

DETECTION OF INVALID LOOP DETECTOR DATA IN URBAN AREAS

W.A.M. Weijermars & E.C. Van Berkum

Centre for Transport Studies

University of Twente

P.O. Box 217

7500 AE Enschede

The Netherlands

Phone: +31 534892449

Fax: +31 534894040

w.a.m.weijermars@utwente.nl; e.c.vanberkum@utwente.nl

Submission date: 4 November 2005

Number of words including abstract (230): 5668

Number of figures: 7 (7*250=1750 words) 1750

Total: 7418

ABSTRACT

Urban traffic data can be used for other purposes besides actuated signal control. It is however important that the data is of sufficient quality. This paper discusses a procedure for the detection of invalid traffic data produced by single loop detectors at signalized intersections. Basic quality checks -based on minimum and maximum flow thresholds- are used for the detection of erroneous data, whilst quality checks based on the principle of conservation of vehicles are introduced for the detection of inaccurate data. In most cases, the principle of conservation of vehicles implies that flow measurements have to be consistent between upstream and monitoring detectors within one intersection.

The basic quality checks are appropriate for the detection of erroneous data. Moreover, quality checks based on the principle of conservation of vehicles are a useful addition to basic quality checks, since they are appropriate for the detection of inaccurate traffic data that was not detected yet by the basic quality checks. However, since flows are inconsistent between upstream detectors mutually as well in some cases, it is not always clear whether the results of this quality check are reliable. These cases were analyzed further by means of field observations. Finally, since the designed control procedure detects more invalid data than the detector station itself it can be concluded that these macroscopic quality checks are a useful addition to the microscopic quality checks.

INTRODUCTION

Data quality is an important issue in traffic data archiving. Quality control techniques for archived data should encompass at least: (a) missing data, (b) suspect or erroneous data, and (c) inaccurate data (1). The difference between erroneous and inaccurate data is in the plausibility of the data. Erroneous data values do not fall within expected ranges or meet established principles or rules, whilst inaccurate data values are systematically inaccurate but within range of plausible values (1). Both erroneous and inaccurate data will be referred to as invalid traffic data in the remainder of this paper.

Data quality checks can be executed at two levels: microscopic and macroscopic (2). The microscopic level concerns tests that are executed on individual vehicle data. Jacobson et al (2) describe three basis microscopic tests that are executed in the field. Chen & May (3) developed a test in which the average vehicle on-time is compared to the average on-time for a station of detectors. Coifman & Dhoorjaty (4) present eight detector validations tests that use (combinations of) the velocity of vehicles, headways, on-times, and vehicle lengths as test statistics.

The macroscopic level refers to quality control algorithms that are executed at the central database on aggregated data like traffic flow measurements and occupancy values. Most macroscopic quality checks that are executed in practice are based on minimum and maximum range value threshold of individual records of traffic flow or occupancy measurements. In the literature, more sophisticated quality checks are discussed.

First of all, traffic data can be checked for implausible combinations of flow characteristics (volumes, occupancies and speeds). Jacobson et al (2) for example developed an error detection algorithm for the Freeway Traffic Management System in the Seattle area that uses the volume to occupancy ratios as a test statistic. Cleghorn et al (5) extended the work of Jacobson et al by tightening the upper bound for the flow-occupancy ratios through the application of traffic flow theory. Turner et al (6) analyzed data from the TransGuide traffic monitoring centre in San Antonio and identified illogical or physically impossible combinations of speed, volume, and occupancy. Secondly, series of traffic flow measurements can be analyzed as a whole. Chen et al (7) for example developed a diagnostics algorithm that examines a series of measurements from an entire day. In that way, not only implausible individual measurements are detected, but also implausible time series of traffic measurements.

Another more advanced quality check compares traffic flow measurements with historical data from the same location. This type of quality check is used in Maryland (3) and for most ITS systems in Europe (8). Moreover, Ishak (9) developed an algorithm that clusters the input space of flows, occupancies and speeds into regions of highly concentrated observations and measures the level of uncertainty in each observation. A problem with this type of quality check is that traffic flows can also deviate from historical values as a result of special or changed circumstances, instead of malfunctioning detectors.

Finally, some quality control procedures compare traffic flow measurements between upstream and downstream loop detectors. Kwon et al (10) developed an algorithm for the detection of configuration errors based on the correlation between measurements. Their method could also be used for the detection of erroneous data. Also quality checks that are based on the principle of conservation of vehicles compare traffic flow measurements between upstream and downstream locations. Until now, this type of

quality check is only applied on highways. The urban transportation network exhibits characteristics that call for an adaptation of the quality check. Moreover, the emphasis of the current literature is on the consistency of flows rather than on the detection of inaccurate loop detectors.

