2012 15th International IEEE Conference on Intelligent Transportation Systems
Anchorage, Alaska, USA, September 16-19, 2012

Automatic inference of map attributes from mobile data

A. Hofleitner, E. Côme, L. Oukhellou, J-P Lebacque and A. Bayen

Abstract— **The development and update of reliable Geographic Information Systems (GIS) greatly benefits Intelligent Transportation Systems developments including real-time traffic management platforms and assisted driving technologies. The collection and processing of the data required for the development and update of GIS is a long and expensive process which is prone to errors and inaccuracies, making its automation promising. The article introduces a method which leverages the emergence of sparsely sampled probe vehicle data to update and improve existing GIS. We present an unsupervised classification algorithm which discriminates between signalized road segments (as having a signal at the downstream intersection) and non-signalized road segments. This algorithm uses a statistical model of the probability distribution of vehicle location within a link, derived from hydrodynamic traffic flow theory. The decision of whether the link has a traffic signal or not is taken according to model selection criteria. Numerical results performed with sparsely sampled probe data collected by the *Mobile Millennium* system in the Bay Area of San Francisco, CA underline the importance of the problem addressed by the article to improve the accuracy and update signal information of GIS. They showcase the ability of the method to detect the presence of traffic signals automatically.**

I. Introduction

Geographic Information Systems (GIS) store various types of geographically referenced data. Accurate and reliable GIS (such as NAVTEQ or OpenStreetMap) are needed for a large number of applications in *Intelligent Transportation Systems* (ITS): an accurate description of the geometry and the features of the road network is necessary to develop real-time traffic information systems [2], [9], routing [15] and driver assistance technologies. The collection and processing of data for GIS is an expensive and time consuming process, making automated techniques desirable. Besides the cost and time inefficiencies, traditional map developments and updates are based on surveying methods and digitizing of satellite images which lead to inaccuracies and systematic errors. The emergence of new technologies opens new possibilities to increase the efficiency in developing and maintaining reliable GIS, in particular in developing countries where the infrastructure is evolving rapidly.

The use of GIS for ITS applications has attracted a lot of interest in the computer vision and robotics community to improve driver assistance technologies. For example, real-time video processing detection algorithms combined with accurate GIS significantly improve the capabilities to infer speed limitations [12]. Similarly, object recognition algorithms based on image processing allow for the real-time detection of traffic signals by intelligent vehicles [5]. These algorithms can update existing map databases as intelligent vehicles travel the network. The distributed nature of the problem makes crowdsourcing approaches appealing as they leverage information collected by a large number of vehicles traveling on the road network.

GPS traces have generated significant interest in the machine learning community, to lower the costs of producing and updating digital maps while improving their accuracy. In particular, GPS traces have been used to learn the map network geometry using clustering and graph inference algorithms [3]. In [14], data-mining approaches are used to process GPS traces and refine existing digital maps to enable safety applications, such as lane-keeping, and convenience applications, such as lane-changing advice. However, high frequency sampling GPS traces remain scarcely available to the public, mainly because of privacy concerns, communication costs and limitation of the battery life of portable GPS devices. Sparsely sampled probe data is still at the present time the main source of geo-location data with the prospect of global coverage in the near future. For these reasons, the present article focuses on the use of sparsely sampled data for digital map learning (vehicles are sampled on average once per minute).

The potential of sparsely sampled probe vehicle data has been demonstrated through the successful implementations of reliable real-time traffic information systems on both the highway and the arterial networks [2], [9]. We investigate the use of sparsely sampled probe data collected by the *Mobile Millennium* system [2] to improve and update existing digital map databases: we study how this data can be used, in combination with to the road network geometry, to automatically detect the presence of traffic signals (traffic lights or stop signs) at each intersection.

