with probability distribution φ, the probability distribution of free flow travel times $y_{f;\,x_1,x_2}$ between locations x_1 and x_2 is computed by scaling φ since $y_{f;x_1,x_2} = p_f(x_1 - x_2)$.
- Derive the probability distribution of travel times y_{x_1,x_2} between locations x_1 and x_2 as the sum of two independent random variables: the delay δ_{x_1,x_2} and the free flow travel time $y_{f;x_1,x_2}$.

We illustrate graphically (Fig. 5) the probability distribution of travel times on a congested link. The pdf of travel times is the convolution between the pdf of free flow travel times and the pdf of delay. On an undersaturated arterial link, some vehicles do not experience delay, their delay can be considered as a random variable with mass probability at 0. The

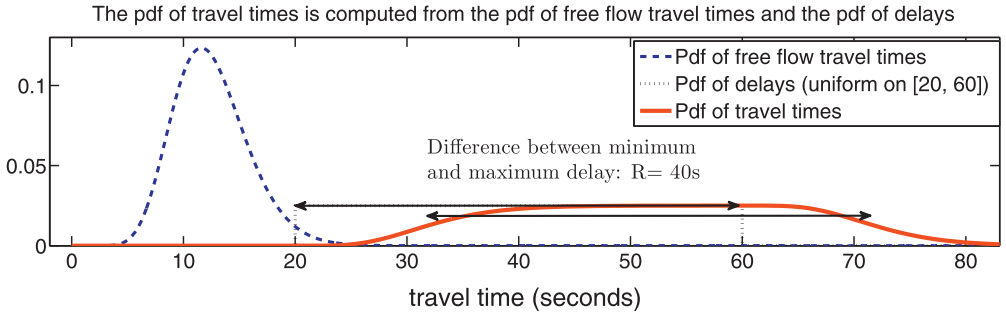

Fig. 5. The pdf of travel times is computed from the pdf of free flow travel times and the pdf of delays. On a congested arterial link, vehicles experience a delay due to the presence of the traffic light. When arrivals are uniform, the delay is uniformly distributed between a minimum value $\delta_{\min}$ and a maximum value $\delta_{\max}$. The convolution of the probability distribution of delays (dotted line) with the probability distribution of free-flow travel times (dashed line) gives the probability distribution of travel times on an arterial link (solid line). The illustration is computed for $\delta_{\min} = 20$ s, $\delta_{\max} = 60$ s, the free flow pace is taken to be a random variable with Gamma distribution, a mean of 1/8 s/m and a standard deviation of 1/30 s/m.

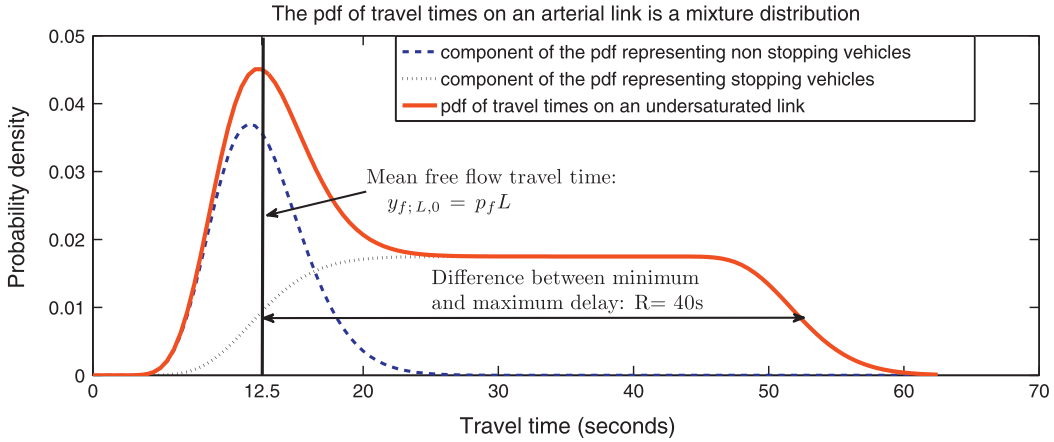

Fig. 6. On an undersaturated link, the probability distribution of travel times (solid line) is a mixture distribution with two components. The first component (dashed line) represents the vehicles that do not stop on the link (zero delay), the second component (dotted line) represents the vehicles that experience delay on the link. Because of the uniform arrivals, the delay is uniform between 0 s and the duration of the red time R. The illustration is computed for $\eta = 0.7$, $R = 40$ s, the free flow pace is taken to be a random variable with a Gamma distribution, a mean of 1/8 s/m and a standard deviation of 1/30 s/m. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

remainder of the vehicles experience delay because of the presence of a traffic light. This delay is uniformly distributed on $[0,R]$. The fraction of stopping vehicles is derived from traffic theory as $\eta = \frac{R}{C} + (1 - \frac{R}{C})\frac{\xi}{\xi_s}$. The pdf of delays on an arterial link is thus given by

$$h^u(\delta_{0,L}) = (1-\eta)\mathbf{Dir}_{\{0\}}(\delta_{0,L}) + \eta\frac{1}{R}\mathbf{1}_{[0,R]}(\delta_{0,L}),$$

where $\mathbf{1}_{[0,R]}(\cdot)$ is the indicator function of the interval $[0,R]$ and $\mathbf{Dir}_{\{0\}}(\cdot)$ is the Dirac Delta distribution. The probability distribution of delays, and, because of the linearity of the convolution, the probability distribution of travel times is a mixture distribution in which each component represents a class of vehicles (stopping or not stopping). Each component of the mixture is computed as the convolution between the probability distribution of delay and the probability distribution of free flow travel times. Fig. 6 shows the probability distribution of travel times for an undersaturated arterial link.

In both the undersaturated and the congested regime, the analytical expression and the shape of the probability distribution of travel times depends on the locations x_1 and x_2 and can be expressed as a finite mixture distribution, illustrating the different delays experienced by the vehicles.

In the following, quantities are indexed by i (and sometimes t) to indicate that they refer to link i (and to time interval t). For a link i and a time interval t, the resulting travel time probability distribution between any two points on the link are parameterized by the network parameters $(R^i, C^i, \xi_s^i, p_f^i, \theta_p^i)$ and the points on the link (x_1 and x_2). The probability distribution of travel time y_{x_1,x_2} between x_1 and x_2 is conditioned on the traffic state $\xi^{i,t}$ and denoted $g^i(y_{x_1,x_2}|\xi^{i,t})$. The dependency on the network parameters R^i, C^i, ξ_s^i and θ_p^i is implicit and only reminded by the indexing of g by i.

3.2. Modeling the spatio-temporal dependencies: transition probabilities

The spatio-temporal dependencies between the links of the network are modeled with a transition probability on the state of each link i at time $t+1$ given the state of the neighbors at time t. For link i, this transition probability is

parameterized by the turn probabilities and intensities of the Poisson processes for the arrival vehicles, as presented in the following sections.

In this article, we assume that all the lanes of a link follow the same dynamics. In particular, each lane of link i is in state $\xi^{i,t}$ during time interval t. The red time R^i, the cycle time C^i, the saturation number of vehicles ξ_s^i and the parameters of the free flow pace θ_p^i are the same for each lane of the link. We denote by κ^i the number of lanes of link i. Note that the derivations can readily be extended if we consider different queue lengths for each lane, corresponding, for example, to dedicated lanes for turn movements. Similarly, a straightforward generalization of the model would consider that the link and intersection parameters (excepted the cycle time) are lane-dependent and would more accurately model the different phases of the signal and turning movements.

3.2.1. Number of vehicles leaving a link in a cycle

The derivations in this section are valid for any link i of the network at any time interval t.

In a congested regime, there are more vehicles on the link than can exit during a cycle. The number of vehicles that exit the link during a cycle within time interval t is $n_{\text{out}}^{i,t} = \kappa^i \xi_s^i$.

In an undersaturated regime, we define the *red phase* during which the light is red and no vehicle goes through the intersection (duration R^i), the *clearing phase* (introduced in Section 2.2, with duration $\tau^{i,t}$) and the *free-flowing phase* during which the vehicles go through the intersection without stopping. Note that the duration of the clearing time (and of the free flowing phase) depends on the time interval t since it depends on the state of the link $\xi^{i,t}$.

The duration of the free flowing phase is the remaining duration of the cycle after the red phase and the clearing phase, with duration $C^i - (R^i + \tau^{i,t})$. The number of vehicles exiting the link during a cycle is the sum of the vehicles exiting the link after stopping in the triangular queue ($\kappa^i \xi^{i,t}$) and the vehicles exiting during the free-flowing phase. For an arrival density $\rho_a^{i,t}$, we have

$$n_{\text{out}}^{i,t} = \kappa^i \left(\xi^{i,t} + \rho_a^{i,t} v_f^i (C^i - (R^i + \tau^{i,t})) \right). \tag{4}$$

In each lane, $\xi^{i,t}$ vehicles stop in the triangular queue. They exit during the clearing time ($\tau^{i,t}$) at the maximum flow ($q_{\max}^i = v_f^i \rho_c^i$), so we have

$$\xi^{i,t} = v_f^i \rho_c^i \tau^{i,t}. \tag{5}$$

Using Eq. (2), we derive the ratio between the arrival and the critical density for each lane of the link

$$\frac{\rho_a^{i,t}}{\rho_c^i} = \frac{\tau^{i,t}}{\tau^{i,t} + R^i}. \tag{6}$$

Combining Eqs. (5) and (6) in Eq. (4), the number of vehicles that leave a link in a cycle C^i is

$$\begin{aligned} n_{\text{out}}^{i,t} &= \kappa^i \left(\xi^{i,t} + \rho_c^i v_f^i \tfrac{\tau^{i,t}}{\tau^{i,t}+R^i} (C^i - (R^i + \tau^{i,t})) \right) && \text{using Eq. (6)}, \\ n_{\text{out}}^{i,t} &= \kappa^i \xi^{i,t} \tfrac{C^i}{\tau^{i,t}+R^i} && \text{using Eq. (5)}. \end{aligned} \tag{7}$$

The number of vehicles leaving the link during time interval t (of duration Δ_t) is derived from (7) as $N_{\text{out}}^{i,t} = n_{\text{out}}^{i,t} \frac{\Delta_t}{C^i}$. Incorporating the equation of $\tau^{i,t}$ from (3), we have for both regimes,

$$N_{\text{out}}^{i,t} = \kappa^i \min\left(\xi^{i,t}, \xi_s^i\right) \frac{\Delta_t}{R^i + (C^i - R^i) \frac{\min\left(\xi^{i,t}, \xi_s^i\right)}{\xi_s^i}}. \tag{8}$$

3.2.2. Dynamic evolution of the state

Each vehicle arriving from link i at an intersection k is assigned to an outgoing link $j \in L_{\text{out}}^k \cup k_{\text{out}}$ with probability $v^{i,j}$ (possibly leaving the network through the sink k_{out}). Each vehicle is assigned independently from the other ones. According to this model, the random vector $\left(N_{\text{in}}^{i,j,t}\right)_{j \in L_{\text{out}}^k \cup k_{\text{out}}}$ of vehicles assigned to the different outgoing links of the intersection has a multinomial distribution with parameters $N_{\text{out}}^{i,t}$ and $(v^{i,j})_{j \in L_{\text{out}}^k \cup k_{\text{out}}}$ such that,

$$\mathcal{P}\left(N_{\text{in}}^{i,j,t} : j \in L_{\text{out}}^k \cup k_{\text{out}}\right) = \begin{cases} \dfrac{N_{\text{out}}^{i,t}!}{\prod_{j \in L_{\text{out}}^k \cup k_{\text{out}}} N_{\text{in}}^{i,j,t}!} \prod_{j \in L_{\text{out}}^k \cup k_{\text{out}}} (v^{i,j})^{N_{\text{in}}^{i,j,t}} & \text{if} \quad \sum_{j \in L_{\text{out}}^k} N_{\text{in}}^{i,j,t} = N_{\text{out}}^{i,t}, \\ 0 & \text{otherwise}. \end{cases}$$

