Vehicle Trajectory Reconstruction Using Conditional Random Fields

Deepthi Mary Dilip and Saif Eddin Jabari

This is the author's version of a work that was presented at the 97th Annual Meeting of the Transportation Research Board, Washington D.C., 2018. No. 18-03053 TRID Record

VEHICLE TRAJECTORY RECONSTRUCTION USING CONDITIONAL RANDOM FIELDS

Deepthi Mary Dilip, Corresponding Author
Division of Engineering
New York University Abu Dhabi
Saadiyat Island, P.O. Box 129188
Abu Dhabi, United Arab Emirates
Tel: +971-2-628-4558; Email: dmd15@nyu.edu

Saif Eddin Jabari
Division of Engineering
New York University Abu Dhabi
Saadiyat Island, P.O. Box 129188
Abu Dhabi, United Arab Emirates
Tel: +971-2-628-4823; Email: sej7@nyu.edu

Word count: 5000 words text + 10 tables/figures x 250 words (each) = 7500 words

July 31, 2017

ABSTRACT

This paper presents a probabilistic approach to reconstruct vehicle trajectories from GPS probe data on arterials. By combining car-following concepts with machine learning algorithms, we overcome the drawbacks of pure statistical modeling to investigate the question of adequate probe penetration levels on single-lane roads. Although the parameters of the traffic state estimation model are learned from historical data, the proposed algorithm is found to be robust to unpredictable conditions. The estimation algorithm is tested using a vehicle trajectory dataset generated using microsimulation software. The results highlight the need to take into account the randomness of the spatio-temporal coverage associated with probe data for reliable state estimation algorithms.

Keywords: Conditional Random Fields, Trajectory Reconstruction, Probe Vehicles, Cellular Automata

INTRODUCTION

As connected and autonomous vehicles begin to penetrate vehicle fleets throughout the world, probe vehicles become a valuable source of real-time traffic information. Probe vehicles act as mobile data sensors by continuously broadcasting their position and speed in real-time, providing Lagrangian data measurements. Fused with stationary sensing data obtained from traditional monitoring devices such as inductive-loop detectors, comprehensive datasets are obtained for traffic monitoring and state estimation (*1, 2, 3*). In urban road networks, where the deployment of stationary detectors is usually limited and traffic lights govern the link dynamics, a higher number of probes may be necessary to accurately characterize traffic conditions. Motivated by the wide spatio-temporal coverage offered by fused traffic data, we address the adequate levels of probe penetration at a microscopic scale in this paper, focusing on the reconstruction of vehicle trajectories over a single arterial roadway.

A number of modeling techniques have been proposed in the recent years, to estimate traffic flows, densities (*4*), speeds (*5*), travel times (*6*), and travel time distributions (*7,8*) from vehicular sensor data. These techniques have been formulated either using traffic flow theory in a model-driven approach (*9,10,11*) or historical traffic patterns in a data-driven approach (*12*). To account for the variability of arterial traffic, a statistical approach using Coupled Hidden Markov Models was proposed by Herring et al. (*13*) to estimate the traffic state from sparse probe data. The limitations of purely statistical approaches were overcome by Hofleitner et al. (*14*), where a hybrid modeling framework combining machine learning with the hydrodynamic traffic flow theory was proposed, to predict arterial travel times from streaming GPS probe data. On the other hand, Papathanasopoulou and Antoniou (*15*), proposed a data-driven car-following model to capture longitudinal interaction among vehicles. We propose a probabilistic approach for the spatio-temporal reconstruction of the traffic state from sparse probe data, wherein the traffic patterns are learned from historical data using Conditional Random Fields (CRFs). By modeling the vehicle interaction potential to reflect the local traffic information (such as spacing), our estimation models seamlessly combine the heuristic car-following model theory (*16*) with statistical patterns to capture the microscopic traffic dynamics.

Research in the field of traffic state estimation from probe data has been focused on network modeling and the reconstruction of traffic states on missing road links (*17*). At a finer scale, Herrera and Bayen (*9*) reconstructed the traffic density on a freeway section by modifying the LWR PDE with a correction term to nudge the model estimate towards the GPS probe measurements. The techniques proposed did not require the knowledge of on and off-ramp detector counts for the density estimation. The tradeoff between probe vehicle and inductive loop velocity data was studied by Mazaré et al. (*18*) (with the goal of predicting travel times on a roadway stretch), who acknowledged the inherent difficulty of specifying, a priori, the probe penetration rates which are dictated by the total traffic flow. Moreover, if the probe sampling requirements are not adequately met due to technical or privacy issues, the observed probe data may be sparse and non-uniformly distributed. Taking into account this randomness of the spatio-temporal coverage of probe vehicles, we investigate the following question in this paper: 'What is the lowest probe penetration rate by which we can reliably capture the traffic dynamics on a single-lane road link?'.

The remainder of the paper is organized as follows: in Section 2, we introduce car-following rules that provide the framework for the graphical modeling approach to vehicle trajectory estimation. In this section, the CRF model that predicts the traffic state on a single-lane road from probe vehicle data is described, and validated by testing the model capability to identify unforseen incidents. Having validated the model, in Section 3 we implement the

Markovian approach for trajectory reconstruction and analyse the impact of the probe vehicle distribution on the estimation. The conclusions and future scope of the work are presented in Section 4.

PROBABILISTIC MODELING FRAMEWORK

The spatio-temporal reconstruction of all vehicle trajectories on a single-lane road link is formulated as a discrete Markov process, modeling the microscopic traffic dynamics using Cellular Automata (CA); see (*19*). CA models are discrete mathematical models of the microscopic dynamics, where vehicle movement is governed by an interaction potential that describes the ("energy profile of") local traffic conditions. These models have been employed for to study interesting traffic phenomena like 'synchronized' traffic at ramps and 'stop-and-go' regimes (*20*). Given the initial and boundary conditions, CA models update the traffic state in discrete time-steps, based on the past state through the potential function. The aim of this study is to exploit the information provided by the probe vehicles, by capturing the spatial dependencies between successive vehicles through a CRF model, which is based on the assumption that the speed of any vehicle α is influenced by its leader. Using appropriate probabilistic inference methods (*21,22*), and modeling the non-probe (or non-instrumented) vehicles as 'hidden', the CRF model predicts the velocity field at every discrete time step. Thus, the traffic state update is carried out sequentially, by augmenting the past (temporal) information (provided by the previously estimated traffic state) with the spatial dependencies (provided by probe vehicle information) in the current time step.

