Data Acquisition, Interfacing and Pre-processing of Highway Traffic Data

T. Bellemans, B. De Schutter, B. De Moor

K.U.Leuven, Belgium & Delft University of Technology, The Netherlands

Synopsis

Data acquisition and pre-processing are important steps in any traffic data collection, analysis or simulation project. In this paper we present the results of a case study we have done on traffic simulation for a stretch of the E17 highway near Antwerp, Belgium. We will especially focus on the data collection and processing part.

The objective of the work presented here is to illustrate how traffic data can be collected and processed in such a way that it can be used for modeling and prediction purposes. The data collection and processing process consists of three steps: data acquisition, interfacing, and pre-processing. During this process a central database with raw sensor data is created. In general the raw sensor data will contain errors or missing values due to sensor failures, data link errors, biases, and other measurement errors. Therefore, we introduce a pre-processing step in order to deal with data that is known to be corrupt or for which the value is temporarily not available (many simulation packages require data records that are complete and contain no gaps). After the three steps have been carried out, we have a database of "clean and complete" traffic data that can be used for further analysis or that can be used in other applications such as highway traffic simulation or traffic forecasts.

Authors' Biographical Details

Tom Bellemans received the degree in Electrical Engineering in 1998 at K.U.Leuven, Belgium. He is currently a research assistant funded by the Belgian federal government (Federal Office for Scientific, Technical and Cultural Affairs–OSTC) and is working towards a doctoral degree in Applied Sciences at the ESAT-SISTA research group of K.U.Leuven. The topic of his research is modeling and control of traffic flows on highways.

Bart De Schutter received the degree in electrotechnical-mechanical engineering in 1991 and the doctoral degree in Applied Sciences in 1996, both at K.U.Leuven, Belgium. Afterwards, he was a senior research assistant of the FWO (Fund for Scientific Research–Flanders) at the ESAT-SISTA research group of K.U.Leuven. Currently, he is assistant professor at the Control Lab of Delft University of Technology, The Netherlands. Bart De Schutter has received the Richard C. DiPrima Prize and the Robert Stock Prize in Exact Sciences for his PhD thesis. His current research interests include hybrid systems control, traffic flow control, multi-agent systems, and optimization.

Bart De Moor is a senior research associate of the FWO (Fund for Scientific Research–Flanders) and associate professor at the Department of Electrical Engineering (ESAT) of K.U.Leuven, Belgium. He received his doctoral degree in Applied Sciences in 1988 at K.U.Leuven. Bart De Moor has published more than 200 papers in international journals and conference proceedings and received several national and international awards for his work. He is also one of the founders of two spin-offs: ISMC NV (specialized in modeling

and control of industrial processes), and Data4S NV (active in data-mining and services in bio-informatics and telecommunications). His research interests include numerical linear algebra, system identification, control theory, and data-mining.

Data Acquisition, Interfacing and Pre-processing of Highway Traffic Data

T. Bellemans*, B. De Schutter†, B. De Moor*

K.U.Leuven, Belgium & Delft University of Technology, The Netherlands

1 INTRODUCTION

Due to the ever growing influence of traffic congestion on the economy and the environment, there is an increasing need for techniques to model, predict and control traffic flows. In order to accurately predict future traffic situations the traffic models that are used for prediction and control require measurements of the current traffic situation. In this paper, we give an account of a case study project in which we processed and collected raw sensor data from traffic sensors on a highway in such a way that it was usable to run microsimulations of the traffic flows on that highway. This process consists of three steps (see Figure 1): data acquisition, interfacing, and pre-processing.

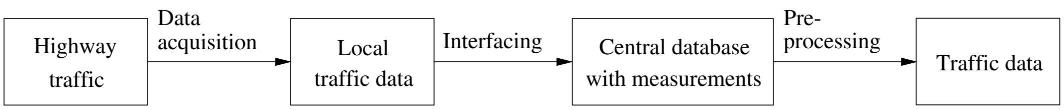

Figure 1: Schematic presentation of the traffic data collection and processing procedure.