This paper describes a control procedure for the detection of invalid loop detector data for the urban transportation network. Basic quality checks, based on minimum and maximum flow thresholds, are used for the detection of erroneous data, whilst the principle of conservation of vehicles is used for the detection of inaccurate data. The next section discusses the principle of conservation of vehicles. Next we describe the available traffic data. Then the control procedure is designed and subsequently evaluated. Finally some conclusions are drawn.

PRINCIPLE OF CONSERVATION OF VEHICLES

The principle of conservation of vehicles states that vehicles cannot be created or lost along the road (11). In other words, the total number of cars counted by an upstream detector should be counted by the downstream detector at some future time (12). The flow conservation law can be represented by (1)

$$-\frac{\partial q}{\partial x} = \frac{\partial k}{\partial t} \quad (1)$$

Where q = traffic flow (vehicles/hour)
 k = density (veh/km)
 x = location (km)
 t = time (hour)

The principle of flow conservation is used in a broad field of research, for example in OD-estimation using traffic counts (13) and travel time estimation (14). In this research, the principle of flow conservation is used for the detection of inaccurate data. Previous research about this topic focused on the adjustment or correction of observed value to comply with the principle of conservation of vehicles (12, 15, 16).

Kikuchi & Miljkovic (15) applied the concept of fuzzy optimization. The method attempts to find the adjusted value as close to the observed value as possible while satisfying the conservation of flow at every point in the network. Vanajakshi & Rilett (16) apply a constrained nonlinear optimization approach, using the generalized reduced gradient method. Also in their study, the resulting traffic flow estimates follow the conservation principle with the least change of the original data. In both studies the emphasis is thus on fulfilling the requirement of conservation of vehicles. When inconsistencies between upstream and downstream traffic flow measurements are detected, all counts are adjusted to meet the requirement of conservation of vehicles. As a consequence it is unknown whether the traffic flow measurements at the individual detectors accurately present the actual flows at these detectors.

Wall and Dailey (12) follow a different approach. They first select a reference station that produces accurate counts. Next, stations adjacent to the reference station are compared with the reference station. If errors are present, the bias in the station is calculated and a correction time series is applied. After correction, the station is used as

the reference station for the next adjacent station calibration. Regarding to correction time series, it is determined how many cars must be counted by the station before the algorithm adds or subtracts an additional car. Subsequently, this correction factor is used for the determination of a corrected time-series. This approach works well on highways, where traffic cannot be created or lost along the route and traffic flow measurements are available for every 20 seconds.

On the urban transport network, traffic is only detected at signalized intersections. Between these intersections, traffic can be generated or lost at for example unsignalized intersection, gas stations and parking lots. Therefore, traffic flows are not necessarily consistent between upstream and downstream intersections and the principle of flow conservation cannot be applied for the detection of inaccurate traffic data. However, at some intersections, traffic flows are measured twice and at most intersections, also other types of detectors are present. The principle of flow conservation can be applied between these locations.

Moreover, Wall & Dailey had 20-second traffic flow measurements at their disposal and accounted for the travel time between two stations by means of a time lag based on the correlation between the traffic flow measurements of both locations. The traffic data we use is available at 5-minute intervals. Moreover, as a consequence of traffic light cycles, no average time lag could be calculated.

Because of the discussed differences between highways and the urban environment, a different approach is chosen for the detection of inaccurate traffic data produced by urban loop detectors. The next section describes the data collection system and the available traffic data.

DATA COLLECTION SYSTEM

In most Dutch cities traffic actuated signal control is implemented. Traffic data necessary for this actuated signal control is collected by means of single loop detectors. This traffic data can also be used for other purposes, like urban traffic monitoring and traffic management.

For this research, traffic data from Almelo – a medium sized city in the east of the Netherlands- is used. Single loop detectors are located on all approaching lanes of signalized intersections. In most cases, a lane contains multiple detectors. In general, two kinds of detectors can be distinguished, short detectors for the detection of vehicles and long detectors for the detection of queues. In Almelo, four types of detectors are distinguished. Type 1 detectors concern short detectors that are used for traffic monitoring. These detectors will be referred to as monitoring detectors (*m*). Type 4 detectors also concern short detectors, these detectors are however only used by the controller and not for traffic monitoring. Type 2 and type 3 detectors concern long detectors. The exact configuration differs by intersection and by lane. Figure 1 shows some of the configurations that occur in Almelo.

The data collected at loop detectors is processed into traffic flow and occupancy measurements and stored in the database of a traffic information system called Viacontent. The data is stored at 5-minute, 15-minute or 30-minute intervals, depending on the type of controller. Only the traffic flow measurements that are stored in the ViaContent database are used for this research. Raw traffic data is not available.