CA Model for Traffic State Update

A single-lane roadway is modeled as a one-dimensional uniform lattice L. The spatial coordinates of each vehicle α on the roadway is discretized such that each cell can be occupied by at most one vehicle, which is achieved by setting the cell length to an appropriate value, e.g., 7.5 m (*23*). The state of each occupied cell at a discrete time k is completely specified by the discretized velocity v_α^k, which can take integer values between 1 and v_{max}, where v_{max} is the maximum number of cells that can be crossed in one time-step. Thus, an order parameter $\sigma^k(l) \in \{0, 1, 2, \ldots, v_{max}\}$ can be defined for each cell $l \in L$ at time k, where 0 represents free cells. The traffic state update is traced in discrete time steps to determine the $\sigma^{k+1}(l)$ according to the update rules in Algorithm 1 below. The vehicle-interaction potential is modeled to capture response of a driver as a function of speed and spacing to the (lead) vehicle ahead. Hence, the state (velocity) of the vehicle α at time-step $k+1$ is a function of the gap g_α^k to the lead vehicle, the vehicle speed v_α^k and the speed of the leader $v_{\alpha+1}^k$ in the previous time-step, k.

Graphical Modeling Approach

Let $X = \{X_s \mid s \in N\}$ be a discrete valued random field with probability mass function (pmf), $p(x)$ defined on N random variables. X is defined as a Markov Random Field (MRF) if it satisfies the Markovian property that, for all $s \in N$,

$$P(X_s = x_s | X_t, t \neq s) = P(X_s = x_s | X_{N_s}), \tag{1}$$

where N_s denotes the neighbours of s. These (conditional) independence assumptions between the variables X_s can be encoded by a graph $G = (V, E)$ where X is indexed by the vertices V such that $X = (X_s)_{s \in V}$ and edges $E \in V \times V$. Defining the vehicles on the roadway at any given time as V and encoding the spatial dependencies in the velocity through edges E (represented by the bold lines in Figure 1) between successive vehicles, the condition in Equation 1 implies that

ALGORITHM 1 Car-following rules for CA model

Input
Length of road (cells) - N, Total simulation time - T, time step ∂k,
Discretized velocity $v_i \in \{1,..,v_{max}\}$,
Interaction potential parameters, $\boldsymbol{\lambda} = [\lambda_1,..,\lambda_i,..\lambda_{\boldsymbol{K}}]$, $\boldsymbol{K} = |\{1,...,v_{max}\}|$
Initialize
Initial traffic state $\sigma^0(l)$, Arrival density - p_1, Probability of Slow-down - p_2
Define
$\psi_i^{k+1} \equiv \mathrm{P}\,(v_\alpha^{k+1} = v_i)$ as the probability of assuming velocity state v_i in time $k+1$
Iterate
Compute for each α, car-following input set}, $\boldsymbol{Y} = [v_\alpha^k \; v_{\alpha+1}^k \; g_\alpha^k]$
Velocity Update
(Unnormalized) probability, $\hat{\psi}_i^{k+1} = e^{\boldsymbol{Y}\cdot\lambda_i}$
Sample v_α^{k+1}according to normalized potential, ψ_i^{k+1}
$u_1 \sim$ Uniform(0,1)
IF $u_1 < p_2$
$v_\alpha^{k+1} = v_\alpha^{k+1} - 1$
END IF
Position Update
Compute vehicle positions s_α^{k+1} in succession, moving in the upstream direction
$s_\alpha^{k+1} = \min\,(\max(s_\alpha^k, s_\alpha^k + v_\alpha^{k+1})\,, s_{\alpha+1}^{k+1} - 1)$ Ensures forward movement without overtaking
Traffic State Update
$\sigma^{k+1}(s_\alpha^{k+1}) = v_\alpha^{k+1}$
Boundary Conditions
$u_2 \sim$ Uniform(0,1)
IF $u_2 < p_1$
$\sigma^{k+1}(1) = v_i$ New vehicle enters with random velocity v_i
END IF

---

the velocity of any vehicle is independent of the traffic state $\sigma(l)$ given the local velocity field. We employ a first order Markov model with the assumption that a vehicle response is influenced by only the leader vehicle. The MRF model with a chain-like structure, employed to predict the velocity field at time $k+1$ given the probe vehicle velocities, is depicted in Figure 1. The dependence of the traffic state on the past can be modeled in two ways (a) through directed temporal edges between $\sigma^k$and σ^{k+1} or (b) by setting $Y = f(\sigma^{k+1})$ as an input feature and conditioning the MRF on Y. Adding temporal edges results in a loopy Markov network with directed and undirected edges, increasing the model complexity. On the other hand, the second approach extends the (unconditional) MRF model to a linear-chain CRF model (*24*), by conditioning the vehicle-interaction potential on an input feature space. By suitably defining the feature set Y, the CRF model has the flexibility to capture the response of a vehicle to its local traffic conditions, as in the CA model. Formally in the CRF model the temporal dynamics are captured by the node (association) potential, Ψ_s, representing the probability of each node (or vehicle) assuming a particular state, and Ψ_{st} is the interaction (or edge) potential that represents the dependencies between neighboring vehicles. Defining single and pairwise cliques (subset of

V that are mutually adjacent) over the node and edges respectively, the conditional PMF over the chain-graph G is

$$p(\boldsymbol{x}|\boldsymbol{y}) = \frac{1}{Z(\boldsymbol{y},\Theta)} e^{\left(\sum_{s\in N} \Psi_s(x_s,\boldsymbol{y},\Theta) + \sum_{s\in N}\sum_{t\in N_s} \Psi_{st}(x_s,x_t,x_s,\Theta)\right)}, \tag{2}$$

where Θ is a parameter vector and Z is a normalization term.

