Multi-Source Data Fusion For Urban Traffic State Estimation: A Case Study Of New York

Conference Paper - October 2018

3 authors:

Abdullah Kurkcu
New York University
24 PUBLICATIONS 105 CITATIONS

Kaan Ozbay
New York University
308 PUBLICATIONS 3,071 CITATIONS

Jingqin Gao
New York University
13 PUBLICATIONS 21 CITATIONS

Some of the authors of this publication are also working on these related projects:

Spatiotemporal Modeling of Transportation-related Data View project

RE-CAST: REsearch on Concrete Applications for Sustainable Transportation View project
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 & Progress (CUSP)
Tandon School of Engineering
New York University
Six MetroTech Center, Room 404, Brooklyn, NY, 11201
http://c2smart.engineering.nyu.edu/
Tel (NYU CUE): 646.997.3691
Email: 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: 6,027 words text + 2 tables x 250 words (each) = 6,527 words
Submission Date: August 1st, 2018

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

Data fusion techniques are often used to enhance traffic state estimations. The objective of this paper is to evaluate and validate the applicability of three different data-driven fusion techniques over a non-trivial urban transportation network. Simple weighted, machine learning and evidence theory based approaches are applied to generate estimate for 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 in the study. A case study is provided to illustrate an application of the fusion techniques with data extracted in 2017 for two weeks. Ground truth information collected from real-time camera feeds are used for validation. The evidence theory based method considering temporal evidence reliability outperforms the other methods in terms of the cross validation of the 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.

Keywords: Data fusion, Traffic state estimation, Virtual Sensors, Evidence theory
INTRODUCTION AND MOTIVATION

With the rapid growth of Intelligent Transportation Systems (ITS), and sensing technologies along with increases in 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 the features from different datasets equally are usually not useful under many of the 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 main motivation is to test 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. Next, they build 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 for finding 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). These models were based on the idea of optimizing the working of transmission gears to fuse data collected from different sensors with different location and features at any time interval which would be tested. This data fusion system consisted of multiple measurements for the input data including data source effectiveness, consistency, variation measurement. The key advantage of the method is its ability to accept any types of sensors including events, 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 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 to 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).
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 in day-to-day traffic operations, monitoring, and evaluation. The recent development of 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 certain locations or longer road segments. The second type of data comes from, for instance, GPS-equipped buses that the data coverage is wider and has higher-resolution information that can be aggregated at a street block level. However, this type of data might be noisy in nature. For instance, bus travel time can be biased at street blocks that have bus stations. The key 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 question to be answered is to whether Virtual Sensors, with no additional cost to the agency, are accurate enough to be served as an alternative data sources for an existing system such as MIM. These issues lead us to further investigate effective methodologies to answer the 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 from MIM, Virtual Sensor, MTA Buses and field (Ground truth) for the reference street block. The raw data from each data source is aggregated at a 5-minute interval.
FIGURE 4 Mean traffic speed from MIM, Virtual Sensor, MTA Buses and field (ground truth) for a reference street block (Timestep 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 assistant from multiple 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 at no additional cost. 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 of the data sources. The weights are generated based on the training 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 \( S_1 \) to \( S_5 \). 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 introduced into the problem of traffic state estimation in the last two decades (13, 29, 30). It is a powerful tool employed to fuse online or offline 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:

\[ \Theta: \Theta = \{ \theta_1, \theta_2, \theta_3, \ldots, \theta_n \} \]  

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

\[ P(\Theta) = \{ \mathcal{A} | \mathcal{A} \subseteq \Theta \} \]  

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 \( \mathcal{A} \), \( m(\mathcal{A}) \) 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 \( \mathcal{A} \) and the plausibility function (pls) represents the sum of masses committed to those subsets that do not discredit \( \mathcal{A} \) (13). They are defined as:

\[ bel(\mathcal{A}) = \sum_{\emptyset \neq B \subseteq \mathcal{A}} m(B) \quad \forall \mathcal{A} \subseteq \Theta \]  

\[ pls(\mathcal{A}) = \sum_{\mathcal{B} \neq \emptyset} m(\mathcal{B}) \quad \forall \mathcal{A} \subseteq \Theta \]  