The contributions of the article are as follows. We develop an algorithm based on hydrodynamic theory [11], [13] to derive the probability distribution (pdf) of the location of a vehicle on an arterial road segment. The presence of traffic signals leads to the creation and dissolution of queues upstream of signalized intersections. The model formalizes the following intuition: vehicles are more likely to be located where they experience delay, *i.e.* in the queue upstream of a signalized intersection. We develop a statistical model to learn the parameters of the pdf of the location of vehicles and present an unsupervised classification algorithm which

Ph.D. student, Electrical Engineering and Computer Science, UC Berkeley, CA and UPE/IFSTTAR/GRETTIA, France. Corresponding author, e-mail: aude@eecs.berkeley.edu.
Researcher UPE/IFSTTAR/GRETTIA, France
Senior researcher UPE/IFSTTAR/GRETTIA, France
Senior researcher UPE, Director of IFSTTAR/GRETTIA, France
Associate Professor, Electrical Engineering and Computer Science and Civil and Environmental Engineering, UC Berkeley, CA.

identifies whether there is a signal at the upstream end of each road segment in the network. We define a road segment (link) as the stretch of road between consecutive intersections.

The rest of the article is organized as follows. In Section II, we present the statistical model derived from hydrodynamic theory and derive the pdf of the location of vehicles upstream of a signalized intersection. Section III describes the unsupervised classification algorithm based on model selection information criterion. We analyze the signal detection potentials of our algorithm in Section IV using data collected from the *Mobile Millennium* system in an arterial network in San Francisco, CA including over 1,000 links. In the reminder of this article, a *link* refers to a road segment between successive intersections (signalized or not).

II. Modeling the distribution of vehicles upstream of a signal

The presence of traffic signals leads to the formation of queues, which results in a non-homogeneous density of vehicles along the link. The density of GPS measurements received from probe vehicle sampled uniformly in time (even with a low frequency) is expected to be higher close to signalized intersections than far from the intersections. From hydrodynamic theory, we derive a model which represents the distribution of vehicles upstream of a signalized intersection and use this model to detect signal locations.

A. Flow model for arterial traffic

We model vehicular flow as a continuum and represent it with macroscopic variables of flow, $q(x,t)$, density, $\rho(x,t)$ and velocity, $v(x,t)$. The conservation of vehicles leads to a relation between these variables: $q(x,t) = \rho(x,t)v(x,t)$. Experimental results show another relationship, referred to as the *fundamental diagram* [11], [13]. We make the standard assumption of a triangular fundamental diagram. Upstream of the traffic signal, queues form and dissipate periodically, creating areas of the link of higher density, when considering the temporal average. We define two discrete traffic regimes: *undersaturated* and *congested*, depending on the presence (resp. the absence) of a remaining queue when the signal turns red. Both regimes are illustrated Figure 1.

Undersaturated regime: The queue fully dissipates within the green time. When the flow arriving on a link is constant (corresponding to a density ρ_a), the speeds of formation and dissolution of the queue are constant and denoted v_a and w respectively. The spatio-temporal region where vehicles are stopped on the link is called the *triangular queue* (because of its triangular shape). Its length is called the maximum queue length, $l_{\max}$.

Congested regime: There exists a part of the queue downstream of the triangular queue called *remaining queue* with length l_r corresponding to vehicles which have to stop multiple times before going through the intersection.

Remark 1: The undersaturated regime is a special case of the congested regime in which the remaining queue l_r has length zero. We present the derivations of the model for the congested regime as it also includes the derivations for the undersaturated regime.

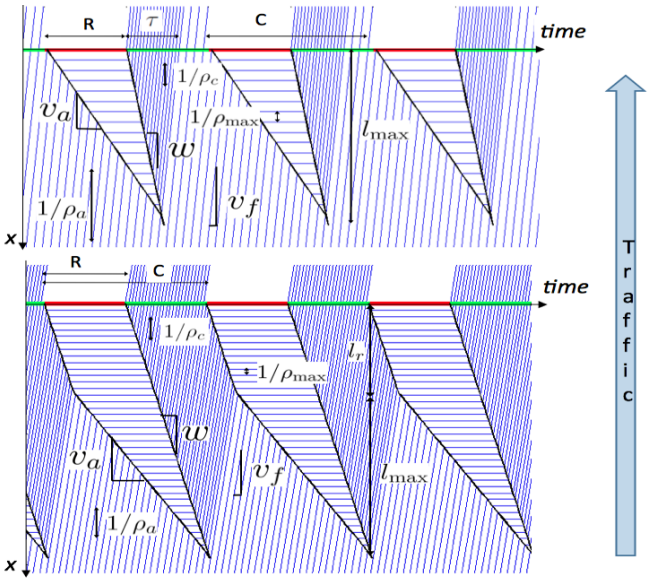

Fig. 1. Space time diagram of vehicle trajectories with uniform arrivals under an undersaturated traffic regime (top) and a congested traffic regime (bottom).

