Decentralized Data Fusion and Active Sensing with Mobile Sensors for Modeling and Predicting Spatiotemporal Traffic Phenomena

Jie Chen[†], Kian Hsiang Low[†], Colin Keng-Yan Tan[†], Ali Oran[§], and Patrick Jaillet[‡]
Department of Computer Science, National University of Singapore, Republic of Singapore[†]
Singapore-MIT Alliance for Research and Technology, Republic of Singapore[§]
Dept. of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, USA[‡]
{chenjie, lowkh, ctank}@comp.nus.edu.sg[†]
aoran@smart.mit.edu[§], jaillet@mit.edu[‡]

ABSTRACT

The problem of modeling and predicting spatiotemporal traffic phenomena over an urban road network is important to many traffic applications such as detecting and forecasting congestion hotspots. This paper presents a decentralized data fusion and active sensing (D^{2}FAS) algorithm for mobile sensors to actively explore the road network to gather and assimilate the most informative data for predicting the traffic phenomenon. We analyze the time and communication complexity of D^{2}FAS and demonstrate that it can scale well with increasing number of observations when the number of sensors is large. We provide a theoretical guarantee on its predictive performance to be equivalent to a sophisticated centralized approximate Gaussian process prediction model. This result implies that the computational load of the centralized model can be distributed among the mobile sensors, thereby achieving efficient and scalable prediction. Empirical evaluation on a real-world traffic phenomenon dataset over an urban road network shows that our D^{2}FAS algorithm is significantly more time-efficient and scalable (i.e., in the number of observations and sensors) than existing state-of-the-art algorithms while achieving comparable predictive performance.

1. INTRODUCTION

Knowing and understanding the traffic conditions and phenomena over road networks has become increasingly important to the goal of achieving smooth-flowing, congestion-free traffic, especially in densely-populated urban cities. According to a 2011 urban mobility report [28], the traffic congestions in the USA have caused 1.9 billion gallons of extra fuel, 4.8 billion hours of travel delay, and $101 billion of delay and fuel cost. Such huge resource wastage can be potentially mitigated if the spatiotemporally varying traffic phenomena (e.g., speeds and travel times along road segments) are predicted accurately enough in real time to detect and forecast the congestion hotspots; network-level (e.g., ramp metering, road pricing) and user-level (e.g., route replanning) measures can then be taken to relieve these congestions, so as to improve the overall efficiency of road networks.

In practice, it is non-trivial to achieve real-time, accurate prediction of a spatiotemporally varying traffic phenomenon because the quantity of sensors that can be deployed to observe an entire road network is cost-constrained. Traditionally, static sensors such as loop detectors [9, 34] are placed at designated locations in a road network to collect data for predicting the traffic phenomenon. However, they provide sparse coverage (i.e., many road segments are not observed, thus leading to data sparsity), incur high installation and maintenance costs, and cannot reposition by themselves in response to changes in the traffic phenomenon. Low-cost GPS technology allows the collection of traffic data using passive mobile probes [35] (e.g., taxis/cabs). Unlike static sensors, they can directly measure the travel times along road segments. But, they provide fairly sparse coverage due to low GPS sampling frequency (i.e., often imposed by taxi/cab companies) and no control over their routes, incur high initial implementation cost, pose privacy issues, and produce highly-varying speeds and travel times while traversing the same road segment due to inconsistent driving behaviors. A critical mass of probes is needed on each road segment to ease the severity of the last drawback [30] but is often hard to achieve on non-highway segments due to sparse coverage. In contrast, we propose the use of active mobile probes [33] to overcome the limitations of static and passive mobile probes. In particular, they can be directed to explore any segments of a road network to gather traffic data at a desired GPS sampling rate while enforcing consistent driving behavior.

How then do the mobile probes/sensors actively explore a road network to gather and assimilate the most informative observations for predicting the traffic phenomenon? There are three key issues surrounding this problem, which will be discussed together with the related works:

Models for predicting spatiotemporal traffic phenomena. The spatiotemporal correlation structure of a traffic phenomenon can be exploited to predict the traffic conditions of any unobserved road segment at any time using the observations taken along the mobile sensors' paths. To achieve this, existing Bayesian filtering frameworks [2, 34, 35] utilize various handcrafted parametric models predicting traffic flow along a highway stretch that only correlate adjacent segments of the highway. Hence, their predictive performance will be compromised when the current observations are sparse and/or the actual spatial correlation spans multiple segments. Their strong Markov assumption further exacerbates this problem. It is also not demonstrated how these models can be generalized to work for arbitrary road network topologies and more complex correlation structure. Existing multivariate parametric traffic prediction models [8, 18] do not quantify uncertainty estimates of the predictions and impose rigid spatial locality assumptions that do not adapt to the true underlying correlation structure.

In contrast, we assume the traffic phenomenon over an urban road network (i.e., comprising full range of road types like highways, arterials, slip roads, etc.) to be realized from a rich class of Bayesian non-parametric models called the *Gaussian process* (GP) (Section 2) that can formally characterize its spatiotemporal correlation structure and refine it with growing number of observations [21]. More importantly, GP can provide formal measures of predictive uncertainty (e.g., based on variance or entropy criterion) for directing the mobile sensors to explore highly uncertain areas of the road network. The work of [9] used GP to represent the traffic phenomenon over a network of only highways and defined the correlation of speeds between highway segments to depend only on the geodesic (i.e., shortest path) distance of these segments with respect to the network topology. Different from the work of [9], we further improve the correlation structure of GP by enabling it to exploit road segment features (e.g., length, number of lanes, direction, speed limit) for differentiating road types, which is not found in the works described above.

Data fusion. The observations are gathered distributedly by each mobile sensor along its path in the road network and have to be assimilated in order to predict the traffic phenomenon. Since a large number of observations are expected to be collected, a centralized approach to GP prediction cannot be performed in real time due to its cubic time complexity.

To resolve this, we propose a decentralized data fusion approach to efficient and scalable approximate GP prediction (Section 3). Existing decentralized and distributed Bayesian filtering frameworks for addressing non-traffic related problems [3, 4, 20, 26, 32] will face the same difficulties as their centralized counterparts described above if applied to predicting traffic phenomena, thus resulting in loss of predictive performance. Distributed regression algorithms [7, 22] for static sensor networks gain efficiency from spatial locality assumptions, which cannot be exploited by mobile sensors whose paths are not constrained by locality. The work of [5] proposed a distributed data fusion approach to approximate GP prediction based on an iterative Jacobi overrelaxation algorithm, which incurs some critical limitations: (a) the past observations taken along the mobile sensors' paths are assumed to be uncorrelated, which greatly undermines its predictive performance when they are in fact correlated and/or the current observations are sparse; (b) when the number of robots grows large, it converges very slowly; (c) it assumes that the range of positive correlation has to be bounded by some factor of the communication range. Our proposed decentralized algorithm does not suffer from these limitations and can be computed exactly with efficient time bounds.

Active sensing. The mobile sensors have to actively gather the most informative observations for minimizing the uncertainty of modeling and predicting the traffic phenomenon. Existing centralized [13, 14, 15] and decentralized [12, 31] active sensing algorithms scale poorly with increasing number of observations and/or mobile sensors. We propose a decentralized active sensing algorithm that overcomes these issues of scalability (Section 4).

