C2SMART
CONNECTED CITIES WITH SMART TRANSPORTATION

Multi-Source Data Fusion for Urban Traffic State Estimation: A Case Study of New York City

Jingqin Gao, Kaan Ozbay, Abdullah Kurkcu

This is the author's version of a work that has been accepted for presentation at the Transportation Research Board's 98th Annual Meeting, Washington D.C., 2019. The final version is available here: http://amonline.trb.org/68387-trb-1.4353651/t0005-1.4505752/1214-1.4506969/19-04253-1.4494307/19-04253-1.4506990

MULTI-SOURCE DATA FUSION FOR URBAN TRAFFIC STATE ESTIMATION: A CASE STUDY OF NEW YORK CITY

Jingqin Gao, M.Sc. (Corresponding author)
Graduate Research Assistant, C2SMART Center,
Department of Civil and Urban Engineering,
Tandon School of Engineering, New York University
Six MetroTech Center, 4th Floor, Brooklyn, NY 11201, USA
Tel: (646) 717-3652
E-mail: jingqin.gao@nyu.edu

Kaan Ozbay, Ph.D.
Professor & Director, C2SMART Center (A Tier 1 USDOT UTC),
Department of Civil and Urban Engineering &
Center for Urban Science and Progress (CUSP),
Tandon School of Engineering, New York University (NYU)
Fifteen MetroTech Center, 6th floor, Brooklyn, NY 11201, USA
Tel: 1-(646) 997-3691; E-mail: kaan.ozbay@nyu.edu

Abdullah Kurkcu, Ph.D.
Research Associate, C2SMART Center (A Tier 1 USDOT UTC),
Department of Civil and Urban Engineering &
Center for Urban Science and Progress (CUSP),
Tandon School of Engineering, New York University (NYU)
370 Jay Street, 12th Floor, Brooklyn, NY 11201, USA
Tel: 1-(646)-997-0538
Email: ak4728@nyu.edu

Word count: 6,001 words text + 2 tables x 250 words (each) = 6,501 words
Resubmission Date: November 15th, 2018

Paper resubmitted for Presentation in the
Transportation Research Board's 98th Annual Meeting, Washington, D.C., 2019

ABSTRACT

Data fusion techniques are often used to enhance traffic state estimation. The objective of this paper is to evaluate and validate the applicability of three different data fusion techniques on a non-trivial urban transportation network. Simple weighted, machine learning and evidence theory based approaches are applied to generate estimates of traffic states. Multiple data sources including information collected from electronic toll collection tag readers, Global Positioning System-equipped probe vehicles, and crowdsourcing map applications are utilized as primary data sources in the paper. A case study is provided to illustrate an application of the proposed data fusion techniques with data extracted in 2017 for a period of two weeks. Ground truth information collected from real-time camera feeds are used for validation purposes. The evidence theory based approach considering temporal evidence reliability is found to outperform other methods of cross validation in model accuracy. In addition, this study proves that the information extracted from web-based map services using "virtual sensors" can be an excellent supplementary data source for current travel time monitoring systems at no additional cost for the installation and maintenance of traditional infrastructure-based sensors.

Keywords: Data fusion, Traffic state estimation, Virtual sensors, Evidence theory

INTRODUCTION AND MOTIVATION

With the rapid growth of Intelligent Transportation Systems (ITS), sensing technologies and the variety and volume of data collected, providing an enhanced and accurate interpretation of monitored traffic state is becoming a major challenge for public agencies and private companies *(1)*. In the era of big data, researchers are faced with a diversity of information from different data sources that have different representations, scales, and density. Recent technology developments also make it possible to use different acquisition methods to obtain various heterogeneous traffic data *(2)*. Such data includes, but is not limited to information from microwave sensors, real-time cameras, Wi-Fi/Bluetooth tracking devices *(3, 4)*, and crowdsourcing mobile applications. For example, Morgul, Ender, *et al*. *(5, 6)* developed a "virtual sensor" methodology using open traffic data sources from web-based map providers to measure travel time from probe vehicles.

However, the traditional way of using a single data source without cross-validation cannot achieve good performance in data mining *(7)* because the single data source may not be a good representative of the whole population. Simply treating features of different datasets equally are usually not useful under many circumstances *(7)*. Advanced techniques that can fuse the knowledge from different but potentially connected datasets in an intelligent way are needed to provide a unified and global view of our transportation system. This can also enable a reliable interpretation of the monitored traffic situation.

As a crucial component in traffic operations and planning, Traffic State Estimation (TSE) usually plays an important role in day-to-day traffic monitoring and evaluation. TSE refers to "the process of inference of traffic state variables, such as traffic flow or speed on road segments using partially observed and noisy traffic data" *(2)*. For example, link travel times collected from Electronic Toll Collection (ETC) tag readers are used to estimate traffic states and reveal prevailing congestion levels in New York City (NYC). The information is fed into a real-time adaptive signal control system, named "Midtown in Motion" (MIM) *(8)*, to manage traffic congestion. However, traffic states are not observed at all locations all the time, and single-source measurements are usually noisy *(2)*.

Seizing the potential of 'big data' and discovering smart ways to integrate them would provide new insights in estimating dynamic traffic states of an urban road network, but how to fully make use of multi-source data to produce a better inference remains a major challenge. The objectives of our study include: 1) Assess weaknesses and strengths of different data fusion approaches using reliable ground truth data over a non-trivial traffic network, 2) conduct real-time traffic state estimation by using this fused traffic information from multiple sources to demonstrate the usefulness of a robust data fusion approach, 3) validate the virtual sensor approach with frequently used ITS data collection technology and evaluate whether it can be used as a good supplementary data source. The focus of this research is concentrated in testing the validity of three data-driven data fusion techniques that rely on historical-data and statistical or machine learning methods and their applicability on a complex transportation network in Midtown NYC. Real-time Automatic vehicle location (AVL), travel time, and speed information collected through in-vehicle GPS devices or ETC tag readers from three different data sources, namely, the MIM system, virtual sensors, and 1,336 buses *(9)*, are extracted and fused. The fused data are compared with "ground truth" information collected from video cameras.

