Transportation Research Part C 58 (2015) 325–342

Contents lists available at ScienceDirect

Transportation Research Part C

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

How much GPS data do we need?

Anthony D. Patire [a,*], Matthew Wright [a,d], Boris Prodhomme [a], Alexandre M. Bayen [b,c]

[a] ITS/PATH, University of California, 3 McLaughlin Hall #1720, Berkeley, CA 94720-1720, USA
[b] Department of Electrical Engineering and Computer Science, University of California, Berkeley, USA
[c] Department of Civil and Environmental Engineering, Systems Engineering, University of California, Berkeley, USA
[d] Department of Mechanical Engineering, University of California, Berkeley, USA

ARTICLE INFO

Article history:
Received 28 May 2014
Received in revised form 17 January 2015
Accepted 16 February 2015
Available online 26 March 2015

Keywords:
Data fusion
Probe data
GPS
Travel time
Estimation

ABSTRACT

With the rapid growth of communications technologies, GPS, and the mobile internet, an increasing amount of real-time location information is collected by private companies and could be marketed for retail. This body of data offers transportation agencies potential opportunities to improve operations, but it also presents unique challenges. This article investigates the question of how much GPS data is needed to power a traffic information system capable of providing accurate speed (and thus travel time) information. A hybrid data framework is proposed to use real-time, GPS-based, point-speed data from mobile sources to augment previous investments in existing fixed sensors. In addition, a systematic analysis of the performance trade-offs among a menu of data sources is described.

The results presented in this article were generated from the first procurement of streaming probe data from the private sector conducted on behalf of the California Department of Transportation and executed by UC Berkeley. Third-party data were incorporated with loop detector data and travel times were estimated within the bounds of driver variability. This achievement was repeated over multiple weeks and multiple congested freeway sites.

Our findings indicate that penetration rates for GPS-based probe data are now suitable for travel time estimation on selected corridors. Data fusion makes possible the effective use of data from multiple sources or providers; when data from multiple sources are fused, superior results are obtained. On a freeway that is already instrumented with loop detectors, better travel time performance may be achieved by fusing a relatively small amount of probe data than by doubling the number of loop detectors. Finally, the answer to how much GPS data is needed must address issues of data quality in terms of sample rate and penetration rate.

1. Introduction

1.1. Motivation

Actionable information is the lifeblood of effective transportation management. Traffic data can be used to estimate current traffic conditions so that travelers and agencies can make better decisions about how to use and manage the

* Corresponding author.
E-mail addresses: adpat@berkeley.edu (A.D. Patire), mwright@berkeley.edu (M. Wright), boris.prodhomme@polytechnique.edu (B. Prodhomme), bayen@berkeley.edu (A.M. Bayen).

http://dx.doi.org/10.1016/j.trc.2015.02.011
0968-090X/© 2015 Elsevier Ltd. All rights reserved.

transportation network. As the challenges of traffic and congestion increase, particularly in urban areas and freight-heavy intercity corridors, real-time traffic information becomes steadily more important.

Currently, transportation agencies capture traffic data primarily from fixed sensors, such as loop detectors, that are relatively expensive to install and maintain. However with the recent growth of communications technologies, GPS, and the mobile internet, an increasing amount of real-time location information is collected and distributed by private companies and even marketed for retail to public agencies such as state *departments of transportation* (DOTs).

This body of data offers transportation agencies a potential opportunity to improve operations, but it also presents unique challenges such as a scarcity of precedent for its procurement and use, and loss of direct quality control. Further, the data collected from GPS devices is limited to position information (from which velocities can be computed), while typical system control strategies, such as ramp metering, normally require density or occupancy data. Thus, while the results presented in this article are very promising for traffic information, more work is needed to establish similar results for traffic management and operations.

However, using GPS-based mobile probe data to complement existing detectors has real viability. This article demonstrates a *hybrid* traffic data system that effectively combines both mobile and fixed sensor data in a way that outperforms what is possible with only one type of data. This article shows that for freeways already instrumented with loop detectors, better travel time estimation performance may be achieved by fusing a relatively small amount of probe data than by doubling the number of loop detectors. These additional probe data may correspond to penetration rates on the order of only 0.2%. In addition, the question of how much GPS data are needed for travel time estimation cannot be answered without thorough investigation of its composition in terms of sample rate and penetration rate. This article illustrates quantitatively the relationship between travel time estimation performance and these measures of data quality.

1.2. Related work

Within the broad field of estimation numerous approaches have been developed, which include sequential estimation, variational data assimilation, statistical filtering, etc. A complete discussion of this work is beyond the scope of this review. For surveys of data fusion literature in a broader context of intelligent transportation systems, see Faouzi et al. (2011) and Zhang et al. (2011).

This review focuses on studies that fuse two or more types of data such as flows or occupancies from loop detectors, point-speeds from GPS sources, or travel times based on some type of re-identification technology, such as Bluetooth. Previous work in the milieu of multisensor data fusion for the purpose of traffic speed estimation can be categorized into four main groups depending on whether experimental or simulated data are used and whether or not the algorithmic approach employs a physical model of traffic dynamics.

Simulated data and no physical model: In Geisler et al. (2012), microsimulation is used to generate mock GPS-based data that are then map-matched to the road network. A classification algorithm is used to identify four levels of traffic: free, dense, slow, and congested. In Bachmann et al. (2013), microsimulation is used to generate synthetic loop and Bluetooth-like probe data that are then used to estimate average link speeds. Performance is evaluated for several statistical techniques that do not explicitly model the physics (Bachmann et al., 2013).

Simulated data and physical model: Physical notions from traffic flow theory appear in the data fusion approach explored in Van Lint and Hoogendoorn (2010) wherein the so-called extended generalized Treiber–Helbing filter is proposed. This method takes as input simulated data in the form of point-speeds, fixed sensor data, and travel times. Estimates of link speeds are generated; performance measures are calculated against a filtered version of the simulated ground truth.

Experimental data and no physical model: In Bachmann et al. (2012) a set of statistical techniques is employed to estimate travel times on a 20.8 km stretch of freeway in Canada. These techniques make use of both data from loops along the freeway as well as data from two Bluetooth devices installed at the most upstream and most downstream ends of the site (Bachmann et al., 2012). One day of GPS-based floating car data are used as the validation set. Another study uses Dempster–Shafer inference to classify travel times into one of four categories using loop and toll collection data on a motorway in France (Faouzi et al., 2009).

Experimental data and physical model: In a previous study on freeways (Mazare et al., 2012), tradeoffs between loop data and GPS-based probe data are investigated for estimating travel times using a data fusion framework based on ensemble Kalman filtering (Evensen, 2003) and a traffic model based on the Godunov scheme (Work et al., 2010). Employed are real data from loops, as well as GPS point-speeds from Mobile Century – a controlled experiment in which drivers were hired to circulate along a fixed route on I-880 in California (Herrera et al., 2010; Patire et al., 2010). Utilizing the Mobile Century data set, probe data were generated at specific virtual trip lines (Virtual trip lines, or VTLs, are a privacy-aware spatial sampling scheme (Hoh et al., 2008) that obviates the need for map-matching). Results in Mazare et al. (2012) are limited to the experimental data over a portion of one day and one set of travel times for validation over a single route on a freeway.

As mentioned above, the main sources of data on freeways are typically fixed sensors. Data assimilation methods that have been designed for, or demonstrated with, fixed sensor data to estimate traffic conditions on freeways may also have potential application to fuse fixed and mobile probe data. Examples of these methods include extended Kalman filtering in combination with a second-order traffic model (Wang and Papageorgiou, 2005), particle filtering in combination with a second-order traffic model (Mihaylova et al., 2007), and the Fourier–Galerkin projection method with minimax estimation

(Tchrakian and Zhuk, 2014). The present work explores the use of ensemble Kalman filtering with a first-order traffic model as in Work et al. (2010).

Beyond the specific topic of data fusion, recent literature is rich with studies that make use of GPS-based data, especially on arterials where the fixed sensing infrastructure has lagged behind that of freeways. For an exhaustive survey of the mining of taxi GPS traces, see Castro et al. (2013) in which this literature is classified into studies of social dynamics, traffic dynamics, and operational dynamics. Traffic-focused studies typically employ GPS-based data to estimate travel times or queue lengths. Statistical methods to estimate travel times on arterials have been targeted to make efficient use of sparsely sampled GPS-based data and assume no prior knowledge of signal timing plans. For example, techniques using Coupled Hidden Markov Models have been studied in Herring et al. (2010). A statistical traffic flow model based on hydrodynamic theory to implicitly model horizontal queues is demonstrated in Hofleitner and Bayen (2011). Studies that attempt to estimate queue lengths on arterials typically require high penetration rates on the order of 20% or more and sometimes assume knowledge of signal timing plans. For examples, see Feng et al. (2014) and Sun and Ban (2013).

At the intersection of crowdsourcing and the growing pervasiveness of GPS-enabled smartphones there has been an explosion of mobile apps such as: (1) BostonStreetbump (StreetBump, 2015) to crowdsource pothole locations for road maintenance in Boston; (2) Green Light Optimal Speed Advisory (Koukoumidis et al., 2011) to advise drivers the optimal speed at which to approach a light to avoid having to brake; (3) RasteyRishtey (Sen, 2014) to assist with organizing ad hoc meet-ups in developing countries; and, (4) WreckWatch (White et al., 2011) to automatically detect vehicular accidents using accelerometer and contextual data.

A number of private companies collect and aggregate traffic-related data from sources (such as fleets, navigational devices, and users of smartphone routing apps) that include GPS-based point speeds. These companies use proprietary systems and algorithms to estimate traffic conditions and to package this information for retail in the form of link speeds updated at a rate of about one minute. One study compared link speeds from the company INRIX with loop data over a two-month period to assess its quality. Kim and Coifman (2014) reports issues of latency and repeated reported speeds that are crucial considerations for any future real-time traffic management applications that may depend on third-party data. As distinct from Kim and Coifman (2014), the present work investigates a sample of non-aggregated, GPS-based, point speeds representative of what may be available as source inputs used by these third-party suppliers of traffic data.

Issues of oversight and quality control for third-party data to be procured from the private sector are increasingly receiving greater attention. One notable study (Schneider IV et al., 2010) performed for the Ohio DOT generated travel time data service evaluation procedures to be used in contract provisions. The effectiveness and accuracy of floating car studies was compared with what is possible with the deployment of Bluetooth technology (Schneider IV et al., 2010). The I-95 Corridor Coalition employs a method in which average speeds for validation (typically determined by Bluetooth re-identification) are compared against speeds from probe data calculated over TMC (traffic messaging channels). An *average absolute speed error* (AASE) as well as a *speed error bias* (SEB) is calculated for a set of speed bin categories (I-95 Corridor Coalition, 2010).