B. Probability distribution of vehicle locations

We define the average density $d(x)$ at location x as the temporal average of the density $\rho(x,t)$ at location x and time t.

$$d(x) = \frac{1}{T}\int_0^T \rho(x,t)\,dt$$

In practice, flow is never perfectly periodic, but we will assume that the above averaging over a duration T is a good proxy as long as T is of higher order of magnitude than the cyclic dynamics imposed by the presence of a signal.

The density at location x and time t takes one of the three following values, denoted $(\rho_i)_{1\leq i\leq 3}$: ρ_1 is the density in the queue, *i.e.* the maximum density $\rho_{\max}$; ρ_2 is the density in the queue release, *i.e.* the critical density ρ_c and ρ_3 is the arrival density ρ_a. The average density at location x is $d(x) = \sum_{i=1}^{3}\beta_i(x)\rho_i$ where $\beta_i(x)$ represents the fraction of the cycle time during which density is equal to ρ_i at x. Upstream of the queue, the average density is the arrival density ρ_3. As the speed of formation and dissolution of the triangular queue are constant, the average density increases linearly in the triangular queue. Vehicles arrive and leave the remaining queue at the maximum flow, corresponding to the critical density. The speed of formation and dissolution of the remaining queue are both equal to w. The average density is constant upstream of the maximum queue length (see [8], [6] for details).

When vehicles are sampled uniformly in time, the pdf of observing a vehicle at location x is proportional to the average density $d(x)$, with the proportionality constant given by $Z = \int_0^L d(x)\,dx$. The shape of the distribution is fully determined by three independent parameters: the

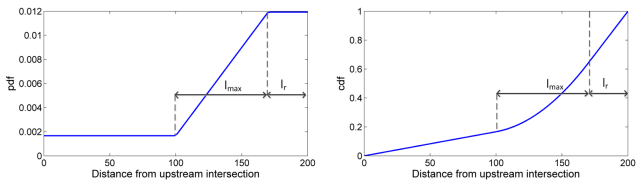

Fig. 2. Distribution of vehicle locations on the link as a function of the distance from the upstream intersection (left: pdf, right: cdf).

remaining queue length l_r, the triangular queue length $l_{\max}$ and the normalized arrival density $\tilde{\rho}_a = \rho_3/Z$. The pdf of vehicle location, denoted $\phi_L(x; \tilde{\rho}_a, l_{\max}, l_r)$, is illustrated in Figure 2, together with the cumulative distribution function (cdf). It can be derived explicitely (see [8], [6] for details):

$$\begin{aligned}
&\phi_L(x; \tilde{\rho}_a, l_{\max}, l_r) = \tilde{\rho}_a && \text{if } x \geq l_{\max} + l_r \\
&\phi_L(x; \tilde{\rho}_a, l_{\max}, l_r) = \tilde{\rho}_a + \tfrac{(l_r + l_{\max}) - x}{l_{\max}} \Delta_{\tilde{\rho}} && \text{if } x \in [l_r,\, l_{\max} + l_r] \\
&\phi_L(x; \tilde{\rho}_a, l_{\max}, l_r) = \tilde{\rho}_a + \Delta_{\tilde{\rho}} && \text{if } x \leq l_r
\end{aligned}$$

$$\text{with } \Delta_{\tilde{\rho}} = \frac{1 - \tilde{\rho}_a L}{l_{\max}/2 + l_r},$$

where L denotes the length of the link. The expression of $\Delta_{\tilde{\rho}}$ above, can be obtained easily by noticing that $\int_0^L \phi_L(x; \tilde{\rho}_a, l_{\max}, l_r)\, dx = 1$.