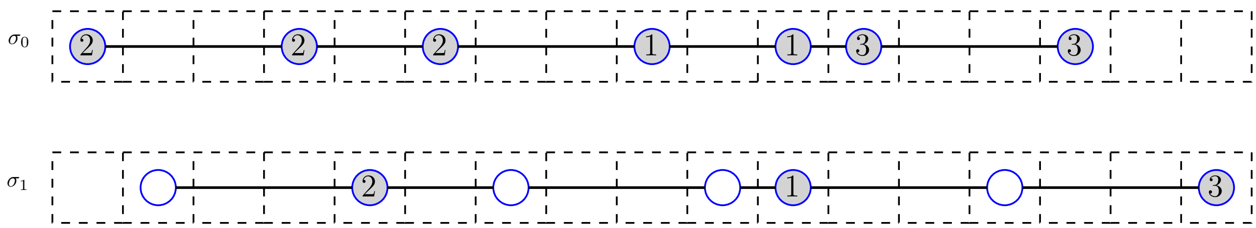


FIGURE 1 Markov Chain for Traffic State Update (The filled (grey) circles are random variables corresponding to the probe vehicles X_p while the clear circles represent the hidden variables)

Formulation of Potentials

The spatio-temporal evolution of the velocity field is carried out sequentially by the CRF model in discrete time-steps. This is achieved by formulating the potential (node and interaction) functions at each time step $k+1$, given the temporal information in k and the spatial information from the probes in $k+1$ (Figure 2). As observed in Figure 1, the spatial dependencies between adjacent vehicles is encoded by an edge, implying that all successive vehicles are neighbours irrespective of the spacing between them. By modeling the spatial correlation in the speeds between two neighbouring nodes (vehicles) as a function of the gap between them (in the previous time), we can ensure that vehicles that are sufficiently separated will behave as free-flowing traffic. This is achieved by defining the edge feature $\boldsymbol{Y}^E = [g_\alpha^k, v_\alpha^d]$ where $v_\alpha^d = |v_{\alpha+1}^k - v_\alpha^k|$ is the absolute speed difference. For instance, assume that the velocity field is discretized into 3 states, i.e $v_\alpha \in \{1, 2, 3\}$. The edge potential is modeled as

$$\Psi_{\alpha,\alpha+1}^{k+1} = \begin{pmatrix} e^{\boldsymbol{Y}^E\cdot\boldsymbol{\theta}_{1,1}^E} & e^{\boldsymbol{Y}^E\cdot\boldsymbol{\theta}_{1,2}^E} & e^{\boldsymbol{Y}^E\cdot\boldsymbol{\theta}_{1,3}^E} \\ e^{\boldsymbol{Y}^E\cdot\boldsymbol{\theta}_{2,1}^E} & e^{\boldsymbol{Y}^E\cdot\boldsymbol{\theta}_{2,2}^E} & e^{\boldsymbol{Y}^E\cdot\boldsymbol{\theta}_{2,3}^E} \\ e^{\boldsymbol{Y}^E\cdot\boldsymbol{\theta}_{3,1}^E} & e^{\boldsymbol{Y}^E\cdot\boldsymbol{\theta}_{3,2}^E} & e^{\boldsymbol{Y}^E\cdot\boldsymbol{\theta}_{3,3}^E} \end{pmatrix}, \tag{3}$$

where $\boldsymbol{\theta}_{i,j}^E$ is the edge parameter set defining the spatial correlation between the speed states of $v_{\alpha 1} = i$ and $v_{\alpha 2} = j$. The potential function in Equation 3 is reminiscent of the Potts model (*25*) with expressive potentials, i.e. states are not interchangeable. In other words, CRF model is trained to learn that the response of a fast-moving vehicle to a slow leader (as a function of their spacing) will not be the same as that of a slow vehicle to fast leader.

To capture the temporal dependencies of v_α^{k+1} (as in the CA model), the feature set for the node potential is set to $\boldsymbol{Y}^V = [g_\alpha^k, v_\alpha^k, v_{\alpha+1}^k]$. Now the node potential Ψ_α^{k+1} can be expressed as

$$\Psi_\alpha^{k+1} = \left[e^{\boldsymbol{Y}^V\cdot\boldsymbol{\theta}_1^V} \quad e^{\boldsymbol{Y}^V\cdot\boldsymbol{\theta}_2^V} \quad e^{\boldsymbol{Y}^V\cdot\boldsymbol{\theta}_3^V}\right], \tag{4}$$

where $\boldsymbol{\theta}_i^V = \left[\theta_i^{f=1}\ \theta_i^{f=2}\ \theta_i^{f=3}\right]$ are the node parameters (for node α at $k+1$) corresponding to the f (node) features and i states. For both the node and edge potential, we include a feature that is always set to 1, introducing an intercept term to account for the probability of a vehicle assuming a particular velocity that is independent of the features.

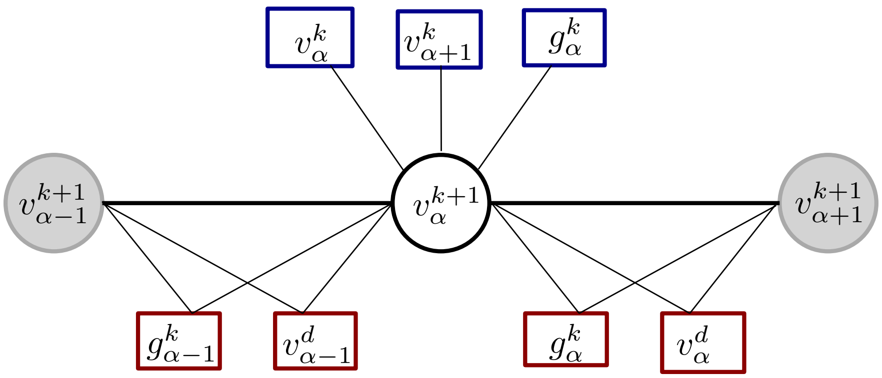


FIGURE 2 CRF Model (The circles represent the nodes, the blue rectangles (above) represent the node features for the hidden nodes and the red rectangles (below) represent the edge features)