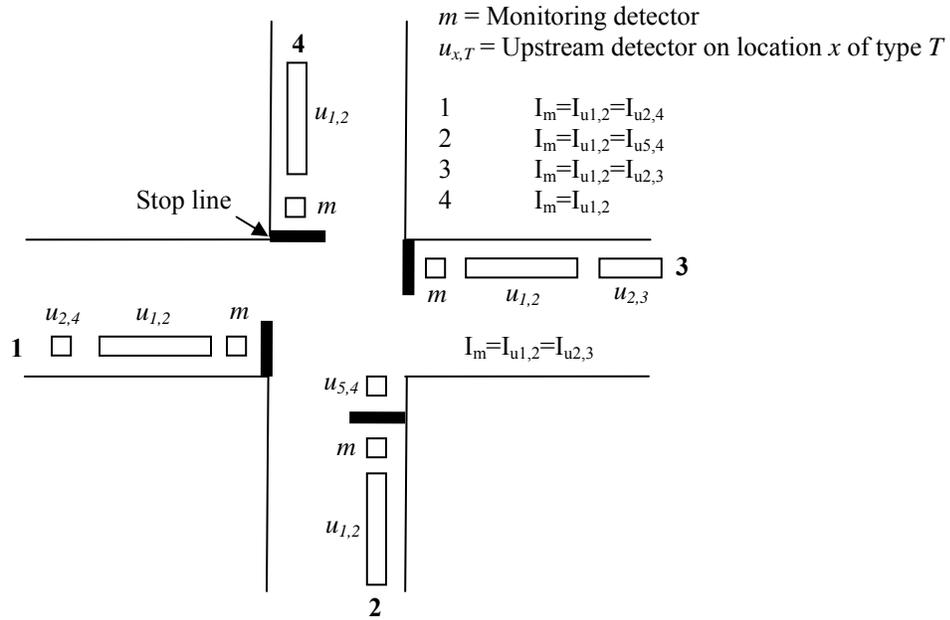


FIGURE 1 Examples of detector configurations that occur in Almelo.

Loop detector stations themselves also execute some quality checks. A record is flagged when (a) the current in the loop is interrupted or (b) occupancy is 0 for a certain time interval or (c) occupancy is 100% for a certain time interval. In case of occupancy of 100% the length of the time interval is fixed and set on 10 minutes for most links and 30 minutes for very congested links. In case of zero occupancy, the length of the time period is based on the occupancy during the previous period.

BASIC QUALITY CHECKS

The basic quality checks, designed for the detection of erroneous data are based on minimum and maximum flow threshold. Since it is agreed with Chen et al (7) that a control procedure that analyses a series of traffic counts is preferred over a control procedure that analyses individual measurements only, all quality checks are executed on a series of traffic flow measurements that are reported on a certain location on a certain day. Therefore, let us define R_{md} as a record of traffic flow measurements on monitoring location m and on day d . Q_i is defined as the i^{th} quality check

$$Q_i(R_{md}) = \begin{cases} 0 & \text{when quality check is not passed} \\ 1 & \text{when quality check is passed} \end{cases} \quad (2)$$

A record is removed from further analysis when one or more of the quality checks are not passed.

Maximum Flow Thresholds

Regarding the maximum flow threshold, traffic flows are bounded by the capacity of the measurement location and by the capacity of upstream locations. Naturally, the capacity is not the same for all locations. Moreover, the capacity varies in time as a result of different circumstances (e.g. weather). For reasons of simplicity, a fixed upper bound is used that is the same for all locations and under all circumstances. Since the only purpose of this upper limit is to detect implausible traffic flows, a rough estimate of the actual capacity is acceptable.

In literature 250 vehicles per 5 minutes (i.e. 3000 vehicles per hour) is considered to be an appropriate upper limit (1, 17). This threshold is however only valid for links. On signalised intersections, only during green time, traffic can flow. Therefore, measurements above 1000 vehicles per hour are flagged to be suspicious, whilst measurements above 3000 vehicles per hour are flagged to be erroneous. Suspiciously high traffic flows are further investigated on the basis of the daily traffic profile. When traffic flows are alternately very high and very low or are very high for a lot of consecutive time intervals, records are assumed to contain erroneous traffic data. When records show traffic flows above 1000 vehicles for some intervals during peak periods, records are assumed to be valid. The resulting algorithm for the quality check based on the maximum threshold can be represented by (3)

$$Q_1(R_{md}) = \begin{cases} 0 & \text{when } \exists_{t \in d} (I_{mt} > 3000 \vee (1000 \leq I_{mt} \leq 3000 \wedge R_{md} \text{ shows abnormalities})) \\ 1 & \text{otherwise} \end{cases} \quad (3)$$

