Data fusion methodology for imperfect traffic data based on sensor accuracy, data reliability, and data consistency

Ruth Isabel Murrugarra Munares, Universidad Adolfo Ibañez ruth.murrugarra@uai.cl

ABSTRACT

The development of Intelligent Transportation Systems (ITS) has made the collection of traffic data in real time easier and less expensive. Usually, different types of sensors are used for data collection. Each of them has different strengths and different sources of errors. However, despite their differences all sources provide data about the traffic conditions, being some of the information redundant but some complementary when fused together. There are still several issues that make the data fusion process challenging; the majority coming from the data itself. In this paper we develop a simple data fusion strategy that addresses data uncertainty, imprecision, and incompleteness, conflicting information, and outliers. It is based on data consistency, data reliability, and sensor accuracy and used to estimate freeway and arterial link travel times for real-time transportation management purposes.

Key words: consistency, data fusion, reliability

1. INTRODUCTION

The development of Intelligent Transportation Systems (ITS) has made the collection of traffic data in real time easier and less expensive. Usually, different types of sensors are used for data collection. Data collected from the various sensors provide different traffic measures which are usually used for transportation management. Each of them has different strengths and different sources of errors. For example, single inductive loop detectors are the most reliable and mature technology to obtain volume counts, but they provide aggregated data, and when calculating speed, they assume a constant vehicle length. Acoustic-based, radio frequency-based, and cell phones-based sensors are subjects to audio, radio frequency, and electromagnetic interference respectively. GPS probe vehicles provide very reliable location and time data, but based on a sample of the total population (Klein, 2001).

However, despite their differences, all sources provide data about the traffic conditions, being some of the information redundant but some complementary when fused together. The collection of techniques that combine data from different sources having different characteristics is called multi-sensor data fusion. The objective is to combine data in a way that provides a more complete picture of network conditions than the one that could be extracted from each source individually.

Although attention to data fusion has undergone rapid growth since the late 1980s, there are still several issues that make the data fusion process challenging. Data fusion methodologies have to deal with three basic problems (Goodman et al, 1997). The first problem is related to the input data itself, such as uncertainty in the measurements, missing data, outliers, conflicting data, data correlation, data dimensionality, among others (El Faouzi et al, 2011; Khaleghi et al, 2013). The second problem is related to the sensors, which have different noise levels, different aggregation periods, or have different collection algorithms. Each sensor is influenced by different factors and therefore has its own accuracy. Typical factors that affect each sensor technology are: large variance of population, light conditions, temperature, high volume traffic, weather, electromagnetic interferences, and acoustic noise. And the final basic problem is how to estimate the performance of a data fusion algorithm, or how to compare the performance of two data fusion algorithms. The proposed data fusion methodology takes into account the first two basic problems, and the proposed measure of performance takes into account the final problem.

Different data fusion methodologies have been used in the transportation field for traffic management purposes. The most common data fusion techniques used for transportation management can be grouped into three different approaches (Hall and McMullen, 2004): probabilistic-based approach, such as Bayesian theory (Choi and Chung, 2002; El Faouzi, 2006) and Dempster-Shaffer inference (Klein et al, 2002; El Faouzi, 2006); model-free approach, such as artificial neural networks and voting logic (Xie et al, 2001; Xie et al, 2004; Wei and Lee, 2007); and state-space model approach, such as Kalman filter (Wang and Xiao, 2010).

Independent of the data fusion algorithm used, most of the papers noted base their fusion weights estimation on the variance of each data source, assigning high weights to low variance data, implying that low variance means high accuracy. In this paper we develop a simple data fusion strategy based on the variance of the measurements and the consistency of the data that addresses data uncertainty, imprecision, and incompleteness, conflicting information, and outliers. It is used to estimate freeway and arterial link travel times for real-time transportation management purposes.

2. METHODOLOGY

The proposed data fusion methodology is based on the variance of the measurements and the consistency of the data. The algorithm objectives are to minimize data variance and sensor bias, and maximize consistency.

The variance of the measurements depends on the data reliability and the sensor accuracy. The quality of the data collected is affected by the sensor accuracy or by the error that the sensor adds to the measurements. Also, data should be consistent with basic notions of traffic theory, such as the relationship among volume, occupancy, speed, and travel time (see Figure 1).