One theoretical study proposes an information measure to quantify the value of heterogeneous traffic measurements (Deng et al., 2013). Cumulative flow count is used as the state variable, and scenarios are created that involve the insertion of an additional fixed sensor, addition of vehicle re-identification at the upstream and downstream ends of the modeled freeway, or insertion of GPS-based probe data. For each scenario the uncertainty reduction in the internal traffic state is calculated. Trajectory data from NGSIM (2013) are used to generate synthetic data to implement the experiments in this framework. As with Deng et al. (2013), the present work also quantifies the value of heterogeneous traffic measurements, but in the context of strictly experimental data on three different sites, over a period of six weeks.

A trade-off analysis between penetration rate and sampling rate of GPS-based probe data is provided by Bucknell and Herrera (2014). Indifference curves are calculated to illustrate the marginal value of modifying the composition of the probe data source with respect to velocity state estimation error. Once again, NGSIM trajectories are used to generate synthetic probe data. A second evaluation is performed using trajectories from a microsimulation. As with Bucknell and Herrera (2014) the present work also quantifies the impact of penetration rate and sampling rate. However, the present work addresses these issues in the context of strictly experimental data and with respect to travel time estimation performance.

Unlike previous studies, the present work addresses practical challenges of using commercially available GPS-based probe data. For example, commercially available, unaggregated GPS point-speed data of today do not use the VTL scheme, but instead are time-sampled at rates between 0.5 and 60 times per minute. As a result, substantial processing is required to project the latitude and longitude information from GPS point-speed data onto the road network. This study employs the *Path Inference Filter* of Hunter et al. (2013) to process the raw probe data feeds that were procured as a part of the hybrid data project described in Bayen et al. (2013b) and Bayen et al. (2013).

One limitation of many other studies is that GPS-based probe data is assumed to have homogeneous sample rates. In practice, experimental data collected from multiple providers features heterogeneous characteristics, such as having a mixture of high- and low-frequency probe data. Due to nonlinearities of traffic dynamics, it is not obvious how to extend idealized theoretical results to the inherent heterogeneities of “Big Data.”

The present work provides a theoretical framework for data fusion, an algorithmic foundation to evaluate the relative usefulness of non-aggregated, GPS-based, point-speeds and fixed sensor deployments. A fusion engine is demonstrated that can incorporate third-party data to estimate travel times within the bounds of driver variability, and this achievement is repeated over multiple weeks and multiple congested freeway sites. In addition, the present work provides a quantitative evaluation of the additional predictive power that probe data can bring to a transportation information system that may already use data from existing sensing infrastructure (such as loops) and thus informs near-term data procurement strategies for transportation agencies.

1.3. Structure of the article

This article proceeds with a brief summary of the developed approach. Following is a description in greater detail of both the methods and the framework necessary for its implementation. Next, travel time estimation results are presented on two urban test sites using loop and probe data and one rural test site using only probe data. Finally, the article concludes with a summary of findings and their implications for transportation agencies.

2. Developed approach

In this study, travel time was chosen as the metric of interest. Travel time is intuitive, and of high interest to the traveling public and public agencies. Furthermore, it is a quantity that can be measured effectively and reliably using various technologies, in particular Bluetooth (Haghani et al., 2010; Martchouk et al., 2011).

Fig. 1a illustrates the data flow for data fusion used in our approach. Raw data feeds providing loop or probe data require substantial filtering to convert the data into a suitable form. Subsets of available data are fused and the output consists of a reconstruction of traffic state in the form of spacetime velocity maps. By themselves, the velocity maps are informative, but require further processing to make an appropriate comparison with independently obtained travel times.

Fig. 1b illustrates the additional data flow for calibration and for performance evaluation. An independent source of travel times obtained from Bluetooth sensors also requires considerable filtering. Vehicle trajectories consistent with the velocity maps are generated and used to estimate travel times along routes between the Bluetooth sensors. Means and standard deviations are calculated over fixed intervals for both the travel times generated from the velocity maps and for the travel times generated by the filtered Bluetooth measurements. These statistics are then compared for calibration and for performance evaluation.

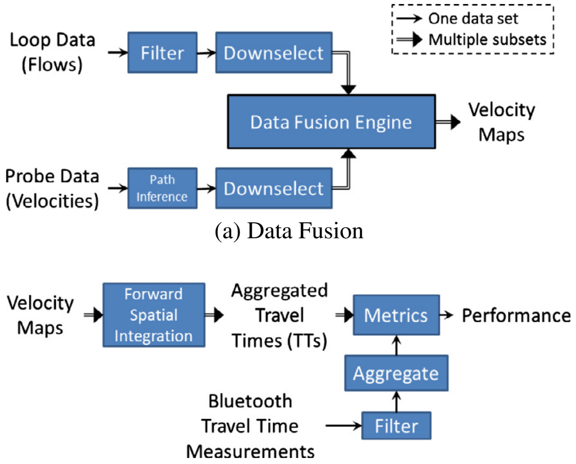
The method introduced in Fig. 1 is appropriate for both real-time and historical applications.

2.1. Data sources

As described in detail in Bayen et al. (2013b) and Bayen et al. (2013), real-time probe feeds were established for 90 days. Each anonymized GPS point-speed datum included at a minimum the following fields: temporary ID, timestamp, latitude, longitude, and velocity.

A real-time loop feed was established via PeMS (2013), providing timestamped values of occupancy and counts over fixed time intervals (e.g., 30 s).

Bluetooth sensors were installed in two-week deployments along three sites of interest. Each raw datum consisted of timestamp and unique MAC address. MAC addresses were hashed to alleviate privacy concerns. During daytime hours, the Bluetooth flow fraction was about 5% of total flow measured by nearby loops along two sites of interest (Bayen et al., 2013b).



(a) Data Fusion

(b) Calibration and performance evaluation

Fig. 1. Schematic diagram of developed method to illustrate the data flow between subsystems.

2.2. Filters

In this article, the term "filtering" is used to refer to any number of computational steps to prepare data obtained from a "raw" feed for use with the data fusion engine. These computational steps are roughly sorted into one of three categories. The first category has to do with identifying and discarding invalid data. Data might be invalid because it is a duplicate, it is missing required fields, its values are outside the range of physical possibility, etc. The second category has to do with filtering noise or measurement error. In addition, data that might be valid might also correspond to a known failure mode of the underlying detector or source; these data are filtered out. Finally, the third category has to do with aligning data with existing infrastructure. One example of this third category is map-matching sparsely sampled probe data to a probable path through the road network.

The loop data filter inherited from *Mobile Millennium* (Bayen et al., 2011) was used to calculate velocity from the provided occupancies and counts as well as to remove irregular or clearly erroneous reports. This filter uses a heuristic method to identify known failure modes of loops, and the loop feed.

The *Path Inference Filter* (PIF) is used to project GPS probe points onto a road network, and to distinguish points on the freeway of interest from nearby points on an overpass or parallel road (Hunter et al., 2013). PIF takes in a series of latitude, longitude, and time data associated with a particular device and reconstructs a list of possible trajectories, each with associated probabilities. Only PIF mappings that contain more than 70% of the probability are accepted for downstream processing.

Bluetooth data are filtered to handle multiple detections by the same detector and to remove devices with outlier travel times. For example, it is common during congestion for any particular Bluetooth device to stay within the range of a Bluetooth detector for multiple scans. This situation might result in a series of multiple timestamps for the same device at both the beginning and at the end of a route between two detectors. For this study, travel time was calculated from the average timestamp at the beginning of the route to the average timestamp and the end of the route. The interested reader is referred to Bayen et al. (2013b) for a detailed description of the analysis justifying this choice.

2.3. Data downselection

A systematic downselection method for the data was employed to determine the performances achievable for combinations of available data sources, and to determine how marginal increases in available data can improve the accuracy of travel time estimation.

Loop data were downselected by assigning loops to sets of roughly equal size and progressively removing the least marginally informative loops. This pattern models a deployment strategy consistent with first installing detectors in locations with greatest informational content.

The downselection strategy for probe data needed to reflect the fact that, during any time interval, the information content from having one data point from each of ten different vehicles is different from the information content from having ten data points all from the same vehicle.

Two data quality measures facilitate discussion: *Sampling rate* is the average rate at which any device reports its position and velocity. Any data set will have a distribution of devices with a range of sampling rates (typically between 0.5 and 60 reports per minute). *Penetration rate* is the flow fraction of vehicles (unique devices) reporting to the probe data set as compared to the total flow of vehicles along a road.

Probe data is downselected in two different ways. Downselecting *by bulk* is performed to approximate the effect of fewer datapoints in any discretized region of spacetime. This is done by limiting the total number of *reports* allowed into the subset within any spacetime bin, regardless of the IDs of the devices associated with each report. Conversely, downselecting *by unique ID* is performed to approximate the effect of changing the market penetration of the data source. This is done by limiting the number of unique devices allowed into the subset per 30 s sample period, regardless of the sample rates of selected devices.

2.4. Calibration

The calibration framework developed as part of this project and in collaboration with Linköping University uses an algorithm based on the Complex Method (Box, 1965) to fine tune a predetermined parameter set. It is an iterative scheme that uses heuristics to search the parameter space, and works well in practice. A complete description can be found in Allström et al. (2014). For this application poor calibration is easily detectable; it results in incongruous occurrence and extent of traffic congestion when compared with Bluetooth measurements (as described below).

The parameters tuned during the calibration procedure are the *split ratios*, which define the proportions of vehicles that take each possible direction in a given junction; the *fundamental diagram parameters*, which define the relationship between the traffic density and flow (or speed); the *source capacities*, which define the input flows from the on-ramps and upstream origin; and the *Kalman Filter noise parameters*, which are required parameters for the filtering system used in data assimilation (fusion). One limitation of the implementation (not the method) was that only time-invariant parameters were used.

2.5. Data fusion

The data fusion approach in the present work is to integrate measurements into a flow model using an *Ensemble Kalman Filter* (EnKF) from Evensen (2009). This approach, commonly used in fields such as oceanography, is effective when measurements are relatively sparse and physical dynamics can be well approximated but not linearized. Highlights are summarized below, while more mathematical details are provided in Work et al. (2010).

The objective of the proposed data fusion engine is twofold. The first objective is to make effective use of heterogeneous data types, in this case both loop detectors and GPS-based probe data. The second objective is to infer traffic states even in places where direct measurements are unavailable. The inputs to data fusion are velocity measurements from loops or probes, and the outputs are velocity estimates everywhere along the freeway.

The physical model is taken as the common LWR partial differential equation from Lighthill and Whitham (1955) and Richards (1956)

$$\frac{\partial\rho(x,t)}{\partial t}+\frac{\partial q(x,t)}{\partial x}=0 \tag{1}$$

where ρ and q represent the density and flow of vehicles along the freeway of interest. Partial derivatives are taken with respect to time, t, and a single dimension of space, x, corresponding to a single-pipe model.