Where I_{mt} = traffic flow measurement at monitoring detector m during time interval t

Minimum Flow Thresholds

Three minimum flow thresholds are used. First of all, negative traffic flow measurements are removed from the database by (4)

$$Q_{2a}(R_{md}) = \begin{cases} 0 & \text{when } \exists_{t \in d} I_{mt} < 0 \\ 1 & \text{otherwise} \end{cases} \quad (4)$$

Secondly, traffic flows may be zero for one or more 5 minute intervals on quiet locations and during the evening and night, but traffic counts of zero vehicles for multiple consecutive time intervals are suspicious. Daily traffic flows cannot be zero (except in case of road works or other special circumstances). Besides, hourly traffic flows of zero vehicles are suspicious, but might occur. Therefore measured hourly flows of zero vehicles are further examined. When present, upstream detectors are used for the verification of zero flow measurements. Because of the travel time between upstream and downstream flow measurements, low traffic flows can be measured at upstream detectors in case of zero traffic flows at a well-functioning monitoring detector. Therefore, upstream hourly flow measurements have to be larger than a certain threshold to report a monitoring detector to be malfunctioning. On the basis of a preliminary analysis, the threshold value is set on 20 vehicles. In cases where no upstream detectors are available, records with hourly flow measurements of zero vehicles are further investigated by

examination of the daily flow profile. When flow measurements are zero for multiple consecutive hours or alternately zero and rather high, a detector is assumed to be malfunctioning. The algorithms that check for zero traffic flows can be represented by (5) and (6).

$$Q_{2b}(R_{md}) = \begin{cases} 0 & \text{when } I_{md} = 0 \\ 1 & \text{otherwise} \end{cases} \quad (5)$$

$$Q_{2c}(R_{md}) = \begin{cases} 0 & \text{when } \exists_{h \in d} (I_{mh} = 0 \wedge ((\exists_u I_{umh} > 20) \vee (n = 0 \wedge R_{md} \text{ shows abnormalites}))) \\ 1 & \text{otherwise} \end{cases} \quad (6)$$

Where $I_{mh} = \sum_{j=1}^p I_{m(h-1)*p+j}$ With p =number of measurement intervals in an hour

$$I_{md} = \sum_{h=8}^{19} I_{mh}$$

I_{umt} , I_{umh} , I_{umd} = reported traffic flows at location u upstream of m during time intervals t , h and d

n = number of upstream detectors

QUALITY CHECK BASED ON THE PRINCIPLE OF FLOW CONSERVATION

As discussed in the second section, quality checks developed for highways are not appropriate for the urban road network. First of all, vehicles can be generated or lost between signalized intersections at for example unsignalized intersections and parking lots. Therefore, the principle of flow conservation can only be applied between locations within one intersection. Moreover, since traffic data is only available on 5-minute level and the time lag is influenced by the traffic light cycle, quality checks are executed on a higher aggregation level. For a fixed time interval, it is investigated whether the percentage difference between traffic flow measurements of two locations within an intersection are within a certain threshold.

The difference in the number of detected vehicles between two locations should be corrected for the number of vehicles between these locations. The more aggregate the data, the less influence the vehicles between the detectors have on the percentage difference in the number of vehicles. Therefore, hourly and daily traffic flow measurements are used for the quality checks. The maximum difference due to storage of vehicles depends on the distance between the detectors (dx) and could be calculated using the jam density (k_j). However, for reasons of simplicity, the maximum allowable difference due to storage is set on the same value for all locations. The more aggregate the data, the less influence of the stored vehicles on the percentage difference between two traffic flow measurements. Therefore, also for the quality checks based on the principle of conservation of vehicles, hourly and daily traffic flow measurements are used. Moreover, a correction factor is used to correct for differences due to storage of vehicles. For reasons of simplicity, the correction factor is set on 20 vehicles for all locations.

When $L1$ and $L2$ are two locations between which the principle of flow conservation is applied and T is the threshold for the percentage difference between these detectors, the

general algorithms for the quality control check on the basis of the principle of flow conservation can be represented by (7) and (8)

