Traffic and mobility data collection for real-time applications

2010 13th International IEEE
Annual Conference on Intelligent Transportation Systems
Madeira Island, Portugal, September 19-22, 2010

Traffic and Mobility Data Collection for Real-Time Applications

J. Lopes, J. Bento
Instituto Superior Técnico,
Brisa Auto-Estradas de Portugal
jlopes, jbento@brisa.pt

E. Huang, C. Antoniou, M. Ben-Akiva
Intelligent Transportation System Laboratory
MIT, Cambridge MA, USA
enyang, costas, mba@mit.edu

***Abstract* - Successful development of effective real-time traffic management and information systems requires high quality traffic information in real-time. This paper presents the state-of-the-art of traffic and general mobility sensory technology and a suite of methods for data pre-processing and cleaning for real-time applications. We propose a suite of methods and techniques to be applied from traffic data acquisition, preprocessing, transformation and integration until data advanced processing and transfer. Next, we detail some techniques for data preprocessing and integration, or fusion, phases. Even though the comprehensive use of historical traffic data and assignment models to support the most part of online services and operations, real-time data is extremely important to promote models' accuracy and, therefore, the reliability of information and outputs derived from data fusion and processing. Together with techniques and theoretical formulas we present a case study applied to the Portuguese Brisa's A5 motorway, a 25 km inter-urban highway between Lisbon and Cascais. Traffic on this motorway heading to Lisbon in the morning rush hours typically experiences high levels of congestion. Brisa, the motorway operator company, has equipped A5 with a variety of traffic sensors to be used in a real-time multi-purpose way, either for traffic management and control or for traveler information and third-part applications.**

Keywords: Traffic data acquisition, Data preprocessing, Data transformation and fusion

I. INTRODUCTION

Continuing deployment of a large number of sensors, telemetry and telematics devices, and other on-line traffic and mobility data collection tools, pushed by road traffic engineering, such as traffic monitoring, signal control, automatic incident detection and recovery, traffic forecasting and traveler information, has increased hugely the amount of time-series traffic data [Antoniou et al., 2008]. The range of new data generated provides remarkable opportunities for enhancing decision-making in areas such as transport systems management and control, traveler information services, and urban and transportation planning. However to manage and make efficient use of such amounts of heterogeneous data sets in a comprehensive way, it's also a big challenge, and even bigger for real-time use cases and applications.

The foremost data-collection task is therefore an essential component on the process chain, and definitively does not end on the data acquisition. Together with data preprocessing and fusion, implements a silent but crucial mission to provide the essential raw material for upcoming applications and systems: stable, coherent and efficient traffic data. Figure 1 presents a schematic overview of the process chain from data acquisition until transfer for third party applications and services.

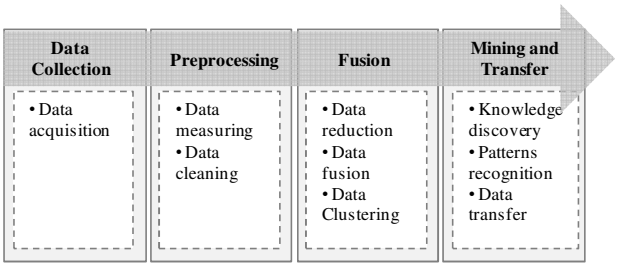

Figure 1: Schematic workflow of traffic data collection and processing

Since all data available is accurately collected from the transportation system, or it is related with it, the ability to use all datasets combined in the most appropriate way is a key issue for the information systems solution providers and a challenge for the data processing systems and methods. On top of those pressing solutions, traffic and congestion estimation and forecasting methods aims to maximize the use of the pool of versatile data, through the development of algorithms and techniques to fuse those seemingly unrelated data into high level information which is readily usable by predicting systems.

This paper is concerned with the traffic data collection, preprocessing and fusion chain to support real-time applications, whereas data completeness, consistence, performance and reliability promote an equilibrium equation for effective implementations.

The paper is organized as follows: State-of-the-art sensory technology for traffic and generic mobility data collection is presented in Section II. Then some descriptions and figures about the case study area in Section III. This is followed by the introduction of methods and techniques for data preprocessing and cleaning, along with some experimental results in Section IV. Finally, in the Section V we discuss results and introduce some works and guidelines for future research.

II. TRAFFIC AND MOBILITY SENSORY TECHNOLOGY

For many years, under growing pressure for improving traffic management and control, traditional on-road sensors, (e.g. inductive loop detectors), were massively installed and collecting methods have been evolving to obtain, compute

and transfer traffic data. [Klein, 2006] Along the roadway, such *in-situ* technologies, either based on intrusive or non-intrusive techniques, revealed necessary but not sufficient because of their limited coverage and expensive costs of implementation and maintenance. In the last years, several alternatives technologies and data sources emerged, driven by innovative methods and models. This section presents a comprehensive overview of sensor technology applied to transport engineering.