The fundamental diagram is expressed as a velocity function $v=V(\rho)$. As in Work et al. (2010), the Smulders fundamental diagram is chosen:

$$v=V(\rho)=\begin{cases} v_{\max}\left(1-\frac{\rho}{\rho_{\max}}\right) & \text{if}\,\rho\leqslant\rho_c \\ -w_f\left(1-\frac{\rho_{\max}}{\rho}\right) & \text{otherwise} \end{cases} \tag{2}$$

where $v_{\max},\ \rho_{\max},\ \rho_c$ and w_f are the maximum velocity, maximum density, critical density for transition from free-flow to congestion, and the backward-propagating wave speed, respectively.

Recall that our goal is to estimate travel times by assimilating velocity data. For simplicity in implementation, and also for conceptual clarity, the entire model is also converted to the velocity domain. This conversion is possible because the Smulders fundamental diagram is invertible.

The Godunov scheme (Godunov, 1959) is employed to discretize Eq. (1). Rewriting the result with velocity as the state variable yields the following one-step *velocity cell transmission model* (v-CTM) as derived in Work et al. (2010):

$$v_i^{n+1}=V\left(V^{-1}(v_i^n)-\frac{\Delta T}{\Delta x}\left(\widetilde{G}(v_i^n,v_{i+1}^n)-\widetilde{G}(v_{i-1}^n,v_i^n)\right)\right) \tag{3}$$

where v_i^n is the velocity associated with Godunov cell i at timestep $n,\ \Delta T$ is the time step, and Δx is the spatial step. Once again, $V(\cdot)$ denotes the Smulders fundamental diagram from Eq. (2), and $V^{-1}(\cdot)$ denotes its inverse. To ensure numerical stability the time and spatial steps are constrained by the CFL condition (Courant et al., 1967) such that: $v_{\max}\frac{\Delta T}{\Delta X}\leqslant 1$. In the case of the Smulders fundamental diagram, the transformed Godunov velocity flux, $\widetilde{G}(\cdot,\cdot)$, is a function of two variables, the velocity state of the sending cell, v_1, and the velocity state of the receiving cell, v_2:

$$\widetilde{G}(v_1,v_2)=\begin{cases} v_2\rho_{\max}\left(\frac{1}{1+\frac{v_2}{w_f}}\right) & \text{if}\,v_c\geqslant v_2\geqslant v_1 \\ v_c\rho_{\max}\left(1-\frac{v_c}{v_{\max}}\right) & \text{if}\,v_2\geqslant v_c\geqslant v_1 \\ v_1\rho_{\max}\left(1-\frac{v_1}{v_{\max}}\right) & \text{if}\,v_2\geqslant v_1\geqslant v_c \\ \min\left(V^{-1}(v_1)v_1,\ V^{-1}(v_2)v_2\right) & \text{if}\,v_1\geqslant v_2 \end{cases} \tag{4}$$

where v_c is the critical velocity, the velocity at the critical density ρ_c.

The junction model of Piccoli (Work et al., 2010) is employed to extend the velocity model to a road network represented by a directed graph, thus allowing freeway onramps and offramps to be modeled. Details are explained in Work et al. (2010). The state update algorithm, summarized as $\mathcal{M}(\cdot)$, involves two main steps:

1. For all junctions, compute the maximum admissible inflows and outflows, and Piccoli junction flows using the procedure of Work et al. (2010).
2. Compute the one-step velocity update as shown in Eq. (3).

For the purpose of estimation, uncertainty is explicitly added to the model

$$v^n=\mathcal{M}(v^{n-1})+\eta^n \tag{5}$$

where η^n represents "process" noise taken from a zero-mean, Gaussian distribution. Measurement data at each time step, y^n, are modeled as noisy versions of the true velocity state,

$$y^n = \mathbf{H}^n v^n + \chi^n \tag{6}$$

where $\mathbf{H}^n$ is a linear observation matrix and χ^n represents measurement noise taken from a distribution with zero mean and covariance matrix $\mathbf{R}^n$.

The EnKF is a sequential technique shown schematically in Fig. 2. At each time step the physical model, $\mathcal{M}(\cdot)$, is used to calculate the next forecast state, v_f^n. The state error covariance is represented by the ensemble covariance, $\mathbf{P}^n_{\text{ens},f}$. The difference between the forecast state, v_f^n, and measurements from the field, y^n (wherever they exist as specified by $\mathbf{H}^n$), is called the innovation. This innovation is scaled by the Kalman gain, $\mathbf{G}^n_{\text{ens}}$, and used to calculate an improved state estimate, v_a^n, the analyzed state. The process is as follows:

1. *Initialization*: Draw K ensemble realizations $v_a^0(k)$ (with $k \in \{1, \cdots, K\}$) from a process with a mean speed $\bar{v}_a^0$ and covariance $\mathbf{P}_a^0$.
2. *Forecast*: Update each of the K ensemble members and update the ensemble mean and covariance according to:

$$v_f^n(k) = \mathcal{M}(v_a^{n-1}(k)) + \eta^n(k) \tag{7}$$

$$\bar{v}_f^n = \frac{1}{K}\sum_{k=1}^{K} v_f^n(k) \tag{8}$$

$$\mathbf{P}^n_{\text{ens},f} = \frac{1}{K-1}\sum_{k=1}^{K}\left(v_f^n(k) - \bar{v}_f^n\right)\left(v_f^n(k) - \bar{v}_f^n\right)^T. \tag{9}$$

3. *Analysis*: Obtain measurements, compute the Kalman gain, and update the network forecast:

$$\mathbf{G}^n_{\text{ens}} = \mathbf{P}^n_{\text{ens},f}(\mathbf{H}^n)^T\left(\mathbf{H}^n\mathbf{P}^n_{\text{ens},f}(\mathbf{H}^n)^T + \mathbf{R}^n\right)^{-1} \tag{10}$$

$$v_a^n(k) = v_f^n(k) + \mathbf{G}^n_{\text{ens}}\left(y^n - \mathbf{H}^n v_f^n(k) + \chi^n(k)\right). \tag{11}$$

4. Return to step 2.

One contribution of the present work was in applying the calibration procedures outlined in Section 2.4 to fine-tune the measurement noise for each of the data sources employed. An expanded version of Eq. (6) can be written as

$$\begin{bmatrix} y_l^n \\ y_a^n \\ y_b^n \end{bmatrix} = \begin{bmatrix} \mathbf{H}_l^n \\ \mathbf{H}_a^n \\ \mathbf{H}_b^n \end{bmatrix} v^n + \begin{bmatrix} \chi_l^n \\ \chi_a^n \\ \chi_b^n \end{bmatrix} \tag{12}$$

with subscripts l, a, and b corresponding to the loop, probe A, and probe B data sources, respectively. Characteristics of probe sets A and B are provided in Section 3.2. Each of χ_l^n, χ_a^n, and χ_b^n represent measurement noise taken from distributions with covariance matrices $\mathbf{R}_l^n$, $\mathbf{R}_a^n$, and $\mathbf{R}_b^n$, respectively. The specification of the right covariance to use depends greatly on which mix of data sources are to be fused.

Note that noise is explicitly added to the measurements in Eq. (11). At each time step n, and for each ensemble k, a zero-mean noise vector $\chi^n(k)$, constructed according to Eq. (12), is added to the measurement vector. In this way, each physical measurement is represented as an ensemble with a mean corresponding to the actual measurement and a covariance equal to the assumed measurement error covariance.

Data are assimilated through the analysis step as they would arrive in a real-time system (here, 30 s intervals for loop data and one probe data feed, and 1 min intervals for the other probe data feed). *Data fusion* occurs in several steps. The first

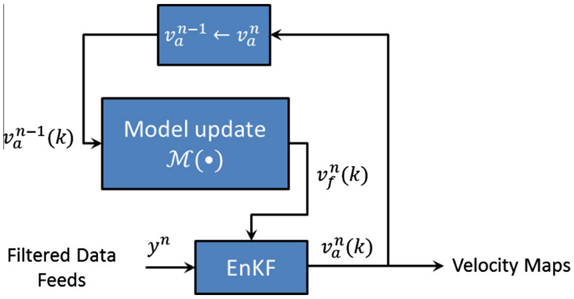


Fig. 2. Schematic representation of the data fusion engine. The model update, $\mathcal{M}(\cdot)$, summarizes equations for the Godunov scheme and the junction model. It takes as input the state vector from the analysis step of the EnKF. It outputs the next state forecast to be assimilated with available measurements. Note that the model is invoked for each member of the ensemble with $k \in \{1, \cdots, K\}$.

step is by constructing y^n according to Eq. (12) with velocity measurements as resulting from the probe data sources, y_a^n and y_b^n, as well as those resulting from the loop data filter, y_l^n. The next step occurs in the analysis step of Eq. (11). If no measurement data at all is available at a certain forecast time step, n, then a small amount of zero-mean process noise is added to the state estimate. If (partial) measurement data are available, then the estimate consists of a weighted average of the measurement and the forecast of the model. Note that even in cells where no measurement data are available, the estimate depends on the state of statistically correlated cells as determined by the correlation matrix, $\mathbf{P}^n_{\text{ens},f}$. In this application, the forecast step may correspond to multiple iterations of the Godunov scheme. For all results in this article, time is discretized into 6 s intervals and space is discretized into roughly 200 m sections.

2.6. Travel time calculations

The two sources for travel time data in this study are (1) forward spatial integration of the estimated velocity maps, and (2) field data from Bluetooth vehicle re-identification, as shown in Fig. 1b. The following subsections explain how these sources are then processed to calculate travel time statistics for performance evaluation.

2.6.1. Average travel times from Bluetooth re-identification

The piecewise-linear trajectories illustrated in Fig. 3 are intended to correspond to assumed trajectories of individual vehicles (real, not simulated) based on Bluetooth re-identification as discussed in Section 2.2. The trajectories are drawn as straight lines to convey the limited granularity of the field data. The different slopes correspond to the fact that different vehicles realize different travel times.

The rectangular bins in Fig. 3 pertain to Bluetooth detectors deployed on the study site with edges in the spatial dimension corresponding to Bluetooth sensor locations, and indexed by j. The bins have a time duration of $T = 15$ min, and are indexed by i along the time axis. Note that the bins described here are completely independent from the Godunov scheme discretization in Section 2.5.