The first step consists of the acquisition of measurements of the desired parameters using sensors. We will describe the sensor technologies used in our project and the traffic parameters that can be measured with them. In the interfacing part we discuss the transport of the data from the sensors to a central computer system. After storage of the data on the central computer, we have a central database with raw sensor data for a certain time span. In general the raw sensor data will contain errors or missing values due to sensor failures, measurement errors, data link errors, etc. Furthermore, due to measurement errors derived data (such as average speeds or vehicle counts) may have values that are physically impossible such as, e.g., negatives values for vehicle counts or speeds. However, some applications (such as microsimulation) require data records that are complete and contain no gaps. Therefore, we also describe a way to efficiently estimate appropriate values for missing data. Note that the techniques presented in this paper can easily be extended to fit other situations.

*ESAT/SISTA, K.U.Leuven, Kardinaal Mercierlaan 94, B-3001 Leuven (Heverlee), Belgium. email: {Tom.Bellemans,Bart.DeMoor} @esat.kuleuven.ac.be

†Control Lab, Faculty of Information Technology and Systems, Delft University of Technology, P.O.Box 5031, 2600 GA Delft, The Netherlands. email: b.deschutter@its.tudelft.nl

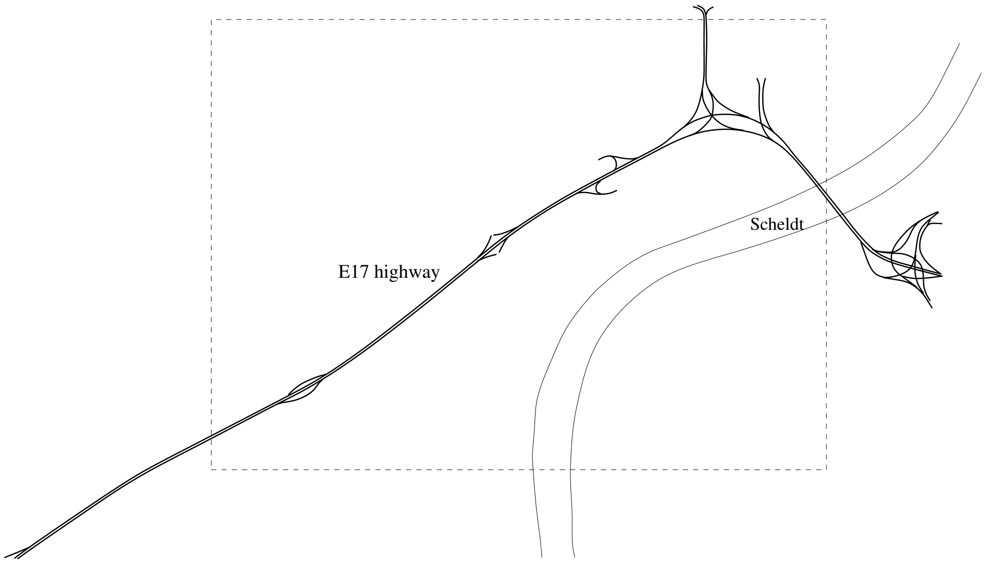

Figure 2: Layout of the set-up of our case study: an 8 km stretch (enclosed by the dashed box) of the E17 highway between Ghent and Antwerp.

2 CASE STUDY: SET-UP

The highway we have considered in our project is the E17 highway between the cities Ghent and Antwerp in Belgium. This highway is one of the main arteries leading to Antwerp seaport. This causes high flows of both cars and lorries on the highway, which leads to structural congestion during rush hours. As a case study, we focus on a stretch of 8 km of highway in the driving direction of Antwerp (see Figure 2) where structural congestion occurs during the morning rush hour. As most important highways in Belgium, the E17 highway is equipped with inductive loop detectors before each off-ramp and after each on-ramp. The combination of these measurements gives us an estimate of the nett number of vehicles leaving (or entering) the highway. The detectors measure the traffic parameters for the traffic flows on the highway and not for the traffic flows on the off-ramps and on-ramps. So if we want to know the exact number of cars leaving the highway via the off-ramp and the exact number entering via the on-ramp, we need at least one additional measurement per on-ramp/off-ramp complex. These additional measurements are in general not implemented on the Belgian highways out of cost considerations. However, for the highway stretch studied in this project, there were additional cameras placed every 500 meters in the scope of a European project. These extra measurements allow us to detect the number of cars using the on-ramps and off-ramps.

