A Scenario-Oriented approach for Noise detection on Traffic Flow data

Mahsa Faizrahnemoon, Francesco Alesiani, *Member, IEEE*, Luis Moreira-Matias, *Member, IEEE*

***Abstract*— Road transport solutions depend on the quality of the measurements of the underlying traffic state. This paper introduces quality indicators that aim at identify the presence of traffic measurement anomalies. The proposed method seeks inconsistency in the traffic measures by statistically evaluating the variability of measures. The computation of this indicator set is mainly based on bootstrapping. Each one of them was developed to address a distinct scenario. Experiments conducted using world traffic data shows promising results.**

I. Introduction

Today, data driven methods are impacting across many industries worldwide. Transportation is one industry that has benefited most of such approaches (e.g.: demand modeling [1], bus schedule planning [2] , agent scheduling for taxi dispatch centers [3] and intelligent parking lots [4]). Recently, a survey on improving public transportation networks indicated an emergent trend to increase such data dependency [5]. Consequently, **reliable** data collection systems are crucial for the successful deployment of Intelligent Transportation Systems (ITS) - independently of their final application.

Road traffic management is a complex task. It includes problems in different domains such as traffic safety, preventive maintenance, congestion, traffic signal deployment or traffic flow modeling and optimization [6]. Traffic flow data provides a good baseline to build data driven solutions to such problems. However, the existing data collection systems for this purpose (e.g. loop induction, video surveillance) are known to add a large amount of **noise** to the output data. This phenomenon may have multiple causes such as human configuration errors, communication outage, electromechanical failures and technological limitations, among others. It has been limiting the real-world deployment of ITS systems for this purpose in the last two decades [7]–[11].

In this paper, we introduce a scenario-oriented framework to handle the presence of such noise in traffic flow data. This framework leverages a set statistical indicators which aim to illustrate the quality levels of the traffic flow data collected on a single road section. They are defined on a percentage-wise fashion to ease their interpretation. The main advantage of employing these indicators is knowing how reliable are our ITS applications (e.g. traffic flow prediction) by evaluating the quality of the data input on a continuous manner - independently of the current traffic state. Experimental results conducted using real-world data demonstrated the high potential of this contribution.

The remainder of the paper is structured as follows: Section 2 revises the existing literature on quality of data in transport domain. Section 3 formally presents the methodology employed. Section 4 firstly describes how the dataset was preprocessed with some descriptive statistics. Section 5 details how the methodology was tested in a real scenario: firstly, the experimental setup and scenarios used to evaluate the model are introduced; then, the results obtained are presented, followed by a brief discussion on their outcome. Finally, conclusion are drawn and topics for future work are outlined.

II. Related Work

The evaluation of the traffic data reliability depends on the existence of a baseline to compare your data with. There are mainly three types of methodologies to evaluate the noise level on traffic data [12]: (1) historical consistency, (2) fundamental consistency and (3) network consistency.

The (1) historical consistency methods evaluate the data regarding the structure of their historical trends (e.g. a Tuesday morning have a typical flow behavior which does not change much on a weekly basis). This trend structure is usually mined from the data using time series analysis techniques. The work in [13] developed a framework leveraged on flow and occupancy rate measurements from one single sensor which employ such type of techniques to detect noise. It outputs four different types of errors: zero-occupancy, zero-flow, steady-occupancy and randomness.

On the other hand, fundamental consistency (2) methods take advantage of traffic flow theory formulations to express the expected flow behavior in a road section. A flow conservation approach is introduced by Vanajakshi and Rilett [14]. The authors introduce a nonlinear optimization model where the objective function aims to express the flow dependency between two consecutive sections on a road segment.

Finally, the network consistency (3) methods also aim to establish dependency relationships between the flow data measured by different sensors. Yet, the structure of these relationships are established on a data driven fashion by mining spatio-temporal patterns. A simple correlation method is proposed in [15] by suggesting an alternative measurement technology to acts as a reference. A more straightforward approach was recently introduced by Yin *et al.* [16]. These authors explore the relationships in place between a given sensor and the remaining ones on the same station by employing linear regression over the flow data. They also

Manuscript submitted May 01, 2015 and revised June 30, 2015.