Traffic data collection is primarily categorized in three major methods: site data, floating car data and wide-area data. Each of these methods has different technical characteristics and principles of operation, including structure of data collected, accuracy of the measurements and network coverage. [Antoniou et al., 2008] An alternative classification establishes such sensor types based on their functionality as point, point-to-point, and wide area. The third and under research methodology, *wide-area survey* aims to get automated traffic condition snapshots from the road network.

Site data: refers to traffic data measured by the means of sensors located along the roadside with diverse technologies and application techniques. Some have been employed for many years such as inductive magnetic loops, pneumatic road tubes, piezoelectric loops arrays and microwave radars. With the recent technology developments, new sensors for roadside sites came out powered by flexibility, multi-purpose and cost effectiveness. Examples of those new sensors are ultrasonic and acoustic sensor systems, magnetometer vehicle detectors, infrared systems, LIDAR – light detection and ranging, and video image processing and detection.

Floating car data (FCD) also defined as vehicle-based detection: refers to mobility data collection by locating and recognizing vehicles at multiple points in a network, where specific detectors are real or virtually installed. [Huang, 2010] Some of those point-to-point sensors provide complete transversal of the travel path, providing excellent and confident information for route choice analysis and OD estimations. Examples of FCD sensors are: license plate recognition (LPR), automatic vehicle identification (AVI) transponders including probe vehicles and electronic toll tags. Varying from previous methods, global positioning system (GPS) combined with wireless communication service GSM/GPRS, provides mobility data from dynamic segments where equipped vehicles, in terms of delay and congestion level.

More recently, new technologies arisen both from road infrastructure and vehicle side, in the first instance established for vehicle-to-infrastructure and vehicle-to-vehicle cooperation. Those identifiable vehicles, equipped with wireless communications devices, are able share mobility data each other and with roadside devices, as shown in the Figure 2. Finally wireless cell-phone tracking also provides important mobility data about moving mobile-phones intensity and times along the road network.

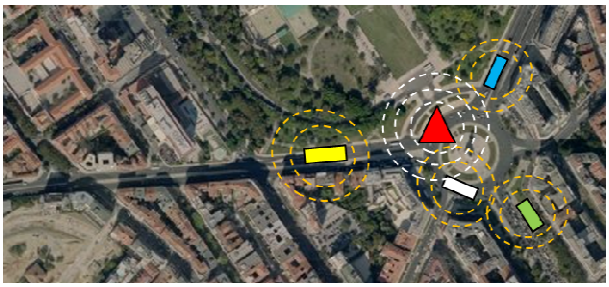

Figure 2: Probe vehicles systems for vehicle-vehicle and vehicle-infrastructure cooperation

Wide-area: under research method and technology aims to carry out wide-area traffic flow monitoring capabilities based on multi-sensor tracking options such as photogrammetric processing, video analysis, sound recording, and space-based radar.

Table I classifies the various traffic data collection methods and technologies by their network and vehicle coverage, and traffic data collected. As previously defined, based on the network coverage, collection methods can be limited to a particular site location; can be stretched to fixed road segments or trips defined by identifying sensors; or can be extended to a wide-area network through autonomous on-board equipped vehicles and airborne sensors.

TABLE I

TRAFFIC AND MOBILITY DATA COLLECTION METHODS AND TECHNOLOGIES

Network Coverage	Collection Method	Traffic and Mobility Data				
		Speed, Volume, Density	Classification	Travel Times	OD Flows	Sub-path flows
Site-based	Inductive loops	X	X			
	Road tubes	X	X			
	Piezoelectric	X	X			
	Microwave Radar	X	X			
	LIDAR/ Infrared/ Acoustic	X	X			
	Video processing	X	X(1)			
	Toll plazas	X	X			
Road segments	LPR			X	X	
	Transponders			X	X	
	Wireless devices			X	X	X

Wide-area	GPS			X	X	X
	Cell-phone tracking			X	X	X
	Airborne sensors			X	X	X

(1)Classification of the vehicles into two classes: regular cars and lorries, from vehicles volume and estimated length; not compatible with axels classification methods.

First conclusion from methods and technologies comparison is the highest complementarity between solutions, in terms of functionalities and data gathered from transportation system.