CRF Model Inference

The CRF model is fully specified by its potential functions and the corresponding parameter vector Θ. Before carrying out the probabilistic inference of the hidden states, the CRF model is trained by estimating the parameters according to labeled data pairs $\mathcal{D} = \{\boldsymbol{y}_m, \boldsymbol{x}_m\}_{m=1}^M$ such that the loglikelihood $\sum_{m=1}^M \log p(\boldsymbol{x}|\boldsymbol{y})$ is maximized. Since the objective function is convex, any gradient-based optimization approach can be adopted for the maximum likelihood estimation of the parameters. Once trained, the models can be applied for two kinds of probabilistic inference problems; see (*21*) for details. Given a subset of known variables $\boldsymbol{x}_p$ (i.e probe vehicle velocity), infer (a) the marginal probabilities of the unknown variables $\boldsymbol{x}_h$ using the sum-product algorithm and (b) the most likely configuration of the states (MAP estimate) obtained by

$$x^* = \arg\max_x P(\boldsymbol{x}|\boldsymbol{y}). \tag{5}$$

For linear-chain CRFs and Markov chain models, the MAP inference can be efficiently performed by dynamic programming (the Viterbi algorithm) over the hidden variables $\boldsymbol{x}_h$.

Markov Chain Model for Position Update

In the car-following model (Algorithm 1), the new position of each vehicle $s_\alpha^{k+1}$was assumed to be a function only of v_α^{k+1}. However, an analysis of the groundtruth simulated to model real-world conditions (see Section below) indicates otherwise. Hence, to update the position of the vehicles in space in a more realistic setting, the CRF estimated velocity field is plugged into a simple Markov chain model (formulated at each time step k). In this Markov chain, the nodes represent all the lattice cells, while the state of the nodes denotes the number of cells moved by the vehicle (if present) in each time-step. Let $c \in \{-1,0,1,..,c_p,...,C\}$ be the set of states where $c = -1$ implies the absence of a vehicle in the cell and C is the maximum number of cells that can be crossed in one time-step. The value of C, which is determined by the maximum velocity and the cell size, defines the (asymmetric) neighbourhood system i.e the number of neighbouring cells (in the downstream direction) in the Markov chain. As with the speed, setting the features to $\boldsymbol{Y}^V = [g_\alpha^k, v_\alpha^k, v_{\alpha+1}^k]$, the node potential at time-step $k+1$ for cell s (if a vehicle is present) is modeled as

$$\Psi_s^{k+1} = \left[P(c=0)\ldots\quad P\left(c=c_p\right)\ldots\quad P(c=C)\right], \tag{6}$$

where the probability of each state c_p is calculated as

$$P\left(c=c_p\right) = \frac{1}{1+\sum_r \beta_r \cdot \boldsymbol{Y}^V}\,. \tag{7}$$

The regression coefficients β_r associated with the state c_p can be estimated using MLE techniques as in a multi-class logistic regression model. Rather than using a simple logistic regression model for the position update, we formulate a Markov chain model with exclusion rules (that discourage overtaking in a single-lane model) as well as to ensure that the position predicted by the model does not coincide with the (known) position occupied by the probe vehicle in the next time-step. This is achieved by appropriately modeling the edge potentials. For instance, assuming that a vehicle can only move 1 cell (or remain in its previous position), the edge potential between cells s and $s+1$ can be set to

$$\Psi_{s,s+1}^{k+1} = \begin{pmatrix} c_{\{-1,-1\}}=1 & c_{\{-1,0\}}=1 & c_{\{-1,1\}}=1 \\ c_{\{0,-1\}}=1 & c_{\{0,0\}}=1 & c_{\{0,1\}}=1 \\ c_{\{1,-1\}}=1 & c_{\{1,0\}}=0 & c_{\{1,1\}}=1 \end{pmatrix}, \tag{8}$$

where $c_{\{1,0\}}=0$ implies that if the leader vehicle advances $c_p=0$ cells (i.e., it remains in $s+1$) and a follower is present in s, the follower is restricted from moving $c_p=1$ cells.