This work was supported by MOBiNET, a project co-funded by EC under DG Connect (FP7-ICT-2011-6.7)-N.318485.

Francesco Alesiani and Luis Moreira-Matias are with NEC Laboratories Europe, Kurfürsten-Anlage 36, 69115 Heidelberg, Germany (phone: 0049-6221-4342261; e-mail: {Francesco.Alesiani,Luis.Matias}[at]neclab.eu).

Mahsa Faizrahnemoon is with NEC Laboratories Europe, Kurfürsten-Anlage 36, 69115 Heidelberg, Germany (e-mail: mahsa.faizrahnemoon.2013[at]nuim.ie).

978-1-4673-6596-3/15 $31.00 © 2015 IEEE
DOI 10.1109/ITSC.2015.32

include the different lane occupancy ratio in order to feed the weights of a Kernel Regression [17] model for the same purpose.

Whereas the previous methods seek to identify situations where the data is locally inconsistent, the present work identifies inconsistency in a statistical sense, by comparing the local statistics with either the temporal and/or the spatial ones. By the use of a general tool, i.e. bootstrapping, the proposed method is able to identify statistical relevant changes on different noise oriented scenarios.

III. Methodology

This section formally introduces a set of three quality indicators to detect noise on traffic data on a single section. Its range is $[0, 1]$ where 1 means a total absence of noise and 0 stands for the opposite scenario. An schematic illustration for their computation is introduced by Fig. 1.

One of the most well-known methods for quantifying the quality of estimators is **bootstrapping** [18]. In statistics, it is commonly applied to derive data driven quality indicators. The basic idea behind this concept is to infer statistical properties about a population by sub-sampling from the original training data. Then, the reliability of each novel sample estimation is evaluated through such properties (e.g. probability distributions and/or descriptive statistics) - as samples extracted from the same population should share such characteristics.

The statistical properties to employ may vary with different domain applications. Hereby, we propose to employ the **arithmetic mean** as the main statistical property to evaluate the data noise levels. This metric is used together with bootstrapping to accomplish such evaluation. By doing so, we can monitor the data quality using *meta-statistics*, such as the standard deviation of the arithmetic means themselves. The data noise levels are expected to increase/decrease along with such meta-statistics.

Each one of the three proposed indicators address a particular scenario. The first and the second indicators are meant to evaluate the data reliability on a single time step and in a time interval, respectively, on multiple days. The third one performs such evaluation within a given time window. These indicators are formally presented below.

Let $x_s(i, d)$ define a traffic measure on section s on time instant i on day d. This measure can be composed on flow, vehicle density, speed and travel time denoted by q, ρ, v and tt, respectively. Let $\mu_s(i)$ denote the traffic measure mean as the arithmetic mean with respect to the days of the same type. It can be computed as

$$\mu_s(i) = \frac{1}{N_d} \sum_{d=1}^{N_d} x_s(i, d) \tag{1}$$

The deviation computed using the bootstrapping around this mean or its extension is used on the core of each one of the indicators, as described in the rest of this section. The higher the variability of the input data in the considered interval, lower its reliability.

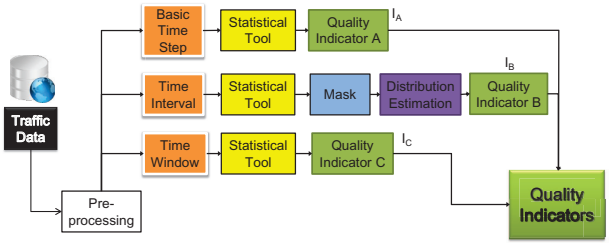

Fig. 1: An illustration of the indicators computational schema.

A. Indicator of Single Time Step

The indicator of the data of time step i is defined based on the mean and standard deviation of the input data, μ_{bi} and σ_{bi}, after applying bootstrap (subscript $_b$) on the data of the time step i along the days (see Fig. 2). The indicator is then built as

$$I_{A,i} = \frac{\mu_{bi}}{\mu_{bi} + \sigma_{bi}} \tag{2}$$

However, this indicator cannot detect noise when there is constant error in everyday data of the time step i, if the sensor is not working properly and is giving invalid measurement values for time step i, σ_{bi} will not detect such issue as the mean is stable. Consequently, a interval-based indicator is needed. It is properly introduced below.