If the intersection has a source k_{in}, we assume that vehicles arrive to the outgoing links j of the intersection according to a Poisson process of intensity λ^j. The probability that $N_{\text{in}}^{k_{\text{in}},j,t}$ vehicles arrive to link j from the source during Δ_t is

$$\mathcal{P}\left(N_{\text{in}}^{k_{\text{in}}j,t}\right) = \frac{\left(\Delta_t \lambda^j\right)^{N_{\text{in}}^{k_{\text{in}}j,t}} e^{-\Delta_t \lambda^j}}{N_{\text{in}}^{k_{\text{in}}j,t}!}.$$

Given the number of vehicles arriving to link j from the incoming links of intersection k $\left(\left(N_{\text{in}}^{i,j,t}\right)_{i\in L_{\text{in}}^k \cup k_{\text{in}}}\right)$, and the state of link j at time t ($\xi^{j,t}$), we can compute the state $\xi^{j,t+1}$ of link j at time $t+1$: (i) we compute the balance of vehicles between the incoming and the outgoing vehicles at time t and (ii) we update the state of the link for time $t+1$ accordingly. The details of this transition are as follows:

- *Balance of vehicles on link j at time interval t:* During a time interval Δ_t, there are $N_{\text{out}}^{j,t}$ vehicles exiting link j and $N_{\text{in}}^{j,t}$ vehicles arriving in link j, which corresponds to a balance of $\Delta N^{j,t} = N_{\text{in}}^{j,t} - N_{\text{out}}^{j,t}$ additional vehicles. Note that a negative number represents a decrease in the number of vehicles on the link. We assume that if link j has several lanes, the increase or decrease in the number of vehicles is the same for all lanes. This can be adapted for a model with lane-specific link and intersection parameters.
- *Update of the state at time interval t + 1:*
  - *Undersaturated regime with arrival flow inferior to the capacity*: At time t, link j is undersaturated ($\xi^{j,t} \leqslant \xi_s^j$) and the number of vehicles arriving per cycle is less than the maximum throughput per cycle ($n_{\text{in}}^{j,t} \leqslant \kappa^j \xi_s^j$). These two conditions imply undersaturated conditions for link j during time intervals t and $t+1$. The queue fully dissipates by the end of each light cycle and the outflow at time $t+1$ equals the inflow at time t ($N_{\text{out}}^{j,t+1} = N_{\text{in}}^{j,t}$). We can invert Eq. (8) to have the expression of the state at $t+1$. Note that in this case, Eq. (8) is simplified since the number of vehicles in the queue is less than the saturation number of vehicles $\left(\min\left(\xi^{j,t+1}, \xi_s^j\right) = \xi^{j,t+1}\right)$.

$$\xi^{j,t+1} = \frac{N_{\text{out}}^{j,t+1}\, R^j \xi_s^j}{\kappa^j \Delta_t \xi_s^j - (C^j - R^j) N_{\text{out}}^{j,t+1}} = \frac{N_{\text{in}}^{j,t}\, R^j \xi_s^j}{\kappa^j \Delta_t \xi_s^j - (C^j - R^j) N_{\text{in}}^{j,t}}$$

  - *Other transitions:* If the regime was congested or if the number of vehicles arriving on the link per cycle is greater than the maximum throughput of the link, there is a constant increase (or decrease) in the number of vehicles on the link through the time period t. The number of vehicles stopping in the queue for time interval $t+1$ is given by the balance of vehicles:

$$\xi^{j,t+1} = \xi^{j,t} + \frac{\Delta N^{j,t}}{\kappa^j}.$$

3.3. Statistical modeling framework

Arterial traffic conditions vary dynamically over space and time. We represent the conditional independencies assumptions of Section 2.1 using a probabilistic graphical model known as a *Dynamic Bayesian Network* (DBN). In this article, the DBN model the stochastic dynamics of the traffic states (number of vehicles stopping in a cycle) of each link in the arterial network. Since we do not observe the state directly, these variables are considered *hidden*. On each link, the travel time

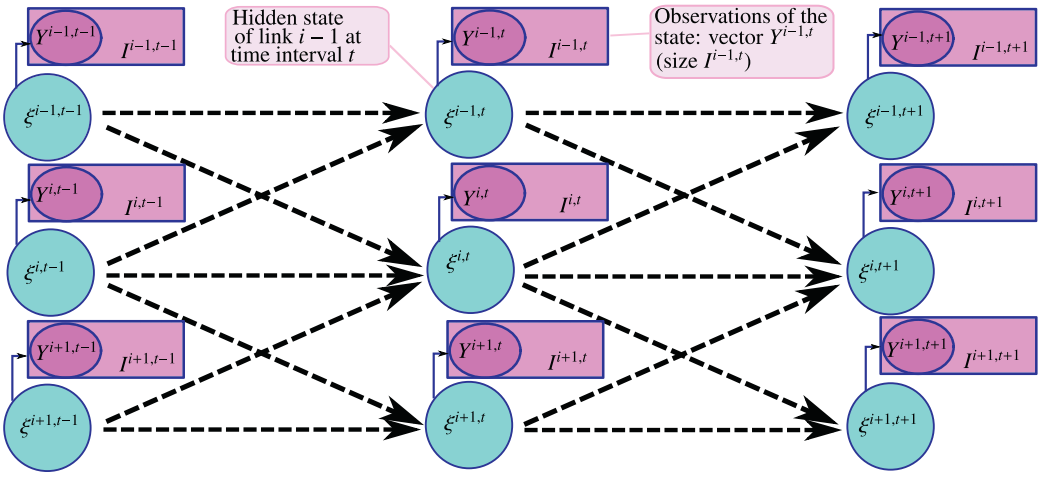

Fig. 7. Spatio-temporal model of arterial traffic evolution represented as a Dynamic Bayesian Network. The circular nodes represent the (hidden) discrete states $\xi^{i,t}$ of traffic for each link i at each time interval t. The rectangular nodes represent the $I^{i,t}$ travel time observations (denoted $Y^{i,t}$) of each link i at each time interval t. The doted arrows represent the stochastic spatio-temporal dependencies between the states. The plain line arrows represent the dependency of the travel time distributions on the *hidden* traffic state.

distribution is conditioned on the (hidden) state of the link. The travel time of the probe vehicles traveling through the arterial network provide sparse observations of the state variables. Fig. 7 illustrates the model representation of link states and probe vehicle observations. Each circular node in the graph represents the state of a link in the road network. The forward arrows indicate the local spatial dependency of links from one time period to the next. Each square node in the graph represent probe vehicle observations on the link to which it is attached. The number of observations for a time interval t and a link i is denoted $I^{i,t}$ From the observable sequence of outputs (path travel time observations) we want to infer the most likely distribution of the state variables (queue lengths) as well as their dynamical evolution. For more background on DBNs, please refer to Murphy (2002).

The observations are successive GPS measurements of vehicle trajectories (approximately one per minute). The issues of filtering the noise of the GPS to estimate the most likely location of the vehicle when the measurement was generated and inferring the path taken by the vehicle are not addressed in this article. There are multiple approaches to solving this problem including using statistical filtering (Hunter et al., 20111; Thiagarajan et al., 2009). In the remainder of this article, we assume that we are given the most likely measurement locations on the road network as well as the most likely path of the vehicle. To completely specify the DBN model, we have to estimate:

- The probability of the state ξ at the start of the experiment. For each link, it is denoted $\pi^i(\xi)$. It represents the probability that link i has ξ stopping vehicles at the initial time.
- The transition probability distribution functions (Section 3.2 and assumption 5c), parameterized by the turn probabilities $v^{i,j}$ and intensities of the Poisson processes λ^j.
- The distribution of travel time g^i on each link i of the network, parameterized by the link parameters $(R^i, C^i, \xi^i_s, \theta^i_p)$ and conditioned on the state of the link.

The traffic state is constant during each time intervals of duration Δ_t, typically chosen between 5 and 15 min (time discretization assumption), and the link and intersection parameters may be assumed constant for several of these time intervals representing specific times of day (*e.g.* morning rush hour, mid-day, afternoon rush hour, evening, night). The present article focuses on the estimation of the parameters for a given traffic period and the dynamic evolution of the state within this traffic period. Future work will address the automatic detection of changes in the network parameters but is not addressed in this article.

We also assume that, given the state of a subset of links, the travel time distributions on these links are independent random variables. In general, travel time distributions across links are not independent (due to light synchronization, platoons, and other factors), although it is a reasonable approximation in many cases. Future work will specifically address the challenge of using correlated distributions, which have the potential to capture more complex dynamics in the arterial road network.

4. Maximum likelihood estimation of the parameters

There is a complex pattern of dependencies among the travel times sent by the probe vehicles, which we want to learn off-line, from historical data, to perform estimation and prediction in real-time. Modeling the dependency between the observations directly is a difficult task because it does not exploit the underlying structure of the dynamical system provided by the conditional independence assumptions. To simplify the learning and estimation task, we introduce the variables $(\xi^{i,t})$, representing the number of vehicles in the queue of each link at each time interval, and model their stochastic dynamic evolution. Since these variables are not observed directly, they are called *latent* or *hidden* variables. The probe vehicle travel times are noisy, sparse observations of these variables. We introduce an *Expectation Maximization* algorithm (EM algorithm) to learn the dependencies among the observations while exploiting the structure of the stochastic dynamic evolution. This choice is supported by the following two realizations: (1) given the parameters of the model and the path observations, we can estimate the most likely state of each link at each time interval and (2) given the state of each link at each time interval, we can compute the parameters of the model (turn probabilities, intensities of the Poisson processes and parameters of the network) which maximize the likelihood of the observations. The EM algorithm iteratively leverages these two realizations and is guaranteed to converge to a local optima of the likelihood function. More detailed information on the EM algorithm can be found in the literature (Dempster et al., 1977) and a short introduction is given in Section 4.2. One challenge of our graphical model approach is that we do not observe link travel times directly, since the probe observations we receive can span several links of the network between two consecutive measurements. This difficulty is addressed by computing the most likely link travel times that make up the path of the probe vehicle (*travel time allocation*), which is described in Section 4.1. We introduce the EM algorithm (Section 4.2) and detail its two iterative steps: Expectation step (E step) in Section 4.3 and Maximization step (M step) in Section 4.4 in the case of traffic estimation.

4.1. Travel time allocation

An observation consists of a travel time over a path consisting of multiple (potentially partial) links. In order to use the graphical model presented in Section 3.3, the total travel time must be decomposed into a travel time for each (partial) link

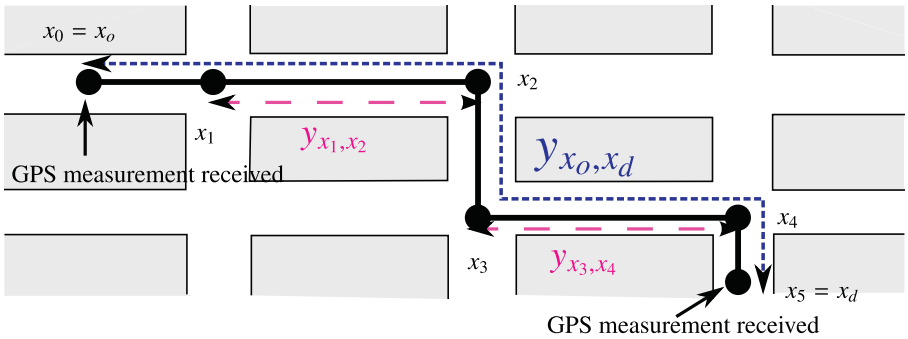

Fig. 8. Illustration of the travel time allocation: decomposition of the path travel time into (partial) link travel times. Along its trajectory, the vehicle sends location measurements successively at x_o and x_d for a travel time y_{x_o,x_d}. This path spans five links (numbered i_1 to i_5). The path only spans a fraction of the first and last links (partial links). We decompose the total travel time y_{x_o,x_d} into five (partial) link travel time $(y_{x_m,x_{m+1}})_{m=0..4}$. These (partial) link travel times correspond to the most likely time spent on each (partial) link given the parameters of the network and the state of traffic. These travel times sum to the total travel time y_{x_o,x_d}.

on the path. The formalization of the intuitive idea that vehicles are more likely to experience delays close to intersections (Hellinga et al., 2008; Hofleitner and Bayen, 2011) shows significant improvements compared to an allocation proportional to the free flow travel time, even though some experiments do not agree with this statement (Zheng and Van Zuylen, 2009). The modeling of vehicle dynamics on an arterial network is promising in order to accurately solve the travel time allocation problem and is recommended by traffic data collection guidelines (Van Zuylen et al., 2010). Given the model of travel time distributions used in this article, we perform *optimal* travel time allocation by maximizing the log-likelihood of the (partial) link travel times for each observation given the model parameters (Hofleitner and Bayen, 2011).