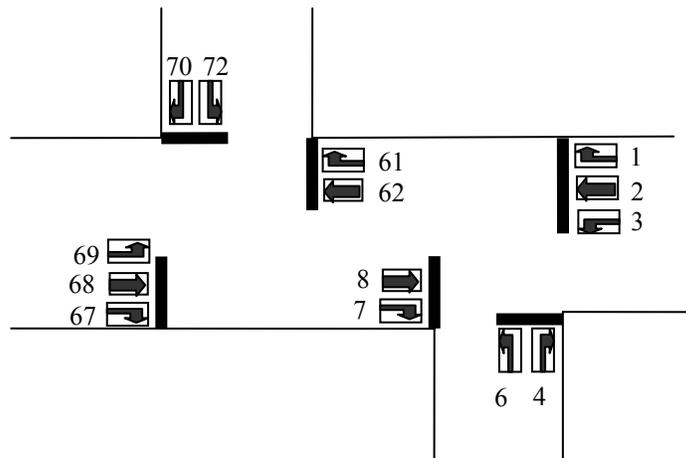
$$q_a(R_{md}) = \begin{cases} 0 & \text{when } (|I_{L1d} - I_{L2d}| - 20/0.5 * I_{L1d} + 0.5 * I_{L2d}) * 100\% > T \\ 1 & \text{otherwise} \end{cases} \quad (7)$$

$$q_b(R_{md}) = \begin{cases} 0 & \text{when } \exists_{h \in \text{day}} (|I_{L1h} - I_{L2h}| - 20/0.5 * I_{L1h} + 0.5 * I_{L2h}) * 100\% > T \\ 1 & \text{otherwise} \end{cases} \quad (8)$$

The maximum allowable percentage difference between two types of detectors is determined by the required data quality and the accuracy of the detectors. The required accuracy depends on the application of the traffic data. For traffic management applications, 10% is a possible accuracy threshold (8, 18). When the detectors are however more accurate, a lower threshold can be adopted. For this study, we determined the maximum acceptable threshold on the basis of regular differences in flows that occur between detectors. The expected inaccuracy should however not be larger than 10%.

Principle of Flow Conservation Between Two Sets of Monitoring Detectors

The principle of flow conservation between two sets of monitoring detectors is shown in figure 2.



Let us define I_k as the flow at detector k . According to the principle of conservation of vehicles the following equations apply:

$$I_1 + I_2 + I_6 = I_{61} + I_{62}$$

$$I_{67} + I_{68} + I_{72} = I_7 + I_8$$

FIGURE 2 Principle of flow conservation between two sets of monitoring detectors.

On the basis of figure 3, 5% is chosen as a threshold for the maximum percentage difference between two sets of monitoring detectors.

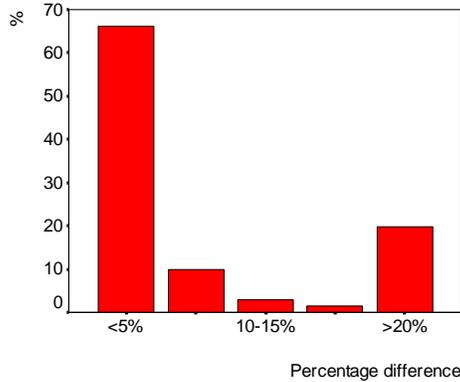


FIGURE 3 Frequency chart percentage difference in daily flow measurements.

Principle of Flow Conservation Between Monitoring and Upstream Detectors

The principle of conservation of vehicles can also be applied between monitoring detectors and upstream detectors. As discussed in the section ‘data collection system’ the number and types of upstream detectors differ by intersection and lane. Moreover, the principle of conservation of vehicles can be applied both on lane and on link level. In this section, only the general control procedure is described. Figure 1 shows the principle of conservation of vehicles for some of the configurations. Figure 4 provides information about the accuracy of different types of detectors.

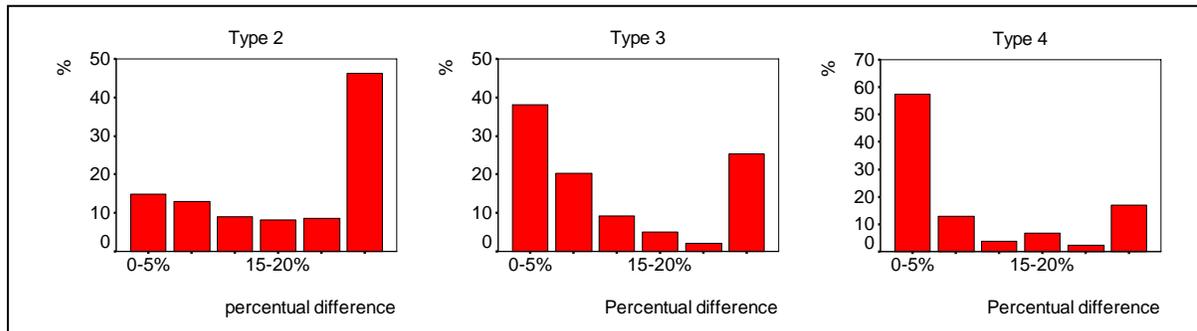


FIGURE 4 Frequency charts ΔI_{dm} between monitoring and upstream detectors

The differences in traffic counts differ between different types of detectors. From the relatively large differences in reported flows between monitoring detectors and type 2 detectors, it can be concluded that type 2 detectors produce less accurate counts than other types of upstream detectors. Moreover, short detectors (monitoring detectors and type 4 detectors) produce more accurate counts than long detectors. This was expected as the main task of long detectors is the detection of queues instead of vehicles. Finally, in general type 4 detectors produce less accurate counts than short detectors that are selected for traffic monitoring. Thus, in case of a difference between a monitoring detector and one upstream detector it is most likely that the upstream detector is producing inaccurate counts. These findings have some implications for the quality check. The quality check is based on the general flow conservation law described by (7) and (8), but is adjusted to the specific situation. First of all, flows measured at monitoring detectors are taken as the base value, i.e. the difference in relation to the flow measured at the monitoring location

