QUATRA

Interim report

Deliverable Nr D 4.1

November 2012

nast consulting

TRANSVER

...

This project was initiated by ERA-NET ROAD.

Project Nr. 832570

Project acronym: QUATRA

Project title:

Software and Services for the Quality Management of Traffic Data

Deliverable Nr D 4.1 - Interim report/Report WP 3 - 4

Due date of deliverable: 01.10.2012

Actual submission date: 16.11.2012

Start date of project: 01.10.2011 End date of project: 30.09.2013

Author(s) this deliverable:

DI Daniel Elias, nast consulting ZT GmbH, Austria

DI Birgit Nadler, nast consulting ZT GmbH, Austria

Dr. DI Friedrich Nadler, nast consulting ZT GmbH, Austria

DI Stefan Gürtler, TRANSVER GmbH, Germany

DI Thomas Heinrich, TRANSVER GmbH, Germany

DI Alexander Kowarik, Technical University of Vienna, Austria

Mag. Bernhard Meindl, Technical University of Vienna, Austria

Dr. DI Matthias Templ, Technical University of Vienna, Austria

Version: draft 01

Executive summary

Main objective of the interim report is the provision of an update on the progress of work packages WP3 "Model Development" and WP4 "Software Development".

Based on the input from the state-of-the-art analysis in WP2 relevant criteria and indicators have been defined for definition of the framework of the software tool and the development of the model. A criteria catalogue (milestone M3.1) for the modelling tool for the freeway application and the urban environment was created as well as the basic model for statistical analyses that can be applied to freeway and urban traffic detection sites. The aim is to identify erroneous data based on statistical estimations. The model can be used for analysis of different parameters such as traffic volumes, traffic densities and average vehicle speeds.

Furthermore local/global/plausibility indicators have been developed that allow data evaluation and detection of inconsistencies. The erroneous data will be flagged and furthermore analysed in order differentiate between detector malfunctions and abnormal road conditions. In addition the project team will try to provide data imputation for erroneous and missing values based on historical and/or actual information.

The primary model development was scheduled for the period of December 2011 until August 2012 under the assumption that test data would be available since beginning of 2012. Due to the late submission of the required test data (June / End of October 2012) the model development is still on going. The first algorithm is currently being tested with German and Austrian motorway data. Due to the late supply of test data that is required for the model development the submission of the interim report was postponed from 01.10.2012 to 16.11.2012.

During the software development the backbone of the software system is being established which will handle the main functions for traffic data quality checks. Main works include the visual appearance of the prototype software, the import of different data formats and file types and a dynamic tool for the visual inspection of the traffic data.

Furthermore intensive contact has been established with road authorities in Austria and Germany. Workshops were held with representatives to define relevant use cases for road authorities and to identify requirements from their perspective.

List of Tables

List of Figures

Table of content

1 Introduction

“ERA-NET ROAD – Coordination and Implementation of Road Research in Europe” was a Coordination Action funded by the 6th Framework Programme of the EC. The partners in ERA-NET ROAD (ENR) were United Kingdom, Finland, Netherlands, Sweden, Germany, Norway, Switzerland, Austria, Poland, Slovenia and Denmark (www.road-era.net). Within the framework of ENR this joint research project was initiated. The funding National Road Administrations (NRA) in this joint research project are Belgium, Switzerland, Germany, Netherlands, Norway and United Kingdom.

Traffic management systems are using traffic data from several different data collection sources for purposes such as visualisation of traffic situation, detection of abnormal road conditions or the generation of appropriate traffic control decisions. Therefore the reliability and plausibility of traffic data needs to be identified and confirmed, faulty data needs to be detected immediately. The reason for these quality assurance measures laid within the quality of traffic control systems itself. Effective control decisions are strongly dependent on the correctness of the underlying traffic data collection.

2 Objective of the work packages WP3 and WP4

Main objective of the work package WP3 "Model Development" is the definition of criteria and indicators for the traffic data assessment which will be carried out with the statistical modelling tool and local/global/plausibility indicators. A basic model is needed that is capable of analysing the data quality of freeway and urban road traffic detector sites.

Main objective of the work package WP4 "Software Development" is the development and implementation of a concept for a software system platform and the corresponding service for data processing and data quality analysis. The two different tools (online-freeway-tool and urban-offline-tool) will provide all the information that is required for manual and automated processing and analysis of traffic data.

3 Development of the strategy for traffic data assessment

Based on the input from the state-of-the-art analysis in WP2 relevant criteria and indicators have been defined for definition of the framework of the software tool and the development of the model. Furthermore the task included the consideration of the capability of various criteria, potential implementation strategies and required data sources for the individual assessment. The analysis and definition of these tasks was carried out during a two stage process. In the first stage the project partners evaluated the results of the literature review for application within the freeways and urban road environment separately. In the second step the interpretations and findings of both project partners were shared and discussed at a round table and a joined approach for definition of quality checks and model development was defined.

The analysis and selection of methods for implementation in QUATRA is strongly depending on available data sources and access points at different hierarchical data exchange levels.

Furthermore one has to differentiate whether the detector itself is sending the wrong information due to physical failure, wrong installation or if the data is abnormal in comparison to historical situations due to for example incidents or road works - for details please refer to chapter 3.1.

The project team has decided to combine local/global/plausibility indicators and statistical model analyses for the evaluation of the freeway and urban traffic data. The local/global/plausibility indicators are based on combinations of different values of the traffic data sets as well as the analysis of neighbouring traffic detection sites. Thus the conservation of traffic flow can be incorporated in the traffic data analysis. Apart from the total number of vehicles balances of single vehicle categories can also be assessed. Details for the local/global/plausibility indicators can be retrieved from the tables 3 to 5.

The statistical model approach is based on the assessment of historical data and prediction of confidence intervals for current road data ranges as well as major anomalies of the data in comparison to standardized statistical parameters. The statistical model is described in detail in chapter 4.2.

The following tables 1 to 7 include detailed information about the various approaches that were identified during the "state-of-the-art" analysis, their relevance for QUATRA, the suitability for assessment of freeway and urban traffic data and potential data sources and their hierarchical level.