LITERATURE REVIEW

Data Fusion Techniques

Generally, data fusion methodologies can be split into three categories: Statistical approaches, probabilistic approaches, and artificial intelligence approaches *(10)*. Statistical approaches include weighted combination, multivariate statistical analysis and so on *(11)*. Among them, the arithmetic mean approach is the simplest method for information combination *(10)*. Probabilistic approaches such as Bayesian approach with Bayesian network, maximum likelihood and Kalman filter based methods *(12, 13)*, evidence theory *(13-16)* are widely used for the multi-sensor data fusion. For example, Kong *et. al* utilized federated Kalman filter and evidence theory to estimate urban traffic states. They assigned dynamic evidence reliability and fused link mean speed from both underground loop detectors and GPS-equipped probe vehicles. Their results showed that the proposed approach could well be used in urban traffic applications on a large scale. However, their approach did not consider the temporal and spatial interactions and dependencies of the data between different links.

Artificial intelligence has also become a very popular approach in recent years. Machine learning techniques such as neural networks, clustering, and artificial cognition are applied in various transportation applications *(7, 17-20)*. For example, Xu *et. al* applied K-means clustering analysis to classify freeway traffic flow into five different states. They built a conditional logistic regression model to study the relationship between traffic states and crash risks. Each traffic state was compared to identify the underlying traffic flow characteristics that made certain states more hazardous than others. Likewise, Bartin *et. al* conducted several studies related to travel time estimation *(21, 22)*. One contribution of their work was the use of k-means clustering of sample space-time vehicle trajectory data. Its purpose is to find the optimal roadway segment configuration to estimate travel times with minimum errors. Moreover, this approach was able to determine the optimum number of segments by using the percent-gain of clustering technique as well.

Furthermore, models inspired by other disciplines such as the Mer-Gesh fusion framework proposed by Zhang *et. al (23)* are also new approaches in data fusion. Their model was based on the idea of optimizing the working process of transmission gears to fuse data that were collected from multiple sensors with different location and features. This data fusion system consisted of multiple measurements for the input data including data source effectiveness, consistency, and variation measurement. The key advantage of this method is its ability to accept any types of sensors including taxi trajectories, bus trajectories, fixed detectors, and cameras.

In short, data fusion techniques appear promising in the context of the estimating traffic states. Nevertheless, several key questions related to data fusion remain to be addressed including how to assess the input data reliability and credibility of fusion system *(24)* and how to fuse them smartly when the input data do not have same representations, scales, or density. There is a need to investigate a different data fusion approach, especially when it comes to a highly complex and congested traffic environment.

DATA PREPARATION

Midtown in Motion (MIM)

Midtown in Motion, as an integral part of the NYC Department of Transportation (NYCDOT) enhanced mobility strategy, has improved travel times on the avenues in Midtown NYC area by 10% since its first implementation in 2011 *(25)*. The MIM system collects traffic flow and occupancy from microwave sensors, travel times from ETC tag readers and uses video cameras

for verification and monitoring the field conditions. By using a unique hierarchical two-level control, traffic engineers at the Traffic Management Center (TMC) are able to quickly respond to congestion issues and smooth the flow of traffic remotely using adaptive signal control *(25)*. Typically, travel times in segments are recorded and aggregated about every two to three minutes. The advantage of the MIM system is that travel times collected from ETC tag readers work well under congested conditions. However, they can only provide traffic states among long segments (typically 8-block segments on north-south avenues). FIGURE 1 illustrates current MIM implementation in Midtown Manhattan.

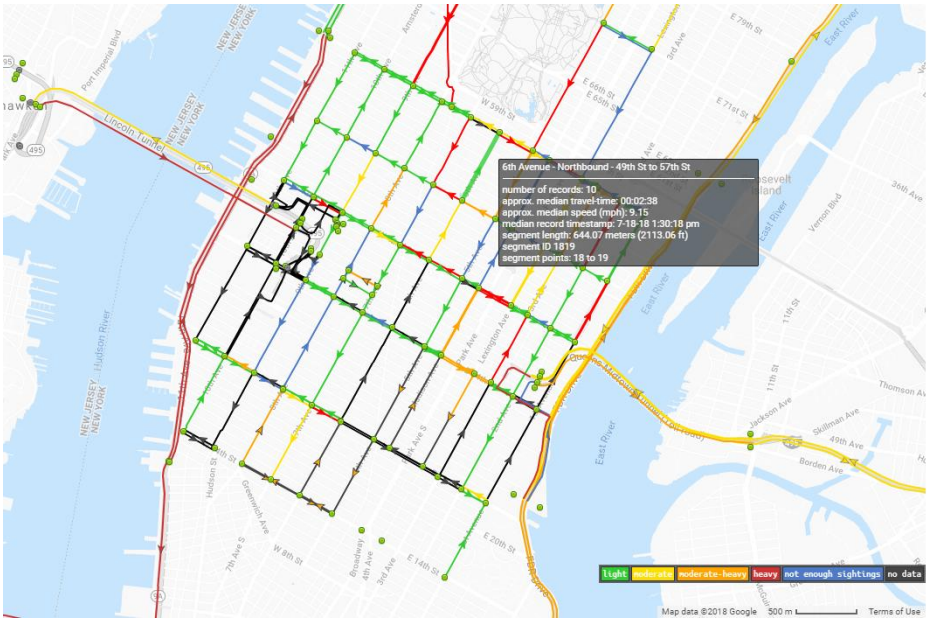


FIGURE 1 Midtown in Motion real time traffic speed (http://nyctmc.org/).

Virtual Sensors

"Virtual sensor" methodology was first purposed by Morgul *et al*. (5, 6) in 2013. Open traffic data sources from web-based map providers namely, Bing Maps™ and MapQuest™ are used to measure travel time from probe vehicles that are already in the traffic stream to estimate traffic conditions in real-time across large networks. FIGURE 2 shows the framework of the virtual sensor concept. After selecting geographical coordinates of origin-destination (OD) pairs for the road segments to be monitored, Bing Maps' REST API and MapQuest Open Data Map API services are utilized for extracting real-time traffic-based routing information. Routing requests for each of the defined OD pairs are sent automatically using the developed Python code and the responses are received from the server in Extensible Markup Language (XML) format every five minutes. In this study, travel times in segments are recorded and aggregated about every five minutes. This virtual sensor methodology comes with almost no additional cost while the quality of obtained data is proved to be quite satisfactory compared with traditional sensors such as loop detectors on highways *(5)*. One of the goals of this study aims to conduct a further validation of Virtual Sensor methodology by comparing it with travel time information collected from NYCDOT's MIM system.