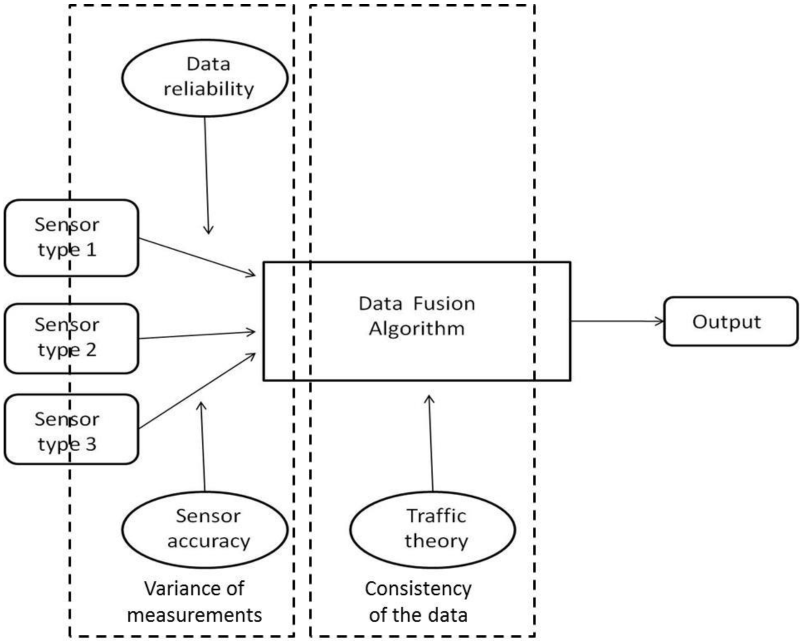


Figure 1 Data fusion scheme

2.1. Variance of measurements:

A measurement is a data point collected by a sensor. The corrected value of a measurement is obtained by adding the error of the sensor to the value of the measurement collected by the sensor.

$$Y_{jt}^{m} = X_{jt}^{m} + e^{m} \tag{1}$$

where
Y_{jt}^{m} is the corrected value of the measurement collected from source m, at link j, and time t
X_{jt}^{m} is the value of the measurement collected from source m, at link j, and time t
e^{m} is the total error from source m

But source errors are not independent of the values of the measurement collected; in fact, the error from source m is a function of the mean of the values of the measurements collected by that sensor.

$$Y_{jt}^{m} = X_{jt}^{m} + f(\bar{X}_{jt}^{m}) \tag{2}$$

To compute the variance of the corrected measurements analytically, the following equation is needed:

$$Var(Y_{jt}^{m}) = Var(X_{jt}^{m}) + Var\left(f(\bar{X}_{jt}^{m})\right) + 2 * Cov(X_{jt}^{m}, f(\bar{X}_{jt}^{m})) \tag{3}$$

The computation of the variance is not straightforward, and because of the different distributions of the data and the dependence between the measurements, an analytical derivation of the variance is difficult to obtain. Hence, a Monte Carlo simulation is used to estimate the variance of the corrected measurements.

The steps to generate each corrected measurement are:

Step 1: Estimate the variance of data reliability according to Equation 4. Data reliability refers to the variation of the measurements.

$$\sigma_{(reliability)}^{2} = \frac{1}{n}\sum_{i=1}^{n}(\bar{O} - O_i)^2 \tag{4}$$

where
$\bar{O}$ is the average of the n observed measurements
O_i is the observed value of measurement i

Step 2: Estimate the variance of sensor accuracy according to Equation 5. Sensor accuracy indicates proximity to the true value.

$$\sigma_{(accuracy)}^{2} = \frac{1}{n}\sum_{i=1}^{n}(T_i - O_i)^2 \tag{5}$$

where
n is the number of measurements
T_i is the ground truth value of measurement i
O_i is the observed value of measurement i

Step 3: Generate random variable X, where X could represent volume, speed, or travel time. This research uses a bimodal normal distribution for volume, Weibull distribution for speed, and inverse Weibull distribution for travel time (Murrugarra, 2011).