3 DATA ACQUISITION

The sensors that are located along the highway can be used to measure traffic flow parameters such as the number of vehicles passing at a certain place during a time interval (traffic intensity), the average velocity of these vehicles, and the occupancy level of the highway, time and distance headways, traffic density, etc. [3]. This information can then be used to estimate the characteristics of the traffic flow at a certain moment, e.g., to detect incidents, to model the traffic, to make estimations of the future traffic situation, and so on.

In our project there are three parameters of special interest: the vehicle count, the speed of the vehicles and the occupancy level of the highway. The most obvious parameter to be measured is the *number of vehicles* passing at a certain point in a certain time interval. The *speed* of the vehicles is another important parameter characterizing the state of the traffic on the highway. A third parameter characterizing the traffic situation is the *occupancy level* of the detector. This is the fraction of the total time that a vehicle is in sight of the detector.

The installation used in this project accomplishes a real-time classification of the vehicles into two classes: lorries and (regular) cars. The classification is done based on the estimated length of the vehicle, which can be derived from the measured speed of the vehicle and the occupancy level of the sensor for the vehicle. The three parameters mentioned above are measured for every class of vehicles separately. Moreover, these parameters are also available for every lane separately in every measurement point.

In our case study we have used three types of measurement devices: inductive loops, cameras and pneumatic sensors. In Belgium, the most widely spread traffic sensor technology is the *inductive loop*. It consists of an electric cable that is put in the road and detection is based on magnetic induction. The main advantage of an inductive loop is the relatively low cost, but the disadvantages are the limited accuracy (especially in situations with little or extreme traffic) and the sensitivity to failures (e.g., due to road surface works). Sometimes the accuracy is improved by using a more expensive double loop configuration.

Cameras along the roadside are a more expensive but also more accurate sensor technology than inductive loops. An additional advantage is that the camera images such as the one shown in Figure 3 can also be used for visual inspection by the operators of the traffic control center in order to detect and diagnose incidents. The images of the cameras are fed to an image processing algorithm, which extracts the necessary traffic parameters.

The last technology used in our project is the *pneumatic sensor*, which basically consists of a hollow tube that is put across the road. The shock waves caused by cars running over the tube are registered and processed. Pneumatic sensors are mainly applied for temporary measurements due to their high portability. Their main disadvantage is the limited accuracy (e.g., it is difficult to tell whether two successive cars or a lorry with four axes passed over the cable).

For more information on other sensor technologies that could be used in traffic data col-

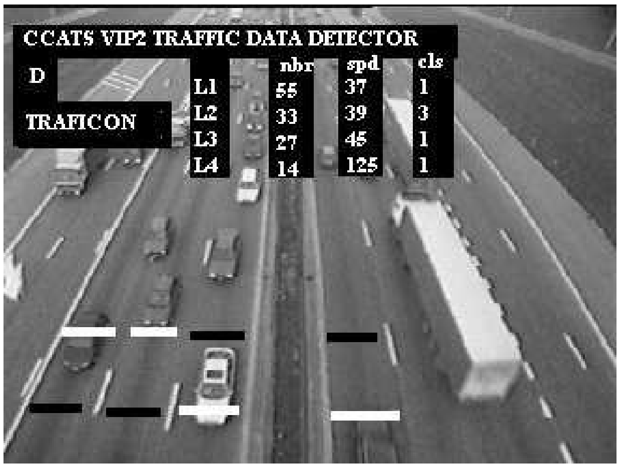

Figure 3: View of a camera along a highway, used to measure the different traffic parameters and to inspect the traffic situation visually. (Photo courtesy of Traficon)

lection such as piezoelectric sensors, ultrasonic sensors, active and passive infrared sensors, microwave sensors, or magnetometers, the interested reader is referred to [2].