Table 1: Strategy development for QUATRA data analysis (1)

item	criteria and indicators	source	description	0 - not suitable, 1 - suitable, 2 - partially suitable			stationary and local data (q,v,k,s, occ...)		global data
				freeway	urban	temp hard shoulder	TLS data sources (loops/radar/TEU)	toll system data, bluetooth, ANPR, DECELL, INRIX	
1.	comparison of volumes and volume/occupancy ratio based on a 20-second intervals	Nihan Wang (1995)	volumes vs. detector utilisation - example: low traffic volumes should be represented in a low detector utilisation rate	1	0	1			
2.	vehicles change lanes physically at the location of the detection (splashover)	Coifman Lee (2011)	detectors on adjacent traffic lanes detect the same vehicle in case of lane change - could be filtered out (IF speed, time, vehicle is same)	2	1	1			
3.	same vehicles are detected twice at the same site (e.g. heavy vehicles with trailers)		testing of time gap at high speed detection, logical enquiry included in item M10	1	0	1			
4.	average vehicle lengths calculated and vehicle distributions estimated and compared with historical data	Turochy Smith (2000)	percentage of vehicle categories on total traffic ok or wrong (differences during time/day/location)	1	1	1	1	0	0
5.	on-times were assessed (ratio of vehicle lengths over speed), vehicle distributions were logically compared with similar vehicle distributions and the average ontime for a time interval e respectively to identify faulty data.	Coifman Lee (2006) Chen May (1987)							
6.	storage rates for time intervals	Nihan (1997)	can data be obtained about storage rates from single vehicle detections? Could be used as indicator for traffic queuing, logical enquiry included in item M19	1	1	1			
7.	minimum and maximum flow thresholds	Weijermars Berkum (2006)	included in items M12-M14						

Source: nast consulting, TRANSVER

Table 2: Strategy development for QUATRA data analysis (2)

item	criteria and indicators	source	description	0 - not suitable, 1 - suitable, 2 - partially suitable			stationary and local data (q,v,k,s, occ...)		global data
				freeway	urban	temp hard shoulder	TLS data sources (loops/radar/TEU)	toll system data, bluetooth, ANPR, DECELL, INRIX	
8.	linear regression of the volumes of neighbouring sites	Chen et al. (2003)	for freeway section this criteria will be incorporated in the statistical model, for the urban road environment the density of traffic detection sites is highly relevant, high coverage is needed in order to get good results (partially included in item M20)	1	2	1	1	1	0
9.	measured flow values on neighbouring data collection sites	Hoops (2002)							
10.	observation of unexpected high jumps of speed level and speed jumps		evaluation of average speeds in comparison to traffic volumes, indicator for traffic queuing	1	1	1	1	0	0
11.	If the speed is unexpected small, it is tested whether the speed is due to a traffic incident or measurement error occurred. The assumption was made that traffic incident in contrast to a measurement error is usually observed on all lanes.			1	0	1	1	0	0
12.	Travel times from v and q		indicator for traffic queuing, can only be obtained in Austria through toll system data, for urban regionss maybe Decell or INRIX data is available	1	2	1	1	0	0
13.	Under the assumption that the travel time of a considered and specified driver group stays stable, the travel time, calculated from the distance between the detectors and the measured speed, has to be similar to the travel time found by correlation analysis.		indicator for traffic queuing, can only be obtained in Austria through toll system data, for urban regionss maybe Decell or INRIX data is available	1	2	1	1	2	
14.	conservation of flow: the test takes time series of different detection sites into account and calculates the quotient of the traffic volumes, estimates the distribution of the quotients and establishes the decision rules for data validation	nast consulting, TU Vienna (2008)	included in statistical model	1	2	1	1	1	0

Source: nast consulting, TRANSVER

Table 3: Strategy development for QUATRA data analysis (3)

item	criteria and indicators	source	description	0 - not suitable, 1 - suitable, 2 - partially suitable			stationary and local data (q,v,k,s, occ...)		global data
				freeway	urban	temp hard shoulder	TLS data sources (loops/radar/TEU)	toll system data, bluetooth, ANPR, DECELL, INRIX	
15.	value occupancy and flow minimum and maximum thresholds	Freudenberger (2001)	included in items M12-14 and M19	1	1	1			
16.	The speed was calculated from q and o due to the correlation of the fundamental diagram. If the speed exceeded the speed limit at the specific detector, the dataset was marked as implausible		logical combination of parameters of fundemtal diagram (can only be evaluated of individual vehicle speeds are available for analysis)	1	1	1	1	0	0
17.	The data was checked regarding recurrence		check, if single vehicles can be detected (data issue)	1	1	1	1	1	1
18.	regression curve second order was placed in a fundamental diagram. Measurement data with a significant high distance to the regression curve were marked as suspect		included in statistical model	1	1	1	1	0	0
19.	The coefficient of determination described how good the scatter plot was represented by the regression curve. If the coefficient of determination was small, there was only a small correlation between o and q und the detector was treated is possibly defect.			1	1	1	1	1	0
20.	conservation of flow: comparison of number of vehicles at neughbouring traffic sites - under consideration of on- and off ramps	nast consulting TRANSVER	included in statistical model	1	2	1	1	1	1
21.	conservation of flow: comparison of number of heavy vehicles at neughbouring traffic sites - under consideration of on- and off ramps		included in statistical model	1	0	1	1	1	0
22.	M1: QKfz = 0 Þ (QLkw = 0 UND QPkw = 0)		total traffic must be the sum of individual vehicle categories. If no vehicles total were counted during time period no car or HV can be registered	1	1	1	1	0	0

Source: nast consulting, TRANSVER

Table 4: Strategy development for QUATRA data analysis (4)