C. Density estimation

The parameters $\tilde{\rho}_a$, $l_{\max}$ and l_r are learned by maximizing the likelihood of the set of location observations (denoted $(x_o)_{o \in O}$) collected from probe vehicles:

$$\begin{aligned}
\underset{\tilde{\rho}_a, l_r, l_{\max}}{\mathbf{max}} \cdot \quad & \sum_{o \in O} \ln(\phi_L(x_o; \tilde{\rho}_a, l_{\max}, l_r)) \\
\mathbf{s.t.} \quad & 0 \leq \tilde{\rho}_a \leq \tfrac{1}{L}, 0 \leq l_r,\, 0 \leq l_{\max}, \\
& l_r + l_{\max} \leq L
\end{aligned} \tag{1}$$

The constraints are derived from the flow modeling: (a) the arrival density is inferior to the average density on the link, (b) the total queue cannot extend beyond the length of the link and (c) the triangular queue and the remaining queue must be non-negative. The constraints on the queue lengths do not limit the generality of the model. Under spill-over conditions (queue length extending beyond the upstream intersection), we consider that the queue length extends up to the upstream intersection, the rest of the queue is accounted for in the upstream links. Bounds on the parameters can be added to limit the feasible set and ensure the quality of the results when little data is available.

Remark 2: The objective function is not concave in the optimization variables. However, the search space is limited (three bounded parameters) and we perform a grid search followed by a local gradient ascent for the B best solutions of the grid-search. We found that a very fine grid was not necessary to provide good results, leading to efficient computation times.

III. Automatic signal detection

We present two algorithms based on the statistical model derived in Section II to automatically identify the presence (resp. absence) of traffic signals at intersections with model selection criteria.

A. Detection based on a single link

The algorithm independently classifies each link of the network as having a traffic signal at the downstream intersection or not. Under the hypothesis that there is a traffic signal, we learn the parameters of the probability distribution by solving (1) from location measurements x_o sent by sparsely sampled probe vehicles. For link i with length L_i and learned parameters $\tilde{\rho}_a^i, l_{\max}^i, l_r^i$, the learned distribution of measurements is denoted

$$\varphi_i^{\text{sig}}(x) = \Phi_{L_i}(x; \tilde{\rho}_a, l_{\max}, l_r).$$

Under the hypothesis that there is no traffic signal, the distribution of measurements is expected to be uniform on the link and we denote

$$\varphi_i^{\text{no sig}}(x) = \frac{1}{L_i} \mathbf{1}_{[0, L_i]}(x),$$

where $\mathbf{1}_{[0,L_i]}$ is the indicator function of interval $[0, L_i]$.

B. Detection based on two consecutive links

Another possible approach to the signal detection is the following. We consider two consecutive links i and j (with link i upstream of link j), connected by an intersection. We are interested in the detection of a signal at the downstream intersection of link i. According to the density modeling of Section II, we train two models representing the pdf of vehicle location under the assumption that there is a signal or that there is no signal at the downstream intersection of link i.

Under the assumption that there is a signal at the downstream intersection of link i, we learn the parameters $\tilde{\rho}_a^i, l_{\max}^i, l_r^i$ using the measurements received on link i and the parameters $\tilde{\rho}_a^j, l_{\max}^j, l_r^j$ using the measurements received on link j. Let x denotes the distance from the upstream intersection of link i on the stretch of road (i, j). The parameters of link i (resp. link j) characterize the shape of the distribution of measurements for $x \leq L_i$ (resp. $x \geq L_i$). The pdf of measurements on the stretch (i, j) is denoted $\psi_{i,j}^{\text{sig}}(x)$ and is given by

$$\begin{aligned}
\psi_{i,j}^{\text{sig}}(x) = {} & \alpha^{i,j} \Phi_{L_i}(x; \tilde{\rho}_a^i, l_{\max}^i, l_r^i) \mathbf{1}_{[0, L_i]}(x) \\
& + (1 - \alpha^{i,j}) \Phi_{L_j}(x - L_i; \tilde{\rho}_a^j, l_{\max}^j, l_r^j) \mathbf{1}_{[L_i, L_i + L_j]}(x)
\end{aligned}$$

The parameter $\alpha^{i,j}$ represents the relative weight of the measurements on links i and j. The maximization of the likelihood with respect to $\alpha^{i,j}$ leads to a closed form formula for $\alpha^{i,j}$: the number of measurements received on link i over the number of measurements received on links i and j combined.