For a vehicle traveling from an origin (first measurement point) x_o to a destination (second measurement point) x_d through M intersections, we decompose the travel time y_{x_o,x_d} as the sum of travel times on each of the links (Fig. 8).

$$y_{x_o,x_d} = \sum_{m=0}^{M} y_{x_m,x_{m+1}}, \tag{9}$$

For $m \in \{1 \dots M\}$, the point x_m represents the most upstream location on the mth link on the paths, $x_0 = x_o$ and $x_{M+1} = x_d$. For $m \in \{0 \dots M\}$, we note i_m the mth link of the path between x_m and x_{m+1}. The function g^{i_m} is the probability distribution of travel times on link i_m. It is parameterized by the link parameters $(R^{i_m}, C^{i_m}, \xi_s^{i_m}, \theta_p^{i_m})$ and conditioned on the state of the link at time interval t, $\xi^{i_m,t}$. The formulation of the travel time allocation problem at time interval t reads:

$$\begin{aligned} &\underset{(y_{x_m,x_{m+1}})_{m=0\dots M}}{\textbf{maximize}} : \quad \sum_{m=0}^{M} \ln(g^{i_m,t}(y_{x_m,x_{m+1}})) \\ &\textbf{s.t.} : \quad y_{x_o,x_d} = \sum_{m=0}^{M} y_{x_m,x_{m+1}}, \end{aligned} \tag{10}$$

The optimization problem in Eq. (10) is solved by computing the solution of a small number of small scale convex optimization programs by using the structure of the travel time distributions, as shown in Hofleitner and Bayen (2011). We denote by $\mathbf{y}^{i,t}$ the set of travel times allocated to link i at time interval t.

4.2. Introduction on EM algorithm

The EM algorithm allows us to exploit the underlying structure of the dynamical model, even though the latent variables $(\xi^{i,t})$ are not observed. It is an iterative algorithm consisting in two steps:

- *The expectation step (E step)* computes the joint probability distribution of the latent variables $\xi^{i,t}$ (number of vehicles in the queue for each link i and each time interval t) given the observed variables $\mathbf{y}^{i,t}$ (allocated travel times for each link i and each time interval t) and the current values of the parameters (signal parameters, turn ratios, driving behavior, saturation number of vehicles). In the Bayesian approach to dynamic state estimation, this computation is known as a *smoothing* step. In practice, the smoothing is replaced by filtering: estimation of the probability distribution of the state at time interval t based on all available measurements up to and including time interval t. Such a filtering step consists of essentially two stages: prediction and update. The prediction uses the transition probabilities to predict the state probability distribution from one time interval to the next. The update operation uses the latest available measurements to modify the state probability distribution using Bayes theorem.

- *The maximization step (M step)* optimizes the parameters (signal parameters, turn ratios, driving behavior, saturation number of vehicles) based on the estimation of the joint probability distribution of the latent variables. This step has the same complexity as if the latent variables were observed.

As illustrated Fig. 7, a dynamic Bayesian network is a directed graphical model, in which each random variable is represented by a node of the graph. Each generic random variable x_i has a set of parents, denoted x_{π_i} such that the joint probability $p(x_1, \ldots, x_n)$ of $x_1, \ldots, x_n$ can be factored as

$$p(x_1, \ldots, x_n) = \prod_{i=1}^{n} p(x_i | x_{\pi_i}),$$

where $p(x_i|x_{\pi_i})$ is the probability of x_i given that its parents (in the directed graph) have the realization x_{π_i}. In this article, the random variables represent the number of vehicles $\xi^{i,t}$ and the travel time observations $\mathbf{y}^{i,t}$ on each link of the network at each time interval. The conditional independence assumptions and the associated directed graphical model representation provide a compact, factored, representation of the joint distribution of these random variables:

$$\mathcal{P}(\xi, \mathbf{y}) = \underbrace{\left(\prod_{t=0}^{T-1} \prod_{i \in \mathcal{I}} \mathcal{P}\left(N_{\text{in}}^{i,j,t} : j \in L_{\text{out}}^{k_i} | \xi^{i,t}\right)\right)}_{\substack{\text{Probability of the assignment of the vehicles from link } i \text{ to the outgoing links of the intersection, for each} \\ \text{link and each time interval excepted the last one which corresponds to the end of the experiment.}}} \times \underbrace{\left(\prod_{t=0}^{T} \prod_{i \in \mathcal{I}} \mathcal{P}(\mathbf{y}^{i,t} | \xi^{i,t})\right)}_{\substack{\text{Probability of the travel time observations } \mathbf{y}^{i,t} \text{ conditioned on the state of} \\ \text{the link } \xi^{i,t}, \text{ for each link } i \text{ and eachtime interval } t.}}$$

$$\times \underbrace{\left(\prod_{i \in \mathcal{I}} \pi_i(\xi^{i,0})\right)}_{\substack{\text{Probability that link } i \text{ is in state } \xi^{i,0} \text{ at the initial} \\ \text{time interval, for each link } i.}},$$

Note that given the state of the links at a time interval, the number of vehicles from link i assigned to the outgoing links j and the number of vehicles entering or exiting the network through the sources and sinks determine the state evolution for all the links of the network. For convenience, we use these probabilities in the expression of $\mathcal{P}(\xi, \mathbf{y})$ instead of referring directly to the probability of the number of vehicles in the queue of link i at time interval $t+1$ given the number of vehicles in the queue of the neighboring links.

If the hidden variables $\xi^{i,t}$ were observed, the likelihood optimization would amount to maximizing $\mathcal{P}(\boldsymbol{\xi}, \mathbf{y})$ with respect to the link and intersection parameters. More commonly, we consider the logarithm of $\mathcal{P}(\boldsymbol{\xi}, \mathbf{y})$, referred to as the *complete log-likelihood* because it corresponds to the log-probability of the complete set of random variables for a given value of the parameters. Given that the variables $\xi^{i,t}$ are in fact not observed, the complete log-likelihood is a random quantity, and cannot be maximized directly. Given a distribution, denoted $q(\xi|\mathbf{y})$, we define a deterministic function of θ, denoted $\langle l_c(\mathbf{y}, \xi) \rangle_q$ and called *expected complete log-likelihood*: It corresponds to the average of the complete log-likelihood, over the realizations of ξ, when $q(\xi|\mathbf{y})$ is chosen as the averaging distribution:

$$\langle l_c(\xi, \mathbf{y}) \rangle_q = \sum_{\xi} q(\xi|\mathbf{y}) \ln(\mathcal{P}(\xi, \mathbf{y}))$$

Using Jensen's inequality, we can show that the log-likelihood can be maximized by iteratively (1) choosing the proposal distribution $q(\xi|\mathbf{y})$ as the joint distribution of the state variables computed by the E step and (2) maximizing on the parameters of the observations $\left(R^i, C^i, \xi_s^i, \theta_p^i, i \in \mathcal{I}\right)$ and of the dynamics (v^{ij},λ^i, for $i \in \mathcal{I}$ and for j outgoing link of i).

4.3. E step: particle filtering

The E step performs filtering given the current values of the parameters and the travel time observations collected from historical data. The dimension of the state space (number of possible configurations of the number of vehicles in the queue in each of the link) grows exponentially with the number of links in the network, making an explicit representation of the probability distributions intractable (Cooper, 1990). To maintain a compact approximation of the state probability distribution, we use particle filtering (also known as bootstrap filtering or the condensation algorithm) (Russell and Norvig, 1995; Arulampalam et al., 2002). Other approximations algorithm include variational methods (Jordan et al., 1999) and belief state simplification (Boyen and Koller, 1998). Particle filtering is a technique for implementing a recursive Bayesian filter algorithm by Monte Carlo Simulations. The idea is to represent the distribution by a set of random samples with associated weights (importance weights). As the number of samples becomes very large, this Monte Carlo approximation tends to the exact optimal Bayesian estimate.

We simulate V particles, where each particle v represents an instantiation of the *time evolution of the traffic state of the network*, *i.e.* a possible succession of traffic states for each link of the network and each time interval. A particle v at time t is represented by a vector of the states of each link and each time interval up to t (denoted $\left(\xi_v^{i,t'}\right)_{i \in \mathcal{I}, t' \in \{0,\ldots,t\}}$) and a weight (or importance weight) ω_v^t, proportional to the probability of having this instantiation of the state evolution of the network given the available data up to time t. We simulate a high number of particles that evolve through the graphical model and explore the possible state space.

4.3.1. Sufficient statistics to compute the expected complete log-likelihood

At time t, the spatio-temporal instantiations $\Xi_v^t = \left(\xi_v^{i,t'}\right)_{i\in\mathcal{I},t'\in\{0,\dots,t\}}$ of the particles and their associated importance weight ω_v^t form an approximation of the joint probability distribution of the state of the links. We denote by $\mathbf{y}^{i,t}$, the set of travel time observations received on link i during time interval t. Given the travel time observations $(\mathbf{y}^{i,t'})_{i\in\mathcal{I},t\in\{0\dots t'\}}$, the probability of observing a state $\Xi^t = (\xi^{i,t'})_{i\in\mathcal{I},t'\in\{0\dots t\}}$ on the network throughout its time evolution is approximated as follows:

$$\mathcal{P}(\Xi^t|\mathbf{y}^{i,t'}, R^i, C^i, \xi_s^i, \theta_p^i : i\in\mathcal{I}, t'\in\{0\dots t\}) \approx \sum_{v=1}^{V}\omega_v \mathbf{1}_{\Xi^t}(\Xi_v^t).$$

where $\mathbf{1}_{\Xi^t}(\Xi_v^t)$ is equal to 1 if the particle has the state instantiation Ξ^t and to zero otherwise. To derive the expected complete log-likelihood, we introduce $a^{i,t}(\xi^{i,t})$ and $b^{i,t}(\mathbf{N}^{i,t})$, $c^{j,t}\left(N_{\text{in}}^{k_{\text{in}}j,t}\right)$ and $d^i(\xi^{i,0})$, referred to as sufficient statistics and defined as follows.