III. Case Study Area

For the experimental analysis of this research study, we use the A5 motorway, a 25 km inter-urban highway connecting Lisbon and Cascais in the West-cost of Portugal, presented in Figure 3. The first stretch of this motorway, linking Lisbon to the National Stadium in Oeiras, with 8 km length, opened to the traffic in 1944 and became the first motorway in Portugal and one of the firsts in the world.

A5 Main Figures	
Geometry (km)	25
Nodes	14
Ramp connections	64
Toll plazas	6
Annual Average daily traffic (AADT)	67,200
AADT near Lisbon	135,400
ETC rate	71%
Light vehicles rate	93%
Occurrences average/day	
Incidents	28
Accidents	4
Obstructions/ Lane closures	6

Figure 3: The Portuguese A5 inter-urban tolled motorway main figures

A5 motorway is widely equipped with telematics systems for tolling and traffic management and control. It includes a variety of sensors for traffic monitoring, for the most part used in a multi-purpose way, according to the Table II. Primarily telematics installations on the roadway regarded toll collection systems for open tolling service, where tolls fees are levied at certain points on the highway, once on the main carriageway and other at interchanges. Since 1991, the high acceptance of Via Verde – the national-wide electronic toll collection service, based on dedicated short-range communications (DSCR) microwave communications, A5 toll plazas, located between kilometer 11 and 19, are equipped with the electronic toll collection (ETC) devices. Due to the ETC use rate, and additional DSRC detectors were installed to collect travel time information between location points. Automatic video processing cameras and microwave radars are now under deployment to dense the sensory infrastructure, as shown in the Figure 4.

Behind roadway telematics infrastructures and systems, Brisa motorways, including A5, are fully equipped with a private high-speed fiber-optic cabling and wireless solutions to enable real-time remote monitoring of the network conditions, but also, traffic management, tolling operations and enforcement.

TABLE II

TELEMATICS SYSTEMS INSTALLED IN THE CASE STUDY AREA

Telematics system	Number of units	Operating functions			
		Tolling	Monitoring	Traffic data collection	Management & control
Toll gates	40	X	X	X	
ETC toll gates	20	X	X	X	
PTZ Video cameras	47		X		X
VMS Variable message Signs	9				X
Loop detectors	3		X	X	
Microwave radars	2		X	X	
Video processing cameras	4		X	X	
Point-to-point DSCR	12		X	X	

Roadway telematic infrastructures are then connected to Brisa's traffic control center (TCC) where, a state-of-the-art suite of systems and applications corporate an advanced traffic management and information system (ATMIS), providing 24-hour automated electronic tolling, traffic monitoring, traveler information and decision support services for traffic control and management. At ATMIS level, all data is managed together and, even though the specific use for specific functions, such as toll tax charging, it is available for multiple purpose applications, from traffic management to road maintenance and planning.

This work is part of a research project to generate real-time anticipatory short-term traffic conditions to be integrated with traffic management and traveler information services, both for recurrent and non-recurrent, unpredictable, incident-based conditions. Although the primarily A5 motorway application, success results will be applied to the complete network.

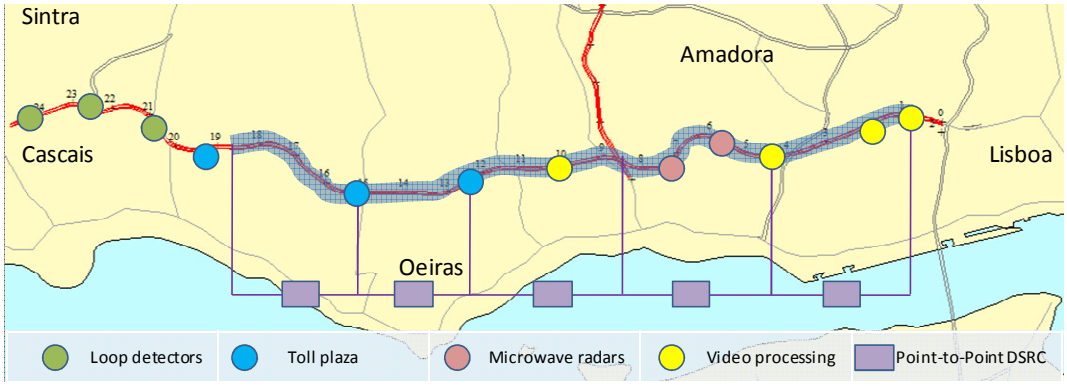

Figure 4: A5 motorway sensory infrastructure

IV. Data Preprocessing and Cleaning

Real-world raw sensor data, technology independently, is highly susceptible to noise, missing values, and inconsistent data due to sensor failures, measurement errors, and data link errors or simply because of their huge size. Low-quality data will lead to low-quality data processing and outcomes, whichever downstream traffic data application is to run. Furthermore, due to measurement errors derived data, such as average speeds or vehicles counts, may have values that are physically impossible such as negative volume counts or vehicles speed, unless wrong-way driving incidents.