Under the assumption that there is no signal at the downstream intersection of link i, we consider links i and j as a

single link of length $L_i + L_j$. We learn the parameters of the distribution of vehicle on this link using the measurements received on both link i and link j. The parameters are denoted $\tilde{\rho}_a^{i,j}$, $l_{\max}^{i,j}$ and $l_r^{i,j}$ and the corresponding pdf is denoted $\psi_{i,j}^{\text{no sig}}(x)$ and given by

$$\psi_{i,j}^{\text{no sig}}(x) = \Phi_{L_i+L_j}(x; \tilde{\rho}_a^{i,j}, l_{\max}^{i,j}, l_r^{i,j})$$

C. Model selection

Whether we consider the distribution of vehicles on a single link i or on two consecutive links i and j, the signal detection problem amounts to a model selection problem.

For the single link approach, we decide whether φ_i^{sig} or $\varphi_i^{\text{no sig}}$ represents the distribution of measurements most accurately. We learn the parameters characterizing φ_i^{sig} using the probe vehicle data received on link i. The model $\varphi_i^{\text{no sig}}$ does not require fitting as it corresponds to a uniform distribution of the measurements over the length of the link. To select the model which explains the best the data, we introduce model selection criteria as described below.

We are interested in choosing between models of different complexity. Note that by model complexity, we refer to the number of parameters required to specify the model. Various criteria have been developed to trade-off between model fit and model complexity. In general, increasing the model complexity leads to a better fit, and thus a higher likelihood but may overfit the available data. In the present article, we use information criteria which *penalize* the number of parameters to compare models with different numbers of parameters.

Remark 3: The uniform distribution is a special case of the parametric distribution function $\Phi_{L_i}(x; \tilde{\rho}_a^i, l_{\max}^i, l_r^i)$ for which $\tilde{\rho}_a^i = 1/L_i$. The parameters $\tilde{\rho}_a^i, l_{\max}^i, l_r^i$ are chosen to optimize the likelihood of the training data under the distribution φ_i^{sig}. The likelihood score for this distribution will necessarily be higher than under the distribution $\varphi_i^{\text{no sig}}$. To choose between the models, it is necessary to account for the number of parameters of each of the proposed models.

For the two links modeling approach, both models require fitting based on the available probe measurements. The model $\psi_{i,j}^{\text{sig}}$ has a higher number of parameters than the model $\psi_{i,j}^{\text{sig}}$. The detection of the presence of a traffic signal requires the use of appropriate model selection criteria.

The selection capabilities of three model selection criteria are compared in the experiments: the *Aikaine Information Criterion* (AIC), its *correction* for finite sample sizes (AICc) and the *Bayesian Information Criterion*. The model selection capabilities of the AIC and the AICc have theoretical motivations from Information theory [1], [4], whereas the derivations of the BIC arise from Bayesian statistics [16]. The different criteria have an analytical expression given by:

$$\begin{aligned} \text{AIC} &= -2\ln(\Lambda) + 2p, \\ \text{AICc} &= -2\ln(\Lambda) + 2p\tfrac{n}{n-p-1}, \\ \text{BIC} &= -2\ln(\Lambda) + p\ln(n), \end{aligned}$$

where Λ is the likelihood of the estimated model, p is the number of model parameters and n is the data size. All the criteria consist of the sum of the opposite of the log-likelihood and a penalization term which depends on the complexity of the model (number of parameters p) and on the size n of the dataset used to train the model.