4 INTERFACING

The traffic parameters are measured at several locations along the highway. In order to combine to data collected at several points and to achieve a manageable system, the measurements are converted and combined in a centralized and standardized database.

The central database is kept on a computer which is networked to the sensors at the different measurement sites. Every minute the computer polls the sensors for the newly measured parameters and stores the values in its memory or on disk.

The database on the central computer has a predetermined standard data format. In general the data generated by different sensors have different formats. Therefore, we process the raw data from the sensors using a data filter in order to convert it to a standard format (see Figure 4). The use of data filters has one major advantage: interchangeability. If a sensor is replaced by another brand or by a sensor that uses a different technology, only the data filter has to be adapted. The database and its components remain unchanged. This allows for a flexible use of different types of sensors. Furthermore, several applications, each possibly using a different data format, can also use the same central database by accessing it through a data filter.

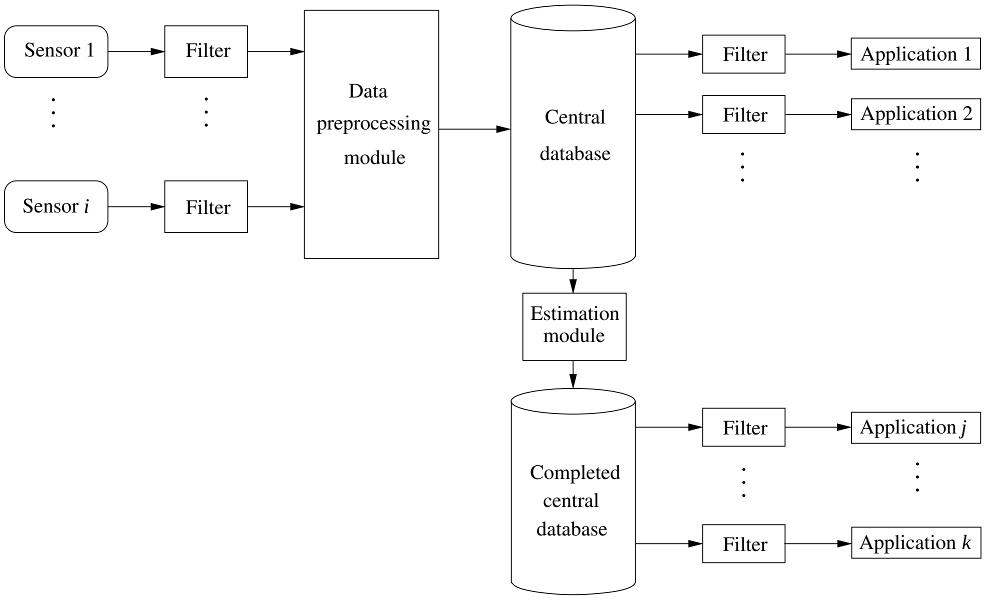

Figure 4: Schematic view of the data flows between the sensors and the database and between the database and the applications. The data filters introduce flexibility into the system.

5 DATA PRE-PROCESSING

Sometimes the sensors or the communication links are out of order or fail to produce accurate values due to power failures, interferences, servicing of the devices, works on the road surface, etc. This results in corrupt or missing data. However, some applications require complete data records without missing values. Therefore, we have included an additional data estimation module to the system as is shown in Figure 4.

5.1 Corrupt data

After it has received the different values from the sensors, the central computer validates this data using some basic assumptions. An example of such an assumption is the non-negativeness of speeds, numbers of cars, and occupancy rates. If a sensor value does not comply with the basic assumptions or if it was not available in the first place, the measurement is labeled as unavailable, using a sentinel element. In our case study the sentinel element -9999 was used; this is a value that cannot occur and that replaces the missing or corrupt value. This way, corrupt data is transformed into missing data.