B. Indicator of Time Interval

We want to evaluate the change in the variation of the measure over short time interval. For this purpose, let F stand for the admissible change rate on the arithmetic mean computed for a time interval. It can be defined as

$$F = \frac{\mu_{b\mu}}{\mu_{b\mu} + \sigma_{b\mu}} \tag{3}$$

where $\mu_{b\mu}$ and $\sigma_{b\mu}$ denote the mean and the standard deviation extracted by bootstrapping from the data arithmetic means measured in the interval(μ_i, μ_{i+1}, ..., μ_{i+K}), where K is chosen such that it captures similar flow on different day, as for example .5 hour. Abrupt changes on the data values will increase the value of $\sigma_{b\mu}$, decreasing F. This methodology is illustrated in Fig. 3. Such bursty changes are often

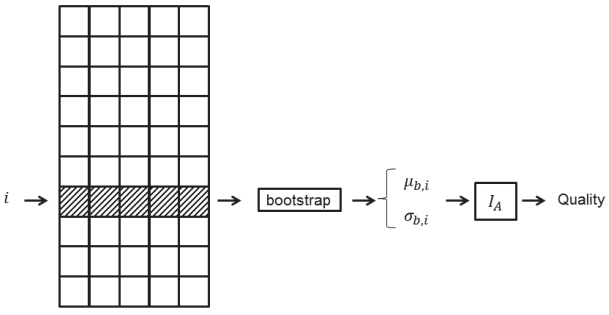

Fig. 2: The quality indicator of data of time step i (y-axis) along the days (x-axis) is defined based on the bootstrap's output.

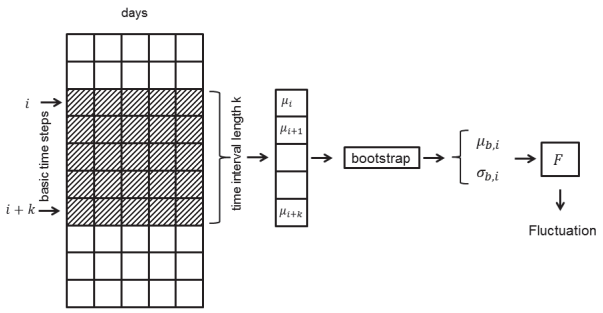

Fig. 3: Illustration of the computation of the indicator I_B.

related to a measurement error or to a physical phenomenon (e.g. heavy rains). These two cases must be distinguished to be able to quantify the quality of the data. To do it so, we use a **data mask** that is build using historical values of $\mu_{b\mu}$ and $\sigma_{b\mu}$. The data mask represents an admissible value range for which we consider any measurement within as a valid one. The next step is to find the probability distribution function (*p.d.f.*) which describes such mask parameters variation. To this scope, we model it by the use of the following *p.d.f.*:

$$f(x) = \begin{cases} \frac{1}{\sqrt{2\pi}\sigma_L} e^{\frac{-(x-\mu_M)^2}{2\sigma_L^2}}, & x \leq \mu_M \\ \frac{1}{\sqrt{2\pi}\sigma_R} e^{\frac{-(x-\mu_M)^2}{2\sigma_R^2}}, & x \geq \mu_M \end{cases} \tag{4}$$

where σ_L and σ_R denote the standard deviation left and the right tails, respectively, and μ_M stands for the arithmetic mean of the averaged fluctuation (eq.3).