The inter-Bluetooth regions are referred to as “routes.” Individual vehicle travel times as measured by Bluetooth detectors are denoted as τ_{jk}, where k is the vehicle start time at the beginning of route j. In each spacetime region indexed by (i,j), average and standard deviations of travel time are calculated:

$$T_{BT}(i,j) = \frac{1}{\sum_{iT \leqslant k<(i+1)T} 1} \sum_{iT \leqslant k<(i+1)T} \tau_{jk} \tag{13}$$

$$S_{BT}(i,j) = \sqrt{\frac{1}{\sum_{iT \leqslant k<(i+1)T} 1} \sum_{iT \leqslant k<(i+1)T} (\tau_{jk} - T_{BT}(i,j))^2} \tag{14}$$

Since trajectories may not always begin and end in the same region, the beginning of the trajectory is used for aggregation purposes.

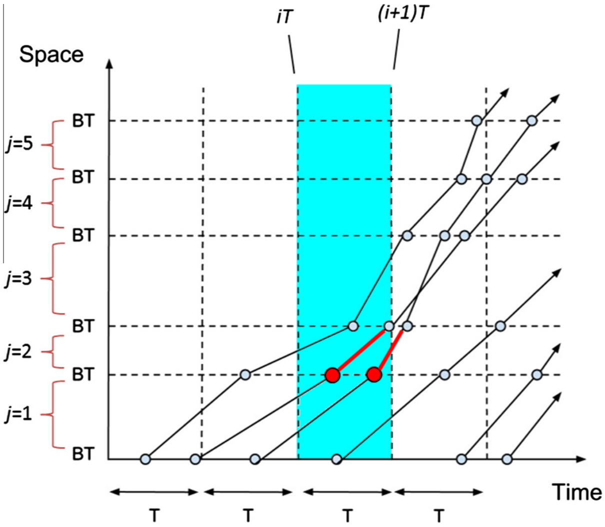


Fig. 3. Schematic representation of vehicle trajectories recovered from Bluetooth field data along a study site. Vehicles move from bottom to top. Bluetooth detectors are labeled on the y-axis and distributed unevenly along the road. Circles represent detections of a vehicle. Spatial segments (also called routes) between successive Bluetooth devices are indexed by j. The time axis is divided into segments of duration $T = 15$ min and indexed by i. The trajectories corresponding to time bin i and route $j = 2$ are colored red.

2.6.2. Average travel times from estimated velocity maps

While the data fusion engine creates an estimate of velocity over time and space (the velocity map), the measured value for comparison via the Bluetooth deployment is vehicle travel times. Therefore, it is necessary to generate comparable travel times consistent with the velocity maps according to the process illustrated in Fig. 1b.

Synthetic trajectories are generated by placing a dummy vehicle at the beginning of each Bluetooth route every 30 s and advancing them through space according to the velocity map at their current locations every 6 s. This process of forward spatial integration results in a set of 30 trajectories (and thus 30 travel times, τ_{jk}) in each 15 min bin. The means $T_{\text{model}}(i,j)$ and standard deviations $S_{\text{model}}(i,j)$ for each bin are then computed according to Eqs. (13) and (14), above.

2.7. Performance metrics

For the purpose of quantifying the performance of the data fusion engine and the accuracy of estimates computed from various data sources, several statistical metrics are described:

2.7.1. Mean Absolute Percentage Error (MAPE)

The *Mean Absolute Percentage Error* (MAPE) is a quantification of the difference between the estimated travel times and the travel time directly measured by the Bluetooth sensor deployment. It is computed as follows:

$$\text{MAPE} = \frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}\frac{|T_{\text{model}}(i,j) - T_{BT}(i,j)|}{T_{BT}(i,j)} \tag{15}$$

where T_{model} is the mean estimated travel time, T_{BT} is the mean Bluetooth travel time, and $i \in \{1,\ldots,N\}$ and $j \in \{1,\ldots,M\}$ are the time aggregation periods, and routes between Bluetooth detectors, respectively.

In congested traffic moving at 20 mph, and calculated over a 12 mile freeway segment, a 5 mph overestimation of speed will result in a MAPE of 20%. The MAPE metric as defined above is very strict in that small over- or under-estimations along M successive routes do not cancel out, they are added up instead.

It is also useful to compute a measure of the variability in traffic travel times as measured by Bluetooth detectors for comparison purposes. The statistical metric used is the BTMAPE:

$$\text{BTMAPE} = \frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}\frac{S_{BT}(i,j)}{T_{BT}(i,j)} \tag{16}$$

where $T_{BT}(i,j)$ and $S_{BT}(i,j)$ are the means and standard deviations from Eqs. (13) and (14). This metric is related to the traditional statistical measurement of coefficient of variation (used as a proxy for the noise floor).

2.7.2. Per Mile Absolute Time Error (PMATE)

The Per Mile Absolute Time Error (PMATE) is used to get an intuitive sense of the error as normalized by the length of the road section:

$$\text{PMATE} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{\sum_{j=1}^{M}L(j)}\sum_{j=1}^{M}|T_{\text{model}}(i,j) - T_{BT}(i,j)| \tag{17}$$

where $L(j)$ is the length of the road section indexed by j. Similarly to MAPE, a BTPMATE metric is computed and used as another proxy for the noise floor:

$$\text{BTPMATE} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{\sum_{j=1}^{M}L(j)}\sum_{j=1}^{M}S_{BT} \tag{18}$$

2.7.3. Congestion Classification Error in Congestion (CCEC)

The Congestion Classification Error in Congestion is a statistical metric used to quantify the accuracy of the data fusion engine in predicting congestion in a binomial sense. CCEC is computed as follows:

$$\text{CCEC} = \frac{1}{\sum_{i=1}^{N_c}\sum_{j=1}^{M_c}1_{\frac{T_{BT}(i,j)}{L(j)}>P}}\sum_{i=1}^{N_c}\sum_{j=1}^{M_c}\left|1_{\frac{T_{\text{model}}(i,j)}{L(j)}>P} - 1_{\frac{T_{BT}(i,j)}{L(j)}>P}\right|1_{\frac{T_{BT}(i,j)}{L(j)}>P} \tag{19}$$

where P is the threshold pace for congestion, and $i \in \{1,\ldots,M_c\}$ and $j \in \{1,\ldots,N_c\}$ index the time aggregation periods, and routes between Bluetooth detectors, respectively, of the spacetime regions *in congestion*. Here, congestion is defined as any spacetime region for which the average speed as measured by Bluetooth is less than 40 mph, corresponding to a threshold pace $P = 1.5$ min/mile. The indicator function $1_{\frac{T_{\text{model}}(i,j)}{L(j)}>P}$ is equal to one whenever the condition $\frac{T_{\text{model}}(i,j)}{L(j)} > P$ is satisfied (the model predicts congestion), and equal to zero otherwise (the model predicts freeflow).

3. Results

Trial experiments were performed for three sites in California. Two urban sites were chosen: one along northbound I-880 and another along northbound I-15 as depicted in Fig. 4a and b, respectively. One rural site was selected along southbound I-15 near Victorville as shown in Fig. 5. The two urban sites had active loop detectors, while the rural site did not (during April of 2012, when data were collected). Summary results are provided here; detailed presentation is available in Bayen et al. (2013b).

3.1. Validation data

Four weeks of validation data for the urban sites are shown in Fig. 6. Displayed are average velocities calculated from average travel times along successive routes between Bluetooth detectors deployed along the highway. Numbered rectangles demark congestion events in spacetime along the I-880 site in Fig. 6a and b, and along the I-15 site in Fig. 6c and d.

The goal in this study was to evaluate the accuracy of travel time estimates from heterogeneous sources of data (loops and probes). During free flow conditions, estimation of travel time is trivial. Therefore, performance metrics were calculated only during the congestion events identified by numbered rectangles in Fig. 6: before congestion, during the onset of congestion, during queue clearance, and immediately after congestion. For each selected congestion event, both probe and loop data sources exhibited typical operational performance, with no unusual outages, thus facilitating a fair comparison of their relative predictive power.

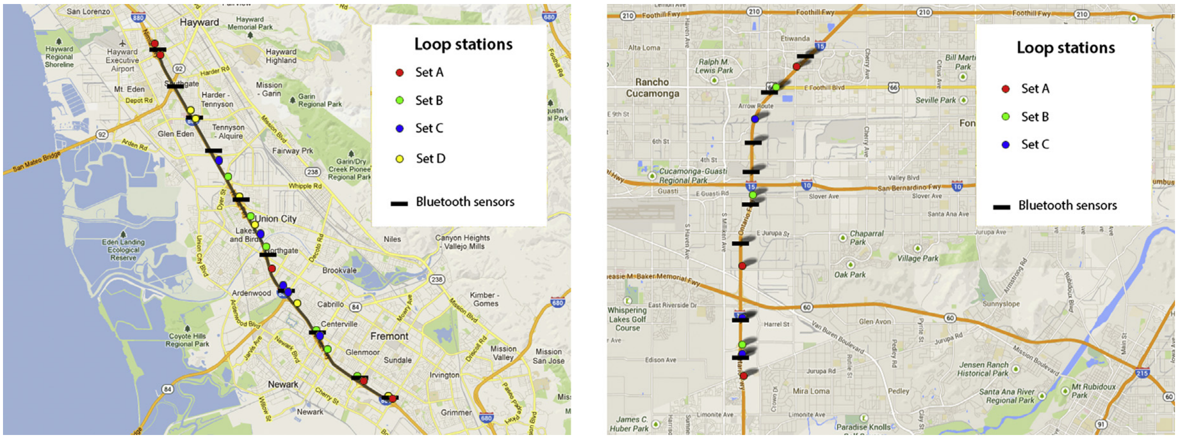
3.2. Probe data characteristics

The probe data are divided into two subsets: *Set A* and *Set B*. Set A had a high sample rate and a low penetration rate whereas set B had a relatively low sample rate and a relatively high penetration rate as shown in Table 1 for the I-880 site. Distributions of sample rates are further illustrated in Fig. 7. Data from individual GPS devices in set A are typically sampled at either 30 or 60 times per minute as shown in Fig. 7a. Most data from individual GPS devices in set B are sampled at less than 2.5 times per minute as shown in Fig. 7b. This division of data facilitated a comparison between data quality and travel time estimation performance.

Daily trends of probe flows and penetration rates along I-880 northbound are explored in Fig. 8. On weekdays, the flow profile of set B probes appears to be shifted forward in time as compared to the flow profile of general traffic as measured by loop detectors and illustrated in Fig. 8a. This pattern results in a penetration rate for set B that falls off quickly during the evening rush as shown in Fig. 8c. On weekends, the flows and penetration rates of probes in set B are significantly lower than those on weekdays, as shown in Fig. 8b and d. Although probes in set A have comparatively lower flows and penetration rates, their values are more stable throughout the day.