5.2 Missing data: an estimation strategy

The system in our case study marks corrupt and missing data by writing a sentinel element in the database. However, for some applications, such as microsimulation of traffic flows, monitoring, or prediction, a complete dataset is required. In this section we illustrate an interpolation-based method to estimate the missing values using historical data. The main advantage of this method is that it is very efficient (so that it can also be used in on-line real-time applications) while at the same time it is also sufficiently accurate.

We make the assumption 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*. The data records for the reference week are constructed as a moving average over several weeks in the past. However, "special" days such as official holidays during the week, or days on which major events take place are not included while constructing the data records for the reference week since on such days the transportation needs and the traffic patterns may differ significantly from those on "regular" days. Note that the moving average technique also allows us to track seasonal changes in the traffic patterns.

In order to obtain a good estimation for the real value of a missing measurement value $x_{\text{data}}(k)$ we use the following formula:

$$x_{\text{data}}(k) = \frac{x_{\text{data}}(k-1)}{x_{\text{ref}}(k-1)} x_{\text{ref}}(k) \ ,$$

where $x_{\text{data}}(k)$ is the (missing) data value at time step k and $x_{\text{ref}}(k)$ is the reference data value at time step k. So the value found in the reference week value is scaled to the traffic intensity level of the time step that precedes the missing measurement. This is done using the factor $\dfrac{x_{\text{data}}(k-1)}{x_{\text{ref}}(k-1)}$.

The main advantage of the method presented above in comparison to more sophisticated methods is that its computational complexity is low so that it can also be used in on-line real-time applications such as on-line prediction of future traffic flows using measurements of the actual traffic situation in combination with traffic flow simulation, as we have done in our case study.

6 TRAFFIC DATA APPLICATIONS

In the previous sections we have discussed the steps undertaken to get a complete database of traffic data. This database can then be used in the following applications:

- Assessment of the current traffic situation, e.g., for government officials, traffic police, ...
- Incident detection.

- Estimation of road parameters such as the critical flow, free flow speed of the cars on the highway, the fundamental diagrams, ...
- Fitting a traffic model to the data.
- Simulation of highway networks in order to assess the effects of traffic controllers and control strategies (see also [1]).
- Prediction or forecasting of future traffic situations.
- ...

7 CONCLUSION

We have described a procedure to collect traffic data and to combine them in a central database, which has a standardized data format. Communication between the database and the sensors and applications is done via data filters which perform the necessary translations between the standardized data format and sensor-specific or application-specific data formats. We have also presented a technique to efficiently obtain good estimations of missing data values. The resulting "complete and clean" traffic database can then be used in applications such as microsimulation, prediction and monitoring of traffic flows.

8 REFERENCES

[1] S. Algers, E. Bernauer, M. Boero, L. Breheret, C. Di Taranto, M. Dougherty, K. Fox, and J.F. Gabard, "Review of micro-simulation models," Rep. SMARTEST/D3, Institute for Transportation Studies, University of Leeds, Leeds, UK, Mar. 1998. See also `http://www.its.leeds.ac.uk/smartest`.

[2] B. Johnson, "Keeping the world flowing," *Traffic Technology International*, pp. 38–40, June-July 1999.

[3] A.D. May, *Traffic Flow Fundamentals.* Englewood Cliffs, New Jersey: Prentice-Hall, 1990.

9 ACKNOWLEDGMENTS

Bart De Moor is a senior Research Associate with the F.W.O. (Fund for Scientific Research-Flanders).
This work is supported by several institutions: the Flemish Government (Research Council K.U.Leuven: Concerted Research Action Mefisto-666; FWO projects: G.0240.99 and

G.0256.97; FWO-Research Communities: ICCoS and ANMMM; IWT projects: EUREKA 1562-SINOPSYS, EUREKA 2063-IMPACT, STWW), the Belgian State, Prime Minister's Office-OSTC (IUAP P4-02 (1997-2001) and IUAP P4-24 (1997-2001); Sustainable Mobility Programme-Project MD/01/24 (1997-2000) "Traffic Congestion Problems in Belgium: Mathematical Models, Simulation, Control and Actions"), and the European Commission (TMR Networks: ALAPEDES and System Identification; Brite/Euram Thematic Network: NICONET).