To perform a sample-based estimation of such *p.d.f.*, a Kernel Smoothing method [19] is employed over a mask-based histogram. σ_L, σ_R can be obtained through computing the values of the variable that are equal $e^{-\frac{1}{2}}$ of the maximum value of the estimated distribution. Finally the quality indicator can be obtained as follows

$$I_B(F) = \begin{cases} 2\int_{-\infty}^{F} \frac{1}{\sqrt{2\pi}\sigma_L} e^{\frac{-(x-\mu_M)^2}{2\sigma_L^2}} dx, & F \leq \mu_M \\ 2\int_{F}^{+\infty} \frac{1}{\sqrt{2\pi}\sigma_R} e^{\frac{-(x-\mu_M)^2}{2\sigma_R^2}} dx, & F \geq \mu_M \end{cases} \tag{5}$$

where the indicator represents the likelihood of F fitting the mask's *p.d.f.*. The chosen *p.d.f.* compactly describes the distribution of data and allow us to derive the indicator defined in eq. (5). Although the eq. (4) is a non-continuous function, the discontinuous point μ_M is never within the integral definition in eq. (5). Consequently, such discontinuity presents no challenge to the computation of I_B.

C. Indicator of Time Window Consistency

The indicator $I_C(x_i)$ evaluates the consistency of the data x_i within a given timespan. It can be defined as

$$I_C(x) = 2Pr(X \leq x) = 2\int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}\sigma_W} e^{\frac{-(X-\mu_W)^2}{2\sigma_W^2}} dX \tag{6}$$

where μ_W and σ_W are the bootstrapping outputs whenever the data within such timespan is provided as input. Hereby, a normal distribution is employ to compute the indicator - as expressed in eq. (6). This process is illustrated in Fig. 4. These indicators were evaluated using a real world dataset collected from a freeway. This case study is briefly detailed in next Section.

IV. Dataset and Preprocessing

The proposed indicators are studied using highway traffic data collected on a major freeway operator in Europe. The area of study is a road stretch including 4 lanes. Sensors are located every 500 meters. The data includes measurements from 99 sensors. The full data set include 8 on/off-ramps. Data is available for each minute of each lane for each day of the week excluding the weekend. Before extracting the quality of the data using the proposed quality indicators, the data is aggregated at road level for two directions. Aggregation at road level avoids measuring the effect of movement of the vehicle among lanes. Each sensor data is processed to remove missing data by averaging the nearby sensor measurements, The preprocessing step uses a smoothing filter of $l = 10$. The formula used to smooth the data is

$$\frac{1}{(\sum_{j=i}^{i+l} a_j) + 1} \sum_{j=i}^{i+l} a_j x_j \tag{7}$$

where $a_j = 1$ if x_j is valid data and 0 otherwise. In the following sections the single indicators are shown graphically, in order to comparing the quality indicators in different traffic measurement condition. Of the whole set of data, we focused on flow data from the 5 days for our specific study on a section that shows more critical behavior and for verification on a section with regular behavior. Procedures were implemented in Matlab™. Fig.5 shows the distribution of the flow for all day and every 6 hours.

V. Experiments

This section details which were the experimental setup followed to conduct our simulations, the results obtained and a brief discussion about them.

A. Experimental Setup

To evaluate the robustness of I_A, it was evaluated on two scenarios: (i) two sensors that have a high discrepancy of

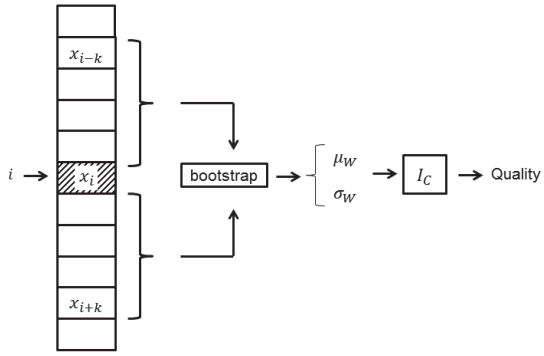

Fig. 4: Data Quality indicator within a predefined timespan.

noise, and (ii) where artificial noise was added to the data for different time periods and week days. Regarding I_B, a mask was constructed using historical data of 5 sensors of interest. These sensors are consecutive sensors and thus the measurements are correlated. The mask was used to test I_B on five sensors of interest. The indicator I_C consider the variability along the days. For evaluation proposes the third day of the week has been selected, in order to compare against the two following and preceding days. For an online system the comparison can be implemented against the last 5 days. In order to validate the approach we selected two sections and all the data of the week. The two sections were selected such that one represents a regular flow (section A), while the second (section B) exhibit relevant changes. Tab.I shows main statistics of the for the two days. For this week we identify a day with particular critical sensor readings. We used this day to shown the effect of the indicators.