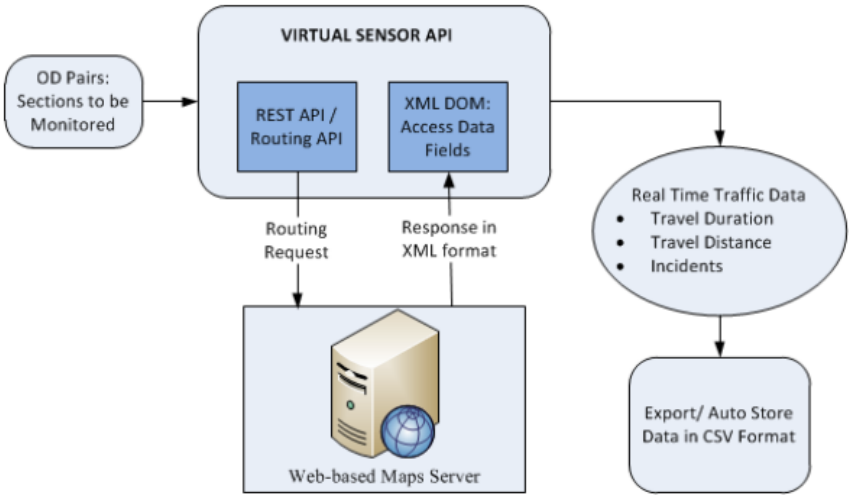


FIGURE 2 Virtual Sensor framework *(5).*

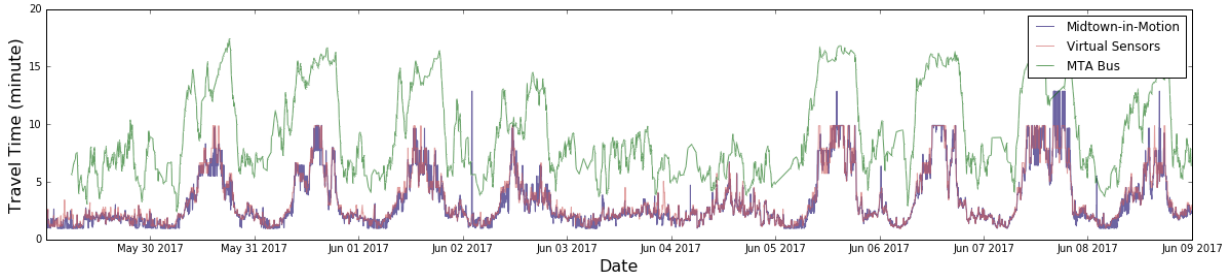
MTA Bus Time

Data from GPS-integrated cellphones or in-vehicle devices becomes a new cost-effective way of monitoring transportation system. Numerous technologies was developed to identify vehicle locations from vehicle-embedded smart devices using roadside sensors such as Bluetooth *(5)*. Many public agencies installed in-vehicle GPS devices to collect information such as spot speed or AVL as well. In NYC, a service called MTA Bus Time by the Metropolitan Transportation Authority (MTA), has been providing massive GPS based information from MTA buses since 2011 *(9)*. Rich data such as vehicle location, expected arrival/departure time, next stop names of all buses are collected at 30 seconds intervals.

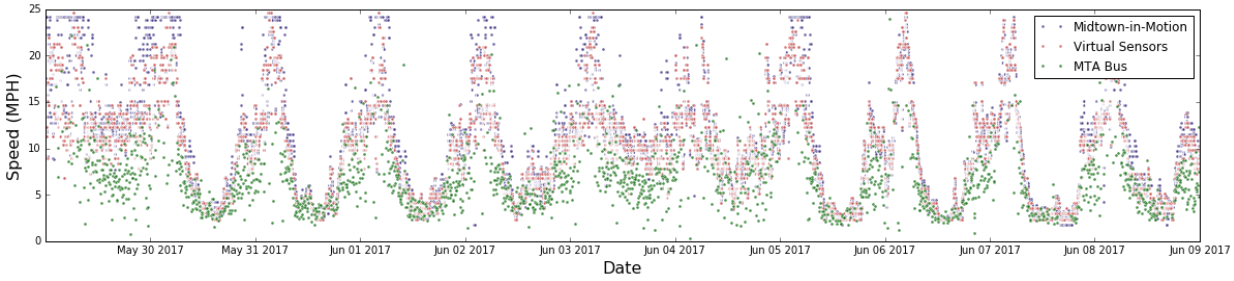
Video Cameras

Traffic video camera recordings also provide a vast volume of information including traffic volume, travel time, driver behavior, incidents, occupancy and detailed traffic operations. It often serves as a tool for traffic monitoring and congestion/incident verification. Open-source video feeds from NYCDOT's closed circuit television (CCTV) cameras *(26)* installed on major arteries allow road users to view real-time traffic movements from frequently updated still images (typically every 3-5 seconds).

To evaluate different data fusion technologies, two-week long data was manually collected for 6th Avenue between 49th Street and 57th Streets from May 29th to June 9th in 2017. Travel time and speed information for this 8-block segment were extracted from the MIM system and virtual sensors. MIM data contains 6,105 records and virtual sensor has 7,920 records. For MTA Bus Time, mean travel time and mean speed for each street block among 6th Avenue between 49th Street and 57th Streets were estimated through haversine estimation (great-circle distance) between two geographic coordination. The raw data has 42,768 data points that record the locations of each individual bus before aggregating them at a 5-minute time interval. Travel time and speed for each individual vehicle on a reference street block on 6th Avenue from 56th Street to 57th Street are recorded by manual processing from CCTV camera videos for AM peak hour (8:00AM – 9:00AM) from May 30th to June 8th in 2017 and are used as "ground truth" information in this study. FIGURE 3 shows travel time and traffic speed information from the MIM system, virtual sensor, and MTA Buses for the same road segment (6th Avenue between 49th Street and 57th Streets).



(a) Travel time



(a) Traffic Speed

FIGURE 3 Travel time and traffic speed information extracted from MIM, virtual sensor and MTA Bus Time during May 29th, to Jun 9th, 2017.