				0 - not suitable, 1 - suitable, 2 - partially suitable			stationary and local data (q,v,k,s, occ...)		global data
item	criteria and indicators	source	description	freeway	urban	temp hard shoulder	TLS data sources (loops/radar/TEU)	toll system data, bluetooth, ANPR, DECELL, INRIX	
23.	M2: QKfz – QLkw = 0 ⇒ (QPkw = 0 UND VPkw = 255)	nast consulting TRANSVER	example: if total traffic = total vehicles category 1 there can't be any amount of vehicles for category 2... - furthermore speed of category 2... must be 0	1	1	1	1	0	0
24.	M3: QLkw = 0 ⇒ VLkw = 255		if no HV has been counted average speed must be 0 or code 255	1	1	1	1	0	0
25.	M4: QPkw = 0 ⇒ VPkw = 255		if no car has been counted average speed must be 0 or code 255	1	1	1	1	0	0
26.	M5: QKfz ≥ QLkw		total traffic must be sum of all vehicles categories	1	1	1	1	0	0
27.	M6: QKfz – QLkw > 0 ⇒ 0 < VPkw		if cars have been counted average car speed must increase	1	1	1	1	0	0
28.	M7: QKfz > 0 ⇒ 0 < VKfz		if vehicles have been counted average speed must be present	1	1	1	1	0	0
29.	M8: QLkw > 0 ⇒ 0 < VLkw		if vehicles have been counted average speed must be present	1	1	1	1	0	0
30.	M9: 0 < t < T		covers item 1 - detector utilization time must be higher 0 and shorter than time interval	1	2	1	1	0	0
31.	M10: QKfz = 0 ⇒ 0 < Vg,Kfz(t) = Vg,Kfz(t-T)		if no vehicle has been counted during time interval, averaged vehicle time must be higher 0 and same as previous time period	1	2	1	1	0	0
32.	M11: VKfz > VGrenz ⇒ B < Bgrenz		If average vehicle speed is high during a time interval the detector occupancy rate has to be below a certain treshold (fundamental diagram)	1	2	1	1	0	0
33.	M12: QKfz,min ≤ QKfz ≤ QKfz,max		traffic volumes during a certain time have to be within a certain range - otherwise there is a disturbance - refer to item 7	1	1	1	1	0	0
34.	M13: QPkw,min ≤ QPkw ≤ QPkw,max		car volumes during a certain time have to be within a certain range - otherwise there is a disturbance - refer to item 7	1	1	1	1	0	0
35.	M14: QLkw,min ≤ QLkw ≤ QLkw,max		HV volumes during a certain time have to be within a certain range - otherwise there is a disturbance - refer to item 7	1	1	1	1	0	0
36.	M15: VKfz,min ≤ VKfz ≤ VKfz,max		average vehicle speed during a certain time has to be within a certain range - otherwise there is a disturbance - refer to item 7	1	1	1	1	0	0

Source: nast consulting, TRANSVER

Table 5: Strategy development for QUATRA data analysis (5)

item	criteria and indicators	source	description	0 - not suitable, 1 - suitable, 2 - partially suitable			stationary and local data (q,v,k,s, occ...)		global data
				freeway	urban	temp hard shoulder	TLS data sources (loops/radar/TEU)	toll system data, bluetooth, ANPR, DECELL, INRIX	
37.	M16: VLkw,min ≤ VLkw ≤ VLkw,max	nast consulting TRANSVER	average HV speed during a certain time has to be within a certain range - otherwise there is a disturbance - refer to item 7	1	1	1	1	0	0
38.	M17: VPkw,min ≤ VPkw ≤ VPkw,max		average car speed during a certain time has to be within a certain range - otherwise there is a disturbance - refer to item 7	1	1	1	1	0	0
39.	M18: Vg,Kfz,min ≤ Vg,Kfz ≤ Vg,Kfz,max		smoothed vehicle speed from during a certain time has to be within a certain range - otherwise there is a disturbance	1	2	1	1	0	0
40.	M19: Bmin ≤ B ≤ Bmax		average car speed during a certain time has to be within a certain range - otherwise there is a disturbance - refer to item 6	1	2	1	1	0	0
41.	M20: VPkw,links > VPkw,rechts		Germany: average car speed in right freeway lane should be below average car speed in left freeway lane (due to driver behaviour this assumption can not be used for Austrian motorways and urban areas)	1	2	1	1	0	0
42.	M21: VAusfahrt < VAusfahrt,grenz		average vehicle speed at on-/off ramps during a certain time has to be within a certain range - otherwise there is a disturbance	1	0	0	1	0	0
43.	M22: QLkw,rechts > QLkw,links		HV volume in left freeway lane should be below HV volume in left freeway lane (problem during overtaking manouvres)	2	2	1	1	0	0
44.	Indication of sources of errors based on previous events of the same type that were saved in the database	DAVINCI	should be included in a follow up project	0	0	0	-	-	-
45.	Peak-Hour-lane occupancy: erroneous data outside of operating hours		outside of operating hours the number of detected vehicles should be below a certrain threshold	1	0	1	1	0	0
46.	Acceptance monitor of speeds (checking whether the traffic control strategy, which was shown on the road, was accepted by the drivers)	BUSCH	not part of core focus of quatra, could be used for evaluation of compliance to dynamic signage	0	0	0	-	-	-

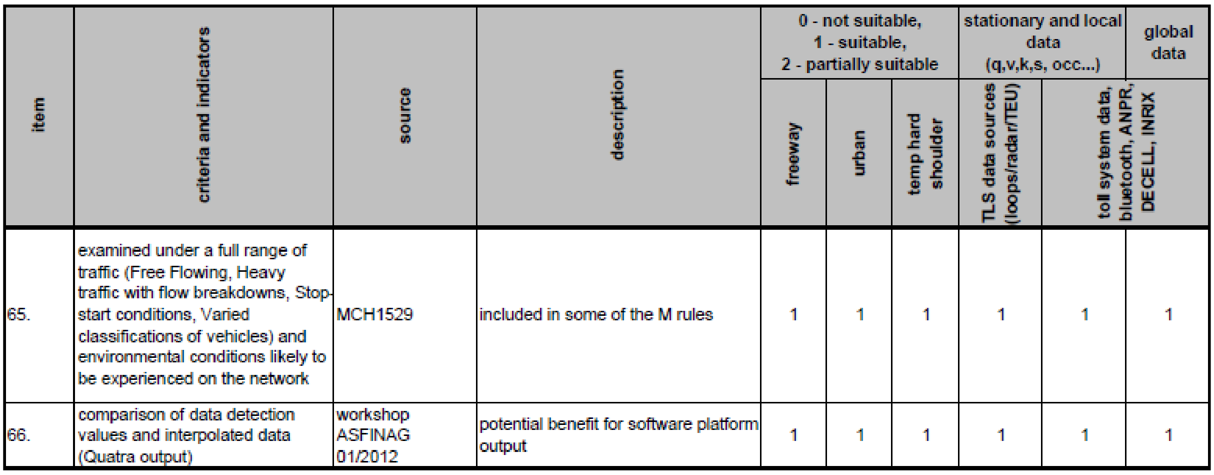
Source: nast consulting, TRANSVER

Table 6: Strategy development for QUATRA data analysis (6)