B. Results

The results using each one of the three proposed noise detection indicators are illustrated in the following subsections.

1) I_A - One time step: Figure 6 illustrates the data quality evaluation example (i) using a single time step along the days of the week using indicator I_A. The two sections exhibit

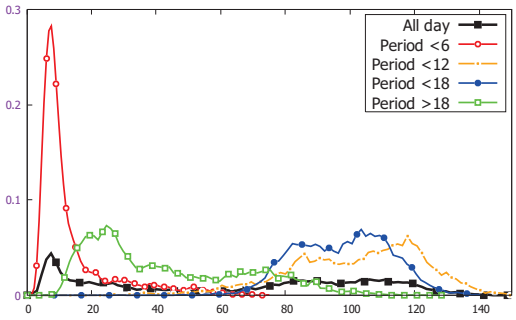

Fig. 5: Flow density distribution on 4 day periods vs time of the day (10 min).

TABLE I: Descriptive Statistics on the present case study for the selected sensors comparing the data set with and without an untypical day.

Flow	Statistics without untypical day			
	Mean	**SD**	**Min.**	**Max.**
allday	64.8761	41.2252	2.9209	141.8512
P1	21.8544	27.5411	2.9209	133.7384
P2	101.0931	**16.8893**	43.4911	141.8512
P3	99.0866	13.3512	61.0634	131.4769
P4	42.6959	23.1849	10.5106	103.8438
Flow	**Statistics with untypical day**			
	Mean	**SD**	**Min.**	**Max.**
allday	64.1308	41.0000	2.8417	143.3421
P1	21.9527	28.0177	2.8417	135.4874
P2	99.9425	**17.6299**	45.2997	143.3421
P3	97.8975	13.4323	61.0297	131.1122
P4	41.8267	23.2002	10.0570	106.8837

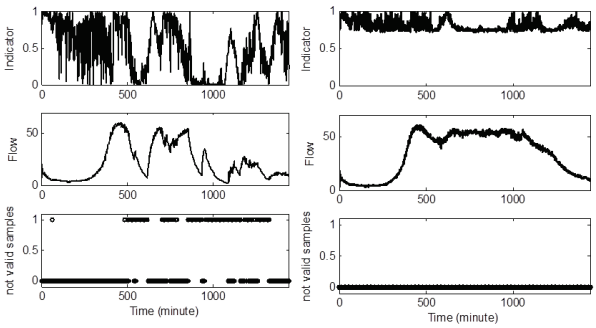

Fig. 6: Data quality evaluation for two different sections (regular on right, not regular on left) using I_A. The negative values represent true positives on the noise detection task.

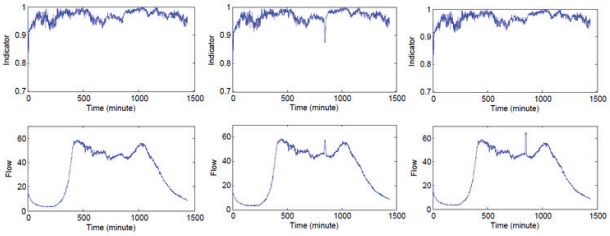

Fig. 7: Data quality evaluation using three different scenarios (organized in columns) with different levels of artificial noise induction.

different behavior, one is performing normally, while the other poorly. Fig. 7 illustrates the evaluation on a scenario with artificial noise. The first column figures shows the noise rate detection over traffic flow with no artificial error. Second column illustrates the results obtained by adding noise to one days of the week, while the third one considers a scenario where noise were artificially added to all days.

2) I_B - Time interval: Fig. 8 illustrates the histogram for the arithmetic means computed through the bootstrapping process at 7:30 time interval on section A. The figure also shows the symmetric gaussian distribution and the kernel approximation. The non symmetric gaussian distribution allows to better approximate the distribution and at the same time simplify the definition of the related indicator of eq.5.