This paper presents a novel *Decentralized Data Fusion and Active Sensing* (D^2FAS) algorithm (Sections 3 and 4) for sampling spatiotemporally varying environmental phenomena with mobile sensors. Note that the decentralized data fusion component of D^2FAS can also be used for static and passive mobile sensors. The practical applicability of D^2FAS is not restricted to traffic monitoring; it can be used in other environmental sensing applications such as precision agriculture, mineral prospecting [16], monitoring of ocean and freshwater phenomena [6, 23, 17] (e.g., plankton bloom, anoxic zones), forest ecosystems, pollution (e.g., oil spill), or contamination (e.g., radiation leak). The specific contributions of this paper include:

- Analyzing the time and communication overheads of D^2FAS (Section 5): we prove that D^2FAS can scale better than existing state-of-the-art algorithms with increasing number of observations when the number of sensors is large;
- Theoretically guaranteeing the predictive performance of the decentralized data fusion component of D^2FAS to be equivalent to that of a sophisticated centralized approximate GP prediction model (Section 3). This result implies that the computational load of the centralized model can be distributed among the mobile sensors, thereby achieving efficient and scalable prediction;
- Improving the correlation structure of GP model by enabling it to exploit road segment features (e.g., length, number of lanes, direction, and speed limit) and the road network topology (Section 2.1);
- Empirically evaluating the predictive performance, time efficiency, and scalability of D^2FAS algorithm on a real-world traffic phenomenon (i.e., speeds of road segments) dataset over an urban road network (Section 6): D^2FAS is more time-efficient and scales significantly better with increasing number of observations and sensors while achieving predictive performance close to that of existing state-of-the-art algorithms.

2. GAUSSIAN PROCESS REGRESSION OVER GRAPH

The *Gaussian process* (GP) can be used to model a spatiotemporal traffic phenomenon over a road network as follows: The traffic phenomenon is defined to vary as a realization of a GP. Let V be a set of road segments representing the domain of the road network such that each road segment $s \in V$ is specified by a p-dimensional vector of features and is associated with a realized (random) measurement z_s (Z_s) of the traffic condition such as speed if s is observed (unobserved). Let $\{Z_s\}_{s \in V}$ denote a GP, that is, every finite subset of $\{Z_s\}_{s \in V}$ follows a multivariate Gaussian distribution [25]. Then, the GP is fully specified by its *prior* mean $\mu_s \triangleq \mathbb{E}[Z_s]$ and covariance $\sigma_{ss'} \triangleq \text{cov}[Z_s, Z_{s'}]$ for all $s, s' \in V$. In particular, we will describe in Section 2.1 how the covariance $\sigma_{ss'}$ for modeling the correlation of measurements between all pairs of segments $s, s' \in V$ can be designed to exploit the road segment features and the road network topology.

A chief capability of the GP model is that of performing probabilistic regression: Given a set $D \subset V$ of observed road segments and a column vector z_D of corresponding measurements, the joint distribution of the measurements at any set $Y \subseteq V \setminus D$ of unobserved road segments remains Gaussian with the following *posterior* mean vector and covariance matrix

$$\mu_{Y|D} \triangleq \mu_Y + \Sigma_{YD}\Sigma_{DD}^{-1}(z_D - \mu_D) \quad (1)$$

$$\Sigma_{YY|D} \triangleq \Sigma_{YY} - \Sigma_{YD}\Sigma_{DD}^{-1}\Sigma_{DY} \quad (2)$$

where μ_Y (μ_D) is a column vector with mean components μ_s for all $s \in Y$ ($s \in D$), Σ_{YD} (Σ_{DD}) is a covariance matrix with covariance components $\sigma_{ss'}$ for all $s \in Y, s' \in D$ ($s, s' \in D$), and Σ_{DY} is the transpose of Σ_{YD}. The posterior mean vector $\mu_{Y|D}$ (1) is used to predict the measurements at any set Y of unobserved road segments. The posterior covariance matrix $\Sigma_{YY|D}$ (2), which is independent of the measurements z_D, can be processed in two ways to quantify the uncertainty of these predictions: (a) the trace of $\Sigma_{YY|D}$ yields the sum of posterior variances $\Sigma_{ss|D}$ over all $s \in Y$; (b) the determinant of $\Sigma_{YY|D}$ is used in calculating the Gaussian posterior joint entropy

$$\mathbb{H}[Z_Y|Z_D] \triangleq \frac{1}{2}\log(2\pi e)^{|Y|}|\Sigma_{YY|D}| \ . \quad (3)$$

In contrast to the first measure of uncertainty that assumes conditional independence between measurements in the set Y of unobserved road segments, the entropy-based measure (3) accounts for their correlation, thereby not overestimating their uncertainty. Hence, we will focus on using the entropy-based measure of uncertainty in this paper.

2.1 Graph-Based Kernel

If the observations are noisy (i.e., by assuming additive independent identically distributed Gaussian noise with variance σ_n^2), then their prior covariance $\sigma_{ss'}$ can be expressed as

$$\sigma_{ss'} = k(s, s') + \sigma_n^2\delta_{ss'}$$

where $\delta_{ss'}$ is a Kronecker delta that is 1 if $s = s'$ and 0 otherwise, and k is a kernel function measuring the pairwise "similarity" of road segments. For a traffic phenomenon (e.g., road speeds), the correlation of measurements between pairs of road segments depends not only on their features (e.g., length, number of lanes, speed limit, direction) but also the road network topology. Therefore, the kernel function should be defined to exploit both the features and topology information, which will be described next.

Let the road network be represented by a weighted directed graph $G \triangleq (V, E, w)$ comprising a set V of vertices that denotes the domain of all possible road segments, a set $E \subseteq V \times V$ of directed edges such that there is a directed edge (s, s') from $s \in V$ to $s' \in V$ iff the end of segment s connects to the start of segment s' in the road network, and a weight function $w : E \rightarrow \mathbb{R}^+$ measuring the standardized Manhattan distance [1] of each directed edge:

$$w((s, s')) \triangleq \sum_{i=1}^{p} \frac{|[s]_i - [s']_i|}{r_i}$$

where $[s]_i$ ($[s']_i$) is the i-th component of the feature vector specifying road segment s (s'), and r_i is the range of the i-th feature. The weight function w serves as a dissimilarity measure between adjacent road segments.

The next step is to compute the shortest path distance $d(s, s')$ between all pairs of road segments $s, s' \in V$ (i.e., using Floyd-Warshall or Johnson's algorithm) with respect to the topology of the weighted directed graph G. Such a distance function is again a measure of dissimilarity, rather than one of similarity, as required by a kernel function. Furthermore, a valid GP kernel needs to be positive semidefinite and symmetric [27], which are clearly violated by d.