However, this straightforward approach to improve data quality becomes ineffective in case of road incidents. For this, extended algorithms for pattern analysis focus precisely on the identification of unexpected significant variations so called incident-affected data or outliers, either in data measured or predicted. Consistent values leads to identify outlying events, correspondent to roadway incidents. This automatic process for incident detection is extremely useful for network operations, to proceed with automatic responses and control. However, new incidents conduct to unexpected traffic patterns commonly not found in the historical database. Therefore, data preprocessing can be used as a parallel process, valid and useful until an unpredicted scenario event occurs and disrupt traffic conditions.

In a short definition, data preprocessing is a combination chain of techniques to be applied to improve data quality through the completeness, consistency and simplification of datasets.

A. *Data properties summarization*

To be successful and effective with real-time data preprocessing it is essential to have a comprehensive, overall picture of existing datasets. It can be based on summarized representative data properties, including highlights of data values to be treated as noise or outliers. [Han, 2006] The technique is to understand the distribution of the data based on descriptive statistics, regarding both central tendency and dispersion of data. It includes mean, median, mode and midrange as measures of central tendency and quartiles, inter-quartile range and variance for data dispersion. Figure 5 shows the correlation relationship between FCD average speed and flow rate in the A5 corridor.

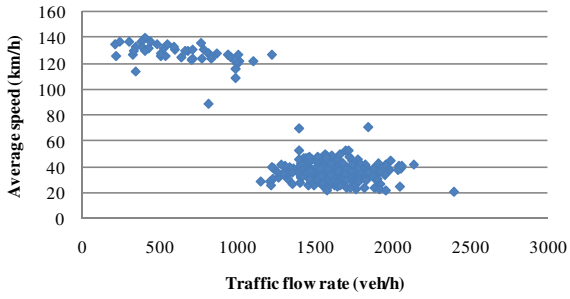

Figure 5: A scatter plot for FCD average speed versus Point flow rate in the A5 corridor

Both central tendency and dispersion properties are obtained as distributive functions, calculated by partitioning the dataset into smaller subsets and computing measures for each subset. The global and unique measure is obtained by merging results.

The end result of this process step is a package of discrete characteristics about traffic data, either for a short sliding window reaching real-time timeline or for the entire dataset for analysis, such as, the complete day. The spatial domain of those characteristics is also stretchy and can be applied to a single measure site, to a set of sites defined individually or within a corridor, or to the network as a whole.

However, for some following pre-processing steps and for the some applications further on, it is essential to know how current traffic conditions compare with past periods in the historical database. Once again the time and space scale to find out the reference period, depends on the application goal.

B. *Data cleaning – Missing values*

Most real-time applications, such as monitoring, simulation, require complete datasets without missing values. Data cleaning techniques attempt to fill in missing values, smooth out noise while identifying outliers, and correct other inconsistencies in the data. Several methods endeavor to

complete missing gaps dynamically, implementing estimation strategies balanced between data accuracy and computational complexity. Beyond this technical approach to estimate wanting data, there is an essential variable prior to the strategy definition: the gap size.

Taking into account the typical interval for roadside data acquisition and aggregation, varying from one to 5 minutes, a missing gap up to 15 minutes, corresponding to a one-step iteration bootstrapping, can be estimated with a straight-line regression analysis, involving a response traffic value, v, and a single predictor time-based period, t. That is,

$$v = b + wt \tag{1}$$

where b and w are regression coefficients, thought as weights to be solved by the method of least squares. So that we can equivalently write,

$$v = w_0 + w_1 t \tag{2}$$

Let D be the set of existing time-based observation values $(t_1, v_1), (t_2, v_2), \ldots (t_{|D|}, v_{|D|})$, the regression coefficients can be estimated using the method with the following equations,

$$w_1 = \frac{\sum_{i=1}^{|D|}(t_i - \bar{t})(v_i - \bar{v})}{\sum_{i=1}^{|D|}(t_i - \bar{t})^2} \tag{3}$$

$$w_0 = \bar{v} - w_1 \bar{t} \tag{4}$$

where $\bar{t}$ and $\bar{v}$ are the respective mean values of existing dataset D. Figure 6 shows an implementation example of missing values estimation using linear regression method, as described in the equations (1) to (4).