The quality indicator I_B of section B is illustrated in Fig. 9, along with the fluctuation function, the reference mask and the flow diagram. The mask function for the evaluation of the indicator as been generated using the other sections.

3) I_C - Time Window Consistency: Time window consistency verify the statistical consistency of the data along the time dimension. Fig. 10 illustrates the evaluation of the indicator I_C using the predefined experimental setup. From Figure 10 it is possible to notice that the flow of the considered day deviates from the week average, such that the consistency indicator decreases. Figure 11 shows the time window consistency indicator for section A, the regular day. In this case the statistics along the week and the data is more consistent and lead to a high indicator output. It is possible to note that the indicator is inverse proportional to

=======

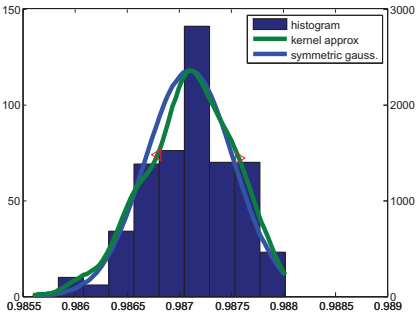

Fig. 8: The kernel approximation of the histogram of the arithmetic means computed through the bootstrapping of I_B at 7:30. Symmetric gaussian approximation, the left and right σ are also shown.

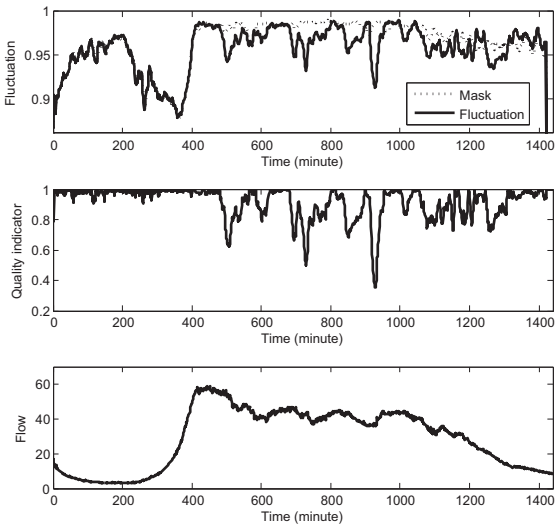

Fig. 9: The quality indicator, I_B, of a sections B. The first diagram shows the fluctuation of the measure, which shows the speed of the variation of the measure. Lower fluctuation values are associated to slowly signal changes. When the fluctuation function has values lower then the mask, as in the case of a inconsistent measure, it consequently generates a low value for the quality indicator. For section A (diagram not shown), the quality indicator does not decrease significantly.

the variance of the measure such that when the measure has lower variability the indicator is more sensitive to changes.

C. Discussion

In the previous section the method and the experimental set up are described. This section highlights the results obtained by the defined quality indicators and summarize the most relevant aspects.

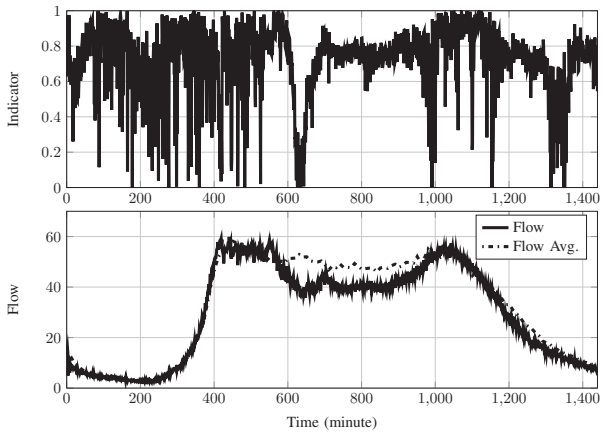

Fig. 10: Window consistency indicator I_C for a critical day. The indicator decreases when flow of the day in substantial different from the average flow of the window.