Step 4: Generate random error e according to the normal distribution with the following parameters $N(0, \sigma^2_{(accuracy)} + \sigma^2_{(reliability)})$

Step 5: Compute corrected variable $Y = X + e$

After having generated the desired number of corrected measurements, the variance of measurements $Var(Y^m_{jt})$ is computed as the sample variance of the generated values.

2.2. Consistency of the data:

Consistency means how plausible the data from one data source is, given measurements from other data sources collected at the same location and time period. Data should be consistent with basic notions of traffic theory, such as the relationship among volume, occupancy, speed, and travel time (McShane et al, 1998).

Because of the different aggregation periods of data and different collection algorithms, it is not possible to compare measurements from different sources in a one-to-one basis. Comparing the average of measurements collected by different sensors for the same time interval results in loss of information because does not take into account the spread of the measurements. Comparing both averages and variances also results in loss of information because does not take into account the shape of the distribution of the data collected during that time interval. This study compares the distributions of the data and assigns levels of consistency depending on the maximum difference between the empirical cumulative distributions.

The degrees of consistency between measurements from sensor types m and n for link j at time t are obtained using the following steps.

Step 1: Transform all measurements to travel time.

Step 2: Identify the minimum and maximum value of travel time from all measurements from all sensors (min and max).

Step 2: Divide the range [min,max] into k bins, where $k = [log_2(n)]$ and n is the number of total measurements.

Step 3: For each sensor m compute the percentage of measurements that fall in the ith bin (p^m_i) for $i = 1, \dots, k$.

Step 4: For each sensor m compute the cumulative percentage of the ith bin

$$P^m_i = \sum_{l=1}^{i} p^m_l \qquad (6)$$

Step 5: For each pair of sensors (m,q) compute the maximum absolute percentage error as

$$e^{m,q} = max_i\left|P_i^m - P_i^q\right| \tag{7}$$

Step 6: The value of the level of consistency for source m is

$$c^m = \frac{\sum_{q\neq m} e^{m,q}}{s-1} \tag{8}$$

where
s is the number of sources (sensors)

Note that the bigger the value of c_{jt}^m, the bigger the maximum absolute percentage error when comparing data from sensor m and any other sensor q, and hence, the more inconsistent the data sets are.

2.3. Data fusion algorithm:

Once the consistency between the measurements is established, and the variance of the data source identified, the selected measurements are composed using a weighted average to estimate link travel times. The weight of each data source is based on their consistency and total variance of the corrected measurements, where measurements with more consistency and less variance have bigger weights.

The estimated link travel time TT_{jt} for link j at time period t is computed by a weighting average of the contributions from all different sources.

$$TT_{jt} = \sum_{m=1}^{s} w_{jt}^m * TT_{jt}^m \tag{9}$$

$$w_{jt}^m = \frac{M * c_{jt}^m}{\frac{Var(Y_{jt}^m)}{max_q\left[Var(Y_{jt}^q)\right]}} \tag{10}$$

$$\sum_m w_{jt}^m = 1 \tag{11}$$

where
TT_{jt}^m is the computed travel time obtained from source m
w_{jt}^m is the weight assigned to sensor m
M is a scale parameter

The flexibility of this data fusion strategy makes it a simple and straightforward process to build upon when including additional data sources.

3. CASE STUDY RESULTS

Data for this study were collected during two consecutive years (2007-2008) during the duration of the New York State Fair (NYSF), including one week prior and one week after the Fair. The obtained data were collected from the roads surrounding the New York State Fairgrounds, located in the town of Geddes, in Syracuse, NY, and the nearby I-690 highway. It was collected continuously during the period of study, including weekdays, weekends, and holidays, and during both peak and non-peak hours.

Data collected include: (1) volume from loops, (2) spot speed from passive acoustic sensors, and (3) path speed and travel time from RFID tag readers, deployed on freeways, ramps, and arterials (Wojtowicz et al, 2008). Volume counts and spot speeds were aggregated in 15-minutes time intervals. Data obtained from the RFID tag readers were recorded every time a vehicle with an E-ZPass® tag passed by a tag reader. Point-to-point travel time and average speed were obtained by identifying and tracking individual vehicles from one tag reader to another. RFID data include attributes such as; encrypted tag ID, reader location, and time stamp. The percentage of vehicles with E-ZPass® during the NYSF has been found to be on average 28%, derived from manual counts. RFID data included unusual trips, such as those that included stopping for gas, meals, etc.