is determined instead of the difference in relation to the average flow. The threshold T is set on 10. The threshold is only calculated when the record meets the following requirements:

- Two or more upstream detectors are available
- At least one of the upstream detectors is of type 3 or type 4.
- All upstream detectors report either higher or lower traffic flows.

EVALUATION OF QUALITY CHECKS

Ideally, the control procedure would detect all measurements that do not represent actual traffic flows. However, invalid data might be undetected by the quality control checks. On the other hand, data might be flagged to be erroneous, whilst in reality the data is of sufficient quality. This section describes the evaluation of the quality checks. First of all, the basic quality checks are evaluated by means of examination of the daily flow profiles of the flagged records. Secondly, the quality checks based on the principle of conservation of vehicles are more thoroughly evaluated. Finally, the results of the described quality checks are compared with the results of the quality checks executed by the detector station itself.

Basic Quality Checks

Table 1 shows the percentage and number of records that were concluded to contain erroneous or suspicious data on the basis of the basic quality checks. From the table can be concluded that only a small percentage (2.8%) of the records showed erroneous data.

TABLE 1 Results of Basic Quality Checks

Quality Check	% of records that failed check	# of rec that failed check	Number of erroneous rec (after further investigation)
$I_{mt} > 3000$	0.5%	111	111
$1000 < I_{mt} \leq 3000$	1.3%	309	103
$I_{mt} < 0$	0%	0	0
$I_{md} = 0$	1.2%	286	259
$\exists_{head} I_{mh} = 0 \wedge I_{md} > 0$	6.0%	1376	157
Total data set	100%	23.213	

The examination of the daily flow profiles of these records showed that all records that contained flow measurements exceeding the maximum threshold showed abnormal daily flow profiles, i.e. flows were alternately very high or very low or very high for a lot of consecutive time intervals. Furthermore, a minority of the cases that showed suspiciously high traffic flow measurements appears to contain erroneous data (33%). For the other cases, the daily flow profile did not show any abnormalities.

In case of zero traffic flows it is verified whether these could be due to road works. This appeared to be the case in 9% of the cases. Moreover, hourly traffic flow measurements of zero vehicles often were due to the absence of traffic on Saturday and Sunday mornings or a late morning peak on quiet locations. Only in 11% of the cases that zero hourly traffic flows were reported, the detector was malfunctioning.

Principle of Conservation of Vehicles

The principle of flow conservation could only be applied in about one third of the cases. For the other cases, the detector configuration did not meet the requirements. Table 2 shows the results of the quality check between two sets of monitoring detectors. Table 3 shows the results for checks between monitoring detectors and upstream detectors.

TABLE 2 Flow Conservation between Two Sets of Monitoring Detectors

	Flow conservation for hourly flows		
Flow conservation for daily flows	Met	Not met	Total
Met	323 cases	33 cases	356 cases (66.2%)
Not met	7 cases	175 cases	182 cases (33.8%)
Total	330 cases (61.3%)	208 cases (38.7%)	538 cases

For almost all cases that traffic flows are inconsistent between two sets of monitoring detectors, one of the detectors did not pass the basic quality checks. In these cases, only traffic measurements of the malfunctioning detector are removed from further analysis.

TABLE 3 Flow Conservation between Monitoring and Upstream Detectors

	Flow conservation for I_h		
Flow conservation for I_d	Met	Not met	Total
Met	5901 cases	539 cases	6440 cases (86.5%)
Not met	87 cases	915 cases	1002 cases (13.5%)
Total	5988 cases (80.5%)	1454 cases (19.5%)	7442 cases

From table 3 can be seen that the principle of flow conservation is not met in 21% of the cases $((1 - 5901/7442) * 100\%)$. Of these cases, only 3.4% was already reported to contain erroneous data. Most cases thus contain inaccurate data that was not detected yet by the basic quality checks.