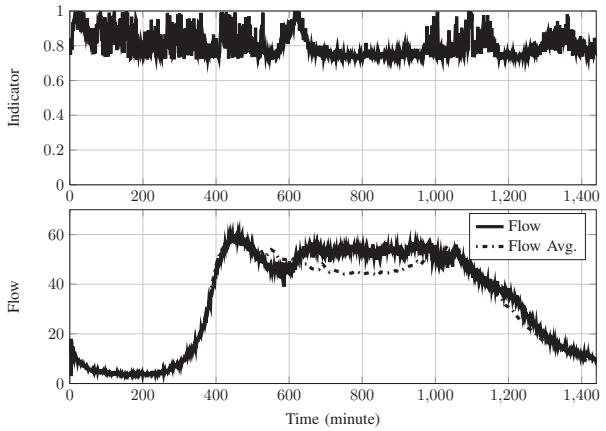

Fig. 11: Window consistency indicator I_C for normal day.

The validation has been performed considering two sections that exhibit regular and irregular measure pattern. While this analysis has been limited to these two section types, it still shows the potential of the methodology.

1) Indicator I_A: The exhibited values clearly illustrates the existing discrepancy between the sensors as well as the noise detection capabilities of this indicator, independently of its nature.

The problem is that adding the constant artificial error to the data of the whole week does not affect the quality extracted by the indicator (2) because it considers the quality of a single minute independent from the other minutes. Therefore, the constant noise added to the data of a specific minute for the whole week is not recognizable by indicator (2); this remark leads to the introduction of the indicator I_B that is focusing on variation that affects specific time window.

2) Indicator I_B: The introduction of the indicator I_B allows for the detection of situations where measures are not consistent both in space and time. This is possible with

the introduction of the mask that includes the effect of the period of the day.

This indicator evaluated with artificial error (not shown here) is able to resolve the injected error presented in figure 7, when the error is affecting the whole week.

The previous defined indicators are able to detect specific inconsistencies independent of the source of inconsistency, but requires the use of historical measurement data. While this in general is not an issue, with the introduction of the indicator I_C it is possible to detect local inconsistency without involving other days.

3) Indicator I_C: The third indicator is able to detect situation where the inconsistency is local on short interval period, allowing to react more quickly to changes in the measure. This indicator completes the possible configuration of the noise presence.

4) Extensions: More involved indicator could be considered by extending the concept, as for example by detecting inconsistency among different sections. This will be considered in future work.

5) Data set and Noisy Data: The data set considered, although small, is representative of a typical highway scenario. A more extensive application of the indicators on larger data set is necessary in order to highlight the performance on rich test scenarios and to identify when the indicators are not able to detect inconsistency. Combining the values of the proposed indicators over the day provides noisy resilient indicators.

A relevant question is the relationship with the underlying cause of the inconsistency. In this work we focused on the traffic volume, but more complex interaction between speed and density are expected to lead to improved performance in both true positive and false negative. The composition of the indicators is also an area of investigation and its compound performance in term of detection capabilities. The last remark is on the ability to distinguish or removing traffic anomaly, i.e. congestion, from measure anomaly. This work does not address this aspect, but situation of traffic anomaly can be removed from the evaluation of the sensor anomaly, indeed the former are characterized by relationship on the traffic measure that are not in general present in the case of sensor anomaly.

Despite the future direction of improvement and research the current work introduces direct and effective tools for the automatic analysis of traffic data in order to identify a class of measure inconsistency in the data.

VI. Final Remarks

This paper point out some innovative schemas to detect noise on traffic flow data. It did it so through a statistical indicator set where each indicator is meant to address distinct real world scenarios on this specific context. The experimental results validated the contributions of such indicator set to turn traffic flow data on a more reliable source.

Notwithstanding its validity, these experiments occurred using data collected through a reduced time period which did not fully exploit to potential rewards of such an approach. Consequently, experiments a more complete dataset would be desirable, Moreover, other type of straightforward methods (e.g. Principal Component Analysis [20]) might also improve the efficiency of this system to address some baseline scenarios. Hence, such hypothesis must be validated on further research.

References

[1] L. Moreira-Matias, J. Gama, M. Ferreira, J. Mendes-Moreira, and L. Damas, "Predicting taxi-passenger demand using streaming data," *IEEE Transactions on Intelligent Transportation Systems*, vol. 14, no. 3, pp. 1393–1402, 2013.