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):

\[ m(\mathcal{C}) = \begin{cases} 0, & B \cap \mathcal{A} = \emptyset \\ \frac{1}{1-K} \sum_{B \cap \mathcal{A} = \emptyset, \forall A, B \subseteq \Theta} m_i(A) \cdot m_j(B), & B \cap \mathcal{A} \neq \emptyset \end{cases} \]  

\[ K = \sum_{B \cap \mathcal{A} = \emptyset, \forall A, B \subseteq \Theta} m_i(A) \cdot m_j(B) \]  

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

Moreover, previous studies have emphasized the importance of “Evidence Reliability” in terms of improving fusing results (13, 24, 32). When the information extracted from the evidence is missing or not totally reliable to result in belief functions, a coefficient \( \alpha \) can be used to discount the belief. When \( \alpha = 0 \), it indicates that the information is completely not reliable; on the contrary, \( \alpha = 1 \) 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 sensor provide segment mean travel time and speed and do not necessarily represent the mean travel time and speed on one of the street blocks within the road segment, they are assigned with lower importance compare to MTA Bus Time data. Moreover, the accuracy of the aggregated mean travel time and speed from GPS-equipped vehicles depends on sample size as well, therefore, the sample size of the evidence before aggregation at each timestep 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):

\[ m_i(A) = \exp(-\alpha D_i^\beta) \]  

(6)

Where \( D_i \) is the distance between the data detected by the evidence and the prototype of the evidence. \( \alpha \) and \( \beta \) are coefficients and the prototype are obtained by historical data.

If we represent each weight with \( v_{i,j} \) and, let \( v_{i,j} \) denote the time-dependent reliability matrix, and \( m_i(A_{i,t}) \) (\( i = 1, 2, \ldots, M \)) represent the BPA extracted from the evidence \( i \) at time \( t \), then the modified time-dependent BPA can be rewritten as:

\[
\begin{align*}
\{ & m_i'(A_{i,t}) = v_{i,j} \cdot m_i(A_{i,t}) \\
& m_i'(\theta) = 1 - \sum_{A_i \subset \theta} v_{i,j} \cdot m_i(A_{i,t}), \quad A_i \subset \theta
\end{align*}
\]

(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 \( m'(C_t) \) of the fusion system at time \( t \) is (13):

\[
m'(C_t) = m'_1(B_{1,t}) \oplus m'_2(B_{2,t}) \oplus \cdots \oplus m'_M(B_{M,t})
\]

\[
= \frac{\sum_{\{i\in M \mid B_{i,t}=C_t\}} \prod_{i=1}^M m'_i(B_{i,t})}{1-\sum_{\{i\in M \mid B_{i,t}=\emptyset\}} \prod_{i=1}^M m'_i(B_{i,t})}
\]

(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 collected data for 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 road segments, Virtual sensors can be a good supplementary data source at no additional cost if it is shown that the information extracted from them is similar 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 Chi-Square Independence Test are used. 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:

\[
D_{KL}(p \parallel q) = \sum_{i=1}^q \ln\left( \frac{p_i}{q_i} \right) p_i, \quad \text{with} \quad D_{KL}(p \parallel q) \text{ equals to zero if } p = q.
\]

(9)

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:

\[
\text{MAPE} = \frac{1}{n} \sum_{i=1}^n \left| \frac{A_i - B_i}{A_i} \right|
\]

(10)

Where \( A_i \) and \( B_i \) are two observations from data source \( A \) and \( B \) at time \( t \).

As KL Divergence and MAPE do not provide enough evidence for temporal information. A graphic representation of absolute difference and a Chi-Square test of independence are applied.
as well. Chi-Square test of independence is a non-parametric tool to analyze group differences and determine if there is a significant relationship between two variables (36). It can be used for unpaired data. The null hypothesis states that there is no association between two variables. Chi-Square statistics can be computed as follows:

\[ \chi^2 = \sum \frac{(O - E)^2}{E} \]

Where \( O \) is the observed data, \( E \) is the expected value.

