Measurement 116 (2018) 193–198

Contents lists available at ScienceDirect

Measurement

journal homepage: www.elsevier.com/locate/measurement

Single loop detector data validation and imputation of missing data

Mohammad Maghrour Zefreh, Adam Torok*

Budapest University of Technology and Economics, Department of Transport Technology and Economics, 1111 Budapest, Stoczek str 2, Building ST, Hungary

ARTICLE INFO

Keywords:
Single loop detector
Loop evaluation
Error detection

ABSTRACT

The data derived from loop detectors are of great importance in terms of traffic monitoring and analysis. These data may contain many holes or incorrect values due to equipment malfunctions and communication faults that may produce unreliable results. These holes (missing samples) or incorrect values (bad samples) might be problematic for any algorithm that uses the data for analysis. In this paper, a method is described that detects bad data samples gathered by the loop detectors and imputes the best available samples in order to fill the holes caused by the bad declared samples. The diagnostics algorithm proposed in this paper is based on the statistical analysis. Unlike the previous approaches, this algorithm considers the time series of many samples, rather than basing decisions on single samples. The imputation algorithm proposed in this paper uses the "good" declared samples from the historical data of the investigated loop detector to fill the holes caused by the bad declared samples. This detection and imputation process allows the algorithms that use loop data to perform analysis without requiring them to compensate for missing or incorrect data samples.

1. Introduction

Collecting better and more informative traffic data plays a vital role in attacking transport problems. This is simply because the decisions for traffic management and transportation planning are based on the quality of traffic data being collected and how well traffic data reflects the actual situations that are occurring [21,6]. There is no doubt that the decisions on traffic management and transportation planning would be compromised without accurate and reliable data collected from traffic flow sensors [17]. Loop detectors provide the most plentiful source of traffic data in most of the cities and the data gathered by them provide a powerful means to study and monitor traffic [3]. Loop data samples often contain many holes (missing samples) or incorrect values (bad samples) due to equipment malfunctions and communication faults (especially in urban areas road works often affect the loop detectors) and require careful 'cleaning' to produce reliable results. The accuracy and reliability of the data obtained from loop detectors are critical for the quality of the applications.

Since the application of electronic surveillance on roadways in the 60s of the last century, researches that evaluate detector output data have continued to be executed. Refs. [22,24] presented that the main causes of malfunction of inductive loop detector are improper installation, inadequate loop sealants, and wire failure. Loop data error has plagued their effective use for a long time [8]. Extensive studies have been conducted to diagnose and correct loop data errors [2,1,23]. In 1976, FHTVA report [19] identified five ways in which detectors can malfunction. This report presented many methods to detect errors in different time intervals based on the volume and occupancy parameters. These methods define some thresholds on minimum and maximum speed, flow, and density, and consider a sample to be invalid if they fail any of the tests. Later, Jacobson et al. [11] developed the previous algorithm by defining an 'acceptable region' in the occupancy-volume plane and declaring the samples to be good only if they fell inside the region. Their algorithm allowed a single detector system to use a surrogate of speed to screen data. This ability adds a dimension to detector error checking that has not been used in the previous algorithm (1990). Ref. [14] presented a method to identify the validity of loop detector measurements in intersections by analyzing the ratio of counts on adjacent lanes. They plotted 5 percentile and 95 percentile of counted volumes and fitted regression lines to these data. Then the loop detector measurements were considered bad if the ratio of the passing vehicles in that lane would not be placed between the two previously mentioned regression line (5 percentile and 95 percentile of counted volumes) considering the adjacent lanes. The main drawback of this method was that only flow parameter was incorporated to the validity test and neither occupancy nor speed was incorporated in the validity test.

Overall, the common errors that may occur in loop detector data are mainly as follows. Two errors will increase the total on time at the loop detector measurement: too early rising edge and too late falling edge. Three errors will decrease the total time at the loop detector measurement: too late rising edge, too early falling edge, and flicker

---

* Corresponding author.
E-mail addresses: mohammad.maghrour@mail.bme.hu (M. Maghrour Zefreh), torok.adam@mail.bme.hu (A. Torok).

https://doi.org/10.1016/j.measurement.2017.10.066
Received 15 November 2016; Received in revised form 29 October 2017; Accepted 31 October 2017

(turning off and back on in the middle of a vehicle). When the loop detector does not detect the vehicle, the missed vehicle will result in no on-time at the loop detector measurement and when it detects a vehicle in its absence the opposite problem will raise. Of course, the real total on-time at the loop detector depends both on vehicle length, which varies from vehicle to vehicle, and vehicle speed which is a function of traffic conditions [5,15,25].

The main aim of this paper is to present an innovative method in order to identify the validity of single loop detector measurements by incorporating all the available data obtained from single loop detectors (e.g. occupancy, traffic volume, speed).

1.1. Description of data

The data used in the current paper represent the entire available data for Monday, May 6, 2013, from 6:00 am to 11:59 pm at location code 710 on Villányi street and Karolina street intersection in Budapest, XI. district, Hungary. The loop detector outputs constitute the number of vehicles crossing the loop detector during a 90-s time interval t (volume, q(t)), and the fraction of this interval during which there is a vehicle above the loop (occupancy, ω(t)). Each pair of the reported volume and occupancy is called a sample in this paper. Uncertainty analysis investigates the uncertainty of variables between observations and models. Uncertainty analysis deals with the quantification of uncertainties of the relevant model parameters. In measurements, uncertainty analysis deals with assessing the uncertainty of measurement. An experiment designed to determine an effect, demonstrate a law, or estimate the numerical value of a physical variable will be affected by errors due to instrumentation, methodology, presence of confounding effects.