3.3. Fusion comparison

The data fusion engine as described in Section 2.5 was used to reconstruct the traffic state over the I-880 study site and two-week time period using filtered loop and probe data. As an example of the inputs, all loop and all probe data (both sets



(a) A 12-mile section on I-880 north bound between Fremont and Hayward. Ten Bluetooth detectors divide the freeway into nine segments. The 22 loop detectors are divided into four sets.

(b) A 9-mile section on I-15 northbound near Ontario, California. Eight Bluetooth detectors divide the freeway into seven segments. The 9 loop detectors are divided into three sets.

Fig. 4. Maps of two urban freeway study sites.

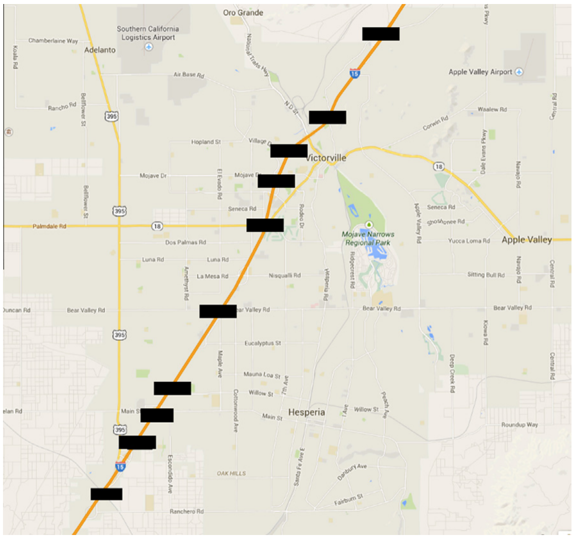


Fig. 5. Map of a rural, 19 mile section on I-15 southbound through Victorville. Ten Bluetooth detectors denoted by black horizontal bars divide the freeway into nine segments. No loop detectors were active during the time of this study.

A + B) available during the 24 h period of March 7, 2012 are illustrated in Fig. 9a and b. Areas colored black in Fig. 9 indicate in space and in time the cells in the v-CTM model for which no direct state measurements are available.

Fig. 10 presents visual representations of the reconstructed state over the same 24 h period based on some or all of the available data. Fig. 10a shows the state as reconstructed from the loop data of Fig. 9a only. Fig. 10b shows the state as reconstructed from the probe data of Fig. 9b only. Fig. 10c shows the state as reconstructed from both loop data of Fig. 9a and probe data of Fig. 9b. Although Fig. 10d shows the Bluetooth measurements as transformed to average velocities for easier visual comparison, the raw data used in computing the statistical metrics defined in Section 2.7 are travel times, not velocities.

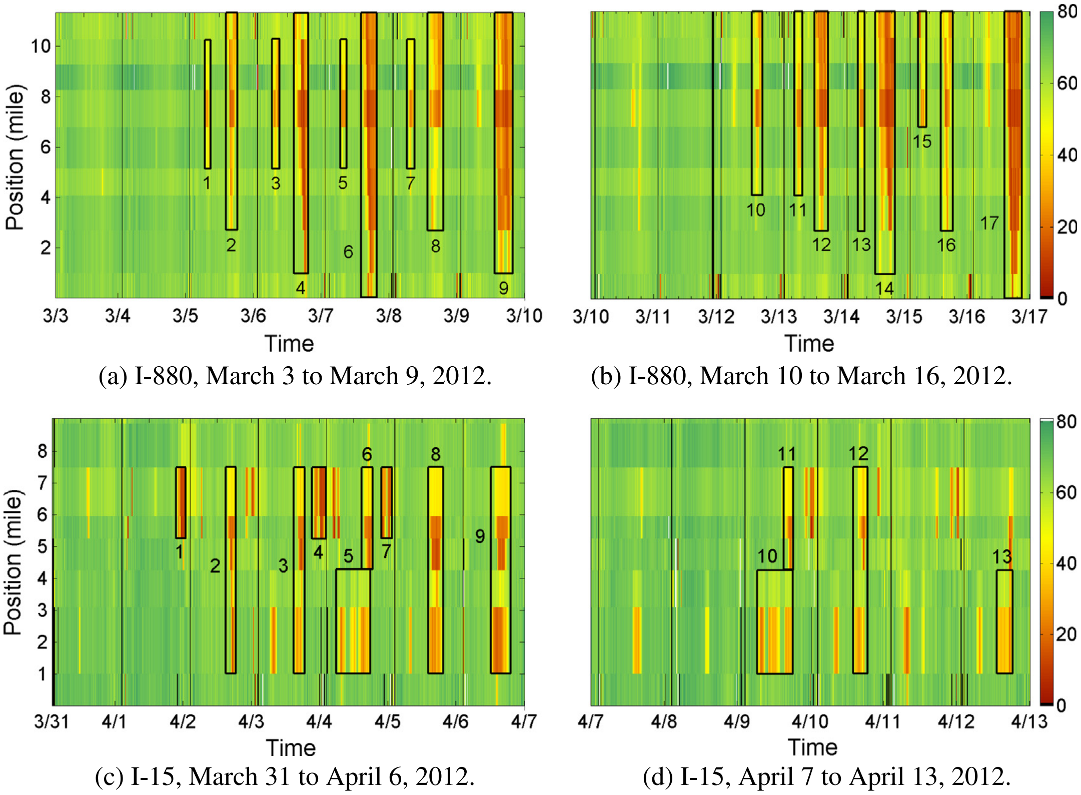
Travel time estimation performance is calculated over three sites for combinations of input data; summary results are presented in Table 2. Results for the I-880 site are calculated over the seventeen congestion regions of Fig. 6a and b. Results for the I-15 site near Ontario are calculated over the thirteen congestion regions of Fig. 6c and d. Use of multiple data sources produced dramatically more accurate estimations as measured by all three metrics, with the metric values dropping below those of the noise floors when all three sources are used. Two notions of noise floor, BTMAPE and BTPMATE, are used in this study to quantify the underlying variability in the validation data as calculated according to Eqs. (16) and (18), respectively.

During the survey period along the rural site of southbound I-15 near Victorville, a significant congestion event occurred on April 22, 2012. Results calculated over an 8 h period on that day are shown on the bottom rows of Table 2. During this time, the combined probe data set contained 2.63 unique devices per 5 min per mile. Unfortunately, there were not enough congestion periods to calibrate the parameters. Results are provided for the default *Mobile Millennium* (Bayen et al., 2011) parameter values. Using only probe data, good agreement was achieved between measured and estimated travel times.

Further investigation of the effects of marginal probe data availability was conducted by using data sets A and B separately to estimate travel times for the I-880 study site. Table 1 quantifies the differences in sampling characteristics between the two sets for this site. Fig. 11 illustrates the reduction in error as additional data is made available to the data fusion engine. In Fig. 11a, data is downselected by unique devices (as described in Section 2.3) and MAPE shows a relatively steady decrease as data from additional unique devices are used. Since set B has a higher penetration rate, it achieves a lower minimum MAPE. In Fig. 11b, however, data is downselected by bulk (as described in Section 2.3). This figure shows that even though set A has several factors more data points available due to the higher average sampling rate, addition of more data points produces diminishing returns in MAPE, since these marginally added data points are reports from the same devices.

This contrast in behavior of severe diminishing returns in Fig. 11b and moderate degradation in Fig. 11a can be explained by the observation that two data points from the same vehicle closely spaced in time should be well correlated. As a result, the marginal value (in terms of information) of the next data point from that same vehicle is reduced. The results of Fig. 11 suggest that the question of how much GPS data is needed for a particular application cannot be answered without thorough investigation of its quality in terms of sample rate and penetration rate.

To summarize: On freeways, a high penetration rate of probes is to be preferred over a high average sample rate. When data from multiple sources are fused, superior results are obtained. This analysis provides a framework for quantifying the usefulness of probe data, as is further described next.



(a) I-880, March 3 to March 9, 2012.

(b) I-880, March 10 to March 16, 2012.

(c) I-15, March 31 to April 6, 2012.

(d) I-15, April 7 to April 13, 2012.

Fig. 6. Validation data set as calculated from Bluetooth travel times and presented as average velocities. Numbered rectangles demark congestion events. Vehicles move from bottom to top; time moves from left to right. Color scale indicates miles per hour. Regions with no data are colored black. Brief suspensions of data collection occurred nightly at 2 a.m. and appear as slender vertical lines. The thick black vertical line in (b) on March 11 is due to daylight savings time.

Table 1
Attributes of probe data sets on I-880 during the selected congested periods of Fig. 6.

Attribute	Set A	Set B
Unique reports per 5 min per mile	1.0	1.6
Total reports per 5 min per mile	17.2	2.0
Average penetration rate by flow	0.1%	0.6%

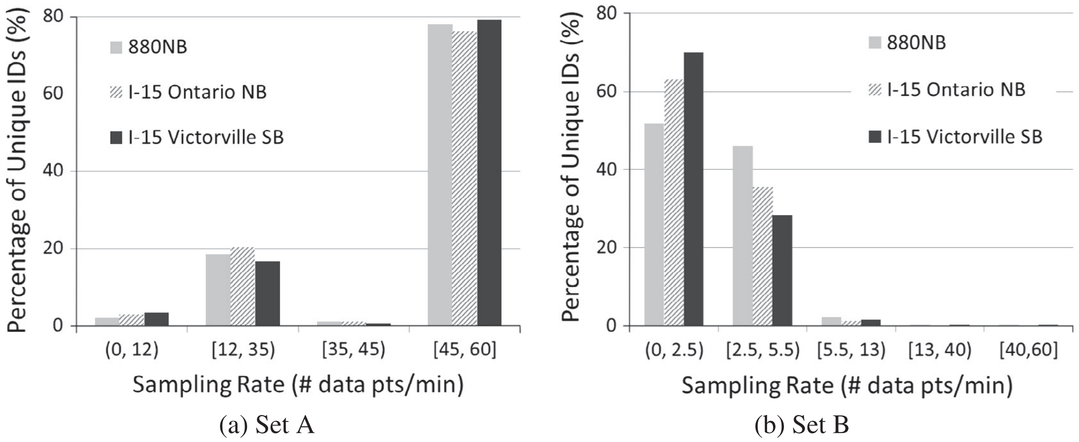
3.4. Performance trade-offs between loops and probes

Fig. 4a and b show the layout of the I-880 and I-15 sites, respectively. The on-site loop detectors (22 on I-880, 9 on I-15) are assigned to subsets as described in Section 2.3. The subsets are used to include different percentages of the loop stations in fusion calculations. In the case of I-880, for example, to include one quarter of the loop stations, only Set A was used. To include half of the loop stations, Sets A and B were used, et cetera.

Table 3 presents MAPE on I-880 illustrating the impact of loop detector spacing and the benefit of small amounts of GPS-based probe data in the proposed data fusion framework. As expected fewer loops correspond with greater error.