To construct a valid GP kernel from d, multi-dimensional scaling [1] is applied to embed the domain of road segments into the p'-dimensional Euclidean space $\mathbb{R}^{p'}$. Specifically, a mapping $g : V \rightarrow \mathbb{R}^{p'}$ is determined by minimizing the squared loss

$$g^* = \arg\min_{g} \sum_{s, s' \in V} (d(s, s') - \|g(s) - g(s')\|)^2 \ .$$

With a small squared loss, the Euclidean distance $\|g^*(s) - g^*(s')\|$ between $g^*(s)$ and $g^*(s')$ is expected to closely approximate the shortest path distance $d(s, s')$ between any pair of road segments s and s'. After embedding into the Euclidean space, a conventional kernel function such as the squared exponential one [25] can then be used:

$$k(s, s') = \sigma_s^2 \exp\left(-\frac{1}{2}\sum_{i=1}^{p'}\left(\frac{[g^*(s)]_i - [g^*(s')]_i}{\ell_i}\right)^2\right)$$

where $[g^*(s)]_i$ ($[g^*(s')]_i$) is the i-th component of the p'-dimensional vector $g^*(s)$ ($g^*(s')$), and the hyperparameters $\sigma_s, \ell_1, \ldots, \ell_{p'}$ are, respectively, signal variance and length-scales that can be learned using maximum likelihood estimation [25]. The resulting kernel function k[1] is guaranteed to be valid.

2.2 Sparse Approximation

Although the GP is an effective predictive model, it faces a practical limitation of cubic time complexity in the number $|D|$ of observations; this can be observed from computing the posterior distribution (i.e., (1) and (2)), which requires inverting the covariance matrix Σ_{DD} that incurs $\mathcal{O}(|D|^3)$ time. If $|D|$ is expected to be large, then GP prediction cannot be performed in real time. For practical usage, we have to resort to computationally cheaper approximate GP prediction.

[1] For spatiotemporal traffic modeling, the kernel function k can be extended to account for the temporal dimension.

A simple method of approximation is to select only a subset U of the entire set D of observed road segments (i.e., $U \subset D$) to compute the posterior distribution of the measurements at any set $Y \subseteq V \setminus D$ of unobserved road segments. Such a sparse *subset of data* (SoD) approximation method produces the following predictive Gaussian distribution, which closely resembles that of the full GP model (i.e., by simply replacing D in (1) and (2) with U):

$$\mu_{Y|U} = \mu_Y + \Sigma_{YU}\Sigma_{UU}^{-1}(z_U - \mu_U) \tag{4}$$

$$\Sigma_{YY|U} = \Sigma_{YY} - \Sigma_{YU}\Sigma_{UU}^{-1}\Sigma_{UY} \ . \tag{5}$$

Notice that the covariance matrix Σ_{UU} to be inverted only incurs $\mathcal{O}(|U|^3)$ time, which is independent of $|D|$.

The predictive performance of SoD approximation is sensitive to the selection of subset U. In practice, random subset selection often yields poor performance. This issue can be resolved by actively selecting an informative subset U in an iterative greedy manner: Firstly, U is initialized to be an empty set. Then, all road segments in $D \setminus U$ are scored based on a criterion that can be chosen from, for example, the works of [10, 11, 29]. The highest-scored segment is selected for inclusion into U and removed from D. This greedy selection procedure is iterated until U reaches a pre-defined size. Among the various criteria introduced earlier, the differential entropy score [11] is reported to perform well [19]; it is a monotonic function of the posterior variance $\Sigma_{ss|U}$ (5), thus resulting in the greedy selection of a segment $s \in D \setminus U$ with the largest variance in each iteration.

3. DECENTRALIZED DATA FUSION

In the previous section, two centralized data fusion approaches to exact (i.e., (1) and (2)) and approximate (i.e., (4) and (5)) GP prediction are introduced. In this section, we will discuss the decentralized data fusion component of our D^{2}FAS algorithm, which distributes the computational load among the mobile sensors to achieve efficient and scalable approximate GP prediction.

The intuition to our decentralized data fusion algorithm is as follows: each of the K mobile sensors constructs a local summary of the observations taken along its own path in the road network and communicates its local summary to every other sensor. Then, it assimilates the local summaries received from the other sensors into a globally consistent summary, which is exploited for predicting the traffic phenomenon as well as active sensing. This intuition will be formally realized and described in the paragraphs below.

While exploring the road network, each mobile sensor summarizes its local observations taken along its path based on a common support set $U \subset V$ known to all the other sensors. Its local summary is defined as follows:

Definition 1 (Local Summary). *Given a common support set $U \subset V$ known to all K mobile sensors, a set $D_k \subset V$ of observed road segments and a column vector z_{D_k} of corresponding measurements local to mobile sensor k, its local summary is defined as a tuple $(\dot{z}_U^k, \dot{\Sigma}_{UU}^k)$ where*

$$\dot{z}_U^k \triangleq \Sigma_{UD_k}\Sigma_{D_kD_k|U}^{-1}(z_{D_k} - \mu_{D_k}) \tag{6}$$

$$\dot{\Sigma}_{UU}^k \triangleq \Sigma_{UD_k}\Sigma_{D_kD_k|U}^{-1}\Sigma_{D_kU} \tag{7}$$

such that $\Sigma_{D_kD_k|U}$ is defined in a similar manner to (5).

Remark. Unlike SoD (Section 2.2), the support set U of road segments does not have to be observed since the local summary (i.e., (6) and (7)) is independent of the corresponding measurements z_U. So, U does not need to be a subset of $D = \bigcup_{k=1}^{K} D_k$. To select an informative support set U from the set V of all possible segments in the road network, an offline active selection procedure similar to that in the last paragraph of Section 2.2 can be performed just once prior to observing data to determine U. In contrast, SoD has to perform online active selection every time new road segments are being observed.

By communicating its local summary to every other sensor, each mobile sensor can then construct a globally consistent summary from the received local summaries:

Definition 2 (Global Summary). *Given a common support set $U \subset V$ known to all K mobile sensors and the local summary $(\dot{z}_U^k, \dot{\Sigma}_{UU}^k)$ of every mobile sensor $k = 1, \ldots, K$, the global summary is defined as a tuple $(\overline{z}_U, \overline{\Sigma}_{UU})$ where*

$$\overline{z}_U \triangleq \sum_{k=1}^{K} \dot{z}_U^k \tag{8}$$

$$\overline{\Sigma}_{UU} \triangleq \Sigma_{UU} + \sum_{k=1}^{K} \dot{\Sigma}_{UU}^k \ . \tag{9}$$

Remark. In this paper, we assume all-to-all communication between the K mobile sensors. Supposing this is not possible and each sensor can only communicate locally with its neighbors, the summation structure of the global summary (specifically, (8) and (9)) makes it amenable to be constructed using distributed consensus filters [20]. We omit these details since they are beyond the scope of this paper.