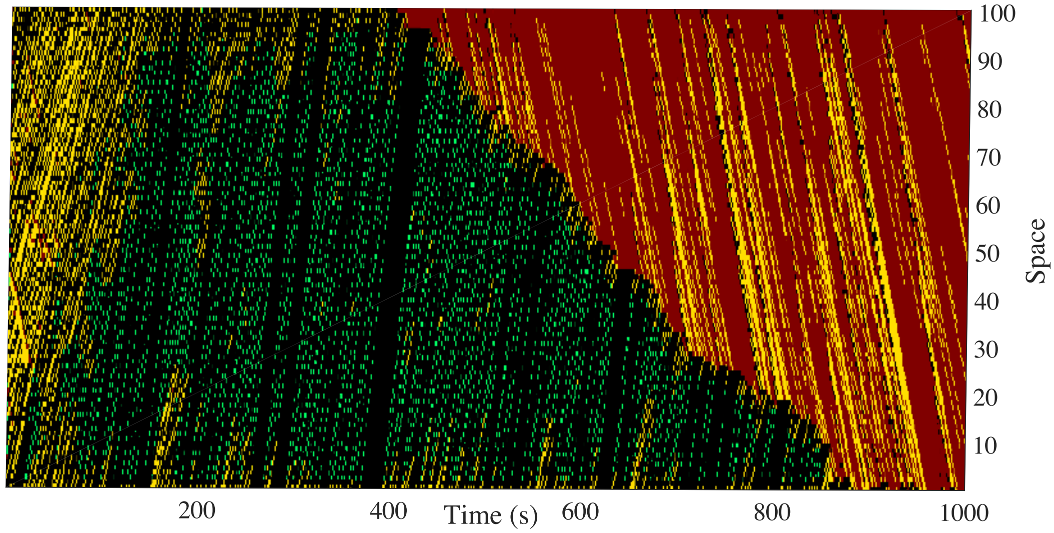
Model Testing and Validation

In this section, we test the CRF model as well as the car-following logic through a numerical example. The velocity is discretized into 3 states respresenting freeflow, synchronized (slow-moving) flow and congested conditions. The traffic state is simulated following the update rules in Algorithm 1; the parameters $\boldsymbol{\lambda}$ are assumed to have been learned from historical trajectory data. The initial distribution of vehicles i.e. σ^0, at time $k=0$ on the roadway is assumed to be completely known. At the upstream boundary, loop detectors provide information about the occupancy and speed of all upstream vehicles, as well as the entry times of new vehicles into arterial section under consideration. This implies that $\sigma^k(1)$ is known $\forall\, k$. We simulate the shockwaves generated in undersaturated stop-and-go conditions by appropriately setting $\sigma^k(N)$ to reflect the red and green signal cycles, i.e., we assume a traffic signal at the downstream end of the road section. By considering the simulated trajectory as our historical dataset, the training data pairs $\mathcal{D}=\{\boldsymbol{y}_m, \boldsymbol{x}_m\}_{m=1}^{M}$ were extracted for every pair of vehicle and leader at all times. For any given pair, α and $\alpha+1$, $\boldsymbol{x}=[v_\alpha^{k+1}]$ while $\boldsymbol{y}$ is the corresponding set of feature vector values.

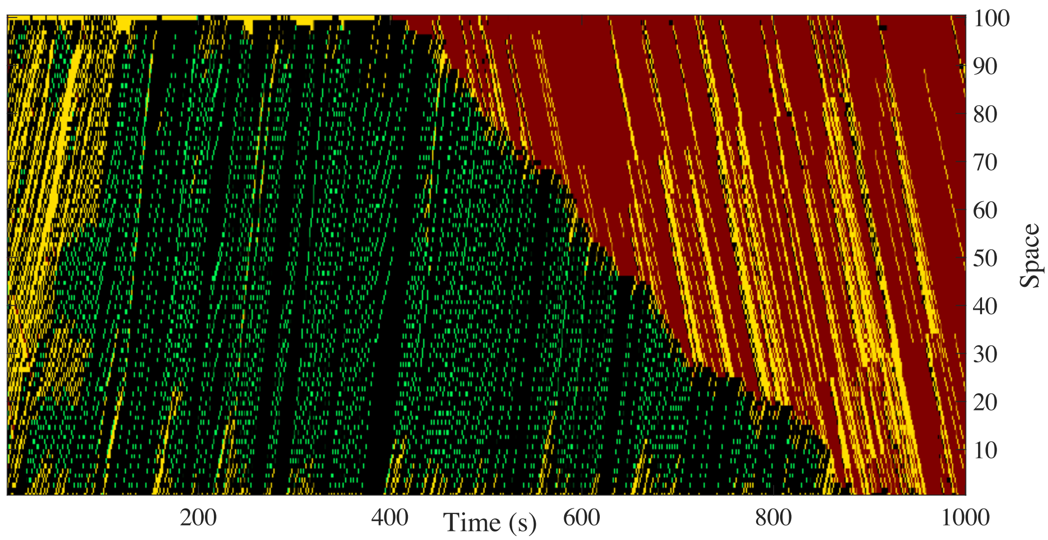
Validation of the CRF Model

In order to validate the model (with out-of-sample data), an incident is assumed to have occurred at the downstream boundary on the road section of length $= N$ cells and $p_1=0.25$. The simulated traffic state is shown in Figure 3a, which provides ground truth data for comparison with the traffic state predicted by the CRF model. A subset of all the simulated vehicles are now chosen randomly to represent the set of probe vehicles. For this study, periodic noise-free updates of the vehicle position s_α (spatial co-ordinates) and speed v_α (derived from successive GPS co-ordinates) are assumed to be available from the probe vehicles at intervals of $\partial k=1$s. The CRF model is used to estimate the velocity field sequentially at discrete time-steps (corresponding to the sampling interval of the probe vehicles), while the position is updated as in

Algorithm 1. The complete vehicle trajectories estimated is depicted in Figure 3b, indicating that a probe penetration rate of 10% is sufficient to capture the backward propagation of the shockwave generated by the incident located downstream.