Table 3 quantifies how dramatically performance can improve with the addition of a small amount of GPS-based probe data. Notice that almost identical MAPE performance is obtained with 1.83 loops per mile and no probe data (21%) as with 0.92 loops per mile and a small amount of probe data (22%). The MAPE reported in the fourth column of Table 3 corresponds to a severe downselection of the total amount of available probe data in this study. For freeways already instrumented with loop detectors, better travel time performance may be achieved by fusing a relatively small amount of probe data than by doubling the number of loop detectors. This additional probe data may correspond to penetration rates on the order of only 0.2%.

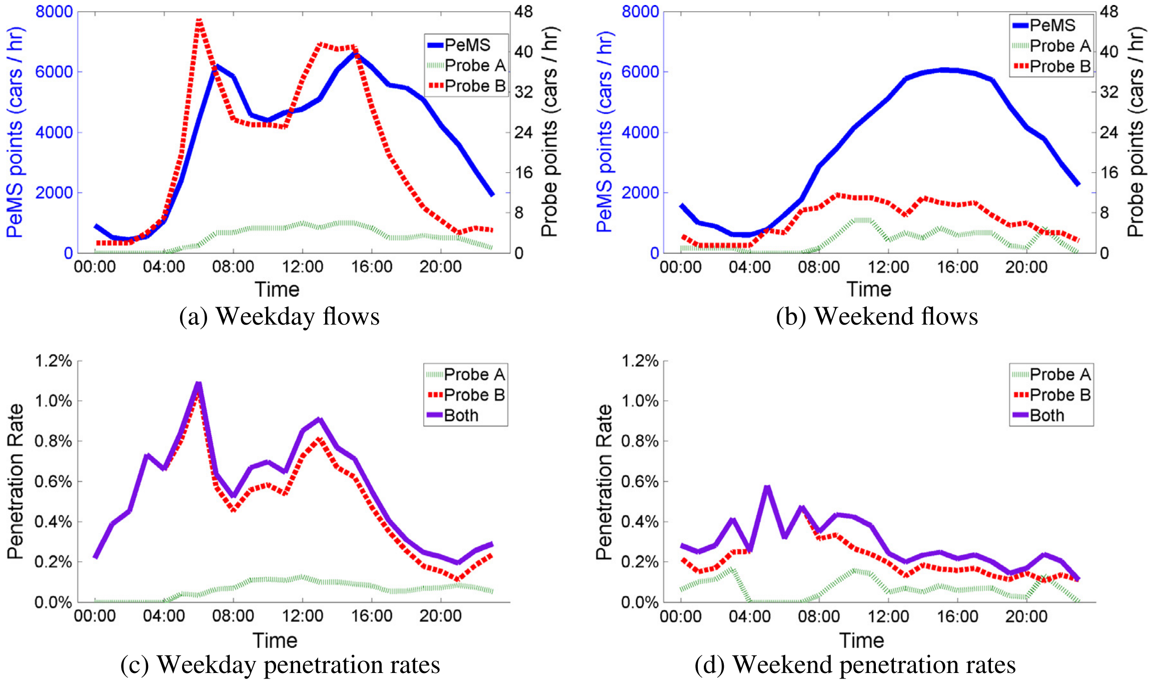
Fig. 12a directly compares the predictive power of loop versus probe data as calculated at the I-880 site. Fig. 12b shows analogous results for the I-15 site in Ontario. Contours are generated by interpolating a grid of metrics over 45 combinations



(a) Set A

(b) Set B

Fig. 7. Histogram of sampling rates for (a) set A and (b) set B probe data. Study periods during 2012 are between March 3 and March 16 for I-880; between March 31 and April 13 for I-15 Ontario; and between April 16 and April 30 for I-15 Victorville.



(a) Weekday flows

(b) Weekend flows

(c) Weekday penetration rates

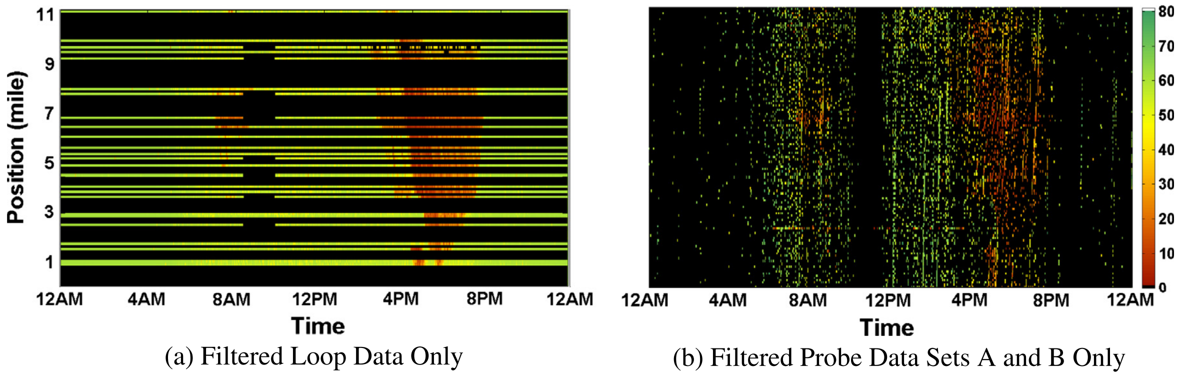
(d) Weekend penetration rates

Fig. 8. Comparison of median vehicle flows and probe penetration rates on I-880 shown separately for weekdays and weekends during the two-week period from Saturday, March 3 to Friday March 16 of 2012.

of data sources. Each figure shows combinations of probe and loop data that can be used to achieve a given MAPE. Reading the values reported on Fig. 12b, a MAPE of 0.20 (20%) is obtained with data from approximately 0.8 loop stations per mile with no probe data, or data from 0.7 probe devices per five minutes per mile with no loop data. Between these two end-points, a range of loop and probe data combinations along the contour could be used to achieve the 0.20 (20%) MAPE.

Upon close comparison of Fig. 12a and b some minor variations are seen along the *x*- and *y*- axes. For example, the interpolated surface along the *y*-axis in Fig. 12a under-reports the MAPE error for a dense deployment of loops greater than 1 loop per mile. This artifact is explained by the dramatic MAPE improvement that occurs when small amounts of probe data are fused with loop data as described above, and shown explicitly in Table 3.

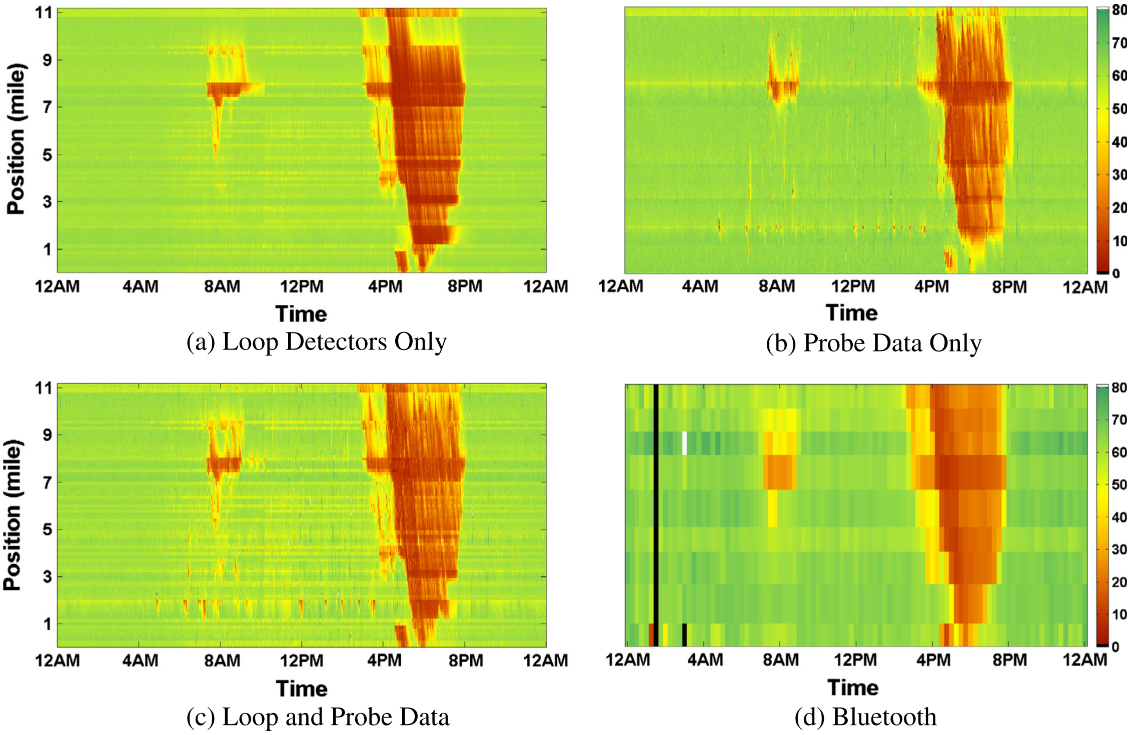
For both sites, having data from one loop per mile in addition to data from two probe devices per five minutes per mile yields MAPE performance of about 13%. Overall, the trends are very similar for both sites. This finding lends confidence in the robustness of the data fusion method, in the repeatability of results, and in the quantification of the relative predictive power of data from loops and probes.



(a) Filtered Loop Data Only

(b) Filtered Probe Data Sets A and B Only

Fig. 9. Data inputs used to reconstruct traffic state over the I-880 study site for the 24 h period on March 7, 2012. Traffic flows from bottom to top; time flows from left to right; traffic velocities are in miles per hour. Regions colored black indicate in space and in time the absence of data. Note the small data outage for some loop detectors around 9 a.m. on this day.



(a) Loop Detectors Only

(b) Probe Data Only

(c) Loop and Probe Data

(d) Bluetooth

Fig. 10. Reconstruction of traffic state over the I-880 study site for the 24 h period on March 7, 2012. Traffic flows from bottom to top; time flows from left to right; traffic velocities are in miles per hour. The first three reconstructions are the output of the data fusion engine for three subsets of data: (a) loop detectors only, (b) probe data only, and (c) both loop and probe data. The last reconstruction consists of the speeds calculated from the Bluetooth travel times and is used as a validation set.

4. Conclusions

4.1. Summary findings

Unaggregated GPS point-speed data are proven to expand the coverage and accuracy of traveler information. Data fusion makes possible the effective use of data from multiple sources or providers, and when data from multiple sources are fused, superior results are obtained. In the context of fusing probe data with loop detector data, probe data are useful even if sparse. With fused data, average travel times can be estimated accurately and reliably.