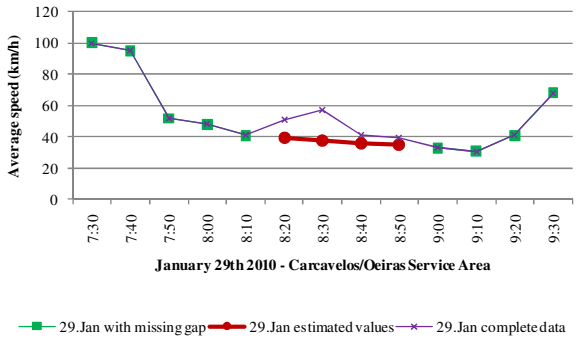

Figure 6: Missing data estimation using linear regression

For larger missing data gaps, (e.g. 45 to 120 minutes), but reasonable to be a punctual system or communications failure, data completion can be estimated using historical data in combination with the most recent observations. For the present research work, gap sizes over 120 minutes are considered input data failure, not able to be estimated from online and historical data sets.

A basic assumption is to consider that the evolution of traffic patterns on a given day of the week is the same as the evolution of the traffic pattern on the corresponding day in a reference week, constructed as a moving average over several weeks in the past. [Bellemans, 2000] This construction requires dealing with scenarios such as "special" days or official holidays on a weekday and days with major events where traffic patterns may differ significantly from regular days.

For instance, to estimate a missing value v for a time step t in the current day d, we use the following equation.

$$v_d(t) = \frac{v_d(t-1)}{v_{rd}(t-1)} \times v_{rd}(t) \tag{5}$$

Where $v_{rd}(t)$ is the corresponding data value at the time t in the reference day rd. Such matching value is then scaled to the traffic value intensity that precedes the missing value, using the factor (6).

$$\frac{v_d(t-1)}{v_{rd}(t-1)} \tag{6}$$

The application of this method is shown with the chart in the Figure 7, where missing values are estimated through the interpolation process (5) using values from the reference day 22.Jan - the regular weekday from previous week and the correction factor.

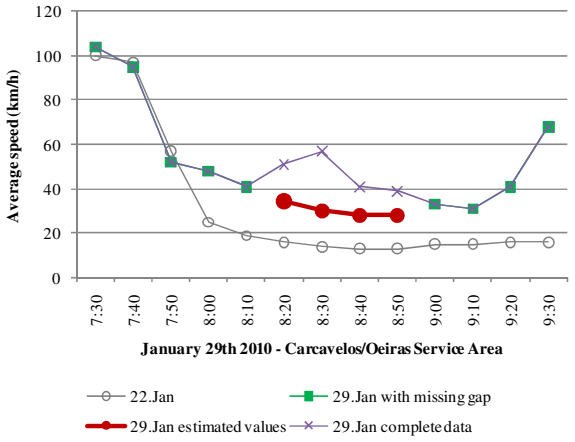

Figure 7: Missing data estimation for a 30 minutes missing gap using reference data interpolation

The big advantage of this method is the simplicity that leads to ease implementation and computation, useful for real-time applications. However it fails thoroughly when traffic patterns suddenly changes because of any roadway incident or demand fluctuations. To handle that situation, keeping the low complexity, we propose a statistical-based analysis to establish dynamically an effective connection supported by evolutionary traffic patterns.

Space-based similarity search

This new process is based in time-series data analysis and aims to find the most similar traffic pattern to the real-time

sequence pattern, or trail, to be used as reference data to complete missing values gaps. Because of the heterogeneity and complexity of the network of sensors and traffic measure database, we implement an experience-based heuristic to optimize patterns recognition and matching. The strategy starts with a space-time based correlation to identify space-time connections among network elements. The objective of this first step is to elect a representative pair of elements, to be used on the next step.

To summarize such linear connections between elements numerically is used the mean and standard deviation of each variable-element separately plus a measure known as the Pearson correlation coefficient. Let P and Q be a pair of n observations from two independent data sources, at time interval $[t, t-n]$, $P = \{ p_t,\ p_{t-1}, p_{t-2}, \dots, p_{t-n} \}$ and $Q = \{ q_t,\ q_{t-1}, q_{t-2}, \dots, q_{t-n} \}$, the correlation coefficient R is given by the formula

$$R(P,Q) = \frac{\sum_{i=1}^{n}(p_i - \bar{p})(q_i - \bar{q})}{\sqrt{\sum_{i=1}^{n}(p_i - \bar{p})^2 \sum_{i=1}^{n}(q_i - \bar{q})^2}} \quad (7)$$

In other words, the coefficient of correlation is obtained from the covariance $COV_{P.Q}$ normalization, by division it by the standard deviation $\sigma_P \times \sigma_Q$ of each variable element.