PROBLEM STATEMENT

As mentioned earlier, traffic state estimation is a crucial component of day-to-day traffic operations, monitoring, and evaluation. The recent developments in ITS and emerging technologies have made two types of travel time/ traffic speed data available to the public. The first type of data is relatively reliable yet has low-resolution. For example, the MIM system has been implemented in the field for 7 years but is only available for limited number of road segments. Each road segment usually contains multiple street blocks. The second type of data which comes from GPS-equipped vehicles like buses that has a wider coverage and higher-resolution and it can be aggregated at a street block level. However, this type of data might be noisy. For instance, bus travel time data obtained from GPS-equipped buses can be biased at street blocks that have bus stations. The key research question is that whether it is possible to reasonably fuse these different data sources to accurately estimate traffic states at the street block level. Another important research question is to test whether crowd sourcing data such as the one obtained by the "virtual sensors" approach proposed by the co-authors of this paper are accurate enough to be served as an additional data source for an existing system such as MIM. The next sections attempt to provide answers to these research questions.

METHODOLOGY

Three data fusion techniques: 1) simple weighted, 2) Random Forest, and 3) evidence theory with improved reliability, are investigated in this study. To compare with available ground truth data collected from video feeds, the three data fusion techniques are applied for the morning peak hours. The data for the implementation of these techniques come from seven weekdays from May 30th to June 8th, 2017 on one reference street block (56th Street to 57th Street on 6th Avenue). Traffic speed is used as the performance measure. FIGURE 4 illustrates mean traffic speed during AM peak hour from MIM and virtual sensor on 6th Avenue from 49th Street to 57th Street, and MTA Buses

and field (Ground truth) for the reference street block on 6th Avenue from 56th Street to 57th Street. The raw data from each data source is aggregated at a 5-minute interval.

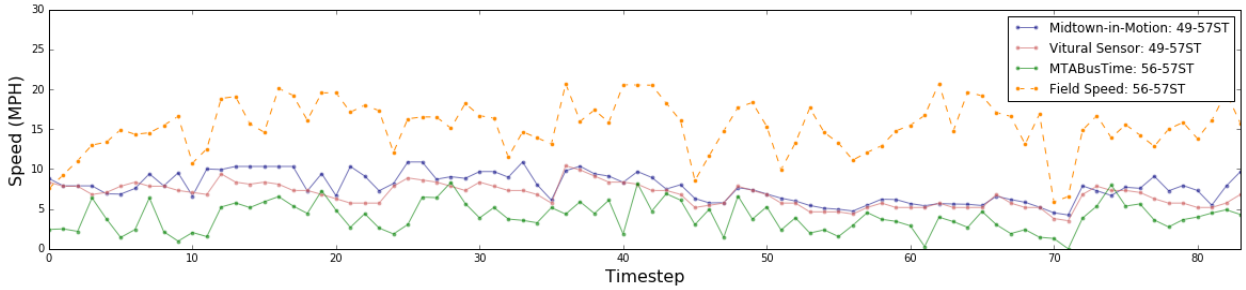


FIGURE 4 Mean traffic speed between 8AM to 9AM from MIM, virtual sensor, MTA Buses and field (ground truth) between May 30th and June 8th, 2017 (Time step is 5 minutes).

Traffic State Estimation

TSE is an important component of traffic control and operation and is often used in emergency response or congestion mitigation. For example, New York City uses estimated traffic states from the MIM system to determine its adaptive signal strategies that helps relieve congestion problems in its central core business area. However, this type of system, such as MIM, can only measure traffic state variables at certain locations or at low resolutions (i.e. MIM typically has a measurement at every 8-block). When it scales down to street block level, such system will need further assistants from additional data sources like occupancy information from microwave sensors or CCTV cameras *(25)*. Unfortunately, these devices are not installed everywhere due to technological or financial limitations. Our study proposes to process the inference of traffic speed information from the three data sources with different resolutions to obtain TSE at street block level. The proposed approaches are tested for the reference street block. Five traffic states {S_1, S_2, S_3, S_4, S_5} are assumed using partition method based on a previous study *(13)*. S_1 to S_5 means "very congested" (<1/6 maximum speed), "congested" (1/6-1/3 maximum speed), "moderate" (1/3-1/2 maximum speed), "smooth" (1/2-3/4 maximum speed), and "very smooth" (>3/4 maximum speed).

Simple Weighted method and Random Forest Classifier

Simple weighted method assigns static weights to each data source. The weights are generated based on the historical data. As a machine learning approach, Random Forest is utilized as a supervised nonlinear classification algorithm in this study. We first label our ground truth mean traffic speed data for the reference link into five traffic states S1 to S5. A 70-30 split is used to generate the training and test data in time order. This train-test split applies for all three methods. A random forest estimator constructs various decision trees on many sub-samples of the dataset. Each decision tree starts with a root node and every non-terminal node has two child nodes. It applies a binary test to every node of the input data and propagates the node to either of the child nodes depending on the test outcome *(27)*. Mean decrease Gini impurity *(27)* is used as the criteria to split the node. Next, the trained model is applied to the test set. Each decision tree of the test set obtains an estimated traffic state and the final traffic state is voted by multiple trees. This allows the predictive accuracy to be improved, and over-fitting problem to be controlled *(28)*. Python Scikit-learn package *(28)* is used with 1000 decision tree estimators, a minimum two samples to split an internal node and bootstrap strategies.

Evidence Theory with Improved Reliability

Information-fusion technology has been employed by several researchers in the context of the traffic state estimation problem in the last two decades *(13, 29, 30)*. It is a powerful tool employed to fuse on-line or off-line data from multiple sources to obtain more accurate and complete traffic state estimation than just using a single source. Based on the literature review, Dempster–Shafer (DS) theory or evidence theory is used because of its advantages in dealing with incompleteness and inaccurateness of traffic data. The initial work introducing DS theory is found in Dempster *(14)* and Shafer *(15)*. The DS theory is a generalization of Bayesian inference and is a great tool for probabilistic reasoning based on a formal calculus for combining evidence *(31)*. Let's denote the DS environment as follows:

$$\Theta: \Theta = \{\theta_1, \theta_2, \theta_3, \dots, \theta_n\} \quad (1)$$

Θ is a set of possible conclusions in which all elements are assumed to be mutually exclusive. Θ is exhaustive. Each subset of Θ can be interpreteds as a possible answer to a question and only one answer is correct. The "frame of discernment" or Power set of Θ, $P(\Theta)$, has all subsets in Θ and is denoted by:

$$P(\Theta) = \{A|A \subseteq \Theta\} \quad (2)$$

In DS Theory, the Degree of Belief in evidence is analogous to the mass of a physical object. The Basic Probability Assignment (BPA) is the evidence that measures the amount of mass and mass is a function that maps each element of the Powerset into a real number in the [0,1] interval: $m: P(\Theta) \rightarrow [0,1]$ *(14)*. For example, mass function of **A,** m(**A**) is the proportion of all evidence that supports this element of the power set. The belief function (*bel*) is the belief measure that represents the sum of masses in all subsets of **A** and the plausibility function (*pls*) represents the sum of masses committed to those subsets that do not discredit **A** *(13)*. They are defined as:

$$\begin{cases} bel(\mathbf{A}) = \sum_{\emptyset \neq \boldsymbol{B} \subseteq \boldsymbol{A}} m(\boldsymbol{B}) \ \ \forall \boldsymbol{A} \subseteq \Theta \\ pls(\mathbf{A}) = \sum_{\boldsymbol{B} \cap \boldsymbol{A} \neq \emptyset} m(\boldsymbol{B}) \ \ \forall \boldsymbol{A} \subseteq \Theta \end{cases} \quad (3)$$

The range [*Bel, Pls*] is called range of belief or evidential interval. In general, $0 \leq Bel \leq Pls \leq 1$. Multiple evidences can be fused using Dempster's combination rules that is expressed in terms of orthogonal sums of the masses of a set and all its subsets *(13)*:

$$\mathrm{m}(\mathbf{C}) = \begin{cases} 0, & \boldsymbol{B} \cap \boldsymbol{A} = \emptyset \\ \frac{1}{1-K} \sum_{\boldsymbol{B} \cap \boldsymbol{A} = C, \forall A, B \subseteq \Theta} m_i(\boldsymbol{A}) \cdot m_j(\boldsymbol{B}), & \boldsymbol{B} \cap \boldsymbol{A} \neq \emptyset \end{cases} \quad (4)$$

$$\mathrm{K} = \sum_{\boldsymbol{B} \cap \boldsymbol{A} = \emptyset, \forall A, B \subseteq \Theta} m_i(\boldsymbol{A}) \cdot m_j(\boldsymbol{B}) \quad (5)$$

where the term K is called the conflict factor between two evidences, which reflects the conflict degree of them.

Moreover, previous studies in transportation and internet of thing have emphasized the importance of "Evidence Reliability" in terms of improving fusion results *(13, 24, 32)*. When the information extracted from the evidence is missing or not totally reliable, a coefficient α can be used to discount the belief. When α = 0, it indicates that the information is completely not reliable; on the contrary, α = 1 denotes that the evidence is absolutely reliable. In this study, two factors are considered for constructing a reliability matrix. The first one is the importance of evidence. Since MIM and virtual sensors both provide segment mean travel time and speed and do not necessarily represent the mean travel time and speed on a single street block segment, they are assigned with a lower importance compared with MTA Bus Time data. Moreover, the accuracy of the aggregated mean travel time and speed from GPS-equipped vehicles depends on the sample size as well, the number of data samples before aggregation at each time step is used as an approximation for reliability.

One of the key challenges in applying D-S theory is how to obtain the mass function *(33)*. Here we construct the mass function using the following negative exponential function proposed by Denoeux *(34)*:

$$m_i(\mathbf{A}) = \exp(-\alpha_i D_i^{\beta}) \quad (6)$$

Where D_i is the distance between the data detected by the evidence and the prototype of the evidence. α and β are coefficients and the prototype are obtained by historical data.

If we let $v_{i,j}$, denote the time-dependent reliability matrix, and $m_i(\boldsymbol{A}_{i,t})$ (i $= 1,2, \dots,$ M) represent the BPA extracted from the evidence i at time t, then the modified time-dependent BPA can be rewritten as:

$$\begin{cases} m_i'(A_{i,t}) = v_{i,j,} \cdot \mathrm{m}_i(\boldsymbol{A_{i,t}}) \\ m_i'(\Theta) = 1 - \sum_{A_{i,t} \subset \Theta} v_{i,j,} \cdot \mathrm{m}_i(\boldsymbol{A_{i,t}}),\ AA_{i,t} \subset \Theta \end{cases} \quad (7)$$

Once all of the BPAs from different evidence for each timestep (every 5 minutes) are obtained, they can be fused using the combining method from DS theory. The integrated result $\mathrm{m}'(\mathbf{C}_t)$ of the fusion system at time t is *(13)*:

$$\mathrm{m}'(\mathbf{C}_t) = \mathrm{m}'_1(\mathbf{B}_{1,t}) \oplus \mathrm{m}'_2(\mathbf{B}_{2,t}) \oplus \cdots \oplus \mathrm{m}'_M(\mathbf{B}_{M,t})$$

$$= \frac{\sum_{\cap_{i=1}^{M} B_{i,t} = C_t} [\prod_{i=1}^{M} \mathrm{m}'_i(\mathbf{B}_{1,t})]}{1 - \sum_{\cap_{i=1}^{M} B_{i,t} = \emptyset} [\prod_{i=1}^{M} \mathrm{m}'_i(\mathbf{B}_{1,t})]} \quad (8)$$

Three evidences E_1, E_2 and E_3 denote the information provided by MIM, virtual sensors and MTA Bus Time. Maximum belief is used as the decision strategy of the output state.

Comparison of MIM and Virtual Sensors

To validate the virtual sensor methodology over a complex and highly congested urban arterial network, we compare it with current-used Midtown in Motion system using the data collected for a period of two weeks. Although both methodologies are using a sample of the true population which may not represent the true travel time or speed among the studied road segments, virtual sensors can be a good supplementary data source at no additional cost if it can be shown that the information they provide is similar in accuracy to that of the MIM system.

