Development and Application of an Evaluation Tool for Data Quality Assessment of Different Sensor Systems—The Case of Munich City

Master‘s Thesis of Yujie Xia

Mentoring:
M.Sc. Mario Ilic
M.Sc. Alexander Kutsch

Introduction

The quality of traffic data has a decisive influence on data management and traffic control procedures, which in turn play a role in traffic safety and traffic efficiency. However, the data quality varies depending on the detection techniques or sensor installation. It can be negatively affected by aging equipment, obstructions to sensors, bad environmental conditions, etc. A reliable statement about the quality of collected traffic data from different sensor systems is, therefore, an essential requirement for many traffic engineering applications.

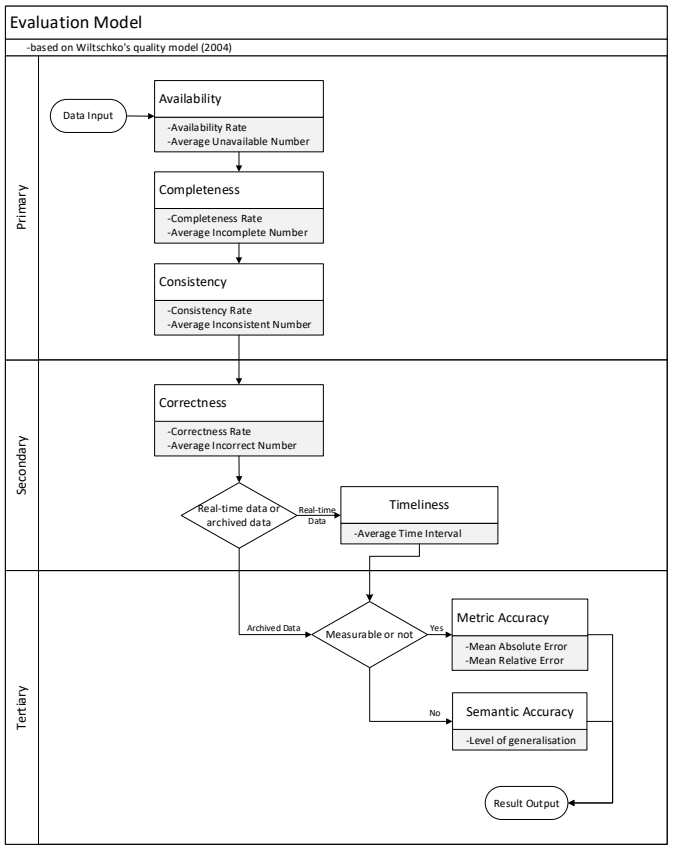


Fig.1 Procedures of the quality evaluation model

Methodology

The thesis adapted a information quality model proposed by Wiltschko in 2004 to a quality evaluation model of traffic data (Fig.1), and based on this developed and applied a quality evaluation tool for loop detector data and floating car data. The procedures of the quality evaluation model were defined to assess the quality of traffic data by quality features, which include availability, completeness, correctness, correctness, timeliness, metric accuracy, or semantic accuracy. The quality parameters were defined accordingly to describe each feature.

Tool Development

The tool was developed in R; the evaluation procedures for loop detector data assess all quality features except correctness, and the parameters for measuring metric accuracy were set as mean absolute error and mean relative error during daytime and nighttime respectively, based on the occupancy-flow model; for floating car data the tool assess correctness and metric accuracy, with accuracy measured by high score rate, while providing a list of the number of low scores for each link.

Tool Application

The thesis applied the developed tool to evaluate the quality of loop detector data and floating car data collected within a certain period, with Munich city as the study area. The tool outputs showed that the loop detector data had good completeness and consistency, while in terms of availability a certain amount of data was lost every month. As for correctness and metric accuracy, there were also a certain amount of inaccurate or implausible data records, but the proportion was small compared to the total amount. The problems with data quality indicated some errors of detectors; those errors were classified and detectors with each type were identified and listed (Fig.2). For floating car data, it was found that real-time data were insufficient on certain road segments and especially at night, so these missing values were replaced and filled with the historical data, showing a low score.

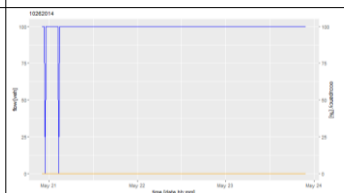
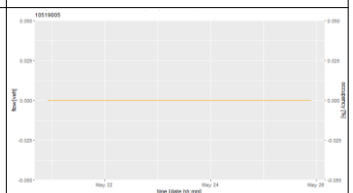
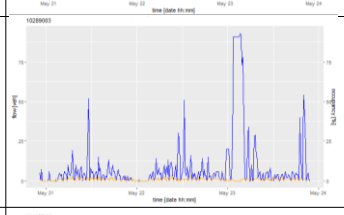
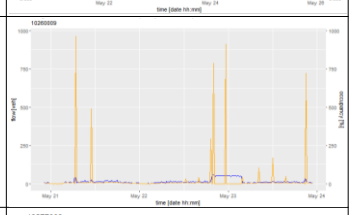
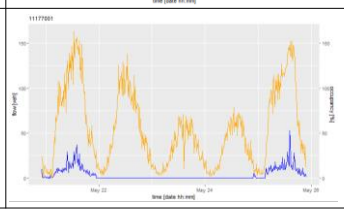
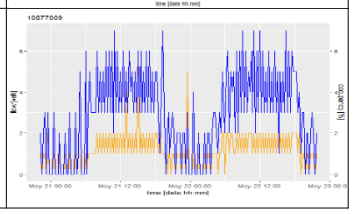
Error types	Time series diagram	Error types	Time series diagram
1. flow=0, occupancy= 100		4. mostly flow=0, occupancy=0	
2. flow=0, 0<occupancy<100		5. flow higher than maximum threshold	
3. flow>0, occupancy= 0		6. constant or fixed values	

Fig.2 Types of error and diagrams of error detectors after tool application

Tool Assessment

With the help of Tool Plus for visualization, by comparing the results of manual checks with the output of the tool, the tool presented a low false alarm rate, while a high detection rate can be achieved with an additional algorithm.

Limitations and Future Work

In terms of the quality evaluation model, other types of traffic data and more quality features should be considered, quality parameters definitions for floating car data should be refined.

Concerning the evaluation tool, an overall assessment procedure and optional ground truth data input and comparison can later be embedded. Evaluation of availability, consistency, and correctness should be implemented for the floating car data. One error type (constant zero occupancy and normal flow) cannot be detected by current plausibility checks; output of the calculated mean errors is currently implemented by an additional algorithm not embedded in the tool; issues concerning infinite or null values caused by zero values of occupancy as denominators are neglected in this tool and should be addressed.