For the one link approach, the parameters of the model with signal are $\tilde{\rho}_a^i$, $l_{\max}^i$ and l_r^i and thus $p = 3$. The model without signal does not have free parameters: the two parameters of the uniform distribution are 0 and the length of the link and are not set based on the data. As for the two link approach, the parameters of the model with signal are $\tilde{\rho}_a^i$, $l_{\max}^i$, l_r^i, $\tilde{\rho}_a^j$, $l_{\max}^j$, l_r^j and $\alpha^{i,j}$, and thus $p = 7$. The model without signal is parameterized by $\tilde{\rho}_a^{i,j}$, $l_{\max}^{i,j}$ and $l_r^{i,j}$ and thus $p = 3$

IV. Results

A. Experimental setup

We apply the signal detection algorithm using data collected by the *Mobile Millennium* [2] system in the Bay Area of San Francisco, CA. The system collects several millions GPS data points per day from probe vehicles reporting their location at a given sampling frequency (typically about every minute). The data used for the specific study presented below comes from a sub-fleet of around 500 probe vehicles collected on Tuesdays from 6 am to 10 am.

Remark 4: Numerical experiments have shown that the assumption of uniform arrival rates and periodicity (Section II) does not limit the decision capabilities of the algorithm. For signal detection, the most important feature is the detection of a queue which characterizes the presence of a traffic signal (traffic light or stop sign). The numerical analysis presented in the present article was also performed on data collected during 15 consecutive days (all times of day from January 1st, 2011 to January 15th, 2011) with very similar conclusions.

B. Automatic signal detection

We filter the GPS measurements and match them on the road network using a map-matching and path inference algorithm [10] which combines models of GPS measurements and drivers' behavior into a conditional random field. The road network *geometry* is given by the NAVTEQ digital map. The derivations of Section II underline that the most important feature of the pdf of measurements is characterized by the presence (resp. absence) of a queue due to the presence (resp. absence) of a signal. The presence of a queue is characteristic of any signalized intersection. So far, the noise in the data and the temporal aggregation of the data prevent the model from distinguishing between traffic lights and stop signs, which both induce queue and delay at the end of the link. We classify the links according to whether they have a signal (light or stop) or not at the downstream intersection. We study a sub-network of San Francisco, CA of 1,172 links. The percentage of signalized links, as indicated in the NAVTEQ database, is 54%. For each link of the network, we use the classification algorithm to identify the presence of a traffic signal. The decision of the algorithm is then

compared with the signal label available in the NAVTEQ database (Quarter 3, 2008).

In Section III, we investigated two different approaches to identify the presence of a signal using data collected on the link or on the link and the downstream consecutive link. We also suggested different model selection criteria to classify each link as being signalized or not. The results of all the proposed approaches are summarized in Table I.

Confusion matrices are 2 by n matrices presenting the counts of good and bad classifications for each class of decision. In the present case, $n = 2$ and the two pobbible decisions are signalized or non-signalized link. At first, the results look disappointing with only around 70% of correct classification. Furthermore, the results show an unexpectedly high false-positive rate (e.g. 30.7% for BIC using two links), *i.e.* a significantly large number of links for which the algorithm detects the presence of a signal while the database does not indicate the presence of a signal. The performance of the one link models and the two links models are similar, and so are the performance of AIC and AICc. We expected the results of the AIC and the AICc to be similar as the difference between the criteria tends to zero as the sample size increases and most links received several hundreds of measurements. The main differences are observed between the AIC type criterion (AIC and AICc) and the BIC criterion. The AIC approach leads to higher false positive rates, which is not unexpected as BIC penalizes more the complex models than AIC does.

The results reported in Table I are based on the validation against a GIS which is also prone to errors. The confusion matrices presented in Table I represent a quantitative comparison between the GIS labels (which contain some label noise) and the labels provided by the algorithm for each of the proposed approach (model with one or two links and different model selection criteria). The analysis of the label noise of the GIS is presented in the following section and underline the importance of the automatic labeling approach.