An important result is that no data collected can be assumed to be ground truth. The variability and the completeness of the computed travel times from volume data, speed data, and travel times collected using RFID tag readers are different. While computed travel times from volume and speed have less variability and register measurements for every time period, travel times from RFID tag readers have several missing values and have greater variability.

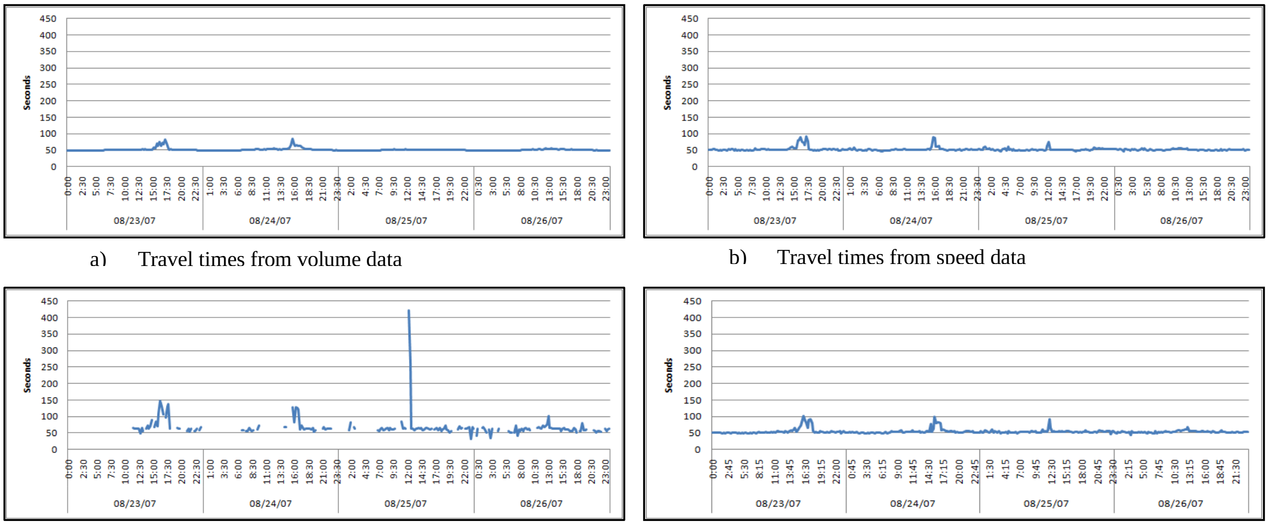
In the absence of ground truth data, this research uses data obtained from a microscopic traffic simulation to validate de data fusion methodology. In addition, information from an incident database, help trucks, variable message signs (VMS), and radio log files to validate it qualitatively.

Two examples are provided to demonstrate the different behavior of the data. Figure 2 shows an example obtained from a freeway and Figure 3 shows an example from an arterial road.

In Figure 2, we notice three major spikes; the first one on August 23rd, 2007 between 15:30 and 17:30 hours, the second one on August 24th, 2007 between 15:30and 17:00, and the third one on August 25th, 2007 between 12:00 and 13:00 hours. Travel times from volume data fail to detect the third spike (see Figure 2a), while RFID travel times seem to overestimate the measurements (see Figure 2c).

The incident database, and help trucks, variable message signs (VMS), and radio log files, indicate no incident occurred on August 25th, 2007 that could have caused an enormous increase of travel times as suggested by the RFID tag readers data. The reason of these overestimated measurements is due various consecutive unusual trips that do not appear to be outliers.

Fused travel times show the three major spikes; all with similar increases of travel time (see Figure 2d).