On freeways, a high penetration rate of probes is to be preferred over a high average sample rate. In other words, better performance is achieved by having less frequent data from a larger number of unique vehicles than by having more frequent

Table 2
Performance metrics for travel time estimation for a range of data sources.

Site	Data	MAPE or BTMAPE (%)	PMATE or BTPMATE (s/mile)	CCEC (%)
I-880	Probe Set A	21.15	28.24	46.66
	Probe Set B	19.58	22.47	34.78
	Sets A + B	15.22	18.77	27.82
	Loops only	20.91	30.39	24.5
	A + B + Loops	9.64	11.47	8.77
	Noise floor	18.74	24.28	na
I-15 Ontario	Probe Set A	20.13	22.46	29.31
	Probe Set B	18.02	19.51	19.45
	Sets A + B	14.84	15.73	13.59
	Loops only	19.64	20.16	12.09
	A + B + Loops	12.88	12.51	7.93
	Noise floor	14.73	14.20	na
I-15 Victorville	Probe Set A	26.0	36.68	7.6
	Probe Set B	17.0	23.5	9.7
	Sets A + B	13.6	18.5	5.5
	Noise floor	12.39	14.88	na

na = not applicable.

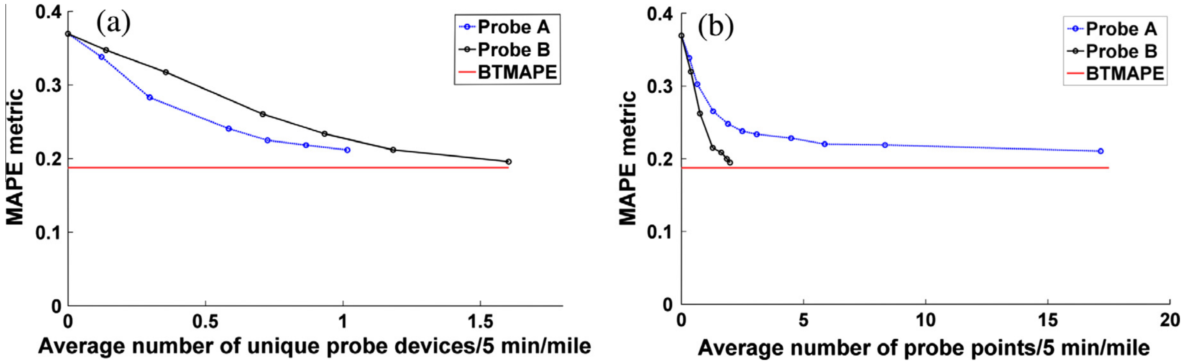


Fig. 11. Comparison of the rates at which errors decrease as a function of the amount of available probe data. (a) Error plotted versus downselection by unique devices. (b) Error plotted versus downselection by bulk.

Table 3
MAPE on I-880 illustrating the impact of loop detector spacing and the benefit of small amounts of GPS-based probe data.

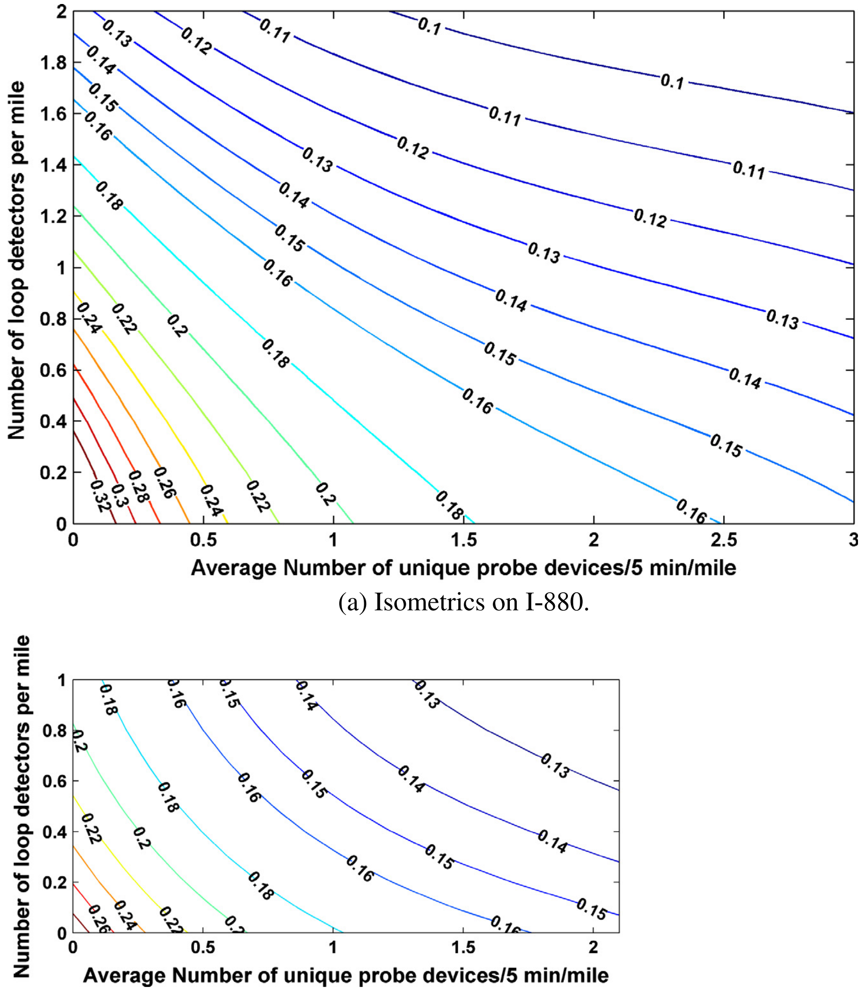
Number of Loops	Loops per mile	MAPE with no probe data (%)	MAPE with 0.125 unique probe devices per 5 min per mile (%)	Incremental MAPE improvement (%)
5	0.42	33	30	3
11	0.92	27	22	5
16	1.33	25	19	6
22	1.83	21	13	8

data from a smaller number of unique vehicles. Moreover, prevailing penetration rates for probe data are now suitable for travel time estimation on selected corridors.

Probe data can be substituted for loop data for travel time estimation. This means that where active loops are not deployed, as in rural areas, operational performance metrics such as travel times can still be obtained by the fusion engine of Fig. 2. Specific to the case of interest, no loops were active on the I-15 Victorville site at the time of this study. Therefore, data fusion of two or more data types was not possible. Nonetheless, the methodology is still applicable and good agreement was achieved between the travel times estimated by the fusion engine (based solely on probe data) and the independent validation data set (Bluetooth measured travel times).

4.2. Estimation accuracy

Although not directly applicable to the v-CTM model presented in this article, the FHWA calibration criteria targets travel times to be within 15% for 85% of cases in a (micro) model (Dowling et al., 2004). The proposed data fusion engine achieved this level of accuracy on all three sites and is therefore appropriate for planning or for traveler information applications. In



(a) Isometrics on I-880.

(b) Isometrics on I-15 in Ontario.

Fig. 12. Contours of constant MAPE performance as a function of both number of loop detectors (*y* axis) and average number of unique probe vehicles per mile per 5 min (*x* axis). The numbers on the contours represent MAPE as calculated in Eq. (15) and expressed as a fraction.

addition, the proposed fusion engine achieves accuracy within the underlying variability of the validation data (as measured by the noise floor calculations described in Section 2.7.

In congested traffic moving at 20 mph, and calculated over a 12 mile freeway segment, a 5 mph overestimation of speed will result in a MAPE of 20%. The MAPE as defined in Eq. (15), and applied in this study, is very strict in that small over- or under-estimations on successive routes along each of the study sites (nine routes on I-880, seven routes on I-15 Ontario, and nine routes on I-15 Victorville) do not cancel out, they are added up instead.

Any real-time traffic management application along a corridor would involve a systems engineering process in which functional requirements (beyond the scope of this article) would be identified following the review of a concept of operations.

4.3. Implications for transportation agencies

The work presented here has several implications for transportation agencies. The possibility of data fusion may impact strategies for the maintenance of existing fixed sensors or the deployment of new ones. Outsourcing the collection of data would be a new way of business, requiring new operational procedures as well as new information systems to handle new complexities.

4.3.1. Detector deployment strategy

Probe data and data fusion offer the opportunity for a new strategy for fixed detectors, reducing the need to deploy them with a very high spatial resolution. Where detectors are currently sparse, mobile data sources could fill the information gap.

On roadway segments where detector density is high, some stations could be allowed to fail and not be repaired or replaced if mobile data sources were available. This could reduce agency costs and enable the maintenance workforce to focus on the most critical detectors. Further, beyond filling gaps between fixed detectors, probe data could be used to expand coverage to rural areas where no fixed infrastructure exists and speed data alone is sufficient.

4.3.2. Outsourced data collection

Purchasing probe data from the commercial sector would be something new for transportation agencies and means, in effect, outsourcing the collection of traffic data. Any such undertaking comes with new risks (such as data quality risk, privacy protection, or business continuity risk) which would need to be carefully managed. A rigorous vetting and data acquisition process would require clear specifications and contractual agreements, and robust data assessment tools, processes, and standards. The framework and methods presented here can be used to determine data quality before a commitment is made to enter an extended contract for probe data procurement. In addition, this work suggests procedures for independent and ongoing quality monitoring during the course of any probe data acquisition.

4.3.3. Redesigned information systems

A crucial consideration for traffic management agencies is that probe data is different from the detector data on which they have built existing traffic management systems. Information systems such as PeMS handle streams of volume, occupancy, and speed data collected at fixed locations (PeMS, 2013). By contrast, data captured from mobile sources may come, as described here, as unaggregated data consisting of single speed observations at arbitrary latitude and longitude positions, and unconnected to any road network. Making use of probe data would therefore require the redesign of information systems that would make it possible to implement the kind of data fusion framework and methods presented in this work and take full advantage of hybrid data in traffic management systems.

Acknowledgements

This work would not have been possible without generous support from the California Department of Transportation under the Connected Corridors program. We acknowledge the strong support of Joan Sollenberger and Dr. Nicholas Compin for the ongoing work on probe data and traffic management, done collaboratively with the Division of Operations at Caltrans. Important assistance in the deployment of Bluetooth equipment was received from the Division of Operations and Division of Maintenance within Caltrans District 4 and District 8. The authors would like to thank California PATH colleagues at all levels and in all capacities who provided energy, dedication, and teamwork during the study.

References

Allström, A., Bayen, A.M., Fransson, M., Gundlegsrd, D., Patire, A.D., Rydergren, C., Sandin, M., 2014. Calibration framework based on Bluetooth sensors for traffic state estimation using a velocity based cell transmission model. Transport. Res. Proc. 3 (0), 972–981, 17th Meeting of the {EURO} Working Group on Transportation, EWGT2014, 2-4 July 2014, Sevilla, Spain. <http://www.sciencedirect.com/science/article/pii/S2352146514002403>.