item	criteria and indicators	source	description	0 - not suitable, 1 - suitable, 2 - partially suitable			stationary and local data (q,v,k,s, occ...)		global data
				freeway	urban	temp hard shoulder	TLS data sources (loops/radar/TEU)	toll system data, bluetooth,	ANPR, DECELL, INRIX
47.	For the current project the possible fault scenarios and sources of error of the individual ways of data collecting described in the reference note are of interest.	FGSV AK 3.5.20	wrong number of vehicles or no vehicle recorded, wrong classification of vehicles, wrong speed, repetition of interval data, included in logical enquiries M	1	1	1			
48.	repetition of interval data with same time stamp	-	indicator for data error	1	1	1	1	1	0
49.	comparison of volumes and average speeds	DVS Netz pilot nast consulting	according to results of research project high chance of queuing in case of high volumes and low average speeds	1	1	1	1	0	0
50.	comparison of volumes and average speeds	DVS Netz pilot nast consulting	in case of high traffic density average speed can't be high due to fundamental diagram	1	1	1	1	0	0
51.	Highway Agencies' requirements for assessment of data quality (target parameter, degree of accuracy)	MCH1529	potential benefit for software platform	1	0	1	1	1	1
52.	fault monitoring/reporting		potential benefit for software platform	1	1	1	-	-	-
53.	Plausibility check of short-term data, differentiation between inflowing and outflowing traffic	FGSV AK 3.5.20	covered in item 16	2	1	0	-	-	-
54.	data quality requirements regarding completeness, availability, veracity, precision, timeliness	QUANTIS	included in some of the M rules				-	-	-
55.	missing data (quantities)	nast consulting TRANSVER	potential benefit for software platform output	1	1	1	1	1	1
56.	missing data (percentage)		potential benefit for software platform output	1	1	1	1	1	1
57.	missing data code 255 (quantities)		potential benefit for software platform output	1	1	1	1	0	0
58.	missing data code 255 (percentage)		potential benefit for software platform output	1	1	1	1	0	0
59.	missing data code 0 (quantities)		potential benefit for software platform output	1	1	1	1	1	1
60.	missing data code 0 (percentage)		potential benefit for software platform output	1	1	1	1	1	1
61.	flagged data sets (resulting out of quality checks - quantities)		potential benefit for software platform output	1	1	1	1	1	1
62.	flagged data sets (resulting out of quality checks - percentage)		potential benefit for software platform output	1	1	1	1	1	1
63.	monitoring levels	TRAFFIC IQ	potential benefit for software platform	1	1	1			
64.	reference systems and data fusion		potential benefit for software platform	1	1	0	1	1	1

Source: nast consulting, TRANSVER

Table 7: Strategy development for QUATRA data analysis (7)



item	criteria and indicators	source	description	0 - not suitable, 1 - suitable, 2 - partially suitable			stationary and local data (q,v,k,s, occ...)		global data
				freeway	urban	temp hard shoulder	TLS data sources (loops/radar/TEU)	toll system data, bluetooth, ANPR, DECELL, INRIX	
65.	examined under a full range of traffic (Free Flowing, Heavy traffic with flow breakdowns, Stop-start conditions, Varied classifications of vehicles) and environmental conditions likely to be experienced on the network	MCH1529	included in some of the M rules	1	1	1	1	1	1
66.	comparison of data detection values and interpolated data (Quatra output)	workshop ASFINAG 01/2012	potential benefit for software platform output	1	1	1	1	1	1

Source: nast consulting, TRANSVER

3.1 *Detector error messaging and abnormal road conditions*

In relation to the detection of data anomalies one has to differentiate whether detectors themselves send wrong information due to physical failure, wrong detector installation and calibration or if the data is abnormal compared to historical traffic conditions due to incidents or road works.

Detector error messaging

Regarding physical or data exchange failures detectors automatically submit error values for different variables. The error reporting varies depending on the TLS[a] used at the detection site.

TLS new standard

q=0 && v=255 → detector reports that no vehicle was detected during the interval

q=0 && v=0 → detector reports disruption

(q - traffic volume, v - average vehicle speed)

TLS old standard

q=0 && v=0 → detector reports disruption OR detector reports that no vehicle was detected during the interval

In order to identify detector disruptions time variation curves need to be assessed to gain information about q=0 && v=0 during a longer time period.

[a] TLS: technical supply conditions used for road stations, used in Austria and Germany

Wrong detector installation and calibration

In order to find out erroneous data based on wrong detector installation and calibration the data of neighbouring traffic detection sites can be compared. Under the assumption of conservation of traffic the balances of these detection sites - for example total number of vehicles, single vehicle categories - can be assessed. Once certain thresholds are exceeded the operator or maintenance crew can be informed about potential installation problems. Of course the conservation of flow also needs to cover traffic that exiting and entering the road network via ramps between neighbouring traffic detection sites.

Abnormal road conditions

Apart from physical disruptions due to the identified sources above detectors can only identify vehicles that are driving through the specified detection field. Once vehicles are passing by outside of the main detection range only a certain percentage of these vehicles can be registered and the vehicle category is likely to be faulty with a high percentage.