[2] J. Mendes-Moreira, L. Moreira-Matias, J. Gama, and J. F. de Sousa, "Validating the coverage of bus schedules: A machine learning approach," *Information Sciences*, vol. 293, no. 0, pp. 299 – 313, 2015.

[3] L. Moreira-Matias, R. Nunes, M. Ferreira, J. Mendes-Moreira, and J. Gama, "On predicting a call centers workload: A discretization-based approach," in *Foundations of Intelligent Systems*, ser. LNCS. Springer International Publishing, 2014, vol. 8502, pp. 548–553.

[4] R. Nunes, L. Moreira-Matias, and M. Ferreira, "Using exit time predictions to optimize self automated parking lots," in *17th International Conference on Intelligent Transportation Systems*. IEEE, 2014, pp. 302–307.

[5] L. Moreira-Matias, J. Mendes-Moreira, J. Freire de Sousa, and J. Gama, "On improving mass transit operations by using avl-based systems: A survey," *IEEE Transactions on Intelligent Transportation Systems*, vol. PP, pp. 1–18, 2015.

[6] K. Gkiotsalitis and A. Chow, "Significance of fundamental diagrams to first-order macroscopic traffic modelling," *International Journal of Transportation*, vol. 2, no. 2, pp. 15–32, 2014.

[7] A. R. Cook and D. E. Cleveland, *Detection of freeway capacity-reducing incidents by traffic-stream measurements*, 1974, no. HS-015 791.

[8] H. Payne and S. Tignor, "Freeway incident-detection algorithms based on decision trees with states," *Transportation Research Record*, no. 682, 1978.

[9] J.-B. Sheu, "A sequential detection approach to real-time freeway incident detection and characterization," *European Journal of Operational Research*, vol. 157, no. 2, pp. 471–485, 2004.

[10] S. Tang and H. Gao, "Traffic-incident detection-algorithm based on nonparametric regression," *IEEE Transactions on Intelligent Transportation Systems*, vol. 6, no. 1, pp. 38–42, 2005.

[11] M. Lippi, M. Bertini, and P. Frasconi, "Short-term traffic flow forecasting: An experimental comparison of time-series analysis and supervised learning," *IEEE Transactions on Intelligent Transportation Systems*, vol. 14, no. 2, pp. 871–882, 2013.

[12] S. T. Waller, K. M. Kockelman, D. Sun, S. Boyles, D.-Y. Lin, M. Ng, S. Seraj, M. Tassabehji, V. Valsaraj, and X. Wang, "Archiving, sharing, and quantifying reliability of traffic data," Tech. Rep., 2008.

[13] C. Chen, J. Kwon, J. Rice, A. Skabardonis, and P. Varaiya, "Detecting errors and imputing missing data for single-loop surveillance systems," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 1855, no. 1, pp. 160–167, 2003.

[14] L. Vanajakshi and L. Rilett, "Loop detector data diagnostics based on conservation-of-vehicles principle," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 1870, no. 1, pp. 162–169, 2004.

[15] X.-Y. Lu, Z. Kim, M. Cao, P. P. Varaiya, and R. Horowitz, *Deliver a Set of Tools for Resolving Bad Inductive Loops and Correcting Bad Data*. California PATH Program, Institute of Transportation Studies, University of California at Berkeley, 2010.

[16] W. Yin, P. Murray-Tuite, and H. Rakha, "Imputing erroneous data of single-station loop detectors for nonincident conditions: Comparison between temporal and spatial methods," *Journal of Intelligent Transportation Systems*, vol. 16, no. 3, pp. 159–176, 2012.

[17] E. Nadaraya, "On estimating regression," *Theory of Probability & Its Applications*, vol. 9, no. 1, pp. 141–142, 1964.

[18] B. Efron, "Bootstrap methods: another look at the jackknife," *The annals of Statistics*, pp. 1–26, 1979.

[19] A. W. Bowman and A. Azzalini, "Applied smoothing techniques for data analysis," 1997.

[20] I. Jolliffe, *Principal component analysis*. Wiley Online Library, 2005.