Bachmann, C., Abdulhai, B., Roorda, M.J., Moshiri, B., 2012. Multisensor data integration and fusion in traffic operations and management. Transport. Res. Rec.: J. Transport. Res. Board 2308, 27–36.

Bachmann, C., Abdulhai, B., Roorda, M.J., Moshiri, B., 2013. A comparative assessment of multi-sensor data fusion techniques for freeway traffic speed estimation using microsimulation modeling. Transport. Res. Part C: Emerg. Technol. 26 (0), 33–48.

Bayen, A., Butler, J., Patire, A., 2011. Mobile Millennium. Tech. Rep. UCB-ITS-CWP-2011-6, CCIT Research Report, UC Berkeley, Institute of Transportation Studies (ITS).

Bayen, A.M., Sharafsaleh, A., Patire, A.D., 2013a. Hybrid Traffic Data Collection Roadmap: Objectives and Methods. Tech. Rep. UCB-ITS-PRR-2013-2, Univ. California, Berkeley, PATH.

Bayen, A.M., Sharafsaleh, A., Patire, A.D., 2013b. Hybrid Traffic Data Collection Roadmap: Pilot Procurement of Third-Party Traffic Data. Tech. Rep. UCB-ITS-PRR-2013-1, Univ. California, Berkeley, PATH.

Box, M.J., 1965. A new method of constrained optimization and a comparison with other methods. Comput. J. 8, 42–52.

Bucknell, C., Herrera, J.C., 2014. A trade-off analysis between penetration rate and sampling frequency of mobile sensors in traffic state estimation. Transport. Res. Part C: Emerg. Technol. 46 (0), 132–150, <http://www.sciencedirect.com/science/article/pii/S0968090X1400120X>.

Castro, P.S., Zhang, D., Chen, C., Li, S., Pan, G., 2013. From taxi gps traces to social and community dynamics: a survey. ACM Comput. Surv. 46 (2), 17:1–17:34, <http://doi.acm.org/10.1145/2543581.2543584>.

Courant, R., Friedrichs, K., Lewy, H., 1967. On the partial difference equations of mathematical physics. IBM J. Res. Dev. 11 (2), 215–234.

Deng, W., Lei, H., Zhou, X., 2013. Traffic state estimation and uncertainty quantification based on heterogeneous data sources: a three detector approach. Transport. Res. Part B: Methodol. 57 (0), 132–157, <http://www.sciencedirect.com/science/article/pii/S0191261513001513>.

Dowling, R., Skabardonis, A., Halkias, J., McHale, G., Zammit, G., 2004. Guidelines for calibration of microsimulation models: framework and applications. Transport. Res. Rec.: J. Transport. Res. Board 1876.

Evensen, G., 2003. The ensemble Kalman filter: theoretical formulation and practical implementation. Ocean Dynam. 53 (4), 343–367.

Evensen, G., 2009. Data Assimilation: The Ensemble Kalman Filter. Springer, New-York, NY, USA.

Faouzi, N.-E.E., Klein, L.A., Mouzon, O.D., 2009. Improving travel time estimates from inductive loop and toll collection data with Dempster–Shafer data fusion. Transport. Res. Rec.: J. Transport. Res. Board 2129, 73–80.

Faouzi, N.-E.E., Leung, H., Kurian, A., 2011. Data fusion in intelligent transportation systems: progress and challenges a survey. Inform. Fusion 12 (1), 4–10, Special Issue on Intelligent Transportation Systems.

Feng, Y., Hourdos, J., Davis, G.A., 2014. Probe vehicle based real-time traffic monitoring on urban roadways. Transport. Res. Part C: Emerg. Technol. 40 (0), 160–178, <http://www.sciencedirect.com/science/article/pii/S0968090X14000229>.

Geisler, S., Quix, C., Schiffer, S., Jarke, M., 2012. An evaluation framework for traffic information systems based on data streams. Transport. Res. Part C: Emerg. Technol. 23 (0), 29–55.

Godunov, S., 1959. A difference method for numerical calculation of discontinuous solutions of the equations of hydrodynamics. Mate. Sb. 89 (3), 271–306.

Haghani, A., Hamedi, M., Sadabadi, K.F., Young, S., Tarnoff, P., 2010. Data collection of freeway travel time ground truth with Bluetooth sensors. Transport. Res. Rec.: J. Transport. Res. Board 2160, 60–68.

Herrera, J.C., Work, D.B., Herring, R., Ban, X.J., Jacobson, Q., Bayen, A.M., 2010. Evaluation of traffic data obtain via GPS-enabl mobile phones: the mobile century field experiment. Transport. Res. Part C: Emerg. Technol. 18 (4), 568–583.

Herring, R., Hofleitner, A., Abbeel, P., Bayen, A., Sept 2010. Estimating arterial traffic conditions using sparse probe data. In: 2010 13th International IEEE Conference on Intelligent Transportation Systems (ITSC), pp. 929–936.

Hofleitner, A., Bayen, A., 2011. Optimal decomposition of travel times measured by probe vehicles using a statistical traffic flow model. In: 2011 14th International IEEE Conference on Intelligent Transportation Systems (ITSC), pp. 815–821.

Hoh, B., Gruteser, M., Herring, R., Ban, J., Work, D., Herrera, J.-C., Bayen, A.M., Annavaram, M., Jacobson, Q., 2008. Virtual trip lines for distributed privacy-preserving traffic monitoring. In: Proceedings of the 6th International Conference on Mobile Systems, Applications, and Services. MobiSys '08. ACM, New York, NY, USA, pp. 15–28.

Hunter, T., Abbeel, P., Bayen, A., 2013. The path inference filter: model-based low-latency map matching of probe vehicle data. In: Frazzoli, E., Lozano-Perez, T., Roy, N., Rus, D. (Eds.), Algorithmic Foundations of Robotics X, Springer Tracts in Advanced Robotics, vol. 86. Springer, Berlin, Heidelberg, pp. 591–607.

I-95 Corridor Coalition, 2010. Validation of Inrix Data: Two-Year Summary Report July 2008 June 2010. Tech. Rep., I-95 Corridor Coalition.

Kim, S., Coifman, B., 2014. Comparing {INRIX} speed data against concurrent loop detector stations over several months. Transport. Res. Part C: Emerg. Technol. 49 (0), 59–72, <http://www.sciencedirect.com/science/article/pii/S0968090X14002940>.

Koukoumidis, E., Peh, L.-S., Martonosi, M.R., 2011. Signalguru: leveraging mobile phones for collaborative traffic signal schedule advisory. In: Proceedings of the 9th International Conference on Mobile Systems, Applications, and Services. MobiSys '11. ACM, New York, NY, USA, pp. 127–140, <http://doi.acm.org/10.1145/1999995.2000008>.

Lighthill, M., Whitham, G., 1955. On kinematic waves. II. A theory of traffic flow on long crowd roads. Proc. Roy. Soc. Lond. Ser. A, Math. Phys. Sci. 229 (1178), 317–345.

Martchouk, M., Mannering, F., Bullock, D., 2011. Analysis of freeway travel time variability using Bluetooth detection. J. Transport. Eng. 137 (10), 697–704.

Mazare, P.-E., Tossavainen, O.-P., Bayen, A., Work, D., 2012. Trade-offs between inductive loops and gps probe vehicles for travel time estimation: a mobile century case study. In: 91st Transportation Research Board Annual Meeting, Washington DC.

Mihaylova, L., Boel, R., Hegyi, A., 2007. Freeway traffic estimation within particle filtering framework. Automatica 43 (2), 290–300, <http://www.sciencedirect.com/science/article/pii/S0005109806003761>.

NGSIM, 2013. Next generation simulation. <http://ngsim-community.org/>.

Patire, A., Bayen, A., Herrera, J.-C., Work, D., Herring, R., Ban, X., Jacobson, Q., Butler, J., 2010. Mobile Century: Using gps Mobile Phones as Traffic Sensors: A Field Experiment. Tech. Rep. UCB-ITS-CWP-2010-4, ISSN 1557-2269, CCIT Research Report, UC Berkeley, Institute of Transportation Studies (ITS).

PeMS, 2013. Freeway Performance Measurement System. <http://pems.dot.ca.gov/>.

Richards, P., 1956. Shock waves on the highway. Oper. Res. 4 (1), 42–51.

Schneider IV, W.H., Turner, S., Roth, J., Wikander, J., 2010. Statistical Validation of Speeds and Travel Times Provided by a Data Services Vendor. Tech. Rep. FHWA/OH-2010/2, University of Akron.

Sen, R., 2014. Rasteyrishtey: a social incentive system to crowdsource road traffic information in developing regions. In: 2014 Seventh International Conference on Mobile Computing and Ubiquitous Networking (ICMU), pp. 171–176.

StreetBump, 2015. Street Bump Mobile Application. <http://www.cityofboston.gov/DoIT/apps/streetbump.asp>.

Sun, Z., Ban, X.J., 2013. Vehicle trajectory reconstruction for signalized intersections using mobile traffic sensors. Transport. Res. Part C: Emerg. Technol. 36 (0), 268–283, <http://www.sciencedirect.com/science/article/pii/S0968090X13001824>.

Tchrakian, T., Zhuk, S., 2014. A macroscopic traffic data-assimilation framework based on the Fourier–Galerkin method and minimax estimation. IEEE Trans. Intell. Transport. Syst. (99), 1–13

Van Lint, J., Hoogendoorn, S.P., 2010. A robust and efficient method for fusing heterogeneous data from traffic sensors on freeways. Comput.-Aid. Civil Infrastruct. Eng. 25 (8), 596–612.

Wang, Y., Papageorgiou, M., 2005. Real-time freeway traffic state estimation bas on extend Kalman filter: a general approach. Transport. Res. Part B 39 (2), 141–167, <http://www.sciencirect.com/science/article/B6V99-4CHRCJW-1/2/37591107fcc45af34e6d750a7bd2e7>.

White, J., Thompson, C., Turner, H., Dougherty, B., Schmidt, D.C., 2011. Wreckwatch: automatic traffic accident detection and notification with smartphones. Mob. Netw. Appl. 16 (3), 285–303, <http://dx.doi.org/10.1007/s11036-011-0304-8>.

Work, D., Blandin, S., Tossavainen, O.-P., Piccoli, B., Bayen, A., 2010. A traffic model for velocity data assimilation. Appl. Math. Res. Exp. 2010 (1), 1–35.

Zhang, J., Wang, F.-Y., Wang, K., Lin, W.-H., Xu, X., Chen, C., 2011. Data-driven intelligent transportation systems: a survey. IEEE Trans. Intell. Transport. Syst. 12 (4), 1624–1639.