- The probability that link i is in state $\xi^{i,t}$ at time t, conditioned on the observations received up to time interval t is approximated using the particles and denoted $a^{i,t}(\xi^{i,t})$. It is computed by summing the weights of all the particles that represent a state instantiation with link i in state $\xi^{i,t}$:

$$a^{i,t}(\xi^{i,t}) = \sum_{v=1}^{V}\omega_v^t \mathbf{1}_{\xi^{i,t}}\left(\xi_v^{i,t}\right), \quad \forall t\in\{0,\dots,T\},\ \forall i\in\mathcal{I}. \tag{11}$$

- For an incoming link i and an outgoing link j of intersection k, we note $\left(N_{\text{in}}^{i,j,t}\right)_v$ the number of vehicles going from link i to link j during time interval t for the particle v. We approximate the probability that $\mathbf{N}^{i,t} = (N_{\text{in}}{}^{i,j,t}\boldsymbol{j}\in L_{\text{out}}\cup k_{\text{out}})$ vehicles from link i are assigned to the outgoing links L_{out}^k and the sink k_{out} using the particles and denote it by $b^{i,t}(\mathbf{N}^{i,t})$. It is computed by summing the weights of all the particles that represent an instantiation of the dynamics in which the assignments of the vehicles from link i to the outgoing links (and the sink) is $\mathbf{N}^{i,t}$:

$$b^{i,t}(\mathbf{N}^{i,t}) = \sum_{v=1}^{V}\omega_v^t\ \mathbf{1}_{\mathbf{N}^{i,t}}\left(\left(N_{\text{in}}^{i,j,t}\right)_v, j\in L_{\text{out}}^k\cup k_{\text{out}}\right), \qquad \forall t\in\{0,\dots,T-1\},\ \forall i\in\mathcal{I}. \tag{12}$$

- For an intersection k with a source, $\left(N_{\text{in}}^{k_{\text{in}}j,t}\right)_v$ is the number of vehicles from the source assigned to each outgoing link j of the intersection. We approximate the probability that $N_{\text{in}}^{k_{\text{in}}j,t}$ vehicles from the source are assigned to link j as $c^{j,t}\left(N_{\text{in}}^{k_{\text{in}}j,t}\right)$. It is computed by summing the weights of the particles for which $N_{\text{in}}^{k_{\text{in}}j,t}$ vehicles originating from the source were assigned to link j:

$$c^{j,t}\left(N_{\text{in}}^{k_{\text{in}}j,t}\right) = \sum_{v=1}^{V}\omega_v^t\ \mathbf{1}_{N_{\text{in}}^{k_{\text{in}}j,t}}\left(\left(N_{\text{in}}^{k_{\text{in}}j,t}\right)_v\right), \qquad \forall t\in\{0,\dots,T-1\},\ \forall j\in\mathcal{I}. \tag{13}$$

- We also define $d^i(\xi^{i,0})$ the probability of the state of link i at the initial time, and approximate it using the particles as

$$d^i(\xi^{i,0}) = \sum_{v=1}^{V}\omega_v^0\ \mathbf{1}_{\xi^{i,0}}(\xi_v^{i,0}). \tag{14}$$

4.3.2. Filtering using a particle filter

The filtering step consists in successive prediction and update steps which lead to the computation of $\xi_v^{i,t}$ and ω_v^t for all the particles v, all the links i and all the time intervals t. The prediction and update steps are performed as follows:

- *Update of the state posterior probability distribution at time interval t*. We compute the *posterior* state distribution using the measurements available at time interval t. The weight ω_v^t of each particle is multiplied by the probability of each travel time measurement received at time interval t given the state $\xi_v^{i,t}$ of the particle. The weights of the particles are normalized so that they sum to one.
- *Prediction of the state at time interval t + 1*. We predict the state distribution using the parameters of the turn movements and of the Poisson processes of the sources. For each incoming link i and each particle v, we compute the number of vehicles leaving link i. Using sampling, these vehicles are randomly assigned to the outgoing links of the intersection (including the sink) according to a multinomial distribution parameterized by the turn probabilities. Similarly, a random number of vehicles (coming from the source of the intersection) is assigned to the outgoing links according to the corresponding Poisson process. This allows for the computation of $\left(N_{\text{in}}^{i,j,t}\right)_v$ and for the simulation of the state of the particle at time interval $t + 1$ according to the dynamic evolution described in Section 3.2.2. This algorithm is known a *Sequential Importance Sampling* (SIS) particle filter.

- *Improvement to prevent degeneracy: the Sequential Importance Resampling (SIR) algorithm.* A common issue of the SIS particle filter is the degeneracy phenomenon, where after a few iterations, all but one particle have negligible weights. It implies that a large computational effort is devoted to updating particles whose contribution to the posterior distribution is almost zero. To reduce the effects of degeneracy, we use *resampling* in the SIS algorithm after computing the importance weights for time interval t. The modified algorithm is known as Sequential Importance Resampling (SIR) or Sampling Importance Resampling. The idea of resampling is to eliminate particles that have small weights at time interval t. To resample the particles, V particles are successively chosen randomly (with replacement) from the original set of particles. Particle v is chosen with probability ω_v^t (the weights sum to 1). Each resampled particle has a weight equal to $1/V$. This set of re-sampled particles is used to perform the prediction step of the state probability distribution at time interval $t+1$.

Algorithm 1. Maximum likelihood estimation of the parameters of the model with an EM algorithm.

Initialize the parameters ($R^i, C^i, \xi_s^i \theta_p^i, v^{i,j}$ and λ^j) and the initial state probabilities $\pi_i(\xi)$
while The algorithm has not converged **do**
E Step [Computation of $a^{i,t}(\xi^{i,t}), b^{i,t}(\mathbf{N}^{i,t}), c^{j,t}\left(N_{\text{in}}^{k_{\text{in}},j,t}\right)$ and $d^i(\xi^{i,0})$] (Section 4.3)
Initialize the E Step: Simulate samples representing the state of the network at the initial time given the initial state probabilities $\pi_i(\xi)$. Each sample has initial weight $\omega_v = 1/V$
for Time interval $t = 0{:}T$ **do**
Allocate the travel times by solving (10) for each probe vehicle path
Update the weight of the particles according to the observations $\mathbf{y}^{i,t}$: $\omega_v = \omega_v \prod_{y_{x_1,x_2} \in \mathbf{y}^{i,t}} g^i\left(y_{x_1,x_2} | \xi_v^{i,t}\right)$
Normalize the weights of the particles: Compute the sum Ω of the weights of the particles and normalize the weight of each particle, $\omega_v = \omega_v / \Omega$
Compute $a^{i,t}(\xi^{i,t}), b^{i,t}(\mathbf{N}^{i,t}), c^{j,t}\left(N_{\text{in}}^{k_{\text{in}},j,t}\right)$ and $d^i(\xi^{i,0})$ using Eqs. (11)–(14)
Re-sample the particles (Arulampalam et al., 2002)
For each link i, randomly assign the vehicles leaving link i to the outgoing links and the vehicles coming from the sources of the network according to the turn probabilities and intensities of the Poisson processes
Update the state of the particles according to the number of vehicles that left and arrived on the link during time interval t. Each particle now represents an instantiation of the state of the network at $t+1$
end for
M Step [Maximization of the expected complete log-likelihood.] (Section 4.4)
Update the initial state probabilities $\pi^i(\xi)$ (17), the turn probabilities $v^{i,j}$ (15), the vehicle creation rates λ^i (16) and the link parameters $\left(C^i, R^i, \xi_s^i, \theta_p^i\right)$ (18)
end while

4.4. M step: update of the parameters

For each link i, the travel time distribution g^i is conditioned on the state of the link and parameterized by the red time R^i, the cycle time C^i, the number of vehicles in the queue at saturation ξ_s^i and the parameters of the driving behavior θ_p^i. To fully characterize the model, we also need to learn the parameters of the dynamics *i.e.* estimate the turn probabilities $v^{i,j}$ and the intensities of the Poisson processes λ^j. The M step uses the sufficient statistics $a^{i,t}(\xi^{i,t}), b^{i,t}(\mathbf{N}^{i,t}), c^{j,t}\left(N_{\text{in}}^{k_{\text{in}},j,t}\right)$ and $d^i(\xi^{i,0})$ to update the value of these parameters by maximizing the expected complete log-likelihood, with respect to these parameters. The factored expression of the complete log-likelihood implies a similar structure for the complete log-likelihood:

$$\langle l(\xi, \mathbf{y}) \rangle = \overbrace{\left(\sum_{t=0}^{T-1} \sum_{i \in \mathcal{I}} \sum_{\mathbf{N}^{i,t}} b^{i,t}(\mathbf{N}^{i,t}) \ln(\mathcal{P}(\mathbf{N}^{i,t}))\right)}^{\substack{\text{Assignments of vehicles exiting link } i \text{ to the outgoing links of the intersection} \\ \text{(and the sink) for each link } i \text{ and each time interval } t \text{ (excepted the last one).}}} + \overbrace{\left(\sum_{t=0}^{T-1} \sum_{j \in \mathcal{I}} \sum_{N_{\text{in}}^{k_{\text{in}},j,t}} c^{j,t}\left(N_{\text{in}}^{k_{\text{in}},j,t}\right) \ln\left(\mathcal{P}\left(N_{\text{in}}^{k_{\text{in}},j,t}\right)\right)\right)}^{\substack{\text{Arrival of vehicles in link } j \text{ from the source of the intersection,} \\ \text{for each link } j \text{ and each time interval } t \text{(excepted the last one).}}}$$
$$+ \underbrace{\left(\sum_{t=0}^{T} \sum_{i \in \mathcal{I}} \sum_{\xi^{i,t}=0}^{\xi_{\max}^i} a^{i,t}(\xi^{i,t}) \left(\sum_{y_{x_1,x_2} \in \mathbf{y}^{i,t}} \ln\left(g^i(y_{x_1,x_2} | \xi^{i,t})\right)\right)\right)}_{\substack{\text{Travel time measurements, for each travel time } y_{x_1,x_2} \\ \text{received on each link } i \text{ at each time interval.}}} + \underbrace{\left(\sum_{i \in \mathcal{I}} \sum_{\xi^{i,0}=0}^{\xi_{\max}^i} d^i(\xi^{i,0}) \ln(\pi^i(\xi^{i,0}))\right)}_{\text{Initial state of each link } i.}.$$

where $\mathcal{P}(\mathbf{N}^{i,t})$ represents the probability (multinomial distribution) of the assignment $\mathbf{N}^{i,t}$ of the vehicles leaving link i to the outgoing links of the intersection (including the sink) and $\mathcal{P}\left(N_{\text{in}}^{k_{\text{in}},j,t}\right)$ is the probability (Poisson distribution) of the arrival of $N_{\text{in}}^{k_{\text{in}},j,t}$ vehicles in link j from the source of the intersection. The factored structure of the complete log-likelihood, and thus of the expected complete log-likelihood allows the learning of the parameters to be performed independently for the turn

probabilities, the intensities of the Poisson processes, the initial state probabilities and for each set of link parameters. We use the values of $a^{i,t}$,$b^{i,t}$,$c^{j,t}$ and d^i computed by the E step (Eqs. (11)–(14)) to update the link and intersection parameters. We detail the derivations of the update of the parameters in the following (Eqs. (15)–(18)).

- The update of the turn probabilities from the incoming link i of intersection k is the solution of the following optimization program:

$$\begin{aligned}
&\underset{v^{i,j}}{\textbf{maximize}}: \quad \sum_{t=0}^{T-1}\sum_{\mathbf{N}^{i,t}} b^{i,t}(\mathbf{N}^{i,t})\left(\sum_{j\in L^k_{\text{out}}\cup k_{\text{out}}} N^{i,j,t}\ln(v^{i,j})\right)\\
&\textbf{subject to}: \quad \begin{cases} v^{i,j} \geqslant 0 & \forall\, j \in L^k_{\text{out}}\cup k_{\text{out}},\\ \sum_{j\in L^k_{\text{out}}\cup k_{\text{out}}} v^{i,j} = 1. \end{cases}
\end{aligned}$$

  where we have ignored the constants that arise when we take the logarithm of the multinomial distribution. It is solved in closed form by writing the *Karush Kuhn Tucker* (KKT) conditions. The values of $v^{i,j}$ which maximize the expected complete log-likelihood are given by

$$v^{i,j} = \frac{\sum_{t=0}^{T-1}\sum_{\mathbf{N}^{i,t}} b^{i,t}(\mathbf{N}^{i,t})N^{i,j,t}}{\sum_{t=0}^{T-1}\sum_{\mathbf{N}^{i,t}} b^{i,t}(\mathbf{N}^{i,t})\sum_{j'\in L^k_{\text{out}}\cup k_{\text{out}}} N^{i,j',t}}. \tag{15}$$

- For each intersection k with a source k_{in}, the update of the intensities of the Poisson processes for the outgoing links $j \in L^k_{\text{out}}$ is done independently for each link j by solving the following optimization program:

$$\underset{\lambda^j\geqslant 0}{\textbf{maximize}}: \sum_{t=0}^{T-1}\sum_{N^{k_{\text{in}},j,t}_{\text{in}}} c^{j,t}\left(N^{k_{\text{in}},j,t}_{\text{in}}\right)\left(N^{k_{\text{in}},j,t}_{\text{in}}\ln(\Delta_t\lambda^j) - \Delta_t\lambda^j\right)$$

  This optimization problem is solved in closed form as follows:

$$\lambda^j = \frac{1}{\Delta_t}\frac{\sum_{t=0}^{T-1}\sum_{N^{k_{\text{in}},j,t}_{\text{in}}} c^{j,t}\left(N^{k_{\text{in}},j,t}_{\text{in}}\right)N^{k_{\text{in}},j,t}_{\text{in}}}{\sum_{t=0}^{T-1}\sum_{N^{k_{\text{in}},j,t}_{\text{in}}} c^{j,t}\left(N^{k_{\text{in}},j,t}_{\text{in}}\right)}. \tag{16}$$