a) Travel times from volume data

b) Travel times from speed data

c) Travel times from RFID tag readers

d) Fused travel times

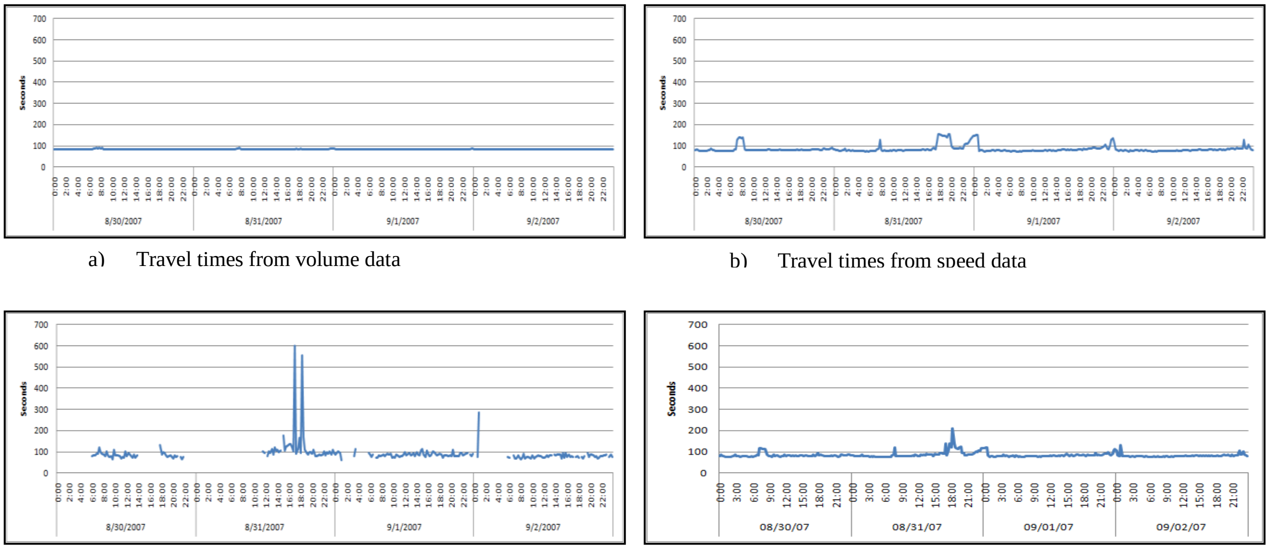
Figure 2 Sample of travel times from Interstate I-690

In Figure 3, according to the incident database and log files, a five car multi vehicle accident occurred on August 31st, 2007 at 17:55 p.m. The traffic was cleared after two hours approximately. Volume data succeeds to capture all peaks but underestimate the travel times during those periods (see Figure 3a). Due to scale reasons it is difficult to recognize them in the plot. Once again RFID data show missing values and the greatest variability of all sources (see Figure 3c). The fused travel times detect the incident and all other major peaks. It shows that during the accident the travel times increased as much as three times the normal values, which agree with road observations made during that period of time (see Figure 3d).

In order to measure how close the fused output and the simulated travel times are, two measure of performance are proposed; the mean absolute percentage error (MAPE) and the maximum absolute percentage error (MxAPE). Three data fusion methodologies are compared; method 1 is the data fusion methodology proposed in this chapter, method 2 is the simple average of the data obtained from different sensors, and method 3 is a data fusion methodology where the weights are based only on the variance of the data.

In addition, it is important to compare the performance metrics between types of roads in order to verify that the data fusion algorithm that is proposed works for any type of road. Data sets 1, 2, 3, and 4 are data collected from freeways, and data set 5 comes from an arterial road.

To verify the superiority of the proposed method, a one-way ANOVA is performed, which test the equality of the means of the errors of the three methodologies.



a) Travel times from volume data

b) Travel times from speed data

c) Travel times from RFID tag readers

d) Fused travel times

Figure 3 Sample of travel time from arterial Hiawatha Boulevard

Table 1 shows a summary of the results. The proposed data fusion methodology outperforms the other two methodologies in most of the cases with the smallest values of MAPE and MxAPE. Only the MAPE value of the proposed methodology from data set 2 is slightly worse than from method 3. Notice that the performance on freeways is better than in arterials, probably because of the presence of pedestrians and traffic lights in the latter type of road.

All one-way ANOVA p-values are zero, which means that the null hypothesis that the three errors are equal is rejected. Hence, for every data set at least one error mean is statistically different from the others. To see which methodology error means are different, a Tukey multiple comparison test is conducted.

In every data set that both values of MAPE and MxAPE are the smallest for the proposed methodology, the Tukey test finds that the error mean is different from the other two methods and the smallest. On data set 2, where the MAPE value is the second smallest after method 3, the Tukey test finds that the difference between methods 1 and 3 is not statistically significant, and both have the smallest values for both criteria.