Finally, the global summary is exploited by each mobile sensor to compute a globally consistent predictive Gaussian distribution, as detailed in Theorem 1A below, as well as to perform decentralized active sensing (Section 4):

Theorem 1. *Let a common support set $U \subset V$ be known to all K mobile sensors.*

A. *Given the global summary $(\overline{z}_U, \overline{\Sigma}_{UU})$, each mobile sensor computes a globally consistent predictive Gaussian distribution $\mathcal{N}(\mu_Y^{\mathrm{D^2FAS}}, \Sigma_{YY}^{\mathrm{D^2FAS}})$ of the measurements at any set Y of unobserved road segments where*

$$\mu_Y^{\mathrm{D^2FAS}} \triangleq \mu_Y + \Sigma_{YU}\overline{\Sigma}_{UU}^{-1}\overline{z}_U \tag{10}$$

$$\Sigma_{YY}^{\mathrm{D^2FAS}} \triangleq \Sigma_{YY} - \Sigma_{YU}(\Sigma_{UU}^{-1} - \overline{\Sigma}_{UU}^{-1})\Sigma_{UY} \ . \tag{11}$$

B. *Let $\mathcal{N}(\mu_{Y|D}^{\mathrm{PITC}}, \Sigma_{YY|D}^{\mathrm{PITC}})$ be the predictive Gaussian distribution computed by the centralized partially independent training conditional (PITC) approximation of GP* *model [24] where*

$$\mu^{\text{PITC}}_{Y|D} \triangleq \mu_Y + \Gamma_{YD} \left(\Gamma_{DD} + \Lambda\right)^{-1} \left(z_D - \mu_D\right) \quad (12)$$

$$\Sigma^{\text{PITC}}_{YY|D} \triangleq \Sigma_{YY} - \Gamma_{YD} \left(\Gamma_{DD} + \Lambda\right)^{-1} \Gamma_{DY} \quad (13)$$

such that

$$\Gamma_{AB} \triangleq \Sigma_{AU}\Sigma^{-1}_{UU}\Sigma_{UB} \quad (14)$$

and Λ is a block-diagonal matrix constructed from the K diagonal blocks of $\Sigma_{DD|U}$, each of which is a matrix $\Sigma_{D_k D_k|U}$ for $k = 1, \ldots, K$ where $D = \bigcup_{k=1}^{K} D_k$. Then, $\mu^{\text{D}^2\text{FAS}}_{Y} = \mu^{\text{PITC}}_{Y|D}$ and $\Sigma^{\text{D}^2\text{FAS}}_{YY} = \Sigma^{\text{PITC}}_{YY|D}$.

The proof of Theorem 1B is given in Appendix A. The equivalence result of Theorem 1B bears two implications:

Remark 1. The computational load of the centralized PITC approximation of GP model can be distributed among K mobile sensors, thereby improving the time efficiency of prediction. Specifically, supposing $|Y| \leq |U|$ for simplicity, the $\mathcal{O}\big(|D|((|D|/K)^2 + |U|^2)\big)$ time incurred by PITC can be reduced to $\mathcal{O}\big((|D|/K)^3 + |U|^3 + |U|^2 K\big)$ time of running our decentralized algorithm on each of the K sensors, the latter of which scales better with increasing number $|D|$ of observations.

Remark 2. We can draw insights from PITC to elucidate an underlying property of our decentralized algorithm: It is assumed that $Z_{D_1}, \ldots, Z_{D_K}, Z_Y$ are conditionally independent given the measurements at the support set U of road segments. To potentially reduce the degree of violation of this assumption, an informative support set U is actively selected, as described earlier in this section. Furthermore, the experimental results on a real-world traffic phenomenon dataset[2] over an urban road network (Section 6) show that D^{2}FAS can achieve predictive performance comparable to that of the full GP model while enjoying computational gain over it, thus demonstrating the practicality of such an assumption for predicting traffic phenomena. The predictive performance of D^{2}FAS can be improved by increasing the size of U at the expense of greater time and communication overhead.

4. DECENTRALIZED ACTIVE SENSING

The problem of active sensing with K mobile sensors is formulated as follows: Given the set $D_k \subset V$ of observed road segments and the currently traversed road segment $s_k \in V$ of every mobile sensor $k = 1, \ldots, K$, the mobile sensors have to select the most informative walks $w_1^*, \ldots, w_K^*$ of length L each and with respective origins $s_1, \ldots, s_K$ in the road network G:

$$(w_1^*, \ldots, w_K^*) = \underset{(w_1,\ldots,w_K)}{\arg\max}\ \mathbb{H}\Big[Z_{\bigcup_{k=1}^{K} Y_{w_k}} \Big| Z_{\bigcup_{k=1}^{K} D_k}\Big] \quad (15)$$

where Y_{w_k} denotes the set of unobserved road segments induced by the walk w_k. Interestingly, it can be shown using the chain rule for entropy that these maximum-entropy walks $w_1^*, \ldots, w_K^*$ minimize the posterior joint entropy (i.e., $\mathbb{H}[Z_{V\setminus\bigcup_{k=1}^{K}(D_k \cup Y_{w_k^*})} | Z_{\bigcup_{k=1}^{K}(D_k \cup Y_{w_k^*})}]$) of the measurements at the remaining unobserved segments (i.e., $V \setminus \bigcup_{k=1}^{K}(D_k \bigcup Y_{w_k^*})$) in the road network. After executing the walk w_k^*, each mobile sensor k observes the set $Y_{w_k^*}$ of road segments and updates its local information:

$$D_k \leftarrow D_k \bigcup Y_{w_k^*}\ , z_{D_k} \leftarrow z_{D_k \cup Y_{w_k^*}}, s_k \leftarrow \text{terminus of } w_k^*\ . \quad (16)$$

Without imposing any structural assumption, solving the active sensing problem (15) will be prohibitively expensive due to the space of possible joint walks $(w_1, \ldots, w_K)$ that grows exponentially in the number K of mobile sensors. To overcome this scalability issue, $Z_{Y_{w_1}}, \ldots, Z_{Y_{w_K}}$ are assumed to be conditionally independent given the measurements at the set $D = \bigcup_{k=1}^{K} D_k$ of observed road segments. Such an assumption is not uncommon: it is often made in order to calculate the widely-used sum of posterior variances (i.e., mean-squared error) criterion (Section 2). In practice, this assumption usually becomes less restrictive when the number $|\mathcal{D}|$ of observed road segments increases to potentially reduce the degree of violation of conditional independence, the correlation of measurements between road segments decreases, and/or the mobile sensors are sufficiently far apart. Using the chain rule for entropy and subsequently the conditional independence assumption, the active sensing problem (15) reduces to

$$\begin{aligned}
&\max_{(w_1,\ldots,w_K)} \mathbb{H}\Big[Z_{\bigcup_{k=1}^{K} Y_{w_k}} \Big| Z_D\Big] \\
&= \max_{(w_1,\ldots,w_K)} \sum_{k=1}^{K} \mathbb{H}\Big[Z_{Y_{w_k}} \Big| Z_{\bigcup_{i=1}^{k-1} Y_{w_i} \cup D}\Big] \\
&= \max_{(w_1,\ldots,w_K)} \sum_{k=1}^{K} \mathbb{H}\Big[Z_{Y_{w_k}} \Big| Z_D\Big] = \sum_{k=1}^{K} \max_{w_k} \mathbb{H}\Big[Z_{Y_{w_k}} \Big| Z_D\Big]\ ,
\end{aligned}$$

which can be solved in a decentralized manner by each mobile sensor k:

$$w_k^* = \underset{w_k}{\arg\max}\ \mathbb{H}\Big[Z_{Y_{w_k}} \Big| Z_D\Big] = \underset{w_k}{\arg\max} \Big|\Sigma_{Y_{w_k} Y_{w_k}|D}\Big| \quad (17)$$

such that the second equality follows from (3) and the posterior covariance matrix $\Sigma_{Y_{w_k} Y_{w_k}|D}$ can be obtained using one of the data fusion methods described earlier, specifically, using (2) of full GP model (Section 2), (5) of SoD (Section 2.2), or (11) of D^{2}FAS (Section 3). If full GP or SoD is to be performed separately on each of the K mobile sensors rather than centrally, then the observations that are gathered distributedly by the sensors have to be fully communicated to every sensor. In contrast, D^{2}FAS only requires exchanging local summaries (Definition 1) between sensors.