C. GIS cleaning and update

We represented the empirical distribution of the measurements and the models corresponding to the hypothesis that the link was signalized or not. This qualitative analysis of the results seemed to indicate that, for a large number of the false-positive decisions, the model which was estimated under the assumption that the link was signalized fit the empirical data very accurately, whereas the model corresponding to the assumption of a non-signalized intersection did not capture the shape of the distribution. Figure 3 illustrates the fitting results on such a link, which the algorithm classified as signalized whereas the database did not indicate the presence of a signal. Figure 3 (left) represents the results obtained with the one-link approach and Figure 3 (right) corresponds to the two-link approach. We validated our intuition that a signal was actually present using *Google Street View*.

From the realization that some of the signalized intersections were not present in the map database, we decided to

TABLE I

CONFUSION MATRIX BETWEEN PREDICTION AND GIS INFORMATION FOR THE TWO APPROACHES (ONE LINK OR TWO LINKS) AND THE DIFFERENT MODEL SELECTION CRITERIA.

AIC - one link		Prediction	
		Signal	No signal
Actual	Signal	508	125
	No signal	241	297
AICc - one link			
Actual	Signal	506	127
	No signal	237	301
BIC - one link			
Actual	Signal	384	249
	No signal	149	389
AIC - two links			
Actual	Signal	551	82
	No signal	276	262
ACCc - two links			
Actual	Signal	548	85
	No signal	269	269
BIC - two links			
Actual	Signal	429	204
	No signal	165	373

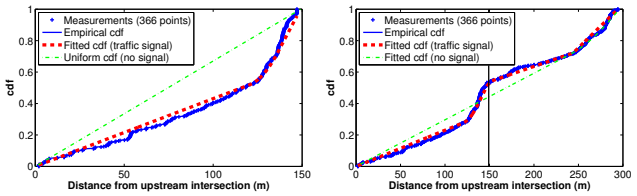

Fig. 3. Pdf of measurement locations. The models learn the parameters of the distribution under the assumption that there is a signal at the end of the link or not. **Left**: One link model. **Right**: Two links model. The position of the intersection of the two links is depicted by the vertical black line. For both the one link and the two links models, all information criteria accurately classify the link as having a traffic signal at the downstream intersection.

do manual checks on the links which were classified (using BIC on two links) as being signalized when the database indicated no signal (false-positives). The manual checks were performed using *Google Street View*. Among the 165 false-positives of the BIC-two links algorithm, 40% actually had traffic signals, 17% had stop signs. More than half of the false positives are in reality true positives. Among the 71 remaining links, 17% had specific features that explain the false-detection such as (a) presence of a pedestrian crossing at the downstream end of the link, even though there is no *actual* signal, (b) complicated intersections or (c) unnels in which GPS reception and cellular communications are nonexistent or inaccurate. We illustrate some of these cases (pedestrian crossing and complicated intersection) in Figure 4. The authors acknowledge that a manual labelling of the entire database would lead to stronger results but this tedious process has not been performed so far.

V. CONCLUSION

The approach proposed in this article leverages the physics of traffic to derive a statistical model representing the distribution of vehicles on a link, depending on the presence

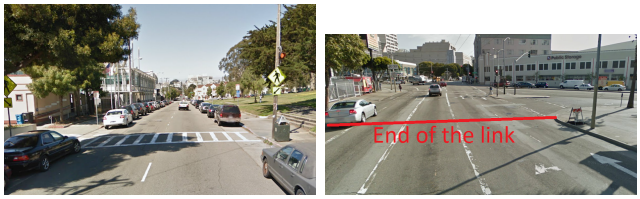

Fig. 4. Illustration of the features of the downstream intersection of some links detected as being signalized by the algorithm (false-positives).

(resp. the absence) of signalization. The model parameters can be estimated using sparsely sampled probe vehicle data which makes it very promising given the emergence of this data at a large scale. The method is a first step towards automate GIS updates for signal location. The algorithm produces interesting result (more than half of the automatic signal detections that were not recorded in the GIS database correspond to real stop or traffic lights). Furthermore, experiments with one week coverage data, were also satisfactory enabling a periodic use of the solution to update and correct a GIS, in particular in areas where the infrastructure evolves rapidly such as developing countries.