- For each link i, we update the initial state probability as

$$\pi_i(\xi) = \sum_{v=1}^{V}\omega_v \mathbf{1}_{\xi^{i,0}_v}(\xi). \tag{17}$$

  To learn this initial state probability, it is important to run the EM algorithm on several days of data (to reduce overfitting due to fitting the initial state probabilities based on a single day of data). In general, it is advised to run the EM algorithm over several days (weeks or months) of data to improve the learning of all the parameters of the model.
- We update the link parameters by maximizing the log-likelihood of the travel time observations $\mathbf{y}^{i,t}$ with respect to these parameters. The travel time allocation enables the optimization problem to decouple into smaller optimization problems, one for each link of the network. The optimization problem for link i is

$$\underset{C^i,R^i,\varsigma^i_s,\theta^i_p}{\textbf{maximize}} \quad \sum_{t=0}^{T}\sum_{\xi^{i,t}=0}^{\xi^i_{\max}} a^{i,t}(\xi^{i,t})\left(\sum_{y_{x_1,x_2}\in\mathbf{y}^{i,t}}\ln(g^i(y_{x_1,x_2}|\xi^{i,t}))\right), \tag{18}$$

  where $g^i(y_{x_1,x_2}|\xi^{i,t})$ represents the probability of observing a travel time y_{x_1,x_2} between x_1 and x_2 on link i given that the state of the link is $\xi^{i,t}$.
  Decoupling the optimization problem for each link of the network (instead of solving a large optimization program over the parameters of the entire network) makes it highly scalable as each of the optimization subproblems can be performed in parallel. If the travel time allocation method is not used, then the resulting optimization problem is coupled across the entire network, resulting in a large optimization problem that may not scale well. Physical constraints may be imposed on parameters of incoming links at an intersection (*e.g.* cycle time is the same for all incoming links of an intersection, the sum of green times of intersecting flows is less than the cycle time). For each intersection k, these constraints couple the optimization problems on the parameters of the incoming links of the intersection, but retains the decoupling of the

network optimization problem into $|\mathcal{K}|$ small optimization problems. We have denoted by $\mathcal{K}$ the set of intersections on the network and $|\mathcal{K}|$ is the number of intersections.

4.5. Real-time estimation and forecast

Estimating and forecasting traffic conditions in real-time can be achieved after the model parameters and turn probabilities have been learned, *i.e* once the Expectation Maximization algorithm has been run on large amounts of historical data. In real time we use the parameters learned by the EM algorithm (which characterize the stochastic dynamics of traffic) to perform inference using data available up to the time when the estimate is produced. This is done by running the particle filter to determine the distribution of traffic states given the available data and the learned value of the parameters. Forecast is done by propagating the particle filter forward from the current time interval. Since there is no available data, the filter only performs prediction steps (no update). For both estimation and forecast, the particle filter runs in real time on large-size networks (our implementation considers a network with over 800 links). However, the EM algorithm needs to run both the particle filter (E step) and the optimization algorithms (M step) for several iterations on large amounts of historical data. For this reason, the EM algorithm is run offline and the model parameters and turn probabilities can be updated periodically (*e.g.* every week or every month).

5. Experimental results

The model presented in this article relies on assumptions made on the dynamics of traffic flows on each link of the network. From these assumptions, we derive an analytical expression of the probability distribution of travel times, parameterized by traffic variables (Section 3.1). The model also relies on assumptions made on the statistical dynamics of traffic flows at intersections and derives a probabilistic model of the traffic dynamics on the network (Section 3.2). Our experimental results first validate the use of the traffic travel time distributions (Section 5.1) and then assess the real-time estimation and short-time prediction capabilities of our model from sparsely sampled probe data. We present the validation methodology of our network estimation model in Section 5.2.1 and report our results that validate the historical learning capabilities (Section 5.2.2) and the real-time estimation and prediction capabilities (Sections 5.2.3 and 5.2.4). We compare our results to a model that estimates only the mean travel time for each link and report that our model shows a 16% improvement over this baseline model to estimate mean link travel times. Our model also possesses several advantages over the baseline model that only estimates mean link travel times. These advantages include the ability to predict traffic conditions into the short-term future, the ability to estimate probability distributions of travel times between arbitrary points on the network (instead of just mean link travel time values), as well as the ability to estimate traffic parameters including signal timing and congestions states (queue lengths).

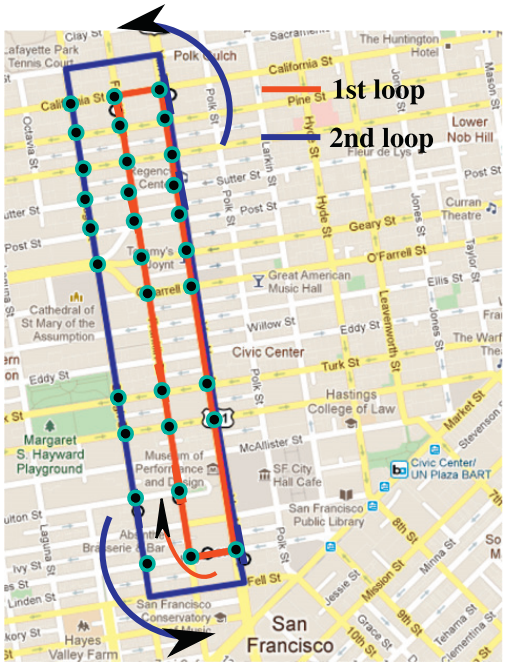

Fig. 9. Routes of the network used for field test validation. The drivers drove around two distinct loops consisting in Van Ness Ave. South bound and Franklin St. north Bound for the first routes and Van Ness Ave. North bound and Gough St. South bound for the second route. Signalized intersections are indicated with a circle.

5.1. Validation of the traffic travel time distributions

To validate the use of the traffic travel time distributions, we use data collected during a *field test experiment* performed during three consecutive days, from the 29th of June to the 1st of July 2010. Twenty drivers, each carrying a GPS device, drove for 3 h (3:15–6:15 pm) around two distinct loops in San Francisco (Fig. 9). The experiments were designed to capture the evening rush-hour congestion. The GPS devices recorded the location of the vehicles every second and provided detailed information on the trajectories of the drivers. Using *Virtual Trip Lines* (Hoh et al., 2008), we down-sample this detailed data *a posteriori* to extract link travel times.

For each link of the network, we compute the maximum likelihood estimates of the traffic parameters using 70% of the link travel times of the drivers, selected randomly. We test the hypothesis H_0: *the link travel times are distributed according to the traffic distributions* on the validation link travel times using the Kolmogorov–Smirnov (K–S) test (Massey, 1951). The K–S test is a standard non-parametric test to state whether samples are distributed according to an hypothetical distribution (in opposition to other tests like the *T*-test that tests uniquely the mean, or the chi-squared test that tests the normality of the data). The test is based on the K–S statistic which is computed as the maximum difference between the empirical and the hypothetical cumulative distributions. The test provides a *p*-value which informs on the goodness of the fit. Low *p*-values

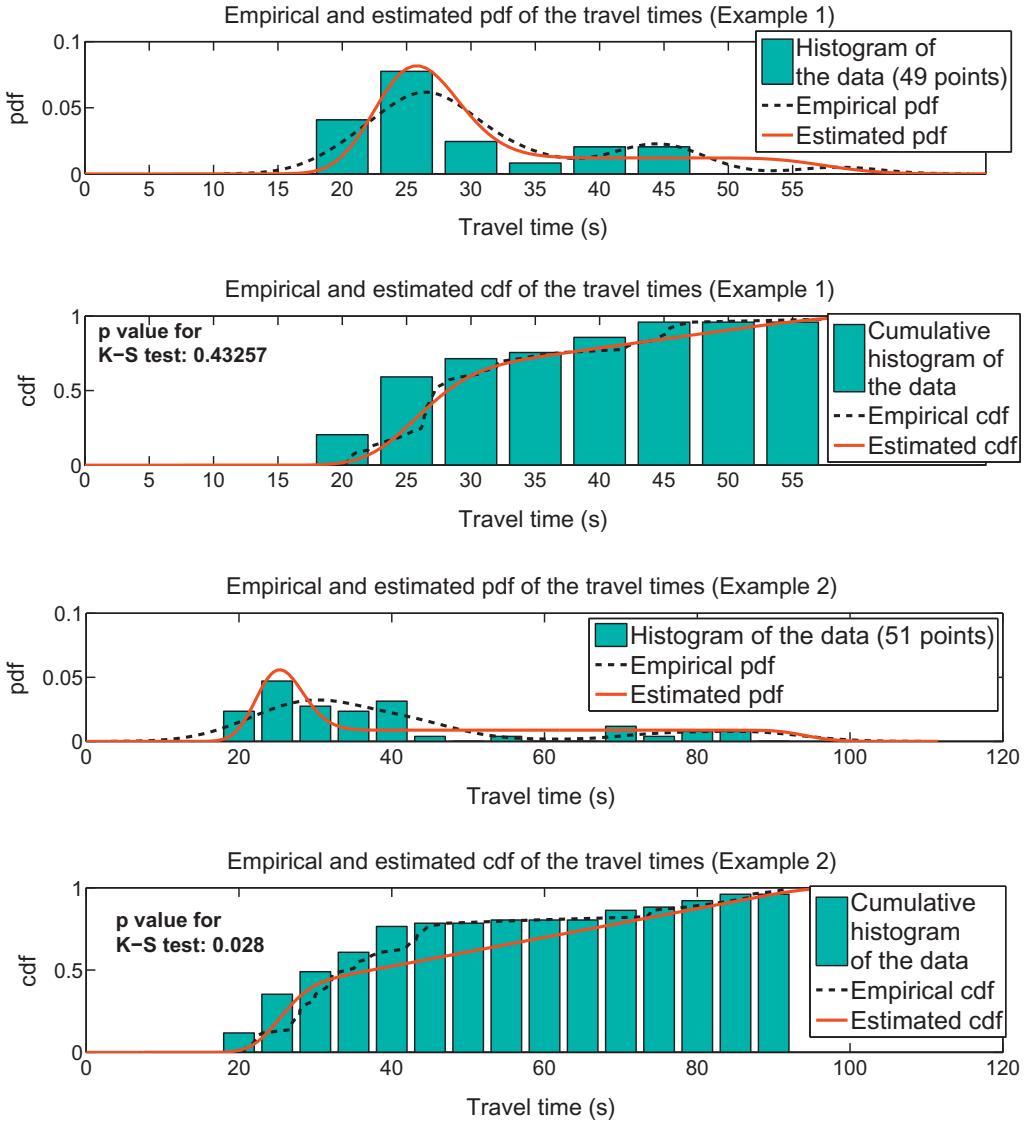

Fig. 10. Empirical and hypothetical probability distribution of travel times on two links of the network. The empirical pdf is computed using a kernel smoothing density estimator on the full data set. The estimated distribution is the traffic travel time distribution learned from 70% of the available data. The remaining 30% are used for validation (histograms and empirical pdf and cdf). (*Top, Example 1*) The link passes the K–S test with average *p*-value and represents how well the distribution of travel times is captured by the traffic model. (*Bottom, Example 2*) The link passes the K–S test for $\alpha = 0.01$ but fails the K–S test for $\alpha = 0.05$. This link illustrates the limitations of the uniform arrivals assumption and the need to extend the model to take into account platoon arrivals. However the uniform arrivals assumption does not prevent the model from estimating the traffic parameters representing the cycle timing and the driving behavior.

indicate that the data does not follow the hypothetical distribution. We reject hypothesis H_0 for p-values inferior to the significance level α. The significance level α corresponds to the percentage of Type-I error allowed by the test (rejecting the null hypothesis when it is actually true). The parameter α is commonly set to 0.05 or 0.01. We find that 95% of the links of the network have a link travel time distribution that follows the traffic assumption when the significance level is set to 0.01 and 84% of the links follow this assumption when the significance level is set to 0.05.

The result of the tests validate the use of the traffic travel time distributions in this article. We show in Fig. 10 the hypothetical and empirical distributions of travel times on two links of the network, one that passes the K–S test with an average p-value (top, Example 1) and another one that fails the K–S test (bottom, Example 2).