From the correlation coefficient R we obtain the coefficient of determination RD, by squaring R. The magnitude of the coefficient of determination RD indicated the proportion of variance in one variable-element, explained from the knowledge of the second variable-element.

$$RD = R^2 \quad (8)$$

Finally we calculate the amount of variation that the two variable-elements have in common - the proportion or percentage of shared variance. There is also the opposite coefficient for free – the coefficient of alienation $1 - R^2$, which defines the measure of non-association or variation between such two variables.

$$shared\ variance = RD \times 100 \quad (9)$$

Therefore, shared variance is the variance accounted for in one element by another element. For each sensor or data source of the network, this space-based similarity process establishes *a reference element*, to be used in the following processing steps.

Time-based similarity search

Traffic conditions can be stated as an evolution chain of values, sequence or trail, established from plain measures, summarized indicators or from a complex, highly elaborated gauge. Each value is the representative of the traffic network for the discrete time t_i. The set of values, behavior as time series sliding window, can be defined as the sequence trail of the network conditions for the time t_n, where n the near real-time index. The trail structure is topologically linear, which can be represented by a vector of measures in the common timeline.

From the real-time traffic trail, this function aims to discover in the historical database, the most similar trail to be defined as the reference trail. The similarity between two trail structures can be measured by the normalized root mean square deviation (NRMSD). Let the real-time traffic trail be $V = \{ v_t,\ v_{t-1}, v_{t-2}, \dots, v_{t-n} \}$ where v_i are the discrete values for the i-th element, and let the historical set of traffic trails for a reference element obtained in (9), be V_h where each V_{hk} is k-th trail index and the v_{hki} is the counterpart i-th element.

Then we can find the minimum value of the NRMSD between the real-time trail V and historical trail set V_h, applying the following formulas,

$$RMSD(V, V_{hk}) = \sqrt{\frac{1}{n}\sum_{i=1}^{n} \|v_i - v_{hki}\|^2} \quad (10)$$

over all selected historical trails, where $\|.\|$ denotes the norm value. To normalize the deviation or error found in terms of percentage, RMSD is then divided by the range of computed values $v_{max} - v_{min}$,

$$NRMSD(V, V_{hk}) = \frac{RMSD(V, V_{hk})}{v_{max} - v_{min}} \quad (11)$$

Hence, the most similar historical trail V_{href} is obtained as the minimum value of the NRMSD between the real-time trail V and the historical trail V_{hk},

$$V_{href} = \min(NRMSD(V, V_{hk})) \quad (12)$$

From here V_{href} will be used as the reference data set to complete missing values in V.

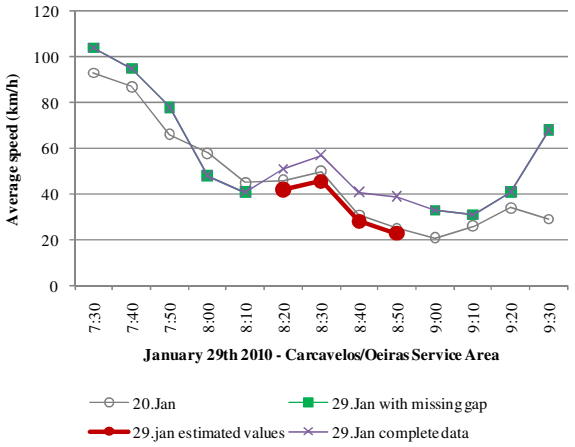

Figure 8: Missing data estimation based on space-time similarity search

The application of this method combination is shown in the Figure 8. For the 29th of January, missing values in the sensor *point-to-point Carcavelos/Oeiras Service Area*, in the time period 8:20 to 8:50 AM, are estimated using data from

20th of January, found in the space-time similarity search. Ultimate values are applying factor (6) where the reference data is V_{href}.

TABLE III

RESULTS COMPARISON FOR 40 MINUTES MISSING GAP

Method	RMSD
Linear regression	13.05
Interpolation with reference data	18.25
Space-time similarity search	12.40

As presented in the Table III, for short missing values periods, both *linear regression* and *space-time similarity search* are satisfactory to estimate missing values. This way promotes the usage of the *linear regression*, due the low computational complexity and easy implementation.

Next we process and present estimation values for a large missing gap, for the same reference day.