The investigation site is shown in Fig. 1.

The rest of the paper is organized as follows. In Section 2, at first diagnostics algorithm is designed based on the uncertainty analysis. Bad samples were diagnosed based on this algorithm and then missing and bad samples were imputed using imputation algorithm. Section 3, represents the obtained results. In Section 4, the results are analyzed and discussed. Section 5, concludes the paper and presents perspectives for future research.

2. Methodology

The main structure of diagnostics algorithm in this paper is based on the fundamental relationship of traffic flow. The fundamental equation of traffic flow:

$$q= k\,v \tag{1}$$

is based on the relation between speed (v) and density (k) which represents descriptive models of traffic behavior. The spatial density (k) of traffic can be estimated from the fraction of time (occupancy, ω) for which a vehicle is present at a fixed detector which can be measured simply by repeated sampling of the state of the detector. The measured occupancy provides a time-based estimate of the proportion of the road surface that is covered by vehicles [9]. This is related to density (k) as follows:

$$\omega = k\,L \tag{2}$$

where (L) represents the mean range of positions of a vehicle for which the detector will be occupied, known as the mean effective length of a vehicle at the detector. Preliminary analysis of Eqs. (1) and (2) will result in the relationship between the quotient (ω/L) and (q). This correspondence is investigated by plotting the values of (ω/L) against those of (q) obtained from the loop detectors in the investigation area. Fig. 2 shows the scatter plots of a sample loop detector for all the measured parameters by loop detector within each 90-s interval.

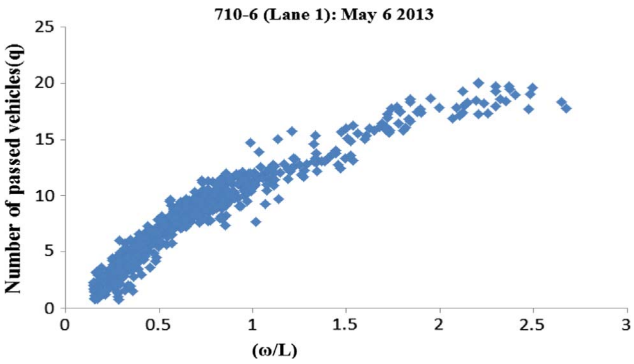

Fig. 2. Plots of flow against occupancy divided by average effective vehicle length from 6:00 am to 11:59 pm (occupancy is in percentage).

Based on the Eqs. (1) and (2) this relationship can be expressed as:

$$\omega = \frac{q*L}{v} => q= \frac{\omega * v}{L} => q= \frac{\omega}{L}*v \tag{3}$$

This fact implies that by fitting a line to the plots, the slope of the line presents speed (v) considering the general linear model. The general linear model is a statistical linear model that might be written as:

$$Y= XB + U \tag{4}$$

According to this equation, Y is a matrix with series of multivariate measurements that can be expressed as a function of a slope (B), known as the regression coefficient which is a matrix containing parameters to be estimated, times the X variable which is a matrix that might be a design matrix and a constant (U), referred to as the intercept, that might be a matrix containing errors [4].

By taking a wide look at Eq. (4), Eq. (3) can be expressed as:

$$q= \left(\frac{\omega}{L}\right) v+ \varepsilon \tag{5}$$

where ε is the error component. It should be noted that this correspondence will show the parameter speed (v) for each interval, that is, parameter B in Eq. (4). In this paper R^2 is the coefficient of determination, as a basis of goodness of linear fitting.

The overall procedure done in this research is shown in Fig. 3 and described in detail in the forthcoming sections.

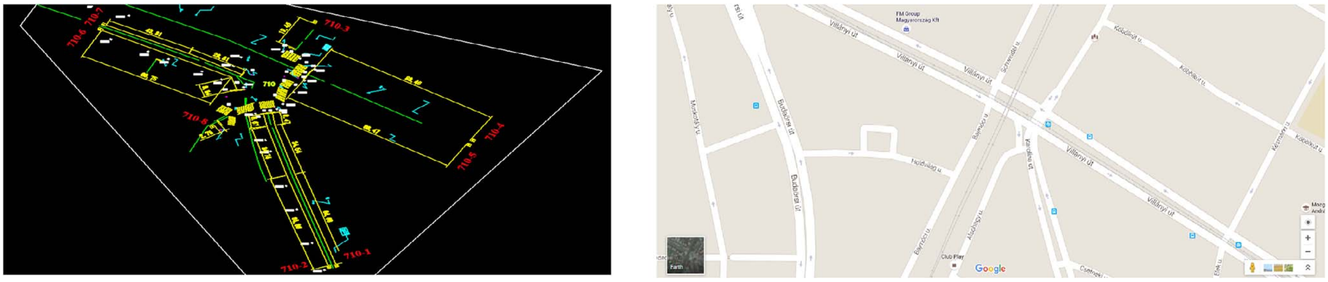

Fig. 1. Investigation site.