Based on the analysis of time variation curves of historical traffic data typical road conditions can be specified according to parameters such as weekday, hour, location. Consequently the current road condition under the assumption of normal traffic flow can be predicted and compared with the observed number of vehicles.

Once the comparison of predicted and observed numbers of vehicles shows high variations the data record (for example one mite interval) can be flagged as abnormal and an automated matching with for example TMC messages and road works databases can be started. If the matching process is successful an automated or manual process can be started that helps to decide whether the flagged data is accurate or needs to be substituted with estimated data from neighbouring traffic sites or historical traffic data. In case no abnormal road condition is identified automatically operators can check via video surveillance or on-site evaluation if the detector is reporting erroneous data.

3.2 Available traffic data sources

For the development of the freeways modelling and software tool data from the following test sections are used:

- Austria: sections on S 1 (city motorway in Vienna) and A 12 (rural motorway in Tyrol)
- Germany: sections on A 8, A 9 and A 99 (motorways in the area of Munich)
  (including a section with temporary hard-shoulder use)

For the development of the urban modelling and software tool data from the following test sections are used:

- Austria: City of Vienna
- Germany: City of Bremen

For the motorway test sections in Germany required offline test data for model and software development was submitted in June 2012. By the end of October 2012 the project team also received input data by ASFINAG for the Austrian motorway test sections A12 and S1.

Concerning the urban traffic data initial limited test data sets for the city of Bremen were submitted by German authorities. For the city of Vienna so far no data could be submitted.

Furthermore test sections in Switzerland were considered in the beginning of the project. So far no test sections could be confirmed. The project team leader and PEB Bruno Mariéthoz tried to contact the representative of ASTRA (Federal Roads Office in Switzerland) several times without success. Therefore works will concentrate on the test sections in Austria and Germany.

4 Development of the modelling tools

The assessment of freeway and urban traffic data is done through combination of local/global/plausibility indicators and statistical model analyses.

4.1 Local/Global/Plausibility indicators

The indicators are based on combinations of different values of the traffic data sets as well as the analysis of neighbouring traffic detection sites and principles of conservation of flow; and are as follows:

Local indicators

- Missing data: number or ratio of missing data sets. Normally, each detector should deliver one data set per minute respectively 1.440 data sets per day - otherwise there is a disturbance
- Failure messages from detector: number or ratio of data sets with failure message “255, 255” generated by the detector itself

Global indicators

- Conservation of flow for cars: comparison (ratio) of number of cars at neighbouring measurement cross sections under consideration of inflow and outflow at ramps. The ratio should be 1 - otherwise there is a disturbance
- Conservation of flow for heavy vehicles (HV): comparison (ratio) of number of HV at neighbouring measurement cross sections under consideration of inflow and outflow at ramps. The ratio should be 1 - otherwise there is a disturbance

Plausibility indicators

M1: $QKfz = 0 \Rightarrow (QLkw = 0 \text{ UND } QPkw = 0)$

total traffic must be the sum of individual vehicle categories. If no vehicles total were counted during time period no car or HV can be registered

M2: $QKfz – QLkw = 0 \Rightarrow (QPkw = 0 \text{ UND } VPkw = 255)$

if total traffic = total vehicles category 1 there can't be any amount of vehicles for category 2... - furthermore speed of category 2... must be 0

M3: $QLkw = 0 \Rightarrow VLkw = 255$

if no HV has been counted average speed must be 0 or code 255

M4: $QPkw = 0 \Rightarrow VPkw = 255$

if no car has been counted average speed must be 0 or code 255

M5: $QKfz \geq QLkw$

total traffic must be sum of all vehicles categories

M6: $QKfz – QLkw > 0 \Rightarrow 0 < VPkw$

if cars have been counted average car speed must increase

M7: $QKfz > 0 \Rightarrow 0 < VKfz$

if vehicles have been counted average speed must be present

M8: $QLkw > 0 \Rightarrow 0 < VLkw$

if vehicles have been counted average speed must be present

M9: $0 < t < T$

covers item 1 - detector utilization time must be higher 0 and shorter than time interval

M10: $QKfz = 0 \Rightarrow 0 < Vg,Kfz(t) = Vg,Kfz(t-T)$

if no vehicle has been counted during time interval, averaged vehicle time must be higher 0 and same as previous time period

M11: $VKfz > VGrenz \Rightarrow B < Bgrenz$

If average vehicle speed is high during a time interval the detector occupancy rate has to be below a certain treshold (fundamental diagram)

M12: $QKfz,min \leq QKfz \leq QKfz,max$

traffic volumes during a certain time have to be within a certain range - otherwise there is a disturbance - refer to item 7

M13: QPkw,min ≤ QPkw ≤ QPkw,max

car volumes during a certain time have to be within a certain range - otherwise there is a disturbance - refer to item 7

M14: QLkw,min ≤ QLkw ≤ QLkw,max

HV volumes during a certain time have to be within a certain range - otherwise there is a disturbance

M15: VKfz,min ≤ VKfz ≤ VKfz,max

average vehicle speed during a certain time has to be within a certain range - otherwise there is a disturbance

M16: VLkw,min ≤ VLkw ≤ VLkw,max

average HV speed during a certain time has to be within a certain range - otherwise there is a disturbance - refer to item 7

M17: VPkw,min ≤ VPkw ≤ VPkw,max

average car speed during a certain time has to be within a certain range - otherwise there is a disturbance - refer to item 7

M18: Vg,Kfz,min ≤ Vg,Kfz ≤ Vg,Kfz,max

smoothed vehicle speed from during a certain time has to be within a certain range - otherwise there is a disturbance

M19: Bmin ≤ B ≤ Bmax

average car speed during a certain time has to be within a certain range - otherwise there is a disturbance

M20: VPkw,links > VPkw,rechts

Germany: average car speed in right freeway lane should be below average car speed in left freeway lane

Austria: due to driver behaviour this assumption can not be used for Austrian motorways and urban areas

M21: VAusfahrt < VAusfahrt,grenz

average vehicle speed at on-/off ramps during a certain time has to be within a certain range - otherwise there is a disturbance

M22: QLkw,rechts > QLkw,links

HV volume in left freeway lane should be below HV volume in left freeway lane (problem during overtaking manouvres)

4.2 Statistical model

The data analysis with a statistical model represents an additional rule for the assessment of the traffic data based on statistical estimation. In addition, erroneous values can also be imputed as well as missing values in the data from historical and/or actual information.