Important insight on the limitations and possibilities of improvement of the model is gained by looking into the empirical and hypothetical distributions directly. In Fig. 10 (bottom), we show the empirical and hypothetical travel time distributions for a link that passes the test for $\alpha = 0.01$ and fails the test for $\alpha = 0.05$ (the p-value is 0.028). The estimation of the traffic parameters captures the traffic conditions on the link (e.g. the mean free flow travel time is 23 s and the red time is 55 s). However, the traffic distribution fails to explain why so few vehicles have a travel time between 45 and 65 s, and why more vehicles have a travel time between 25 and 45 s. A strong hypothesis of the model is the assumption of uniform arrivals, which leads to delays uniformly distributed among the stopping vehicles (Hofleitner and Bayen, 2011). Due to light synchronization, some links have arrivals with platoons. On these links, delays are not uniformly distributed among the stopping vehicles and the derivations of the queuing model have to be adapted. The derivations of traffic travel time distributions with platoon arrivals is the subject of our current research, and a preliminary approach is developed in Bails et al. (2012). In the case of this link, the platoon arrivals lead to more vehicles with short delays (travel times between 25 and 45 s) while fewer vehicles have average delays (travel times between 45 and 65 s).

5.2. Validation of the traffic flow dynamic Bayesian network model

5.2.1. Experiment setup

Beginning in March of 2009, data has been collected from probe vehicles in the San Francisco Bay Area, as part of the *Mobile Millennium* project. One of the available data feeds available through the *Mobile Millennium* system comes from a fleet of over 500 taxis which report their location every minute, along with an identifier and a status (carrying a passenger or not). The status flag allows for the filtering of the taxi stops to load or unload passengers. When a change of status occurs, the

Fig. 11. The subnetwork of San Francisco, CA used for the validation of this model. The network consists of 769 links representing 126 km of roadway.

measurements directly anterior and posterior to this change of status are discarded. In its raw form, the data cannot be used by the algorithm. This is due to several issues.

- Between successive measurements, the vehicle may travel more than one link and we need to reconstruct the path.
- The measurements provide the location of the vehicles and we need to compute the direction of travel.
- The GPS measurement may be noisy and must be mapped onto the road network.

To overcome these difficulties, Hunter et al. (20111) developed a map-matching and path-inference algorithm which provides accurate measurement locations and paths followed by the vehicles. The duration between two successive measurements represents the travel time of the vehicle on its path. The latency in the communication of the location data to our servers is generally less than a minute.

For our study, we focus on a sub-network of San Francisco shown in Fig. 11. This network consists of 769 links representing 126 km of roadway. We validate the performance of our model using error metrics computed on previously unseen data. We report the *Root Mean Squared Error* (RMSE), the *Mean Absolute Error* (MAE) and the *Mean Percentage Error* (MPE).[2]

The Root Mean Squared Error is one of the most widely used metrics to quantify the difference between an estimator and the true value of the quantity being estimated. It measures the average of the squared error. As a result of the squaring of each term, Mean Squared Error heavily weights outliers. For this reason, we also compute the Mean Absolute Error, a common measure of forecast error in time series analysis. Using the convexity of the square function, it is easy to prove that the RMSE is always greater than or equal to the MAE. The Mean Percentage Error computes the average of the percentage error. When the actual values of the process to be estimated vary, this metric allows an equal weighting between the terms, as it is normalized by the actual value of the process.

For comparison, we created a baseline model that estimates mean link travel time. For each measurement in the training data set, the pace of the path is allocated to the links of the path with a weight equal to the proportion of the link that was traveled (1 if the full link is traveled, 0 if the link is not traveled at all). The mean pace of a link in the baseline model is computed as the weighted average of the paces on the paths of the training data set. Note that the baseline model does not provide a statistical distribution of travel times but rather a mean pace. This baseline model was chosen because standard time series statistical techniques (weighted moving average, exponential decay, ARMA) are not applicable to the data source that we receive because the measurement locations are not fixed and the time at which we receive measurements at a particular location is unknown in advance. Thus, it was necessary to develop our own comparison model that we found to be an intuitive method for processing the type of data we receive. In the remainder of this article, we refer to the model developed in this article as *the traffic model*. We refer to the model for comparison as the *baseline model*.

Both models run in a hybrid Matlab/Java environment and takes advantage of the *Mobile Millennium* system infrastructure which provides simple interfaces for accessing a simplified network representation of the roadway. The internal representation of the road network is built using NAVTEQ maps, which provide detailed geometry and attributes of the road network. The system also provides an interface for accessing the data feeds stored in the databases (which are map-matched and filtered in separate processes), and writing the outputs of the model to databases for future use (visualization, air quality related to traffic conditions, routing and so on). The historical learning of the parameters and the real time estimation and forecast run on a laptop for moderate size networks.

5.2.2. Validation of the learning capabilities

The model was trained using data collected on the three first Tuesdays of February 2010 from 3 pm to 6 pm. We use a time discretization Δ_t of fifteen minutes. From all the data collected on these days, we train the model on a randomly chosen subset representing 70% of the data. The training data set is used to estimate the network parameters (cycle time C, red time R, saturation number of vehicle ξ_s and turn proportions) of each link of the network. At each time interval t, the model also estimates the *a posteriori* most likely state of the link $\xi^{i,t}$ using training measurements available up to (and including) time interval t.

The performance of the *learning* capabilities is assessed using the *validation* data set of the training days. The validation data (30% of the full dataset) was previously set aside and not used to train the model. For each path in the validation dataset, we compute the travel time estimation according to the estimated parameters and a posteriori states. We compare this travel time to the true value experienced by the vehicle to compute the error metrics. The results are reported in Table 1. Our model shows an improvement of 16% in terms of RMSE compared to the baseline model. Moreover, the model learns parameters of the network (signal timing, saturation number of vehicles) for which it provides realistic estimates. For example, the duration of signal timings (cycle length) has a mean of 86 s over the network, with a standard deviation of 17 s, a minimum value of 45 s and a maximum value of 120 s.

[2] For a vector of E estimations $\hat{\mathbf{x}} = (\hat{x}_e)_{e=1\ldots E}$ of the true value $\mathbf{x} = (x_e)_{e=1\ldots E}$, the error metrics are defined as follows:

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{e=1}^{E}(x_e - \hat{x}_e)^2}{E}}, \quad \mathrm{MAE} = \frac{\sum_{e=1}^{E}|x_e - \hat{x}_e|}{E} \quad \text{and} \quad \mathrm{MPE} = \frac{1}{E}\sum_{e=1}^{E}\frac{|x_e - \hat{x}_e|}{x_e}$$

Table 1
Error metrics representing the estimation capabilities of the model. The metrics are reported on a validation dataset collected during the training days.

	RMSE	MAE	MPE
Traffic model	25.41	20.23	37.67%
Baseline model	31.56	25.69	46.20%
Improvement (%)	16.32	17.34	16.29

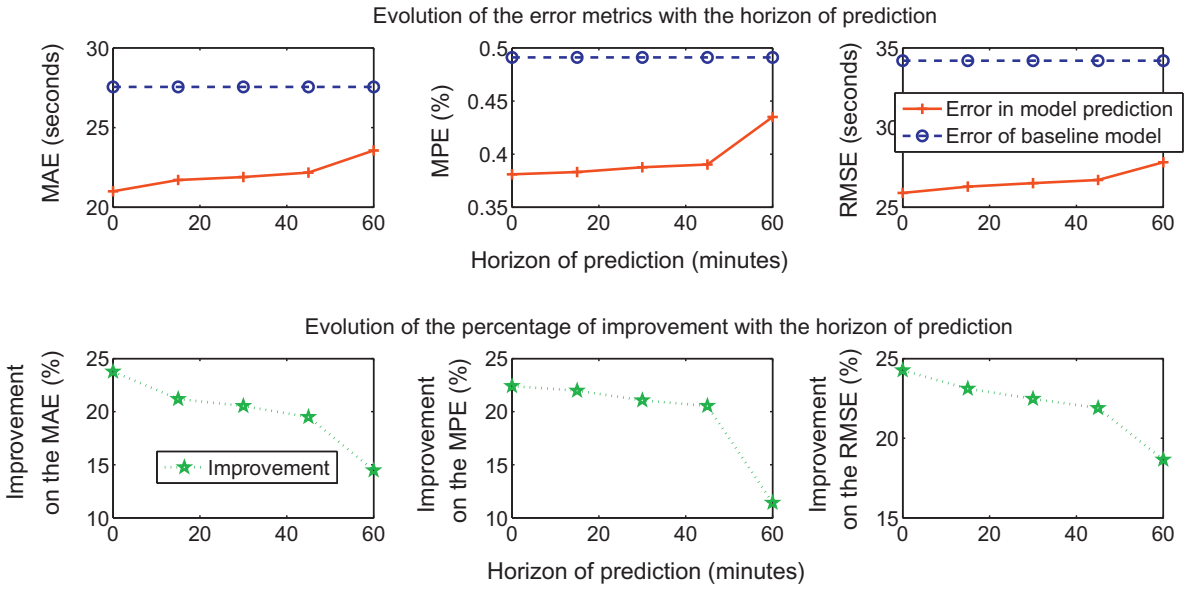

Fig. 12. Error metrics assessing the prediction capabilities of the model. The results show accurate prediction capabilities of the traffic model up to 45 min ahead. The baseline estimates are computed using historical estimates of the mean travel time, computed during the training. The baseline model does not provide prediction capabilities based on the current state of traffic and thus produces the same estimates for all horizons of prediction.

5.2.3. Validation of the real time estimation and prediction capabilities

In real time, the model uses the network parameters and turn probabilities learned historically to estimate and predict the state $\xi^{i,t}$ of each link i at each time interval t. At time interval t, we define the *estimation process* as the computation of the most likely state of the network at time interval t given data received up to and including time interval t. We define the *prediction at q time steps* as the computation of the most likely state of the network at time interval $t+q$ given data received up to and including time interval t. The prediction at 1 time step is also known as *a priori* state estimation. The prediction at 0 time step is identical to the *estimation process*.

The most likely state of the network is computed by performing the E step of the algorithm (particle filter) given the historical values of the network parameters (red time, cycle time, saturation number of vehicles) for each link of the network. For the prediction at time interval $t+q$, no data is available for time intervals posterior to time interval t. The filter is run forward, without weighting the particles (since future data is unobservable). The prediction process is a particular case of missing data in which the data is missing for all the links and all the time intervals after t.

The prediction of the most-likely state at time $t+q$ and the historic values of the link parameters allow for the computation of the travel time distributions of each link of the network at time interval $t+q$. From the travel time distribution, we can extract various information including a mean travel time, a variance, confidence intervals and so on.

The assessment of estimation and prediction capabilities is performed on Tuesday, February 22nd 2010 (Tuesday following the training period) from 3 pm to 6 pm. We report the error metrics for prediction steps ranging from 1 time step (a priori estimation) to 4 time steps (1 h). We compare our estimates with the estimates of the baseline model. For the baseline model, the real-time prediction is computed as the historical average of the pace for each link during the time interval of interest. This means that our prediction for Tuesday, February 22 at 3 pm is the average pace observed at 3 pm from the training data set (the three previous Tuesdays). Therefore, the estimates of the baseline model do not depend on the horizon of prediction. The results are reported in Fig. 12.