(a) Groundtruth

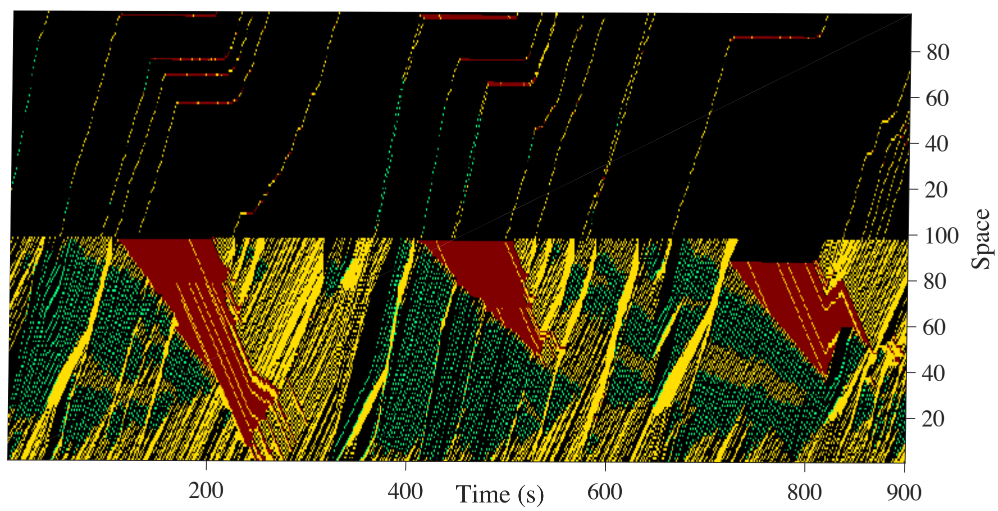


(b) Probe = 10%

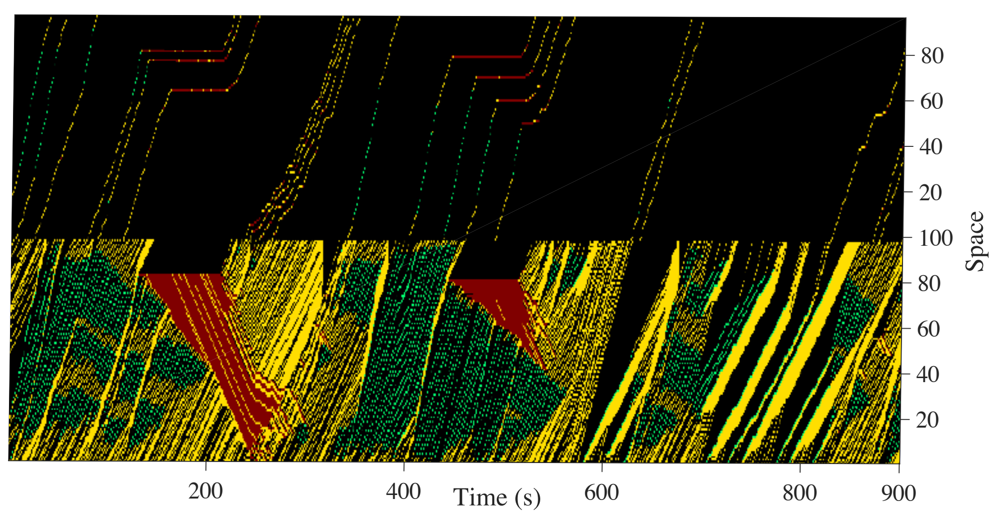
FIGURE 3 Validation of CRF Model: Time-Space Diagram (Velocity)

Randomness in Probe Coverage

As the evolution of the traffic flow is not dictated by the probe vehicles, it is nearly infeasible to select the subset of probe vehicles to be distributed evenly in time and space (*18*). We analyze the effect of randomly distributed probes on the trajectory reconstruction problem by comparing the estimated states for two random distributions of probes with a penetration rate of 5%. Figure 4 depicts the distribution of the randomly selected probes in the upper half and the corresponding estimation trajectories in the lower half. The results demonstrate the need to take into consideration the spatial distribution of the probes for the trajectory estimation problem. The groundtruth was simulated by assuming 3 signal cycles in the time-period of $T = 900s$, with a redtime of $100s$, (arbitarily) fixed a times $k = 100{,}400, 700$. While this information can be easily inferred from a probe level of 5% in Figure 4a, since none of the selected probes pass through the third signal cycle, the estimation algorithm fails to capture the build-up and dissipation of the shochwaves in the time period from 700 to $900s$ in Figure 4b.



(a)



(b)

FIGURE 4 Randomness in Probe Coverage: Time-Space Diagram (Velocity)

This example asserts that specifying a single probe penetration rate to capture the traffic dynamics can be quite misleading, when the unpredictiblity of the probe vehicle arrival times introduces randomness in its spatio-temporal coverage. Hence, the goal of this study can be stated as:

Given the initial state of all vehicles at time t^0, the boundary conditions $\sigma^k(1)$, and the probe vehicle states at all time steps $\{t^1, .., t^k, .. t^T\}$ (where T is a horizon time), determine the smallest probe penetration rate that can predict the trajectories throughout the time interval $[0, T]$, within a relative mean error, ε and with a (specificied) reliability level of r %.

EXPERIMENTAL SETUP

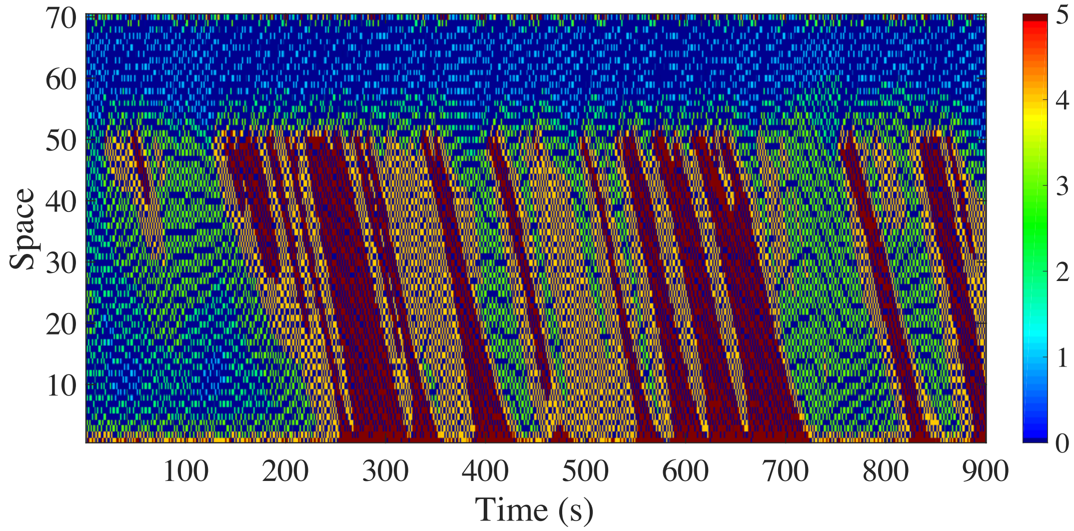
In this section, we simulate real-world conditions on a road section using data generated from microscopic traffic simulation. The simulations are run for 1 hour periods (from 8 am to 9 am) with a 15-minute warm-up period for an arterial link of about 500m length, with a on-ramp located at about 300m downstream. The free-flow speed is 60 km/hr and demand was gradually increased every 15-minute period, from 1200 veh/hr to 2500 veh/hr, to simulate build-up and dissipation of shockwaves. The continuous trajectory data is discretized by dividing the roadway into lattice cells of 7.5m (around 22 feet) in length. The speed, calculated as a difference quotient from the positions, is categorized into the following velocity ranges: < 35 km/hr, 35-50 km/hr, 50-60 km/hr, 60-70 km/hr and >70 km/hr.