Algorithm 1 below outlines the key operations of our D^{2}FAS algorithm to be run on each mobile sensor k, as detailed previously in Sections 3 and 4.

5. TIME AND COMMUNICATION OVERHEADS

In this section, the time and communication overheads of our D^{2}FAS algorithm are analyzed and compared to that of decentralized active sensing coupled with full GP (FGP) or SoD data fusion method to be run on each of the K sensors.

[2] The work of [24] only illustrated the predictive performance of PITC on a simulated toy example.

```
Algorithm 1: $\text{D}^2\text{FAS}(U, K, L, k, D_k, z_{D_k}, s_k)$
while true do
    /* Data fusion (Section 3)                          */
    Construct local summary by (6) and (7)
    Exchange local summary with every sensor $i \neq k$
    Construct global summary by (8) and (9)
    Predict measurements at unobserved road segments by
    (10) and (11)
    /* Active Sensing (Section 4)                       */
    Compute maximum-entropy walk $w_k^*$ by (11) and (17)
    Execute walk $w_k^*$ and observe its road segments $Y_{w_k^*}$
    Update local information $D_k$, $z_{D_k}$, and $s_k$ by (16)
```

5.1 Time Complexity

Our D^{2}FAS algorithm comprises the data fusion and active sensing components. The data fusion component involves computing the local and global summaries and the predictive Gaussian distribution, as shown in Algorithm 1. To construct the local summary using (6) and (7), each sensor has to evaluate $\Sigma_{D_k D_k|U}$ in $\mathcal{O}\big(|U|^3 + |U|(|D|/K)^2\big)$ time and invert it in $\mathcal{O}\big((|D|/K)^3\big)$ time, after which the local summary is obtained in $\mathcal{O}\big(|U|^2|D|/K + |U|(|D|/K)^2\big)$ time. The global summary is computed in $\mathcal{O}\big(|U|^2K\big)$ by (8) and (9). Finally, the predictive Gaussian distribution is derived in $\mathcal{O}\big(|U|^3 + |U||Y|^2\big)$ time using (10) and (11). Supposing $|Y| \leq |U|$ for simplicity, the time complexity of data fusion is then $\mathcal{O}\big((|D|/K)^3 + |U|^3 + |U|^2K\big)$.

The active sensing component involves computing the maximum-entropy walk by (11) and (17). Let the maximum outdegree of G be denoted by Δ. Then, each mobile sensor k has to consider Δ^L possible walks. For each walk w_k, evaluating the determinant of $\Sigma^{\text{D}^2\text{FAS}}_{Y_{w_k}Y_{w_k}}$ incurs $\mathcal{O}\big(L|U|^2 + L^3\big)$ time. The time complexity of active sensing is therefore $\mathcal{O}\big(\Delta^L L(|U|^2 + L^2)\big)$.

Hence, the time complexity of D^{2}FAS is $\mathcal{O}\big((|D|/K)^3 + |U|^3 + |U|^2K + \Delta^L L(|U|^2 + L^2)\big)$. In contrast, the time incurred by decentralized active sensing coupled with FGP and SoD are, respectively, $\mathcal{O}\big(|D|^3 + \Delta^L L(|D|^2 + L^2)\big)$ and $\mathcal{O}\big(|U|^3|D| + \Delta^L L(|U|^2 + L^2)\big)$. It can be observed that D^{2}FAS can potentially scale better with increasing number $|D|$ of observations when the number K of sensors is large. The scalability of D^{2}FAS vs. FGP and SoD will be further evaluated empirically in Section 6.

5.2 Communication Complexity

Let the communication overhead be defined as the size of each broadcast message. Recall from Algorithm 1 (i.e., D^{2}FAS) that, in each iteration, each sensor broadcasts a $\mathcal{O}\big(|U|^2\big)$-sized summary encapsulating its local observations, which is robust against communication failure. In contrast, FGP and SoD require each sensor to broadcast, in each iteration, a $\mathcal{O}(|D|/K)$-sized message comprising exactly its local observations to handle communication failure. If the number of local observations grows to be larger in size than a local summary of predefined size, then our D^{2}FAS algorithm is more scalable than FGP and SoD in terms of communication overhead.

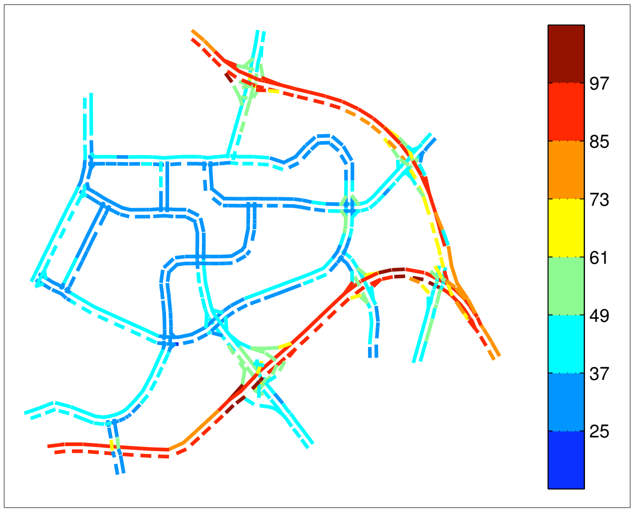


Figure 1: Traffic phenomenon (i.e., speeds (km/h) of road segments) over an urban road network in Tampines area, Singapore during evening peak hours on April 20, 2011. It comprises 775 road segments including highways, arterials, slip roads, etc. The mean speed is 48.8 km/h and the population standard deviation is 20.5 km/h.

6. EXPERIMENTS AND DISCUSSION

This section evaluates the predictive performance, time efficiency, and scalability of our D^{2}FAS algorithm on a real-world traffic phenomenon (i.e., speeds of road segments) dataset over an urban road network, as shown and detailed in Fig. 1. The performance of D^{2}FAS is compared to that of decentralized active sensing coupled with two state-of-art data fusion methods: full GP (FGP) and SoD (Section 2). A network of K mobile sensors is tasked to explore the road network to gather a total of up to 960 observations. To reduce computational time, each sensor repeatedly computes and executes maximum-entropy walks of length $L = 2$ (instead of computing a very long walk), unless otherwise stated. The size of the support set U is set to be 64. The experiments are run on a Linux PC platform with Intel® Core$^{\text{TM}}$2 Quad CPU Q9550 at 2.83 GHz.