Comparing time series data is usually not straightforward. Therefore, three different methods, namely Kullback-Leibler Divergence (KL Divergence), Mean Absolute Percentage Error (MAPE) and graphic representation of absolute difference are used. Graphic representation of absolute difference between MIM and virtual sensors data is first investigated for temporal patterns. Time series data that is originated from diverse natural processes can be conveyed by probabilistic distribution functions (PDFs) *(35)*. KL Divergence from information entropy is often used to measure how much information is lost when comparing two PDFs. For two normalized, discrete PDF p and q, the KL Divergence can be represented as follows:

$$D_{KL}(p \,||\, q) = \sum_{i=1}^{n} \ln(\frac{p_i}{q_i}) p_i \text{, with } D_{KL}(p \,||\, q) \text{ equals to zero if } p = q. \quad (92)$$

Mean Absolute Percentage Error is popular in trend estimation and is applied in this study because of its intuitive interpretation in terms of relative error. Its formulation is as follows:

$$MAPE = \frac{1}{n} \sum_{t=1}^{n} \frac{|A_t - B_t|}{A_t} \quad (13)$$

Where A_t and B_t are two observations from data source A and B at time t.

RESULTS AND DISCUSSIONS

Data Fusion Results

D_i, the prototype of the evidence (equation (9)) is obtained from the training set. After obtaining prototypes, the initial BPA of each evidence for each time step can be constructed. TABLE 1 provides a representation of the initial BPAs for one-hour time-periods (12 time steps). As mentioned before, BPAs are an essential part of the D-S theory since they represents the degree of belief that a single traffic state in the subset is true, given the source of evidence *(33)*. For instance, information extracted from the MIM system at time step 1 indicates a 50.16% degree of belief that the traffic on the reference link is in a traffic state S4 "Smooth" (TABLE 1). Next, the initial BPAs are discounted with the reliability matrix and are combined to obtain the final fusion results. Maximum belief is used as the decision strategy of the final fusion results.

TABLE 1 Example of the Initial BPAs of the the Input Evidence for One Hour

Time step*	Evidence	S1 (very congested)	S2 (congested)	S3 (moderate)	S4 (Smooth)	S5 (very smooth)
1	MIM	0.0428	0.0428	0.0428	0.5016	0.3699
1	Virtual sensors	0.0536	0.0536	0.0536	0.6235	0.2157
1	MTA Bus Time	0.0326	0.0326	0.0326	0.5613	0.3409
2	MIM	0.0906	0.0906	0.0906	0.0906	0.6377
2	Virtual sensors	0.0468	0.0468	0.0468	0.0468	0.8126
2	MTA Bus Time	0.0135	0.0135	0.0135	0.0135	0.9460
3	MIM	0.0614	0.0614	0.0614	0.5763	0.2395
3	Virtual sensors	0.0338	0.0338	0.0338	0.5484	0.3503
3	MTA Bus Time	0.0013	0.0013	0.0013	0.5352	0.4610
4	MIM	0.0376	0.0376	0.0376	0.4797	0.4074
4	Virtual sensors	0.0359	0.0359	0.0359	0.5064	0.3858
4	MTA Bus Time	0.0030	0.0030	0.0030	0.5271	0.4638
5	MIM	0.0344	0.0344	0.0344	0.4934	0.4033
5	Virtual sensors	0.0343	0.0343	0.0343	0.4506	0.4465
5	MTA Bus Time	0.0034	0.0034	0.0034	0.7979	0.1918
6	MIM	0.0360	0.0360	0.0360	0.4500	0.4420
6	Virtual sensors	0.0507	0.0507	0.0507	0.3531	0.4947
6	MTA Bus Time	0.0439	0.0439	0.0439	0.5267	0.3415
7	MIM	0.0457	0.0457	0.0457	0.4564	0.4065
7	Virtual sensors	0.0723	0.0723	0.0723	0.5030	0.2800
7	MTA Bus Time	0.0612	0.0612	0.0612	0.5668	0.2496
8	MIM	0.0336	0.0336	0.0336	0.3740	0.5252
8	Virtual sensors	0.0648	0.0648	0.0648	0.4513	0.3541
8	MTA Bus Time	0.0271	0.0271	0.0271	0.6376	0.2811
9	MIM	0.0327	0.0327	0.0327	0.3880	0.5139
9	Virtual sensors	0.0842	0.0842	0.0842	0.4246	0.3229
9	MTA Bus Time	0.0276	0.0276	0.0276	0.5679	0.3494
10	MIM	0.1187	0.1187	0.1187	0.2892	0.3546

10	Virtual sensors	0.0782	0.0782	0.0782	0.3441	0.4215
10	MTA Bus Time	0.0144	0.0144	0.0144	0.7733	0.1835
11	MIM	0.0381	0.1449	0.0381	0.3345	0.4444
11	Virtual sensors	0.0432	0.1533	0.0432	0.2956	0.4647
11	MTA Bus Time	0.0071	0.0302	0.0071	0.2735	0.6822
12	MIM	0.0374	0.0374	0.1169	0.2875	0.5207
12	Virtual sensors	0.0463	0.0463	0.1347	0.3441	0.4286
12	MTA Bus Time	0.0185	0.0185	0.0185	0.3604	0.5842

*Each time step represents 5 minutes.

TABLE 2 presents the out-of-sample accuracy of the three fusion techniques. Evidence Theory with improved reliability performs best among three approaches tested in this paper. Both the simple weighted and Random Forest models indicate that MTA Bus Time data have the largest impact on estimating accurate traffic state. Simple weighted model assigns the largest weight (7.24) to MTA data and Random Forest model ranks MTA data the top 1 feature contributing to the Mean Decrease Impurity of the model. Evidence theory with improved reliability, even though more computationally complex than a simple weighted model, has much more control over the spatio-temporal features of the traffic characteristics if such information is known. For example, a street block with many truck loading and unloading activities taking place during the mid-day period over weekdays will have an associated reliability matrix with lower values assigned to these time periods for the estimation of traffic states.

TABLE 2 Results of Fusion Accuracy

Method	**Out-of-sample Cross-validate Accuracy**
Simple Weighted	0.62 (Estimated weights: MIM: 0.17, Virtual Sensor: 0.57, MTA Bus Time: 7.24)
Random Forest	0.56 (Feature importance: MIM: 0.30, Virtual Sensor: 0.17, MTA Bus Time: 0.52)
Evidence Theory with improved reliability	0.75

Random Forest has the lowest accuracy among the three proposed approaches. One possible reason is that machine learning models depend on large historical data to learn the underlying patterns. It can be hypothesized that the seven days of data is too small to properly train a Random Forest classifier. This also leads to another challenge when conducting the study: obtaining the ground truth data representing the whole traffic population is extremely challenging and labor intensive. With the developments in machine learning/deep learning and image processing techniques, video cameras have gained importance and potential to be leveraged to automatically extract valuable information via image-processing techniques in the future.