For the *a priori* estimation (prediction at one time step), the error metrics of both the traffic model and the baseline model slightly increase compared to the results presented in Section 5.2.2. This increase in the error metrics accounts for the differences in traffic conditions on a new day and the loss of accuracy between the *a posteriori* and the *a priori* estimates. The

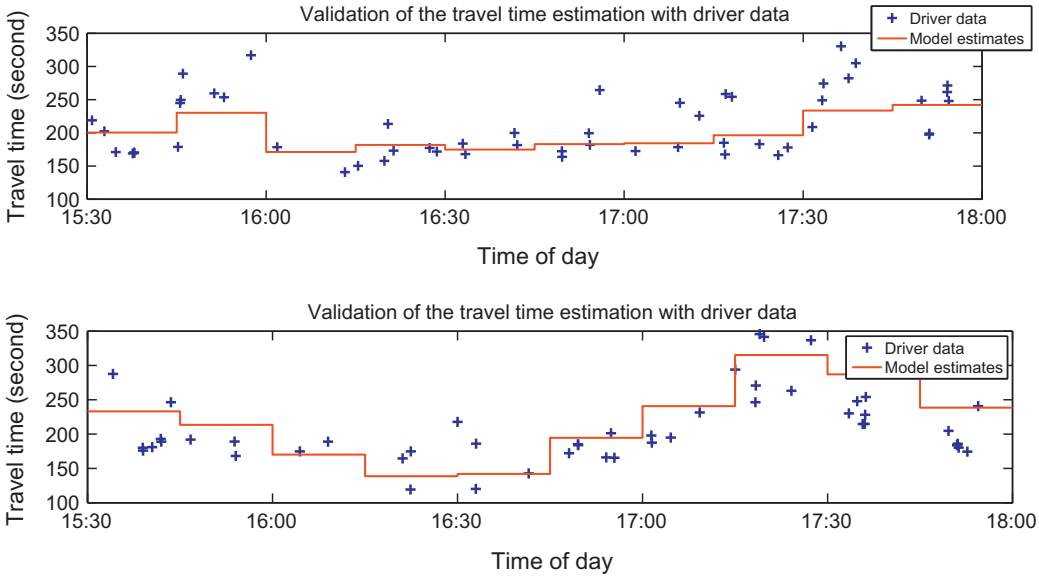

Fig. 13. Comparison of the model estimates with the ground truth route travel times for the north and south bound routes on Van Ness Ave. The red curve represent the average travel time estimate of the traffic model. The blue crosses represent the driver data collected during the field test experiment. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

improvement of the traffic model is higher and shows the capabilities of the model to adapt to slightly different traffic conditions and perform short-term prediction.

As the number of prediction steps increases, the estimation error of the traffic model increases. The modeling of the traffic dynamics ensures a certain regularity in the traffic estimates, and the prediction capabilities of the model remain accurate and represent a significant improvement to the baseline model. The Root Mean Squared Error shows the greatest improvement, which indicates that the traffic model has fewer estimates that differ in a significant way from the true values of the travel times than the baseline model does.

5.2.4. Field test experiment

We use data collected during the field test experiment in San Francisco to provide another validation of the capabilities of the model. We extract route travel times on four different routes of the network (Fig. 9). The north and south end of the routes are respectively California St and Grove St. The four routes consist of Van Ness Ave. north bound, Van Ness Ave. south bound, Franklin St. and Gough St.

In order to assess the validity of the model, we down-sampled the driver data to mimic the kind of data generally available in real-time. The model runs over this sparsely sampled data and we then perform validation by comparing our estimates of the route travel times with the actual route travel times of our drivers. The comparison of the model estimates and the ground truth route travel times are presented in Fig. 13. This data highlights the variability of travel times experienced by vehicles. The travel time estimates closely follow the trend of traffic dynamics. The RMSE of the traffic model on the route travel times of our drivers is 74.42 s, the MAE is 63.62 s and the MPE is 33.24%. The travel times on the routes are significantly higher than the travel times used for validation in Section 5.2.3, hence higher values of the RMSE and MAE. In the computation of the MPE, each estimation error is normalized with the travel time on the path. The MPE is better on longer stretches, as the relative variability of travel times is relatively smaller.

6. Conclusion and discussion

This article presented a statistical model based on the dynamics of arterial traffic flow. These results indicate that the model provides a substantial improvement over a "simple" baseline approach. Besides the improvement of the mean travel time estimation, our model possesses several advantages over the comparison model. It improves the estimation of mean link travel times compared to a baseline model. It estimates the *probability distribution* of travel times (rather than only the mean) between any two location on the network. It *learns* parameters with a physical interpretation (such as fundamental diagram and signal parameters) and also learns turn movement probabilities within the arterial network. Using the learned parameters, real-time estimation and prediction of traffic conditions is performed using a customized particle filter. The model also leverages historical data to estimate traffic conditions in *real-time* throughout the network even *where little or no real-time data is received*. This is due to the model's ability to accurately track flows through the network as well as the relative recurrence of arterial traffic dynamics.

This article presents a general framework to model arterial traffic as a stochastic dynamical system. The presented model can be adapted depending on the sparsity, the noise and the amount of available data. The model could take into account the fact that delays are dependent upon the turn movement through the intersection. The model currently assumes that travel times are independent of the turn movements and this is not true in general. This generalization of the model can be implemented by considering a multi-dimensional state on each link of the network. The dimension of the state on a link of the network would be equal to the number of lanes and each dimension of the state would correspond to the queue length on that lane. Similarly, we can consider several traffic parameters per link representing the different phases of the signal cycle. In a statistical model, one needs to find a compromise in the level of detail and number of parameters chosen for the model depending on the type and the amount of data available. Indeed, a more precise model with numerous parameters is able to fit the training data more accurately and explain more details in the dynamics of the model. However, such a model is more likely to *over-fit* the data when the amount of training data is not sufficient to learn all the parameters. Over-fitting the training data decreases the performance on testing data and thus the capabilities of real time estimation and short-term prediction of the model. In this paper, we consider sparsely sampled probe vehicles (vehicles report their location every minute). The type of data leads us to focus on estimating trends of traffic (estimation every fifteen minutes) rather than fluctuations (variations of queue length and travel time within a traffic cycle). We also decided not to estimate signal phases and lane by lane queue length (even though it can be a natural extension of the model) because we do not have information on protected turn movements. Our information is limited to the number of lanes per link and we assume that all the lanes of a link are in the same state at a given time.

Note that the results presented on this article rely on arterial traffic modeling assumptions that can limit the applicability of the model. In particular, the model assumes uniform arrivals on each link of the network. On controlled arterials, where signal synchronization is important, this hypothesis does not hold and the model does not capture travel time distributions as accurately. We are working on a generalization of the traffic travel time distributions that take into account platoon arrivals and capture more accurately the travel time distributions on these arterials.

This article presents the fundamental concepts needed for performing large-scale estimation of arterial traffic conditions using only low penetration rate GPS probe data. For the next decade, only a small number of municipalities will have the financial resources to equip their entire arterial network with dedicated infrastructure. At the same time, the market of probe data will remain too fragmented to be used in high penetration rate models, forcing traffic engineers to design traffic information systems capable of handling sparse data. The present article is a first step towards this goal, which shows promising results.

Acknowledgments

The authors thank Jean-Patrick Lebacque at the Research Center IFSTTAR-GRETTIA for his valuable contribution and feedback on the model presented in this article. We thank Timothy Hunter from UC Berkeley for providing filtered probe trajectories from the raw measurements of the probe vehicles. We thank the California Center for Innovative Transportation (CCIT) staff for their contributions to develop, build, and deploy the system infrastructure of *Mobile Millennium* on which this article relies and for their help in the logistical planning of the field test described in this article. This research was supported by the Federal and California DOTs, Nokia and the Center for Information Technology Research in the Interest of Society (CITRIS).

Appendix A. Summary of the notations used in the article

1. *Traffic model parameters*The traffic model parameters represent the characteristics of the network. They are specific to a link i of the network. For notational simplicity, the subscript i is omitted when the derivations are valid for any link of the network.

$\rho^i_{\max}$	Maximum density of link i
$q^i_{\max}$	Capacity (maximum flow) on link i
ρ^i_c	Critical density of link i
w^i	Backward shockwave speed of link i
$\xi^i_{\max}$	Maximum number of vehicles that can physically be on link i. This is the number of vehicles when the density is the maximum density. For a link of length L^i, we have $\xi^i_{\max} = \rho^i_{\max} L^i$
v^i_f	Free flow speed of link i
p^i_f	Free flow pace (inverse of free flow speed) of link i. We have $p^i_f = 1/v^i_f$

2. *Traffic signal parameters*
The traffic signal parameters characterize the properties of the traffic signal that condition the traffic dynamics. In this model, we only consider traffic signals in the form of traffic lights. The extension of the model to stop signs will be the subject of future work. As for the traffic model parameters, these variables are specific to a link i of the network. However, the subscript may be omitted.

C^i	Duration of a light cycle on link i
R^i	Duration of the red time on link i
ξ_s^i	Maximum number of vehicles that can exit link i during a light cycle. This variable is related to the ratio of green time and the traffic model parameters

3. *Traffic state variables*

The traffic state variables describe the conditions of traffic that characterize the traffic dynamics on the network. The variables are specific to a link i and a time interval t and represent the dynamic evolution of the traffic state in the different time intervals $t \in \{0 \dots T\}$. The reference to the link or to the time interval may be omitted when the derivations are not link or time specific.

$\rho_a^{i,t}$	Arrival density on link i during time interval t
$v_a^{i,t}$	Arrival shockwave speed on link i during time interval t (speed of growth of the queue due to additional vehicles arrival)
$\tau^{i,t}$	Duration of the clearing time on link i during time interval t
$l_{\max}^{i,t}$	Length of the triangular queue on link i during time interval t
$\xi^{i,t}$	Number of vehicles that stop during each light cycle

4. *Network variables and parameters*

The network variables and parameters characterize the architecture of the road network and describe the flow of vehicles at intersections.

$\mathcal{I}$	Set of the links of the network
$\mathcal{K}$	Set of the intersections of the network
L^i	Length of link i
i,j	Indices of links of the network ($i,j \in \mathcal{I}$). When we refer to an intersection, i refer to a link upstream of the intersection whereas j refers to a link downstream of the intersection
k	Index of an intersection of the network
L_{in}^k	Set of incoming links of intersection k
L_{out}^k	Set of outgoing links of intersection k
k_{in}	Source (if existing) of intersection k
k_{out}	Sink (if existing) of intersection k
$n_{\text{in}}^{i,t}$	Number of vehicles that arrive in link i during a light cycle for time interval t
$n_{\text{out}}^{i,t}$	Number of vehicles that leave link i during a light cycle for time interval t
$N_{\text{in}}^{i,t}$	Cumulative number of vehicles that arrive in link i during time interval t
$N_{\text{out}}^{i,t}$	Cumulative number of vehicles that leave link i during time interval t
$N_{\text{in}}^{i,j,t}$	Cumulative number of vehicles that leave link i and are assigned to link j during time interval t
κ^i	Number of lanes of link i

5. *Particle filter and E Step*

The inference of traffic states on the network given the parameters of the network, of the turn movements and given observed path travel time data is computed using an approximation (for tractability reasons). This approximation relies on particle filtering.

V	Number of particles
v	Index of the particle
$\xi_v^{i,t}$	State of particle v on link i during time interval t
ω_v	Importance weight of particle v
$a^{i,t}(\xi^{i,t})$	Expected probability that link i is in state $\xi^{i,t}$ at time interval t, computed from the approximation of the joint distribution given by the particles and their importance weight
$b^{i,j,t}\left(\xi^{i,t}, N_{\text{in}}^{i,j,t} : j \in L_{\text{out}}^k\right)$	Expected probability that link i is in state $\xi^{i,t}$ at time interval t and that $N_{\text{in}}^{i,j,t}$ vehicles get assigned to the outgoing links of the intersection. It is computed from the approximation of the joint distribution given by the particles and their importance weight

6. *Probabilities*

The model relies on a probabilistic description of the traffic network dynamics, whose notations are summarized in the following table.