It is however not always clear whether the results of this quality check based on the principle of flow conservation between upstream and monitoring detectors are reliable. In some cases that inconsistencies between downstream and upstream flow measurements occurred, upstream detectors also show large mutual differences in traffic flow measurements. These cases were analyzed further by means of field observations and are concluded to be unrightfully removed from further analysis. Moreover, some records possibly unrightfully passed the quality check. In these cases, upstream detectors produced contradictorily results. In 481 cases (6.5% of all records that met the configuration requirements) one of the upstream detectors reported higher flows compared to the monitoring detector, whilst another upstream detector reported lower flow measurements. Also these cases were analyzed further by means of field observations. Some of these cases should have been removed from further analysis, whilst others should not. As a result of the executed field observations some additional records are flagged to contain invalid data, whilst some other -initially flagged records- are reported to contain valid data. After these corrections, 14% of the 7442 records do not meet the principle of flow conservation. Of these records, 5% was already reported to contain erroneous data.

Finally, although the thresholds are chosen after an explorative analysis of the data and are realistic from a traffic theory point of view, they are somewhat arbitrary. Therefore, sensitivity analyses are executed. Regarding the threshold on the maximum number of vehicles queuing between two detectors, the percentage of failures for the flow conservation requirements increased from 22% to 27% when the threshold was set to 10 vehicles and decreased to 20% when the threshold was set to 30 vehicles. Regarding the threshold on the maximum percentage difference between two sets of monitoring detectors, the percentage of failures decreased from 40% to 34% when the difference was set to 10%. Regarding the maximum percentage difference between monitoring detectors and upstream detectors, the percentage of failures increased from 21% to 23% when the difference was set to 9% and decreased to 19% when the difference was set to 11%.

Microscopic Quality Checks vs Macroscopic Quality Checks

The results from the macroscopic quality checks discussed in this paper are compared to the microscopic quality checks executed by the detector station itself. In 8% of the cases, records are concluded to contain erroneous data on the basis of the macroscopic quality checks, whilst the records passed the microscopic checks.

In most of these cases, the principle of conservation of vehicles is not met, so in most cases inaccurate data is not detected by the detector station itself. However, in less than half (47%) of the cases for which basic quality checks report erroneous data, the detector station reports erroneous data as well. A large part of the erroneous data thus remains undetected by the microscopic quality checks.

On the other hand there are only few cases (< 1%) for which the detector station itself reports detector malfunctioning, whilst the macroscopic quality checks are passed. For 40% of the cases for which detector malfunctioning is reported on the basis of microscopic tests executed by the detector itself, the basic quality checks do not report detector malfunctioning. These cases mainly concern records for which detector malfunctioning is reported for one 5-min interval or more consecutive time intervals during the night or evening.

CONCLUSIONS

This paper discusses the design and evaluation of a procedure for the detection of erroneous and inaccurate data, produced by a set of single loop detectors on an urban road network. The basic quality checks, based on minimum and maximum flow thresholds appear to detect erroneous data successfully. Records that contain traffic flow measurements of more than 3000 vehicles per hour show abnormal daily traffic profiles and are thus rightfully flagged to contain erroneous data, whilst records that contain traffic flow measurements between 1000 and 3000 vehicles per hour only show abnormal daily traffic profiles in a minority of the cases. Also hourly traffic flows of zero vehicles are not always due to detector malfunctioning.

Quality checks based on the principle of conservation of vehicles are a useful addition to basic quality checks, since 95% of the invalid data, detected by inconsistencies of flows between upstream and monitoring detectors, was not detected by the basic quality checks. However, it has to be noted that the requirements concerning the detector configuration for applying this quality check are only met in about one third of

the cases. Moreover, in some cases it is unclear whether the results of this quality check are reliable. First of all, in some cases traffic flow measurements at upstream detectors show large mutual differences. From field observations is concluded that these cases do not show inaccurate data. Moreover, there are cases for which one upstream detector reports higher traffic flow measurements than the monitoring detector, whilst another detector reports lower measurements. From field observations is concluded that some of these records should have been removed from further analysis.

From a comparison of the macroscopic quality checks with the quality checks executed by the detector station itself, it is concluded that the detector station itself does not detect all invalid data. Firstly, as expected, most of the inaccurate data (detected by the quality checks based on the principle of conservation of vehicles) is not detected by the microscopic quality checks. Moreover more than half of the erroneous data –detected by the basic quality checks- remains undetected by the microscopic quality checks as well. Therefore, the macroscopic quality checks, described in this paper are a useful addition to the quality checks executed by the detector station itself.

Summarized, the described control procedure has proven to be appropriate for the detection of erroneous and inaccurate loop detector data in urban areas. After detected, invalid loop data can be either removed from the database, ore corrected by means of spatial or temporal regression or other methods proposed in literature (e.g. 12, 15, 16).

ACKNOWLEDGEMENT

This research is part of the Dutch TRANSUMO program. The authors would like to thank the municipality of Almelo and Vialis for providing the traffic data for this research.