6.1 Performance Metrics

The first metric evaluates the predictive performance of a tested algorithm: it measures the *root mean squared error* (RMSE)

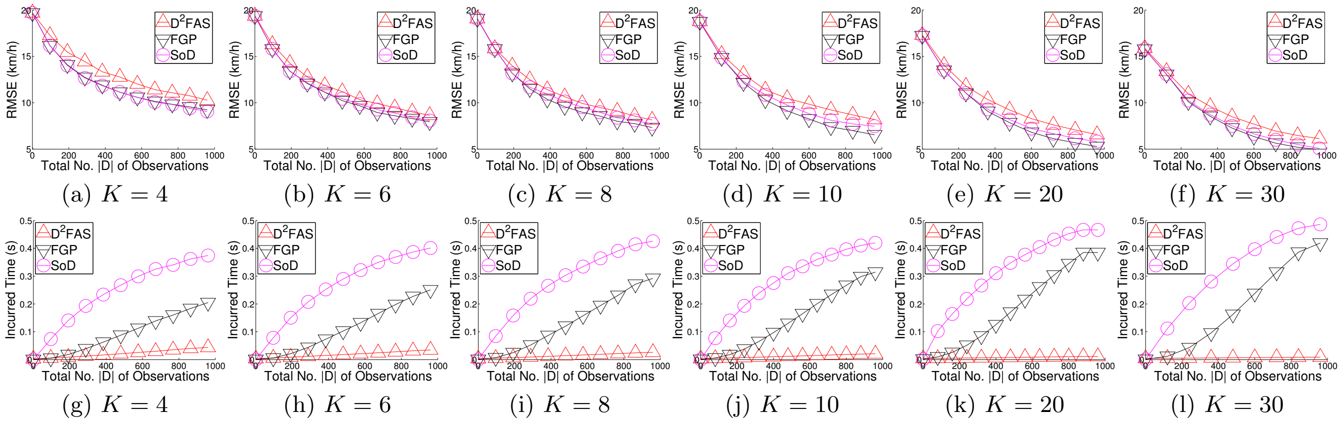
$$\sqrt{\frac{1}{|V|}\sum_{s\in V}\left(z_s - \widehat{\mu}_s\right)^2}$$

over the entire domain V of the road network that is incurred by the predictive mean $\widehat{\mu}_s$ of the tested algorithm, specifically, using (1) of FGP, (4) of SoD, or (10) of D^{2}FAS.

The second performance metric evaluates the time efficiency and scalability of a tested algorithm by measuring its incurred time.

6.2 Results and Analysis

Fig. 2 shows the results of the performance of the tested algorithms averaged over 40 randomly generated starting sensor



(a) $K = 4$ (b) $K = 6$ (c) $K = 8$ (d) $K = 10$ (e) $K = 20$ (f) $K = 30$

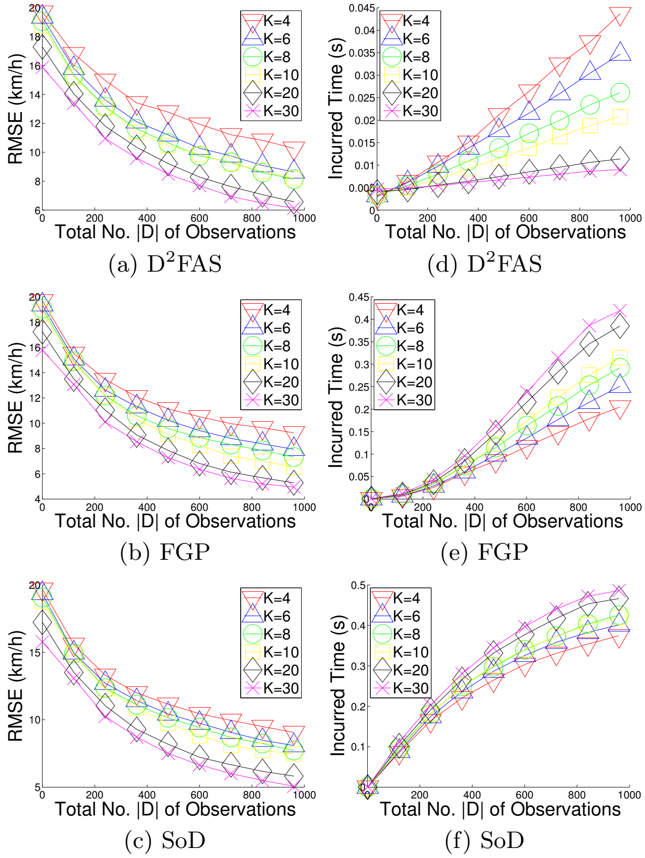
(g) $K = 4$ (h) $K = 6$ (i) $K = 8$ (j) $K = 10$ (k) $K = 20$ (l) $K = 30$

Figure 2: Graphs of (a-f) predictive performance and (g-l) time efficiency vs. total no. $|D|$ of observations gathered by varying number K of mobile sensors.

locations with varying number $K = 4, 6, 8, 10, 20, 30$ of sensors. It can be observed that D^{2}FAS is more time-efficient and scales significantly better with increasing number $|D|$ of observations (Figs. 2g to 2l) while achieving predictive performance close to that of FGP and SoD (Figs. 2a to 2f). Hence, the real-time performance and scalability (i.e., in the number of observations) of our D^{2}FAS algorithm enable it to be used for persistent large-scale traffic modeling and prediction where a large number of observations are expected to be available. The slightly better predictive performance of FGP and SoD are expected since they are able to exploit all collected observations for data fusion. In contrast, D^{2}FAS can only exploit local summaries over the small support set U. As mentioned earlier in Section 3, the predictive performance of D^{2}FAS can be improved by increasing the size of U at the expense of greater time and communication overhead.

Using the same results as that in Fig. 2, Fig. 3 plots them differently to reveal the scalability of the tested algorithms with increasing number K of mobile sensors. It can be observed from Figs. 3a to 3c that the predictive performance of all tested algorithms improve with a larger number of sensors because each sensor needs to execute fewer number of walks and its performance is therefore less adversely affected by its myopic selection (i.e., $L = 2$) of maximum-entropy walks. As a result, more informative unobserved road segments are explored. As shown in Fig. 3d, the time incurred by D^{2}FAS decreases due to its decentralized data fusion component that can distribute the computational load among a greater number of sensors. In contrast, it can be seen from Figs. 3e and 3f that the time incurred by FGP and SoD increase: as discussed above, a larger number of sensors result in a greater quantity of more informative unique observations to be gathered (i.e., fewer repeated observations), which increase the time needed for data fusion. When $K \geq 10$, D^{2}FAS is at least 1 order of magnitude faster than FGP and SoD. Hence, the scalability (i.e., in the number of sensors) of our D^{2}FAS algorithm allows the deployment of a large-scale mobile sensor network to achieve more accurate traffic modeling and prediction.

Fig. 4 shows the results of the performance of our D^{2}FAS algorithm with varying length $L = 2, 4, 6, 8, 10$ of maximum-

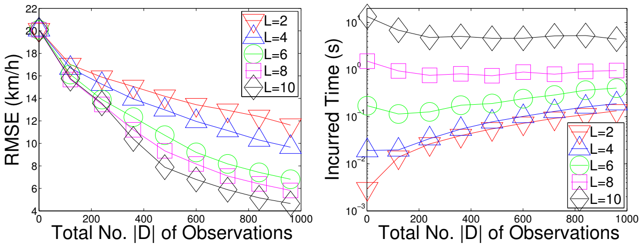


(a) D^{2}FAS (d) D^{2}FAS

(b) FGP (e) FGP

(c) SoD (f) SoD

Figure 3: Graphs of (a-c) predictive performance and (d-f) time efficiency vs. total no. $|D|$ of observations gathered by varying number K of sensors.