RESULTS AND DISCUSSIONS

Data Fusion Results

The training set in evidence theory method is used to obtain the prototype used in equation (9). After having the prototypes, the initial BPAs of each evidence for each time step can be constructed. TABLE 1 provides a representation of the initial BPAs on one-hour time-period (12 timesteps). As mentioned before, BPAs are an essential part of the D-S theory since they represent the degree of belief that one of traffic state in the subset is true, given the source of evidence (33). For instance, information extracted from the MIM system at timestep 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 fusing results.

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

<table>
<thead>
<tr>
<th>Timestep</th>
<th>Evidence</th>
<th>S1</th>
<th>S2</th>
<th>S3</th>
<th>S4</th>
<th>S5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>MIM</td>
<td>0.0428</td>
<td>0.0428</td>
<td>0.0428</td>
<td>0.5016</td>
<td>0.3699</td>
</tr>
<tr>
<td>1</td>
<td>Virtual Sensors</td>
<td>0.0536</td>
<td>0.0536</td>
<td>0.0536</td>
<td>0.6235</td>
<td>0.2157</td>
</tr>
<tr>
<td>1</td>
<td>MTA Bus Time</td>
<td>0.0326</td>
<td>0.0326</td>
<td>0.0326</td>
<td>0.5613</td>
<td>0.3409</td>
</tr>
<tr>
<td>2</td>
<td>MIM</td>
<td>0.0906</td>
<td>0.0906</td>
<td>0.0906</td>
<td>0.0906</td>
<td>0.6377</td>
</tr>
<tr>
<td>2</td>
<td>Virtual Sensors</td>
<td>0.0468</td>
<td>0.0468</td>
<td>0.0468</td>
<td>0.0468</td>
<td>0.8126</td>
</tr>
<tr>
<td>2</td>
<td>MTA Bus Time</td>
<td>0.0135</td>
<td>0.0135</td>
<td>0.0135</td>
<td>0.0135</td>
<td>0.9460</td>
</tr>
<tr>
<td>3</td>
<td>MIM</td>
<td>0.0614</td>
<td>0.0614</td>
<td>0.0614</td>
<td>0.5763</td>
<td>0.2395</td>
</tr>
<tr>
<td>3</td>
<td>Virtual Sensors</td>
<td>0.0338</td>
<td>0.0338</td>
<td>0.0338</td>
<td>0.5484</td>
<td>0.3503</td>
</tr>
<tr>
<td>3</td>
<td>MTA Bus Time</td>
<td>0.0013</td>
<td>0.0013</td>
<td>0.0013</td>
<td>0.5352</td>
<td>0.4610</td>
</tr>
<tr>
<td>4</td>
<td>MIM</td>
<td>0.0376</td>
<td>0.0376</td>
<td>0.0376</td>
<td>0.4797</td>
<td>0.4074</td>
</tr>
<tr>
<td>4</td>
<td>Virtual Sensors</td>
<td>0.0359</td>
<td>0.0359</td>
<td>0.0359</td>
<td>0.5064</td>
<td>0.3858</td>
</tr>
<tr>
<td>4</td>
<td>MTA Bus Time</td>
<td>0.0030</td>
<td>0.0030</td>
<td>0.0030</td>
<td>0.5271</td>
<td>0.4638</td>
</tr>
<tr>
<td>5</td>
<td>MIM</td>
<td>0.0344</td>
<td>0.0344</td>
<td>0.0344</td>
<td>0.4934</td>
<td>0.4033</td>
</tr>
<tr>
<td>5</td>
<td>Virtual Sensors</td>
<td>0.0343</td>
<td>0.0343</td>
<td>0.0343</td>
<td>0.4506</td>
<td>0.4465</td>
</tr>
<tr>
<td>5</td>
<td>MTA Bus Time</td>
<td>0.0034</td>
<td>0.0034</td>
<td>0.0034</td>
<td>0.7979</td>
<td>0.1918</td>
</tr>
<tr>
<td>6</td>
<td>MIM</td>
<td>0.0360</td>
<td>0.0360</td>
<td>0.0360</td>
<td>0.4500</td>
<td>0.4420</td>
</tr>
<tr>
<td>6</td>
<td>Virtual Sensors</td>
<td>0.0507</td>
<td>0.0507</td>
<td>0.0507</td>
<td>0.3531</td>
<td>0.4947</td>
</tr>
<tr>
<td>6</td>
<td>MTA Bus Time</td>
<td>0.0439</td>
<td>0.0439</td>
<td>0.0439</td>
<td>0.5267</td>
<td>0.3415</td>
</tr>
<tr>
<td>7</td>
<td>MIM</td>
<td>0.0457</td>
<td>0.0457</td>
<td>0.0457</td>
<td>0.4564</td>
<td>0.4065</td>
</tr>
<tr>
<td>7</td>
<td>Virtual Sensors</td>
<td>0.0723</td>
<td>0.0723</td>
<td>0.0723</td>
<td>0.5030</td>
<td>0.2800</td>
</tr>
<tr>
<td>7</td>
<td>MTA Bus Time</td>
<td>0.0612</td>
<td>0.0612</td>
<td>0.0612</td>
<td>0.5668</td>
<td>0.2496</td>
</tr>
</tbody>
</table>
TABLE 2 presents the out-of-sample accuracy of the three fusion techniques. Evidence Theory with improved reliability performs the best among three approaches. Both the Simple weighted model and Random Forest model 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, 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 midday period over weekdays will have an associated reliability matrix with lower values assigned to these time periods for the estimation of traffic states.