Comparison of Results of MIM and Virtual Sensors

Two PDFs are constructed for MIM and virtual sensor data using Gaussian kernel estimation and KL Divergence. The result of KL Divergence indicates a small value close to zero (0.01393). Therefore, the information loss when using PDF of virtual sensor data as an approximation of PDF of MIM is small. MAPE test shows that a mean absolute percentage error of 14.23% of two datasets. However, it should be noted that using MAPE to evaluate the overall time period may not be the best approach because errors at night can be much less than the errors during congested

periods. To further investigate the temporal difference between the two datasets, the difference of travel time by date and time was plotted in the following figure. 95.8% of the records have a travel time difference less than one minute and 99.2% of the records have a travel time difference less than two minutes.

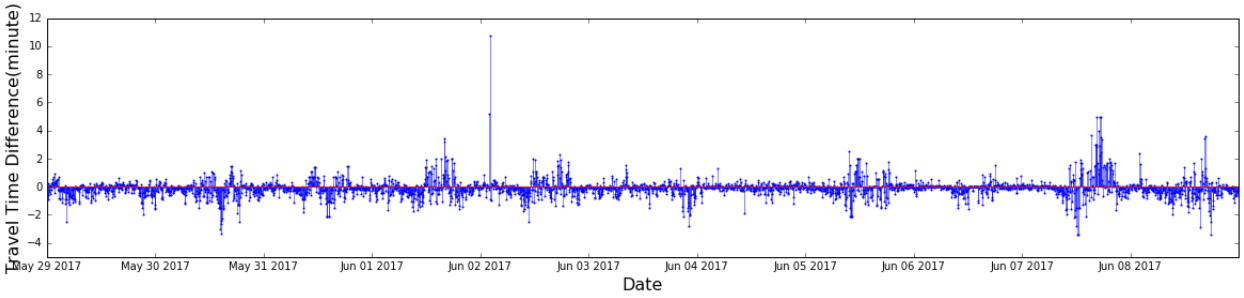


FIGURE 5 Mean travel time difference between MIM and Virtual Sensor.

CONCLUSIONS AND FUTURE WORK

This paper leveraged travel time and traffic speed information from three different data sources, MIM, virtual sensors, and MTA Bus Time, to evaluate multiple fusion technologies on traffic state estimation. Two weeks of data extracted from the three data sources are utilized along with "ground truth" information collected from video cameras.

Virtual sensors have been proven to be a good additional data source over freeways like New Jersey Turnpike as shown in a previous study *(5)*. This paper further validates its usefulness over a complex and highly congested urban arterial network. After applying MAPE, KL Divergence and Chi-Square test of independence, the information extracted from virtual sensors is shown to be a good data source that can improve NYCDOT's current MIM system at no additional cost. Moreover, incidents or extreme weather conditions such as natural disasters have relatively fewer impacts on the virtual sensor methodology since it is based on GPS and map services instead of sensors installed on the road. This may overcome the missing data challenge caused by damaged or malfunctioning infrastructure sensors.

To investigate the applicability of various data fusion techniques, following three methods, namely: 1) simple weighted, 2) Random Forest, and 3) Evidence theory are applied using morning peak hour data for a reference street block in Midtown Manhattan, NYC. Although off-line data is used in the case study, all three techniques can be implemented in real-time. A reliability matrix is estimated for the evidence theory approach to discount the bias caused by small sample size and long road segments with multiple street blocks. The final result indicates that the evidence theory with improved reliability performs the best among all three tested models. The results also show a relatively heavier weight assigned to MTA Bus Time when fusing the information to estimate traffic states, and the underlying reason may be that MTA Bus Time has a better representation on the street block level.

However, the whole approach still has some deficiencies. Firstly, it should be noted that the results are obtained from a specific street block, and they might not be generalized to other urban areas with significantly different demand and infrastructure characteristics. Secondly, the reliability of the evidence is stochastic and has a spatio-temporal dynamic behavior. For example, street blocks with bus stations may have lower reliability in terms of MTA Bus Time data due to loading and unloading of the passengers. Therefore, extending the study area to cover more street blocks is one of the main future research objectives. In addition, the research team is developing an object detection and vehicle tracking application that can help to automate the image processing

of the video feeds to collect more relevant information to improve state estimations described in this paper.

ACKNOWLEDGMENTS

The work in this paper is partially funded C2SMART, a Tier 1 University Transportation Center at New York University. The authors acknowledge Dr. Mohamad Talas from New York City Department of Transportation for sharing Midtown in Motion data. The contents of this paper only reflect views of the authors who are responsible for the facts and do not represent any official views of any sponsoring organizations or agencies.

AUTHOR CONTRIBUTION STATEMENT

The authors confirm contribution to the paper as follows: study conception and design: Jingqin Gao, Kaan Ozbay; data collection: Jingqin Gao, Abdullah Kurkcu; analysis and interpretation of results: Jingqin Gao, Kaan Ozbay; draft manuscript preparation: Jingqin Gao, Abdullah Kurkcu, Kaan Ozbay. All authors reviewed the results and approved the final version of the manuscript.