Table 1 Comparison of data fusion algorithms

Data set	Method	Criteria		One-way ANOVA	
		MAPE	MxAPE	p-value	Tukey´s grouping
1	1	0.0412	0.1266	0.00	C
	2	0.1870	0.4032		A
	3	0.0601	0.2058		B
2	1	0.0918	0.1475	0.00	B
	2	0.1551	0.2075		A
	3	0.0910	0.1477		B
3	1	0.0670	0.1113	0.00	C
	2	0.1375	0.2054		A
	3	0.0722	0.1110		B
4	1	0.0558	0.1044	0.00	C
	2	0.0958	0.1582		A
	3	0.0692	0.1090		B
5	1	0.1746	0.2289	0.00	C
	2	0.2130	0.2965		A
	3	01790	0.2314		B

4. CONCLUSIONS

This research is able to obtain similar values to the simulated ones, even without including information from additional physical or behavioral components of the traffic network. In addition, when a change in the traffic pattern occur, the data fusion algorithm can be implemented in real time without any modification, while the simulation needs first to be calibrated and validated before attempting to analyze the traffic behavior.

The data fusion methodology proves their flexibility to work with different number of available data sources, different reliability, and different variability on each period of time. It is also simple to implement in real time. The proposed fusion technique demonstrates its validity representing successfully changes in traffic pattern, and recognizing incidents and estimating their magnitude accurately.

References

Choi, K. and Y. Chung Y (2002) A data fusion algorithm for estimating link travel time. **Intelligent Transportation Systems**, 7, 235–260.

El Faouzi, N-E. (2006) Bayesian and evidential approaches for traffic data fusion: Methodological issues and case study. In **Proceedings of the 85th Annual Meeting on Transportation Research Board**, Washington, DC, USA.

El Faouzi, N-E., H. Leung, and A. Kurian (2011) Data fusion in intelligent transportation systems: Progress and challenges – A survey. **Information Fusion**, 12, 4 – 10.

Hall, D. and S. McMullen (2004) **Mathematical techniques in multisensor data fusion**. 2nd Ed. Artech House, London

Goodman, I., R. Mahler, and H. Nguyen (1997) **Mathematics of data fusion**. Kluwer Academic Publishers, Norwell, MA, USA.

Khaleghi, B., A. Khamis, F.O. Karray, and S.N. Razavi (2013) Multisensor data fusion: A review of the state-of-the-art. **Information Fusion**, 14, 28 – 44.

Klein, L. (2001) **Sensor Technologies and data requirement for ITS**. Artech House, Norwood, MA.

Klein, L., P. Yi, and H. Teng (2002) Decision support system for advanced traffic management through data fusion. **Transportation Research Record**, 1804, 173-178.

McShane, W., R. Roess, and E. Prasas (1998) **Traffic Engineering**. 2nd Edition. Prentice Hall, New Jersey

Murrugarra, R.I. (2011) **Dynamic estimation and prediction of travel times using multi-sensor and location of sensors**. PhD Thesis, Department of Sciences and Engineering Systems, Rensselaer Polytechnic Institute.

Wang, G. and D. Xiao (2010) Background updating technique in complex traffic scene based on sensor fusion. **Journal of Transportation Systems Engineering and Information Technology**, 10, 4, 27-32.

Wei, C. and Y. Lee (2007) Development of freeway travel time forecasting models by integrating different sources of traffic data. **IEEE Transactions on Vehicular Technology**, 56, 6, 3682-3694.

Wojtowicz J., R. Murrugarra, W. Wallace, B. Bertoli, P. Manuel, W. He and C. Body (2008) RFID technology for AVI: Field demonstration of a wireless solar powered E-ZPass tag reader. **15th World Congress on Intelligent Transportation Systems**, November 2008, New York.

Xie, C., R. Cheu, and D. Lee (2001) An arterial speed estimation model fusing data from stationary and mobile sensors. In **IEEE Intelligent Transportation Systems Conference Proceedings**, Oakland, California, 573-578.

Xie, C., R. Cheu, and D. Lee (2004) Improving arterial link travel time estimation by data fusion. **Proceedings of the 83th Annual Meeting on Transportation Research Board**, Washington, DC, USA.