Results and Analysis

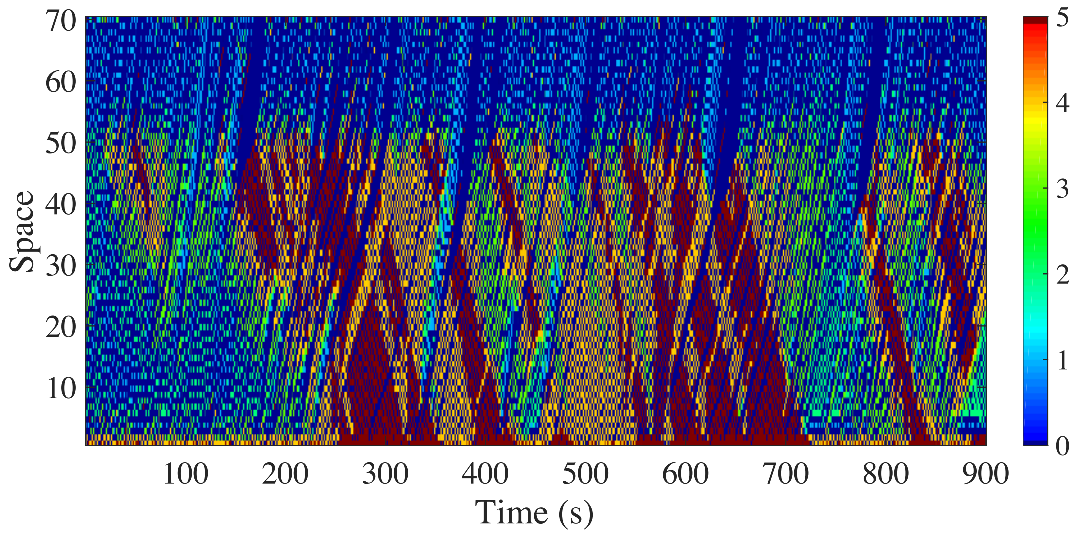
In this study, the performance measure used to investigate probe penetration is the mean absolute percent error (MAPE) defined as

$$\varepsilon_{MAPE} = \frac{1}{n}\sum_{j=1}^{n}\left|\frac{T_j - \hat{T}_j}{T_j}\right|, \quad (9)$$

where n is the total number of vehicles in the system (at time T), T_j and $\hat{T}_j$ are the actual travel times (computed from the ground-truth) and estimated travel times of the j^{th} vehicle, respectively. The traffic state (velocity) estimated for a time-period of $T = 15$ minutes in congested conditions is depicted in Figure 5b. A visual comparison with the ground truth in Figure 5a, as well as the MAPE value of 1.53 % indicates that a probe penetration rate of 30% is sufficient to capture the shockwaves created by the onramp. It should be noted here that although no information regarding the entry times of the on-ramp vehicles was provided to our estimation algorithm, it can be inferred from the output in Figure 5b. Similar estimation studies (*26*) have indicated that probe levels of 2% can capture the shockwaves generated by lane-closure on a freeway. However, on arterial sections where the traffic dynamics is governed by random arrival of the onramp vehicles, it is not surprising that a higher probe penetration rate is required for traffic state estimation in congested conditions. Moreover, as indicated by the validation example in Section 2, when the traffic state is simply in terms of the congested, synchronised and free-flowing phases (i.e 3 levels of velocity discretization), the probe levels as low as 5% were sufficient to capture the shockwaves generated by the traffic signals.



(a)



(b)

FIGURE 5 Vehicle Trajectory Estimation

The spatial distribution of the probe vehicles plays a significant role in the accuracy of estimated results, as observed earlier. To analyse the effect of the randomness introduced by the probe vehicle distribution, we compute the probability distribution of the MAPE by estimating the traffic state for R = 100 simulations. For each simulation, most likely configuration is estimated using Equation 5, by maximizing the joint probability of the CRF model at each time-step. The fitted log-normal distributions obtained for fixed penetration levels of 5%, 10%, 20% and 30% is depicted in Figure 6. The mean value of the MAPE for a probe penetration level of 5% is around 18%, but the high variance indicated by the flattened PDF implies that travel time error can be even higher if the probe distribution if highly random. As the number of probe vehicles increases this variance reduces, as indicated by the narrower probability distributions. The key observation that can made from this analysis is that when the probe level on arterial sections is sparse, it is imperative to ensure that probes are as uniformly distributed as possible for a reliable estimation of the underlying traffic state. In reality, the variability in the driver response to his local surroundings cannot be discounted, and the observed traffic state given the past conditions can deviate from the most likely state predicted by Equation 5. Taking into consideration this randomness we can sample the non-probe vehicle velocities from the marginal probabilities of the hidden variables, computed using the sum-product algorithm. The cumulative distribution function (CDF) of the MAPE is presented in Figure 7. The results imply that when the probe penetration rate is greater than 20%, the probability of obtaining a MAPE value below 5% is significantly high, with $r \approx 90\%$.

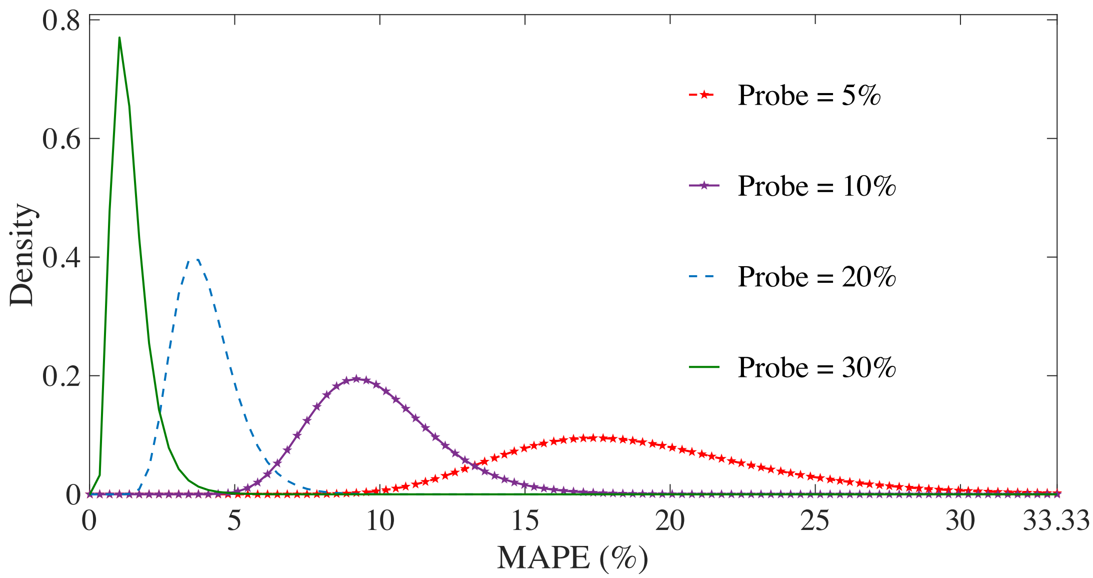


FIGURE 6 PDF of the Mean Absolute Percent Error in travel time at different probe levels