<table>
<thead>
<tr>
<th>Method</th>
<th>Out-of-sample Cross-validate Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simple Weighted</td>
<td>0.62 (Estimated weights: MIM: 0.17, Virtual Sensor: 0.57, MTA Bus Time: 7.24)</td>
</tr>
<tr>
<td>Random Forest</td>
<td>0.56 (Feature importance: MIM: 0.30, Virtual Sensor: 0.17, MTA Bus Time: 0.52)</td>
</tr>
<tr>
<td>Evidence Theory with improved reliability</td>
<td>0.75</td>
</tr>
</tbody>
</table>

Random Forest has an unexpected lowest accuracy among the three proposed approaches. One possible reason is that machine learning models depend on large historical data to learn the underlying pattern. Whereas for this particular study, seven days of data is still 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 Result of MIM and Virtual Sensors

Two PDFs are constructed for MIM and Virtual Sensor data using Gaussian kernel estimation and KL Divergence is computed. 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 a mean absolute percentage error of 14.23% of the 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. Chi-Square test of independence has a Chi statistic of 90.6 and a p-value <0.05 that indicates there is an association between the variables in the two datasets at 95% confidence level. 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.

![Travel Time Difference between MIM and Virtual Sensor](image)

**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 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 alternative data source over freeways like New Jersey Turnpike in previous study (5). This paper further validates it 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 proved to be a good supplementary data source to NYCDOT’s current MIM system at no additional cost. Moreover, incidents or extreme weather conditions such as natural disasters have relatively fewer impacts on Virtual Sensor methodology since it is based on GPS and map services instead of sensors installed on the roadside. This may overcome the missing data challenge caused by destroyed ITS devices.

To investigate the applicability of various data fusion techniques, three methods: 1) simple weighted method, 2) Random Forest, and 3) Evidence theory are applied using 7 days of morning peak hour data for a reference street block in Midtown Manhattan, NYC. Although offline data is used in the case study, all three techniques can be implemented in real-time. A reliability matrix is considered for the evidence theory approach to discount the bias carried from low sample size and long road segment with multiple street blocks. The final result indicates that evidence theory with improved reliability performs the best among all three models. The results also show a relatively heavier weight of using MTA Bus Time when fusing the information to estimate traffic states. The underlying reason may be MTA Bus Time has a better representation on 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 in other areas. Secondly, the reliability of the evidence has stochasticity and is dynamic over not only the time but also location. For example, street blocks with bus stations may have lower reliability in MTA Bus Time due to loading and unloading of the passengers. Therefore, extending the study area to cover more street blocks is one of the 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 as an on-going research effort.

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