Screening rules, often called *edit rules* are often used to determine whether an observation is consistent or not. An example of an logical (balanced) edit is

$$qKfz = qPkw + qLkw$$

where qKfz, qPkw and qLkw is the number of all vehicles, of cars and of heavy vehicles in a given time interval, e.g. measured by detectors on a freeway. The edit above expresses that the amount of trucks and cars should sum up to the total number of vehicles. Such an edit is referred to as a balanced edit.

A non-negative edit is just defined that a value has to be non-negative if it passes the edit. For example, the speed of a vehicle should not be negative.

If a ratio of two variables should be less (or greater) than a certain threshold, then this edit is referred as ratio edit. For example, vKfz2/vKfz1 > 0.8, which means that the speed the second lane at a freeway should be not less than 20% of the speed of the first lane.

In order to construct a set of edits one usually starts with the hard (or logical) edits, which hold true for all correctly observed values. Balance edits are usually referred as hard edits. Hard edits are specified by subject matter specialists. This is also the case within this project, whereas dozen of balanced edits were formulated. During statistical analyses soft edits are set which hold true for a high fraction of correctly observed records but not necessarily for all of them.

Ratio edits can be either hard or soft edits. The threshold related to ratio edits have to be determined carefully, so that on the one hand only few values may violate the edit and that on the other hand erroneous values are detected by these edits. This threshold is either fixed by a subject matter specialist (hard edit) or may vary depending on the input data (soft edit).

To avoid over-editing one should in particular be careful not to specify too many soft edits. In general, users tend to apply more soft edits than necessary to the data.

In former projects - but with the following reasons not applied in this project - possible erroneous values were detected by parametric modelling. The number of cars and the length between the cars were considered as a realization of a Poisson distribution. Theoretical properties of the underlying distributions may be used to define malfunctions. In this case, empirical historic data are modelled by a Poisson distribution. From the theoretical properties of this distribution, a threshold was used to define if an value is suspicious.

Although this method can also be applied on subsets of the data, it is a univariate method that does not consider information on covariates. Therefore, the method is of limited use and not considered in this work. Another problem that is present in traffic data is the missing data problem. Due to malfunction or transmission errors, missing values are introduced.

The imputation of missing values is especially important for traffic data, This has especially consequences for statistical methods using the multivariate data information. The naive approach, namely omitting all observations that include at least one missing value, is not attractive because a lot of valuable information might still be contained in these observations. On the other hand, omitting observations may only lead to non-biased estimates when the missing data are missing completely at random.

The estimation of the missing cells can even introduce additional bias depending on the method used. Valid estimations and inferences can mostly only be made if the missing data are at least missing at random. Not only missing values in the data but also values that violate the edits have to be replaced by reasonable substitutes.

In order to create a statistical model approach the data needs to visualised first in order to gain a better insight about possibly hidden data-structures, relationships and errors. Although these results might be well known by specialists every new data set should be first analyzed by explorative methods.

Boxplots of the number of vehicles (variable qKfz) and the mean speed of vehicles (variable vKFZ) aggregated at hours of the day provide useful information and (robust) key statistics: the box contains the inner 50% of the observations, the line in the box is the median that splits the sorted data by half. Finally, the so called whiskers lay at the last observation outside 1.5 · IQR (interquantile range, which contains the inner 50 % of the ordered observations) measured from the inner box. Values outside the whiskers may be considered as (univariate) outliers. Because of this visual summary statistics of the data - the boxplot, the distribution of the data and possible outliers can be easily seen.

In figure 1 it can be seen that on a typical rural freeway section during the day between 06:00 and 16:00, the amount of vehicles per minute is constantly high for all two lanes. However, during the night, the mean of the amount of Kfz is close to zero for lane 2, expect some values that are far away from the mean. Apart from the time of the day detector sites near on- and off ramps can also show a different picture due to the lane changing and bypassing manoeuvres.

The data values outside of the whiskers are the ones that are of high interest for the statistical analysis because some of them represent abnormal road conditions for the particular time during a day.