(a) (b)

Figure 4: Graphs of (a) predictive performance and (b) time efficiency vs. total no. $|D|$ of observations gathered by 2 mobile sensors running D^{2}FAS with varying length L of maximum-entropy walks.

entropy walks; we choose to experiment with just 2 sensors since Fig. 3d reveals that a smaller number of sensors produce poorer predictive performance and higher incurred time. It can be observed that the predictive performance improves with increasing walk length L because the selection of maximum-entropy walks is less myopic. When L increases to 10, the incurred time increases to about 10 seconds, which is reasonable in practice. By deploying a larger number of sensors, the incurred time is expected to decrease while improving the predictive performance.

7. CONCLUSION

This paper describes a decentralized data fusion and active sensing algorithm for modeling and predicting spatiotemporal traffic phenomena with mobile sensors. Analytical and empirical results have shown that our D^{2}FAS algorithm is extremely time-efficient and scales significantly better with increasing number of observations and sensors while achieving predictive performance close to that of state-of-the-art FGP and SoD. Hence, D^{2}FAS is practical for deployment in a large-scale mobile sensor network to achieve persistent and accurate traffic modeling and prediction. For our future work, we will assume that each sensor can only communicate locally with its neighbors (instead of assuming all-to-all communication between sensors) and develop a *distributed* data fusion approach to efficient and scalable approximate GP prediction based on our D^{2}FAS algorithm and consensus filters [20].

8. ACKNOWLEDGMENTS

This work was supported by Singapore-MIT Alliance Research and Technology (SMART) Subaward Agreement 14 R-252-000-466-592.

9. REFERENCES

[1] I. Borg and P. J. F. Groenen. *Modern Multidimensional Scaling: Theory and Applications.* Springer, NY, 2005.

[2] H. Chen, H. A. Rakha, and S. Sadek. Real-time freeway traffic state prediction: A particle filter approach. In *Proc. IEEE ITSC*, pages 626–631, 2011.

[3] T. H. Chung, V. Gupta, J. W. Burdick, and R. M. Murray. On a decentralized active sensing strategy using mobile sensor platforms in a network. In *Proc. CDC*, pages 1914–1919, 2004.

[4] M. Coates. Distributed particle filters for sensor networks. In *Proc. IPSN*, pages 99–107, 2004.

[5] J. Cortes. Distributed kriged Kalman filter for spatial estimation. *IEEE Trans. Automat. Contr.*, 54(12):2816–2827, 2009.

[6] J. M. Dolan, G. Podnar, S. Stancliff, K. H. Low, A. Elfes, J. Higinbotham, J. C. Hosler, T. A. Moisan, and J. Moisan. Cooperative aquatic sensing using the telesupervised adaptive ocean sensor fleet. In *Proc. SPIE Conference on Remote Sensing of the Ocean, Sea Ice, and Large Water Regions*, volume 7473, 2009.

[7] C. Guestrin, P. Bodik, R. Thibaus, M. Paskin, and S. Madden. Distributed regression: an efficient framework for modeling sensor network data. In *Proc. IPSN*, pages 1–10, 2004.

[8] Y. Kamarianakis and P. Prastacos. Forecasting traffic flow conditions in an urban network: Comparison of multivariate and univariate approaches. *Transport. Res. Rec.*, 1857:74–84, 2003.

[9] A. Krause, E. Horvitz, A. Kansal, and F. Zhao. Toward community sensing. In *Proc. IPSN*, pages 481–492, 2008.

[10] A. Krause, A. Singh, and C. Guestrin. Near-optimal sensor placements in Gaussian processes: Theory, efficient algorithms and empirical studies. *JMLR*, 9:235–284, 2008.

[11] N. Lawrence, M. Seeger, and R. Herbrich. Fast sparse Gaussian process methods: The informative vector machine. In S. Becker, S. Thrun, and K. Obermayer, editors, *Advances in Neural Information Processing Systems 15*, pages 609–616, Cambridge, MA, 2003. MIT Press.

[12] K. H. Low, J. Chen, J. M. Dolan, S. Chien, and D. R. Thompson. Decentralized active robotic exploration and mapping for probabilistic field classification in environmental sensing. In *Proc. AAMAS*, 2012.

[13] K. H. Low, J. M. Dolan, and P. Khosla. Adaptive multi-robot wide-area exploration and mapping. In *Proc. AAMAS*, pages 23–30, 2008.

[14] K. H. Low, J. M. Dolan, and P. Khosla. Information-theoretic approach to efficient adaptive path planning for mobile robotic environmental sensing. In *Proc. ICAPS*, pages 233–240, 2009.

[15] K. H. Low, J. M. Dolan, and P. Khosla. Active Markov information-theoretic path planning for robotic environmental sensing. In *Proc. AAMAS*, pages 753–760, 2011.

[16] K. H. Low, G. J. Gordon, J. M. Dolan, and P. Khosla. Adaptive sampling for multi-robot wide-area exploration. In *Proc. IEEE ICRA*, pages 755–760, 2007.

[17] K. H. Low, G. Podnar, S. Stancliff, J. M. Dolan, and A. Elfes. Robot boats as a mobile aquatic sensor network. In *Proc. IPSN-09 Workshop on Sensor Networks for Earth and Space Science Applications*, 2009.

[18] W. Min and L. Wynter. Real-time road traffic prediction with spatio-temporal correlations. *Transport. Res. C-Emer.*, 19(4):606–616, 2011.

[19] S. Oh, Y. Xu, and J. Choi. Explorative navigation of mobile sensor networks using sparse Gaussian processes. In *Proc. CDC*, pages 3851–3856, dec. 2010.

[20] R. Olfati-Saber. Distributed Kalman filter with embedded consensus filters. In *Proc. CDC*, pages 8179–8184, 2005.

[21] P. Orbanz and Y. W. Teh. Bayesian nonparametric models. In C. Sammut and G. I. Webb, editors, *Encyclopedia of Machine Learning*, pages 81–89. Springer, NY, 2010.

[22] M. A. Paskin and C. Guestrin. Robust probabilistic inference in distributed systems. In *Proc. UAI*, pages 436–445, 2004.

[23] G. Podnar, J. M. Dolan, K. H. Low, and A. Elfes. Telesupervised remote surface water quality sensing. In *Proc. IEEE Aerospace Conference*, 2010.

[24] J. Quiñonero-Candela and C. E. Rasmussen. A unifying view of sparse approximate Gaussian process regression. *JMLR*, 6:1939–1959, 2005.

[25] C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning.* MIT Press, Cambridge, MA, 2006.

[26] M. Rosencrantz, G. Gordon, and S. Thrun. Decentralized sensor fusion with distributed particle filters. In *Proc. UAI*, pages 493–500, 2003.

[27] B. Schölkopf and A. J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond.* MIT Press, Cambridge, MA, 1 edition, 2002.

[28] D. Schrank, T. Lomax, and B. Eisele. *TTI's 2011 Urban Mobility Report.* Texas Transportation Institute, Texas A&M University, 2011.