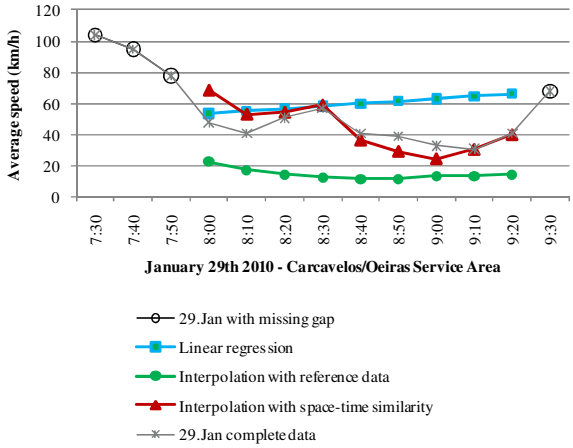

Figure 9: Missing values estimation methods comparison

Using the same package of techniques and methods, we produced a 90 minutes missing values gap for the 29th of January in the same sensor. Figure 9 presents the estimation values results graphical comparison, using precisely the same reference data sets.

TABLE IV

RESULTS COMPARISON FOR 90 MINUTES MISSING GAP

Method	RMSD
Linear regression	21.03
Interpolation with reference data	29.81
Space-time similarity search	9.16

As shown in the Table IV, for large missing values periods, *space-time similarity search* with NRMSD is definitively the most appropriated method to complete dataset.

C. Data cleaning – Outliers

Outlying observations are measures that numerically deviate radically from the rest of the data. [Mendenhall, 1993] defines the term "outliers" to values "that lie very far from the middle of the distribution in either direction". They may be due to sensor noise, acquisition process instability, equipment degradation, computer or communication system fails, or human-related errors. However, with some applications, abrupt changes in the acquisition field may occur and cause fluctuations in upcoming observations from the bulk of values. In case of traffic data measurement, such sudden changes are usually related with traffic congestion caused by accidents, broken vehicles or any other type of incident. For real-time applications, based on online data collection and processing, it is crucial to assure data quality through the identification and isolation of outliers.

In this research work, automated detection of outliers and removal were developed and integrated with automatic incident detection, in order to preserve all data, including such apparently unreliable data. Therefore, this process is made up of two distinct functions: i) identify, isolate and replace outliers; and ii) automatic data-driven incident detection. In this paper and section we focus on the first function, in order to support real-time data applications.

The generation of outliers can be described by the time-series process analysis additive outlier model [Martin, 1986].

$$y_k = x_k + o_k \tag{13}$$

Where y_k are the observations data sequence, x_k is the expected data sequence and o_k represents a sequence of contaminating outliers. The expected sequence x_k is based strictly on the current and past data observations y_{k-j} for $k \geq j \geq 0$, and the median value $\tilde{y}$. For a data point p, the distance d_p is

$$d_p = \left| y_p - \tilde{y}_p \right| \tag{14}$$

Is the distance d_p exceeds some specific threshold $T_p > 0$, and then we declare y_p to be an outlier and replace it with an estimated value obtained from the process to complete missing values, defined previously.

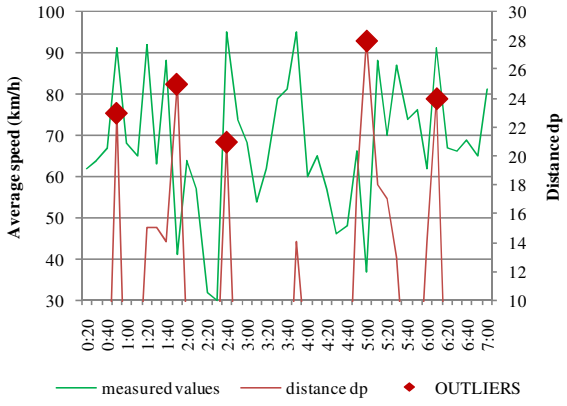

Figure 10: Automated outliers detection and replacement process

The method application is shown in the Figure 10 with the variable *average speed* in the sensor *point-to-point Carcavelos/Oeiras Service Area* for the 29th of January.

D. Data reduction

For conventional surveillance systems, raw data is gathered from a multi-variety of sensing methods either installed on the network operators' infrastructure or on fleeting vehicles, as previously described. This increasing array of data sources leads to an increasing difficulty to accomplish desired results by making use of the whole data. [Huang et al., 2009] Some of the main reasons for that relies on the intrinsic characteristics of data sources and types: i) Are deployed with uneven density over the network; 2) Are heterogeneous in type; 3) Provide highly correlated data; 4) Report at non-uniformed resolution; 5) Report at different frequencies. Data reduction techniques can be applied to work out some of those difficulties by harmonizing data references and dimensions, and bringing down the size and complexity of datasets, maintaining the integrity of the original data.

In the present work we design a two-tier approach for data reduction: i) at the data acquisition level; and ii) at data fusion level. For the first tier, data acquisition process computes raw or elementary data and, along with events detection, proceeds with data aggregation and summarization per regular periods, varying in space, time and measure type. For the second tier, heterogeneous data sets are merged together do obtain a single data platform, and is defined as data integration and fusion.