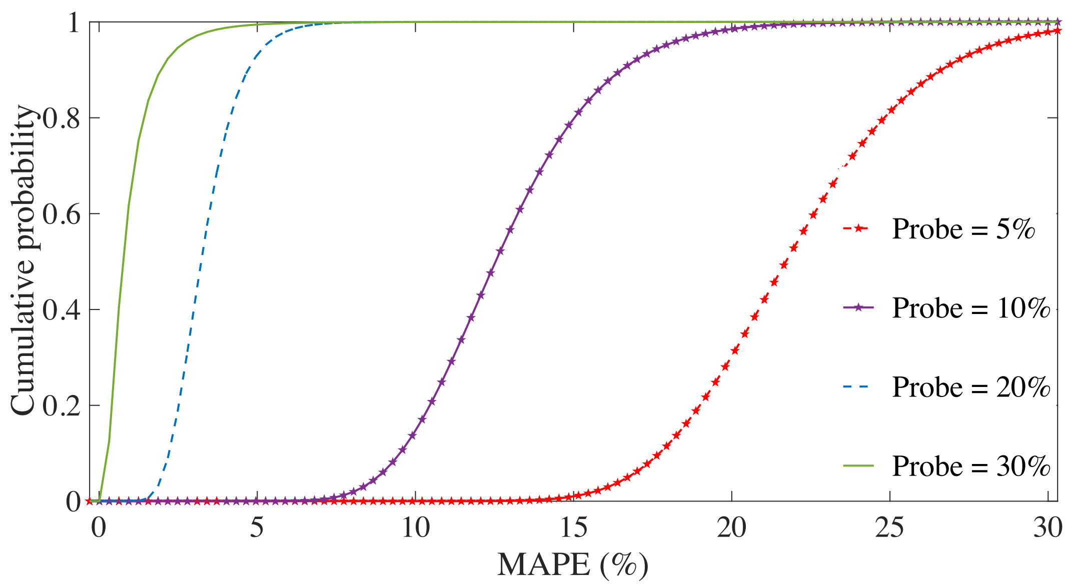


FIGURE 7 CDF of the Mean Absolute Percent Error in travel time at different probe levels

CONCLUSIONS

We present a methodology for traffic state estimation combining car-following theory with probabilistic graphical models to learn the traffic patterns from historical data. We propose a

CRF mainly to model (a) the dependence of the current traffic state on the past, without increasing the complexity of the model by additional edges, and (b) the effect of diminishing influence of the leader vehicle with increasing vehicle spacing. In real-world settings, as the coverage of the probe vehicle information is expected to be random (as well as sparse) and its distribution in the traffic stream cannot be specified apriori, it is not sufficient to specify adequate penetration levels with a single value. To address this randomness, we present a probabilistic approach to examine the probe penetration rate. Position update was modeled as a Markov chain, with multi-class logistic regression used to formulate the node potentials. However, classification accuracy of logistic regression is low, with the probability of incorrectly estimating the updated vehicle position being around 15%. As the predicted vehicle gap is fed into the input vector in the next time step, the Markov chain model error propagates with time. This drawback needs to be addressed with better models of position update. As future research, we propose investigating support vector machines for this purpose. The model can also be extended to multi-lane roads, and the CRF models can be improved by adopting second or higher order Markov models to capture the influence of vehicles further downstream (ahead of the leader), which could improve the estimation accuracy.

15. Papathanasopoulou, V., Antoniou, C. Towards data-driven car-following models. *Transportation Research Part C: Emerging Technologies*, Vol. 55, 2015, pp. 496–509.
16. Treiber, M., Kesting, A. *Traffic Flow Dynamics: Data, Models and Simulation*, Springer-Verlag Berlin Heidelberg, 2013.
17. Furtlehner, C., Lasgouttes, J.M., de La Fortelle, A. A belief propagation approach to traffic prediction using probe vehicles, *IEEE Intelligent Transportation Systems Conference*, 2007, pp. 1022–1027.
18. Mazaré, P.E., Tossavainen, O.P., Bayen, A., Work, D. Trade-offs between inductive loops and gps probe vehicles for travel time estimation: A mobile century case study, *Transportation Research Board 91st Annual Meeting*, 2012.
19. Bham, G.H., Benekohal, R.F. A high fidelity traffic simulation model based on cellular automata and car-following concepts. *Transportation Research Part C: Emerging Technologies*, Vol. 12, 2004, pp. 1–32.
20. Sopasakis, A., Katsoulakis, M.A., 2006. Stochastic modeling and simulation of traffic flow: Asymmetric single exclusion process with Arrhenius look-ahead dynamics. *SIAM Journal on Applied Mathematics*, No. 66, pp. 921–944.
21. Bishop, C.M. *Pattern recognition and Machine learning*. Springer, 2006.
22. Schmidt, M. *UGM: A matlab toolbox for probabilistic undirected graphical models*, 2007.
23. Lárraga, M., Del Rio, J., Alvarez-Lcaza, L., 2005. Cellular automata for one-lane traffic flow modeling. *Transportation Research Part C: Emerging Technologies*, Vol. 13, pp. 63–74.
24. Sutton, C., McCallum, A. An Introduction to Conditional Random Fields. Foundations and Trends in Machine Learning 4, 2012, pp. 267–373.
25. Murphy, K.P. Machine learning: A probabilistic perspective. MIT press, 2012.
26. Bucknell, C., Herrera, J.C., 2014. A trade-off analysis between penetration rate and sampling frequency of mobile sensors in traffic state estimation. *Transportation Research Part C: Emerging Technologies*, Vol. 46, pp. 132–150.