[29] M. Seeger and C. Williams. Fast forward selection to speed up sparse Gaussian process regression. In C. M. Bishop and B. J. Frey, editors, *Proc. AISTATS*, 2003.

[30] K. K. Srinivasan and P. P. Jovanis. Determination of number of probe vehicle required for reliable travel time measurement in urban network. *Transport. Res. Rec.*, 1537:15–22, 1996.

[31] R. Stranders, A. Farinelli, A. Rogers, and N. R. Jennings. Decentralised coordination of mobile sensors using the max-sum algorithm. In *Proc. IJCAI*, pages 299–304, 2009.

[32] S. Sukkarieh, E. Nettleton, J. Kim, M. Ridley, A. Goktogan, and H. Durrant-Whyte. The ANSER project: Data fusion across multiple uninhabited air vehicles. *IJRR*, 22(7-8):505–539, 2003.

[33] S. M. Turner, W. L. Eisele, R. J. Benz, and D. J. Holdener. Travel time data collection handbook. Technical Report FHWA-PL-98-035, Federal Highway Administration, Office of Highway Information Management, Washington, DC, 1998.

[34] Y. Wang and M. Papageorgiou. Real-time freeway traffic state estimation based on extended Kalman filter: a general approach. *Transport. Res. B-Meth.*, 39(2):141–167, 2005.

[35] D. B. Work, S. Blandin, O. Tossavainen, B. Piccoli, and A. Bayen. A traffic model for velocity data assimilation. *AMRX*, 2010(1):1–35, 2010.

APPENDIX

A. PROOF OF THEOREM 1B

We need to first simplify the $\Gamma_{YD}\left(\Gamma_{DD}+\Lambda\right)^{-1}$ term in the expressions of $\mu^{\text{PITC}}_{Y|D}$ (12) and $\Sigma^{\text{PITC}}_{YY|D}$ (13).

$$\begin{array}{l}
\left(\Gamma_{DD}+\Lambda\right)^{-1}\\
=\left(\Sigma_{DU}\Sigma_{UU}^{-1}\Sigma_{UD}+\Lambda\right)^{-1}\\
=\Lambda^{-1}-\Lambda^{-1}\Sigma_{DU}\left(\Sigma_{UU}+\Sigma_{UD}\Lambda^{-1}\Sigma_{DU}\right)^{-1}\Sigma_{UD}\Lambda^{-1}\\
=\Lambda^{-1}-\Lambda^{-1}\Sigma_{DU}\overline{\Sigma}_{UU}^{-1}\Sigma_{UD}\Lambda^{-1}\ .
\end{array}\tag{18}$$

The second equality follows from matrix inversion lemma. The last equality is due to

$$\begin{array}{l}
\Sigma_{UU}+\Sigma_{UD}\Lambda^{-1}\Sigma_{DU}\\
=\Sigma_{UU}+\displaystyle\sum_{k=1}^{K}\Sigma_{UD_k}\Sigma^{-1}_{D_kD_k|U}\Sigma_{D_kU}\\
=\Sigma_{UU}+\displaystyle\sum_{k=1}^{K}\dot{\Sigma}^{k}_{UU}=\overline{\Sigma}_{UU}\ .
\end{array}\tag{19}$$

Using (14) and (18),

$$\begin{array}{l}
\Gamma_{YD}\left(\Gamma_{DD}+\Lambda\right)^{-1}\\
=\Sigma_{YU}\Sigma_{UU}^{-1}\Sigma_{UD}\left(\Lambda^{-1}-\Lambda^{-1}\Sigma_{DU}\overline{\Sigma}_{UU}^{-1}\Sigma_{UD}\Lambda^{-1}\right)\\
=\Sigma_{YU}\Sigma_{UU}^{-1}\left(\overline{\Sigma}_{UU}-\Sigma_{UD}\Lambda^{-1}\Sigma_{DU}\right)\overline{\Sigma}_{UU}^{-1}\Sigma_{UD}\Lambda^{-1}\\
=\Sigma_{YU}\overline{\Sigma}_{UU}^{-1}\Sigma_{UD}\Lambda^{-1}
\end{array}\tag{20}$$

The third equality is due to (19).

From (12),

$$\begin{array}{rcl}
\mu^{\text{PITC}}_{Y|D} &=& \mu_Y+\Gamma_{YD}\left(\Gamma_{DD}+\Lambda\right)^{-1}\left(z_D-\mu_D\right)\\
&=& \mu_Y+\Sigma_{YU}\overline{\Sigma}_{UU}^{-1}\Sigma_{UD}\Lambda^{-1}\left(z_D-\mu_D\right)\\
&=& \mu_Y+\Sigma_{YU}\overline{\Sigma}_{UU}^{-1}\overline{z}_U\\
&=& \mu_Y^{\text{D}^2\text{FAS}}\ .
\end{array}$$

The second equality is due to (20). The third equality follows from

$$\begin{array}{rcl}
\Sigma_{UD}\Lambda^{-1}\left(z_D-\mu_D\right) &=& \displaystyle\sum_{k=1}^{K}\Sigma_{UD_k}\Sigma^{-1}_{D_kD_k|U}\left(z_{D_k}-\mu_{D_k}\right)\\
&=& \displaystyle\sum_{k=1}^{K}\dot{z}^{k}_{U}=\overline{z}_U\ .
\end{array}$$

From (13),

$$\begin{array}{l}
\Sigma^{\text{PITC}}_{YY|D}\\
=\Sigma_{YY}-\Gamma_{YD}\left(\Gamma_{DD}+\Lambda\right)^{-1}\Gamma_{DY}\\
=\Sigma_{YY}-\Sigma_{YU}\overline{\Sigma}_{UU}^{-1}\Sigma_{UD}\Lambda^{-1}\Sigma_{DU}\Sigma_{UU}^{-1}\Sigma_{UY}\\
=\Sigma_{YY}-\Big(\Sigma_{YU}\overline{\Sigma}_{UU}^{-1}\Sigma_{UD}\Lambda^{-1}\Sigma_{DU}\Sigma_{UU}^{-1}\Sigma_{UY}\\
\quad-\Sigma_{YU}\Sigma_{UU}^{-1}\Sigma_{UY}\Big)-\Sigma_{YU}\Sigma_{UU}^{-1}\Sigma_{UY}\\
=\Sigma_{YY}-\Sigma_{YU}\overline{\Sigma}_{UU}^{-1}\left(\Sigma_{UD}\Lambda^{-1}\Sigma_{DU}-\overline{\Sigma}_{UU}\right)\Sigma_{UU}^{-1}\Sigma_{UY}\\
\quad-\Sigma_{YU}\Sigma_{UU}^{-1}\Sigma_{UY}\\
=\Sigma_{YY}-\left(\Sigma_{YU}\Sigma_{UU}^{-1}\Sigma_{UY}-\Sigma_{YU}\overline{\Sigma}_{UU}^{-1}\Sigma_{UY}\right)\\
=\Sigma_{YY}-\Sigma_{YU}\left(\Sigma_{UU}^{-1}-\overline{\Sigma}_{UU}^{-1}\right)\Sigma_{UY}\\
=\Sigma_{YY}^{\text{D}^2\text{FAS}}\ .
\end{array}$$

The second equality follows from (14) and (20). The fifth equality is due to (19).