Data fusion is the process of merging together information gathered from various heterogeneous sensors, into a single data platform. In space-time domain, such as traffic field, data fusion is synonymous with data integration and aims to combine diverse data sets into a unified, or fused, data set which includes all of the data points and time frames from the input data sets. The resulted data set differs from a simple merged superset in that data tuples contains attributes and metadata which might not have been included for these points in the original data set.

TABLE V
TRAFFIC DATA, INFORMATION AND KNOWLEDGE

Data acquisition and preprocessing	**Data Fusion**		**Knowledge and decision making**
Elementary data	Object data	Situation information	Response
- Sensor data - Tolling data - Road segment data - GPS data - Cell-phone data - (…)	- Point speed - Point flow - OD travel times - OD flows - Class-relative density - (…)	- Road point conditions - Link conditions - OD conditions - Driver-choice options - Incidents - (…)	- Traffic control - Driver warning - Congestion pricing - Maintenance - (…)

Table V presents a data architecture overview, concerning the evaluation chain from elementary data to knowledge for decision making support.

V. DISCUSSION AND FUTURE WORKS

Real-world traffic databases are highly susceptible to noise, redundancy and inconsistent data due to their typically huge size and their likely origin from multiple, heterogeneous sources and sensory technologies. Low-quality traffic data will lead to low-quality results processing. For real-time traffic data applications this postulation is even more significant, since low-quality information for decision support will lead to incompetent control and management. Data preprocessing and cleaning defines a set of techniques and methods to analyze databases, identify errors and inconsistencies and proceed with dataset correction, completion and simplification.

This paper presents a suite of combination methods to analyze and summarize real-time traffic data sets, to estimate missing values and to identify and correct outliers in datasets. Either used separately in the simplest way for short missing gaps or in a complex way, combining several methods and time-space based historical datasets, data preprocessing techniques aims to improve data quality through the harmonization and completeness.

Our contribution is the designing and implementation of a systematic methodology to measure, check and repair traffic data in order to enable following steps in the processing chain till decision making and transfer. With this approach, facing data problems in early stage, including data simplification, reduction and integration, we promote the multi-purpose usage of traffic data. However, the key advantage of our process definition goes for real-time data applications, used to manage primitive data with noise and errors inside. Future works in this research program will take advantage of this work, and will focus data fusion and decision making processes.

ACKNOWLEDGMENT

This research work is supported by Brisa Auto-Estradas de Portugal, a leading world toll motorway operator and transportation infrastructures manager. Brisa is also member of MIT-Portugal research program and, with an active collaboration of Brisa Innovation research group and the MIT ITS lab, to develop this work.

REFERENCES

[Antoniou et al., 2008] Antoniou, C., Balakrishna, R., and Koutsopoulos, H. N. (2008). *Emerging data collection technologies and their impact on traffic management applications*. Proceedings of the 10th International Conference on Application of Advanced Technologies in Transportation, Athens, Greece.

[Bellemans, 2000] Bellemans, T., Schutter, B., Moor, B. (2000). *Data acquisition, interfacing and pre-processing of highway traffic data*. Proceedings of Telematics Automotive 2000, Birmingham, UK, vol. 1, pp 4/1-4/7, Apr. 2000

[Han, 2006] Han, J., Kamber, M. (2006). *Data Mining Concepts and Techniques* (2nd Edition). Elsevier, Morgan Kaufmann Publishers.

[Huang et al., 2009] Huang, E., Antoniou, C., Ben-Akiva, M., Lopes, J., Bento, J. (2009). *Real-time multi-sensor multi-source network data fusion using dynamic traffic assignment models*. In proceedings to the 12th International IEEE Conference on Intelligent Transportation Systems.

[Huang, 2010] Huang, E. (2010). *Algorithmic and Implementation Aspects of Online Calibration of Dynamic Traffic Assignment*. Master's thesis, Massachusetts Institute of Technology.

[Liu, 2004] Liu, H., Shah, S., Jiang, W. (2004). *On-line outlier detection and data cleaning*. Journal of Computers and chemical engineering. Pages 1635-1647. Elsevier

[Klein, 2006] Klein, L., Mills, M., Gibson, D. (2006). *Traffic Detector Handbook: Third Edition—Volume I*. Report No. FHWA-HRT-06-108. Federal Highway Administration, USA

[Martin, 1986] Martin R., Yohai, V. (1986). *Influence Functionals for Time Series*. Ann. Statist., Volume 14, Number 3

[Mendenhall, 1993] Mendenhall, W., Reinmuth, J., Beaver, R. (1993). *Statistics for Management and Economics*. Belmont, Duxbury Press