REFERENCES

1. Federal Highway Administration (FHWA). *ITS system architecture and standards, FHWA ruling on architecture standards and conformance of projects to ITS architecture standards.* 2001.
2. Seo, T., A.M. Bayen, T. Kusakabe, and Y. Asakura, Traffic state estimation on highway: A comprehensive survey. *Annual Reviews in Control 43,* 2017, pp. 128-151.
3. Kurkcu, A. and K. Ozbay, Estimating Pedestrian Densities, Wait Times, and Flows with Wi-Fi and Bluetooth Sensors. *Transportation Research Record: Journal of the Transportation Research Board(2644),* 2017, pp. 72-82.
4. Shlayan, N., A. Kurkcu, and K. Ozbay, Exploring pedestrian Bluetooth and WiFi detection at public transportation terminals. In: Proceedings of the ITSC, 2016. pp. 229-234.
5. Morgul, E., H. Yang, A. Kurkcu, K. Ozbay, B. Bartin, C. Kamga, and R. Salloum, Virtual Sensors: Web-Based Real-Time Data Collection Methodology for Transportation Operation Performance Analysis. *Transportation Research Record: Journal of the Transportation Research Board(2442),* 2014, pp. 106-116.
6. Kurkcu, A., M. Eng, E.F. Morgul, and K. Ozbay, Extended Implementation Methodology for Virtual Sensors: Web-based Real Time Transportation Data Collection and Analysis for Incident Management. In the compendium of the Transportation Research Board 94th Annual Meeting, 2015.
7. Bengio, Y., A. Courville, and P. Vincent, Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence 35(8),* 2013, pp. 1798-1828.
8. Xin, W., J. Chang, S. Muthuswamy, and M. Talas, " Midtown in Motion": A New Active Traffic Management Methodology and Its Implementation in New York City. In the compendium of the Transportation Research Board 92nd Annual Meeting, Washington DC, 2013.
9. The Metropolitan Transportation Authority, MTA Bus Time, http://bustime.mta.info/, Accessed March 3rd, 2018.
10. El Faouzi, N.-E., H. Leung, and A. Kurian, Data fusion in intelligent transportation systems: Progress and challenges–A survey. *Information Fusion 12(1),* 2011, pp. 4-10.
11. Han, J., J. Pei, and M. Kamber, *Data mining: concepts and techniques*, Elsevier, 2011.
12. Huang, D. and H. Leung, An expectation-maximization-based interacting multiple model approach for cooperative driving systems. *IEEE Transactions on Intelligent Transportation Systems 6(2),* 2005, pp. 206-228.
13. Kong, Q.-J., Z. Li, Y. Chen, and Y. Liu, An approach to urban traffic state estimation by fusing multisource information. *IEEE Transactions on Intelligent Transportation Systems 10(3),* 2009, pp. 499-511.

14. Dempster, A.P., Upper and lower probabilities induced by a multivalued mapping. *The annals of mathematical statistics,* 1967, pp. 325-339.
15. Shafer, G., *A mathematical theory of evidence*, Princeton university press Princeton, 1976.
16. El Faouzi, N.-E., L. Klein, and O. De Mouzon, Improving travel time estimates from inductive loop and toll collection data with dempster-shafer data fusion. *Transportation Research Record: Journal of the Transportation Research Board(2129),* 2009, pp. 73-80.
17. Hashem, S., Optimal linear combinations of neural networks. *Neural networks 10(4),* 1997, pp. 599-614.
18. Jiang, G.-y., J.-f. Wang, X.-d. Zhang, and L.-h. Gang, The study on the application of fuzzy clustering analysis in the dynamic identification of road traffic state. In: Proceedings of the Intelligent Transportation Systems, 2003. Proceedings. 2003 IEEE, 2003. pp. 1149-1152.
19. Xu, C., P. Liu, W. Wang, and Z. Li, Evaluation of the impacts of traffic states on crash risks on freeways. *Accident Analysis & Prevention 47,* 2012, pp. 162-171.
20. Foerster, G., Traffic state estimation using hierarchical clustering and principal components analysis: a practical approach. In: Proceedings of the the European Conference of Transport Research Institutes Young Reseachers Seminar 2007, Brno, Czech Republic, 2007.
21. Bartin, B., K. Ozbay, and C. Iyigun, Clustering-based methodology for determining optimal roadway configuration of detectors for travel time estimation. *Transportation Research Record: Journal of the Transportation Research Board(2000),* 2007, pp. 98-105.
22. Bartin, B. and K. Ozbay, Determining the optimal configuration of highway routes for real-time traffic information: A case study. *IEEE Transactions on Intelligent Transportation Systems 11(1),* 2010, pp. 225-231.
23. Zhang, S., B. Du, and N. Du, Mer-Gesh: A New Data Fusion Framework to Estimate Dynamic Road Travel Time, *Geo-Informatics in Resource Management and Sustainable Ecosystem*. Springer, 2013, p.^pp. 1-15.
24. El Faouzi, N.-E. and L.A. Klein, Data fusion for ITS: techniques and research needs. *Transportation Research Procedia 15,* 2016, pp. 495-512.
25. O'Connor, A.T., B. Schaller, M. Talas, J. Tipaldo, and S. Muthuswamy, New York City Active Traffic Management, Midtown in Motion, https://www.pcb.its.dot.gov/t3/s130418/s130418_mim_presentation.pdf, Accessed July 30, 2018.
26. NYCDOT, Real-Time Traffic Cameras, http://www.nyc.gov/html/dot/html/motorist/atis.shtml, Accessed July 20, 2018.
27. Hänsch, R. and O. Hellwich, When to fuse what? random forest based fusion of low-, mid-, and high-level information for land cover classification from optical and SAR images. In: Proceedings of the Geoscience and Remote Sensing Symposium (IGARSS), 2016 IEEE International, 2016. pp. 3587-3590.
28. Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, and V. Dubourg, Scikit-learn: Machine learning in Python. *Journal of machine learning research 12(Oct),* 2011, pp. 2825-2830.
29. El Faouzi, N.-E., Data fusion in road traffic engineering: An overview. In: Proceedings of the Defense and Security, 2004. pp. 360-371.
30. Choi, K. and Y. Chung, A data fusion algorithm for estimating link travel time. *ITS journal 7(3-4),* 2002, pp. 235-260.
31. Martin, R., J. Zhang, and C. Liu, Dempster–Shafer theory and statistical inference with weak beliefs. *Statistical Science,* 2010, pp. 72-87.
32. Esposito, C., A. Castiglione, F. Palmieri, M. Ficco, C. Dobre, G. Iordache, and F. Pop, Event-based sensor data exchange and fusion in the Internet of Things environments. *Journal of Parallel and Distributed Computing,* 2018.
33. Ali, T. and P. Dutta, Methods to obtain basic probability assignment in evidence theory. *International Journal of Computer Applications 38(4),* 2012, pp. 46-51.

34. Denoeux, T., A k-nearest neighbor classification rule based on Dempster-Shafer theory. *IEEE transactions on systems, man, and cybernetics 25(5),* 1995, pp. 804-813.
35. Kowalski, A.M., M.T. Martin, A. Plastino, and G. Judge, On extracting probability distribution information from time series. *Entropy 14(10),* 2012, pp. 1829-1841.
36. McHugh, M.L., The chi-square test of independence. *Biochemia medica: Biochemia medica 23(2),* 2013, pp. 143-149.