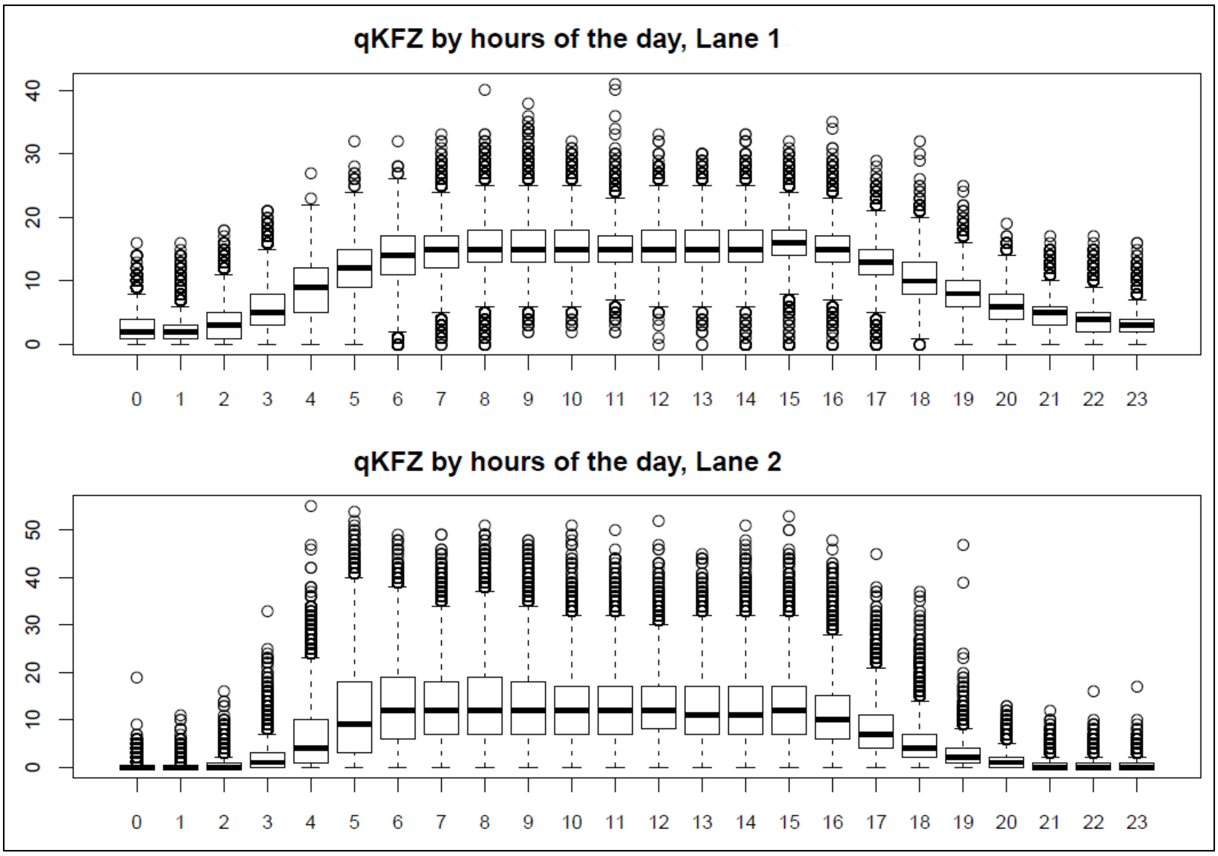


Figure 1: Boxplots of number of KFZ by hours of the day for a typical rural freeway section

Source: nast consulting, Technical University of Vienna

Detection of malfunction and measurement errors in one sector

Data that has been collected or measured generally includes errors due to a variety of reasons. In any case, statistical data editing methods, e.g. checks and corrections, are usually necessary to increase the quality of the available data and to be able to detect malfunctions of measurement units.

First, erroneous values in the data set have to be localized. It is preferable if this problem can be tackled in an automated manner. These localized erroneous values then have to be dealt with. One possibility is to replace faulty measurements by reasonable values using a suitable imputation procedure whenever complete data are needed for further analysis. It is usually not necessary to remove all errors from a data set. Doing automated micro-editing for small errors is often too ambitious and leads to over-editing. However, it is a necessarily to detect systematic errors from measurement units or malfunctions in measurement units because systematic errors do affect results of statistical data analysis.

Of course, a good property of any imputation method is that logical relationships in the data should be preserved. For example, in the case of traffic flow data, the sum of trucks and cars given a specific sector and lane should sum up to the corresponding total number of vehicles also after the imputation process.

The detection of outliers is very important in statistical analysis. Outliers can be considered as atypical observations which deviate from the usual data variability. Since classical statistical models applied to data including outliers can lead to misleading results. In addition to that, measurement errors may also have great influence on aggregates typically published in output tables.

Mahalanobis Distance and Cut-off values

Considering an *n* × *p* data set **X**, the usual measure used in this context is the Mahalanobis distance, a one-dimensional statistic measuring the distance of a data point from a location with respect to a shape. It is defined as

$$d_i = d(\mathbf{x}_i) = \sqrt{(\mathbf{x}_i - \mathbf{t})^T \mathbf{C}^{-1} (\mathbf{x}_i - \mathbf{t})}$$

for an observation **x***i*, *i* = 1, . . . , *n*, and the respective location and covariance estimates **t** = **t**(**X**) and **C** = **C**(**X**). Using the arithmetic sample mean as an estimator for the location **t** and the sample covariance matrix as an estimate for **C**, the resulting Mahalanobis distance is not robust since it depends on estimators which are extremely sensitive to outliers. It can easily be shown that the classical Mahalanobis distance can already be corrupted if the data contain only one single outlier.

If robust statistics such as the median as the location estimate t and a robust estimate for the shape parameter **C** are used, the resulting distance measure is referred to as the robust Mahalanobis distance. If the data is multivariate normally distributed, the squared Mahalanobis distances based on the classical mean and covariance estimates are approximately χ^2-distributed with p degrees of freedom. To classify the points of a data set as regular points or outliers, a cut-o_ value has to be specified, which in practice is usually a certain quantile of the respective distribution of the corresponding distances, e.g. the 97.5% quantile of the χ^2_p-distribution. Data points with distances larger than this threshold are then considered as potential outliers.

Assuming that for one sector, detectors are on two or three lanes installed, the two (or three) dimensional joint distribution have to be considered. Herby, no malfunction is detected when

(a) the classical correlation between observations obtained from different lanes should be large, and

(b) the percentage of outliers (as identified though the Mahalanobis distance) is small,

(c) a dependency on the day time is given, for example, the ratio (measure first lane / measure second lane) has a typical behaviour which is estimated by historical data.

For outlier detection in (b) robust statistical methods have to be applied, since classical methods are itself influenced by outliers. The exact parameters to define a set of rules based on (c) are estimated from historical data while (a-b) can be defined beforehand. The proposed procedures can be applied off- and online.

From the following figure it can be inferred, for example, that the detection and measurement of traffic volumes and average speeds show abnormal road conditions or faulty detection periods (see figure 2). The problem of detecting malfunctions can be viewed as a statistical testing problem. Observations with large multivariate distance the centre of the data are highlighted in red. Clearly, the observations in red in the lower middle and right part of the plot are such situations since the amount of vehicles during free flow conditions cannot be large while the mean speed is very low. The figure shows the potential benefit of the application of robust multivariate distances to detect malfunctions and abnormal road conditions. While the figure (and the related estimations) is only based on two-dimensional data to be able to visually show the concept, the method for detecting outliers is still the same for multivariate data greater equal three dimensions.