φ^i	Probability distribution function of the free flow pace on link i. This function is defined on $\mathbb{R}^+$ and for $p_f \in \mathbb{R}^+$, $\varphi^i(p_f)$ is the probability density that vehicles drive with a free flow pace p_f
θ_p^i	Parameters of the probability distribution function φ^i
$\gamma(\cdot)$	Probability distribution function of a random variable with Gamma distribution
(α^i,β^i)	In the case of a gamma distribution on the free flow pace, the parameters of the distribution are the shape α^i and inverse scale parameter β^i. The Gamma distribution reads $\gamma(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$, where Γ is the Gamma function defined on $\mathbb{R}^+$ and with integral expression $\Gamma(x) = \int_0^{+\infty} t^{x-1} e^{-t}\, dt$, when i is omitted for simplicity
y_{x_1,x_2}	Observation of the random variable representing the travel time between locations x_1 and x_2
$\mathbf{y}^{i,t}$	Set of travel time observations received on link i during time interval t
$I^{i,t}$	Number of travel time observations received on link i during time interval t
$g^{i,t}(\cdot)$	Probability distribution function of travel times on link i during time interval t. The function is parameterized by the traffic model and signalization parameters. It changes over time with the state of the link. The function also takes into account the location of the measurements x_1 and x_2 on link i such that $g^{i,t}(y_{x_1,x_2})$ is the probability density of the travel time observation y_{x_1,x_2}
$v^{i,j}$	Probability that a vehicle leaving link i is assigned to link j
λ^j	Intensity of the Poisson process of vehicles arrival on an outgoing link $j \in L_{\text{out}}^k$ of intersection k, coming from a source k_{out}
$\pi^i(\xi)$	Probability that link i is in state ξ at the beginning of the experiment. These probabilities represent probabilistic initial conditions for the state of link i

7. *Other variables*

t	Index of the time interval
T	Index of the last time interval. By convention, the first interval is numbered 0 so $T+1$ is the number of time intervals
Δ_t	Duration of a time interval
$\mathbf{1}_S$	Indicator function of set S

8. *Probability distributions*

$\mathcal{P}\left(\Xi \middle\vert \mathbf{y}^{i,t}, R^i, C^i, \xi_s^i, \theta_p^i : i \in \mathcal{I}, t \in \{0 \dots T\}\right)$	Probability of observing a state evolution Ξ given the travel time observations
$\mathcal{P}(\xi, \mathbf{y})$	Likelihood of the state evolution of the system, with observations $\mathbf{y}$
$\mathcal{P}(\mathbf{y}^{i,t} \vert \xi^{i,t})$	Conditional probability of the travel time observations $\mathbf{y}^{i,t}$, given that link i is in state $\xi^{i,t}$ during time interval t
$\mathcal{P}(N_{\text{in}}^{i,j,t} : j \in L_{\text{out}}^k \cup k_{\text{out}})$	Probability that $\left(N_{\text{in}}^{i,j,t}\right)_{j \in L_{\text{out}}^k \cup k_{\text{out}}}$ vehicles leave link i and are assigned to link j during time interval t

References

Arulampalam, M., Maskell, S., Gordon, N., Clapp, T., 2002. A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking. IEEE Transactions on Signal Processing 50 (2), 174–188.

Bails, C., Hofleitner, A., Xuan, Y., Bayen, A., 2012. Three-stream model for arterial traffic. In: Proceedings of the 91st Transportation Research Board Annual Meeting, Number 12-1212, Washington, DC (January).

Ban, X., Herring, R., Hao, P., Bayen, A., 2009. Delay pattern estimation for signalized intersections using sampled travel times. In: Proceedings of the 88th Transportation Research Board Annual Meeting, Washington, DC (January).

Bayen, A., Butler, J., Patire, A., et al., 2011. Mobile Millennium final report. Technical report. University of California, Berkeley, CCIT Research Report UCB-ITS-CWP-2011-6.

Bickel, P., Chen, C., Kwon, J., Rice, J., Zwet, E.V., Varaiya, P., 2007. Measuring traffic. Statistical Science 22 (4), 581–597.

Blandin, S., Work, D., Goatin, P., Piccoli, B., Bayen, A., 2011. A general phase transition model for vehicular traffic. SIAM Journal on Applied Mathematics 71 (1), 107–127.

Boyen, X., Koller, D., 1998. Tractable inference for complex stochastic processes. In: Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence, pp. 33–42.

Cooper, G.F., 1990. The computational complexity of probabilistic inference using Bayesian belief networks. Artificial Intelligence 42 (2–3), 393–405.
Daganzo, C., 1994. The cell transmission model: a dynamic representation of highway traffic consistent with the hydrodynamic theory. Transportation Research Part B 28 (4), 269–287.
Dempster, A., Laird, N., Rubin, D., 1977. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society. Series B (Methodological), 1–38.
Evans, L.C., 1998. Partial Differential Equations. Graduate Studies in Mathematics, vol. 19. American Mathematical Society, Providence, RI.
Furtlehner, C., Lasgouttes, J., de la Fortelle, A., 2007. A belief propagation approach to traffic prediction using probe vehicles. In: 10th Intelligent Transportation Systems Conference, pp. 1022–1027.
Geroliminis, N., Daganzo, C., 2008. Existence of urban-scale macroscopic fundamental diagrams: some experimental findings. Transportation Research Part B 42 (9), 759–770.
Geroliminis, N., Skabardonis, A., 2006. Prediction of arrival profiles and queue lengths along signalized arterials by using a Markov decision process. Transportation Research Record 1934 (1), 116–124.
Geroliminis, N., Skabardonis, A., 2010. Queue spillovers in city street networks with signal-controlled intersections. In: Proceedings of the 89th Transportation Research Board Annual Meeting, Washington, DC.
Hellinga, B., Izadpanah, P., Takada, H., Fu, L., 2008. Decomposing travel times measured by probe-based traffic monitoring systems to individual road segments. Transportation Research Part C 16 (6), 768–782.
Herring, R., Hofleitner, A., Abbeel, P., Bayen, A., 2010. Estimating arterial traffic conditions using sparse probe data. In: Proceedings of the 13th International IEEE Conference on Intelligent Transportation Systems, Madeira, Portugal, pp. 929–936 (September).
Herring, R., Hofleitner, A., Amin, S., Nasr, T.A., Abdel Khalek, A., Abbeel, P., Bayen, A., 2010. Using mobile phones to forecast arterial traffic through statistical learning. In: Proceedings of the 89th Transportation Research Board Annual Meeting, Number 10-2493, Washington DC.
Hofleitner, A., Bayen, A., 2011. Optimal decomposition of travel times measured by probe vehicles using a statistical traffic flow model. In: 14th International IEEE Conference on Intelligent Transportation Systems (ITSC), pp. 815–821 (October).
Hofleitner, A., Herring, R., Bayen, A., 2011. A hydrodynamic theory based statistical model of arterial traffic. Technical report. University of California, Berkeley, UCB-ITS-CWP-2011-2 (January).
Hofleitner, A., Herring, R., Bayen A., 2012. Probability distributions of travel times on arterial networks: a traffic flow and horizontal queuing theory approach. In: Proceedings of the 91st Transportation Research Board Annual Meeting, Number 12-0798, Washington, DC (January).
Hoh, B., Gruteser, M., Herring, R., Ban, J., Work, D., Herrera, J., Bayen, A., 2008. Virtual trip lines for distributed privacy-preserving traffic monitoring. In: The Sixth Annual International Conference on Mobile Systems, Applications and Services (MobiSys 2008), Breckenridge, USA (June).
Horvitz, E., Apacible, J., Sarin, R., Liao, L., 2005. Prediction, expectation, and surprise: methods, designs, and study of a deployed traffic forecasting service. In: Twenty-First Conference on Uncertainty in Artificial Intelligence.
Hunter, T., Moldovan, T., Zaharia, M., Merzgui, S., Ma, J., Franklin, M.J., Abbeel, P., Bayen, A.M., 2011. Scaling the mobile millennium system in the cloud. In: Proceedings of the 2nd ACM Symposium on Cloud Computing, vol. 28, SOCC '11, ACM, pp. 1–8.
Jordan, M.I. (Ed.), 1999. Learning in Graphical Models. MIT Press, Cambridge, MA, USA.
Jordan, M.I., Ghahramani, Z., Jaakkola, T.S., Saul, L.K., 1999. An introduction to variational methods for graphical models. Machine Learning 37 (2), 183–233.
Kimber, R.M., Hollis, E.M., 1979. Traffic queues and delays at road junctions. Traffic Systems Division, Traffic Engineering Department, Transport and Road Research Laboratory.
Krause, A., Horvitz, E., Kansal, A., Zhao, F., 2008. Toward community sensing. In: Proceedings of the 7th International Conference on Information Processing in Sensor Networks. IEEE Computer Society, pp. 481–492.
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 (6), 586–606.
Leeuwaarden, J.S.V., 2006. Delay analysis for the fixed-cycle traffic-light queue. Transportation Science 40 (2), 189–199.
Lighthill, M., Whitham, G., 1955. On kinematic waves. II. A theory of traffic flow on long crowded roads. Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences 229 (18), 317–345.
Liu, H., van Zuylen, H., van Lint, H., Salomons, M., 2006. Predicting urban arterial travel time with state-space neural networks and Kalman filters. Transportation Research Record (1968), 99–108.
Liu, H., Danczyk, A., Brewer, R., Starr, R., 2008. Evaluation of cell phone traffic data in Minnesota. Transportation Research Record 2086, 1–7.
Massey, F.J., 1951. The Kolmogorov–Smirnov test for goodness of fit. Journal of the American Statistical Association 46 (253), 68–78.
Murphy, K., 2002. Dynamic bayesian networks: representation, inference and learning. Ph.D. thesis, University of California, Berkeley.
Park, T., Lee, S., 2004. A Bayesian approach for estimating link travel time on urban arterial road network. In: Computational Science and Its Applications, pp. 1017–1025.
Richards, P., 1956. Shock waves on the highway. Operations Research 4 (1), 42–51.
Russell, S., Norvig, P., 1995. Artificial Intelligence – A Modern Approach. Prentice-Hall, Inc., Englewood Cliffs, NJ.
Sun, X., Munoz, L., Horowitz, R., 2004. Mixture Kalman filter based highway congestion mode and vehicle density estimator and its application. In: Proceedings of the 2004 American Control Conference, Boston, MA, pp. 2098–2103.
Sun, S., Zhang, C., Yu, G., 2006. A Bayesian network approach to traffic flow forecasting. IEEE Transactions on Intelligent Transportation Systems 7 (1), 124–132.
Thiagarajan, A., Sivalingam, L., LaCurts, K., Toledo, S., Eriksson, J., Madden, S., Balakrishnan, H., 2009. VTrack: accurate, energy-aware traffic delay estimation using mobile phones. In: 7th ACM Conference on Embedded Networked Sensor Systems (SenSys), Berkeley, CA (November).
Trans Res Board, 2000. Highway Capacity Manual. Washington, DC: Transportation Research Board, National Research Council.
Van Den Broek, M., Van Leeuwaarden, J., Adan, I., Boxma, O.J., 2006. Bounds and approximations for the fixed-cycle traffic-light queue. Transportation Science 40 (4), 484496.
Van Lint, J., Hoogendoorn, S.P., Van Zuylen, H., 2005. Accurate freeway travel time prediction with state-space neural networks under missing data. Transportation Research Part C 13 (5-6), 347–369.
Van Zuylen, H.J., Zheng, F., Chen, Y., 2010. Using probe vehicle data for traffic state estimation in signalized urban networks. In: Barcelo, J., Kuwahara, M. (Eds.), Traffic Data Collection and its Standardization, International Series in Operations Research and Management Science, vol. 144. Springer, New York, pp. 109–127.
Viti, F., Van Zuylen, H.J., 2009. The dynamics and the uncertainty of queues at fixed and actuated controls: a probabilistic approach. Journal of Intelligent Transportation Systems 13 (1).
Webster, F.V., 1958. Traffic Signal Settings. Paper No. 39, Road Research Laboratory, England, HMSO.
Work, D., Blandin, S., Tossavainen, O.-P., Piccoli, B., Bayen, A., 2010. A traffic model for velocity data assimilation. Applied Mathematics Research Express 2010 (1), 1–35.
Zhang, H., Kim, T., 2005. A car-following theory for multiphase vehicular traffic flow. Transportation Research Part B 39 (5), 385–399.
Zhang, K., Taylor, M.A.P., 2007. A new concept and general algorithm architecture to improve automated incident detection. In: Proceedings of the 17th International Symposium on Transportation and Traffic Theory. Elsevier.
Zheng, F., Van Zuylen, H., Xia, L., Chen, Y., 2009. Investigating the feasibility of urban link travel time estimation based on probe vehicle data. In: International Conference on Transportation Engineering 2009, pp. 3387–3392. ASCE.
Zheng, F., Van Zuylen, H., 2010. Reconstruction of delay distribution at signalized intersections based on traffic measurements. In 13th IEEE Intelligent Transportation Systems (ITSC'10), pp. 1819 –1824 (September).