REFERENCES

- (1) Turner, S. *Guidelines for Developing ITS Data Archiving Systems*, Report 2127-3 Texas Transportation Institute, 2001.
- (2) Jacobson, L. N. Nihan & J. Bender. Detecting Erroneous Loop Detector Data in a Freeway Traffic Management System. In: *Transportation Research Record: Journal of the Transportation Research Board, No 1287*, TRB, National Research Council, Washington, D.C., 1990, pp. 151-166.
- (3) Chen, L. & A. May. Traffic detector errors and diagnostics. In: *Transportation Research Record: Journal of the Transportation Research Board, No 1132*, TRB, National Research Council, Washington, D.C., 1987, pp. 82-93.
- (4) Coifman, B. & S. Dhoorjaty. Event Data Based Traffic Detector Validation Tests. In 81th annual meeting of the Transportation Research Board. CD-ROM. Transportation Research Board, National Research Council, Washington, D.C., 2002.
- (5) Cleghorn, D. F. Hall & D. Garbuio. Improved Data Screening Techniques for Freeway Traffic Management Systems. In: *Transportation Research Record: Journal of the Transportation Research Board, No 1320*, TRB, National Research Council, Washington, D.C., 1991, pp. 17-31.
- (6) Turner, S. L. Albert, B. Gajewski & W. Eisele. Archived Intelligent Transportation System Data Quality, Preliminary Analyses of San Antonio TransGuide Data. In: *Transportation Research Record: Journal of the Transportation Research Board, No 1719*, TRB, National Research Council, Washington, D.C., 2000, pp. 77-84.
- (7) Chen, C. J. Kwon, J. Rice, A. Skabardonis & P. Varaiya. Detecting Errors and Imputing Missing Data for Single Loop Surveillance Systems. In 82th annual meeting of the Transportation Research Board. CD-ROM. Transportation Research Board, National Research Council, Washington, D.C., 2003
- (8) Turner, S. Defining and Measuring Traffic Data Quality, White Paper on Recommended Approaches. In: *Transportation Research Record: Journal of the Transportation Research Board, No 1870*, TRB, National Research Council, Washington, D.C., 2004, pp. 62-69.
- (9) Ishak, S. (2003) Fuzzy-Clustering Approach to Quantify Uncertainties of Freeway Detector Observations. In: *Transportation Research Record: Journal of the Transportation Research Board, No 1856*, TRB, National Research Council, Washington, D.C., 2003, pp. 6-15.
- (10) Kwon, J, C. Chen & P. Varaiya. Statistical Methods for Detecting Spatial Configuration Errors in Traffic Surveillance Sensors. In: *Transportation Research Record: Journal of the Transportation Research Board, No 1870*, TRB, National Research Council, Washington, D.C., 2004, pp. 124-132.
- (11) Daganzo, C. *Fundamentals of Transportation and Traffic Operations*. Pergamon, Oxford, United Kingdom, 1997.
- (12) Wall, Z.R. & D.J. Dailey. Algorithm for Detecting and Correcting Errors in Archived Traffic Data. In: *Transportation Research Record: Journal of the*

Transportation Research Board, No 1855, TRB, National Research Council, Washington, D.C., 2003, pp. 183-190.

(13) Van Zuylen, H.J. & L.G. Willumsen. The most likely trip matrix estimated from traffic counts. *Transportation Research Part B*, Volume 14, 1980, pp 281-293.

(14) Nam, D.H. & Drew, D.R. Automatic measurement of traffic variables for intelligent transportation systems applications. *Transportation Research Part B*, Vol. 33, 1999, pp 437-457

(15) Kikuchi, S. & D. Miljkovic. Method To Preprocess Observed Traffic Data for Consistency, Application of Fuzzy Optimization Concept. In: *Transportation Research Record: Journal of the Transportation Research Board, No 1679*, TRB, National Research Council, Washington, D.C., 1999, pp. 73-80.

(16) Vanajakshi, L. & L.R. Rilett. Loop Detector Data Diagnostics Based on Conservation-of Vehicles Principle. In: *Transportation Research Record: Journal of the Transportation Research Board, No 1870*, TRB, National Research Council, Washington, D.C., 2004, pp. 162-169.

(17) Lomax, T. S. Turner & R. Margiotta. *Monitoring Urban Roadways in 2002: Using Archived Operations Data for Reliability and Mobility Measurement*. FHWA-HOP-04-011, FHWA, U.S. Department of Transportation, 2004.

(18) Tarnoff, P.J. Getting to the INFOstructure. In TRB Roadway INFOstructure Conference. CD-ROM. Transportation Research Board, National Research Council, Washington, D.C., 2002.