This first step towards automate GIS updates and cleaning has the prospect to be improved and generalized by taking into account additional features. For example, the graph structure of the network structure can be leveraged to improve the decision results *e.g.* at an intersection with a light, all links are signalized and therefore a global decision by intersection (and not by link) should improve the robustness of the decision. Another possible extension of the methodology regards the discrimination of stop signs from traffic lights. Other information derived from sparsely sampled probe vehicle data such as travel times should be a mean to perform this discrimination more accurately than the average density of measurements.

VI. Acknowledgments

The authors wish to thank Timothy Hunter from UC Berkeley for providing filtered and map-matched locations of probe vehicles from the raw GPS measurements. We are grateful to Xavier Louis and Nadir Farhi from the GRETTIA (IFSTTAR, France) for their valuable input and discussions. We thank the staff from the California Partners for Advanced Transportation TecHnology (PATH) for their contributions to develop, build, and deploy the system infrastructure of *Mobile Millennium* on which this article relies. This research was supported by the Federal and California DOTs and NAVTEQ Inc.

References

[1] H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974.

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

[3] R. Bruntrup, S. Edelkamp, S. Jabbar, and B. Scholz. Incremental map generation with GPS traces. In *IEEE Intelligent Transportation Systems Conference*, pages 574 – 579, September 2005.

[4] K.P. Burnham and D.R. Anderson. *Model selection and multimodel inference: a practical information-theoretic approach*. Springer Verlag, 2002.

[5] R. De Charette and F. Nashashibi. Traffic light recognition using image processing compared to learning processes. In *2009 IEEE International Conference on Intelligent Robots and Systems (IROS)*, pages 333–338, 2009.

[6] R. Herring, A. Hofleitner, P. Abbeel, and A. Bayen. Estimating arterial traffic conditions using sparse probe data. In *Proceedings of the 13th International IEEE Conference on Intelligent Transportation Systems*, pages 929–936, Madeira, Portugal, September 2010.

[7] A. Hofleitner, R. Herring, and A. Bayen. A hydrodynamic theory based statistical model of arterial traffic. *Technical Report UC Berkeley, UCB-ITS-CWP-2011-2*, January 2011.

[8] A. Hofleitner, R. Herring, and A. Bayen. Learning the dynamics of arterial traffic from probe data using a dynamic bayesian network. *IEEE Transactions on Intelligent Transportation Systems*, 2012.

[9] B. Hull, V. Bychkovsky, Y. Zhang, K. Chen, M. Goraczko, A. Miu, E. Shih, H. Balakrishnan, and S. Madden. Cartel: a distributed mobile sensor computing system. In *SenSys '06: Proceedings of the 4th international conference on Embedded networked sensor systems*, pages 125–138, New York, NY, USA, 2006. ACM.

[10] T. Hunter, T. Moldovan, M. Zaharia, S. Merzgui, J. Ma, M.J. Franklin, P. Abbeel, and A.M. Bayen. Scaling the mobile millennium system in the cloud. In *Proceedings of the 2nd ACM Symposium on Cloud Computing*, pages 28:1–28:8. ACM, 2011.

[11] M. Lighthill and G. Whitham. 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(1178):317–345, May 1955.

[12] A.-S. Puthon, F. Nashashibi, and B. Bradai. A complete system to determine the speed limit by fusing a GIS and a camera. In *14th International IEEE Conference on Intelligent Transportation Systems (ITSC)*, pages 1686 – 1691, oct. 2011.

[13] P. Richards. Shock waves on the highway. *Operations Research*, 4(1):42–51, February 1956.

[14] S. Rogers, P. Langley, and C. Wilson. Mining gps data to augment road models. In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '99, pages 104–113. ACM, 1999.

[15] S. Samaranayake, S. Blandin, and A. Bayen. A tractable class of algorithms for reliable routing in stochastic networks. *Transportation Research Part C*, 20(1):199 – 217, 2012.

[16] G. Schwarz. Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464, 1978.