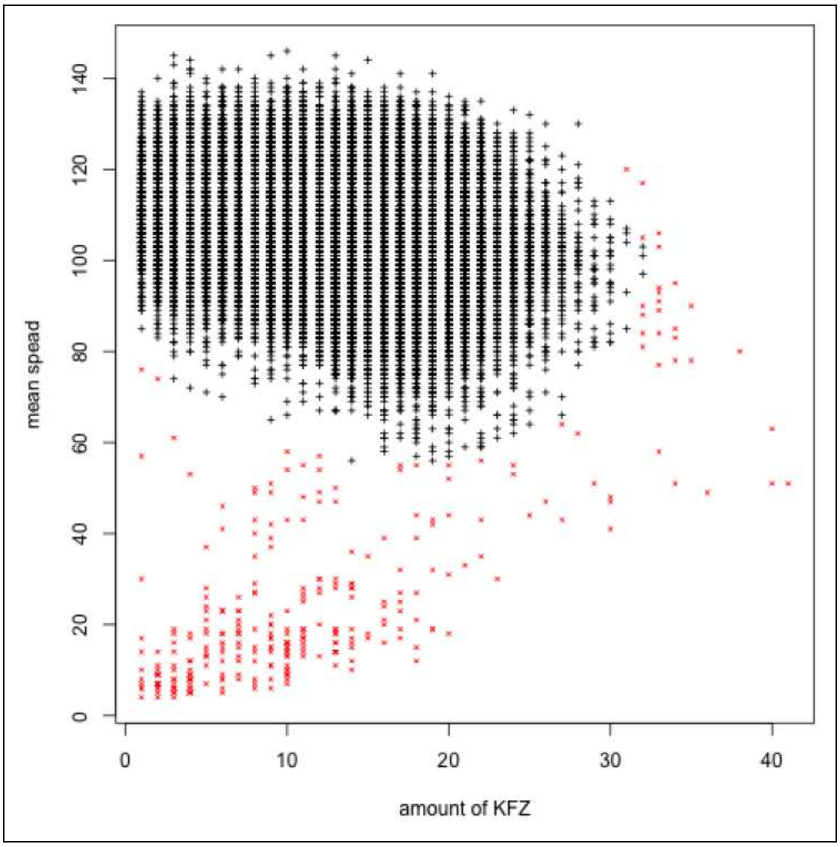


Figure 2: Scatterplot of the amount of vehicles (KFZ) versus the mean speed of vehicles

The colour code red represents observations with large (robust) multivariate distance based on 99.9% tolerance ellipses.

Source: nast consulting, Technical University of Vienna

In the figure above can also be seen that these techniques mathematically described in the previous section can be generally applied to multivariate data in order to take information on several covariates (e.g. day time, lane number, weekday,...) into account. In this case, spatial dependencies can also be considered which is hardly possible using multivariate methods that are based on (robust) distance measures.

Correlation between lanes

If one is interested in the ratios for example of the mean speed of vehicles between lanes at a given sector, one can use again measures based on multivariate distances to detect and identify suspicious ratios/observations. Figure 2 highlights those observations with large (robust) Mahalanobis distances. The observations in red are candidates for outliers and/or erroneous data or abnormal road conditions.

Assuming that for one sector, detectors are on two or three lanes installed, the two (or three) dimensional joint distribution has to be considered. Herby, no malfunction is detected when

(a) the classical correlation between observations obtained from different lanes should be large, and

(b) the percentage of outliers is small

A robust estimation for the correlation can be obtained with the MCD algorithm.

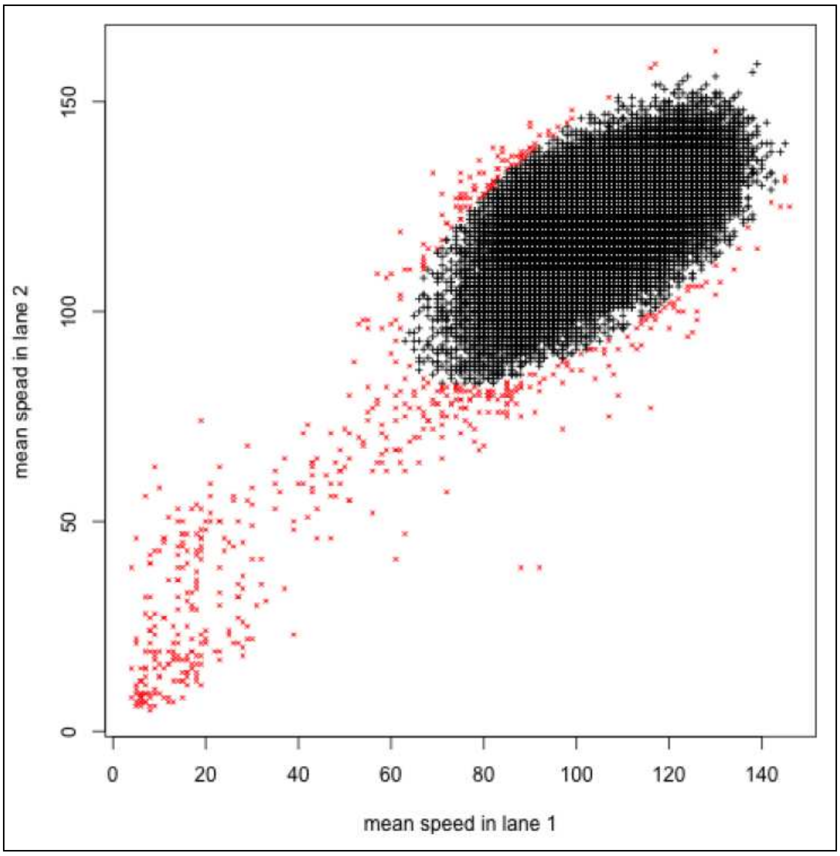


Figure 3: Scatterplot of the mean speed of vehicles (KFZ) in adjacent lanes

The colour code red represents observations with large (robust) multivariate distance based on 99.9% tolerance ellipses.

Source: nast consulting, Technical University of Vienna

4.2.1 Statistical approach

The main idea of the statistical approach is the use of the information of both space and time dependent sectors as well as historical data to formulate a general regression model. In order to decide whether given new measurements (offline or online) of detectors are likely to be faulty, not only prediction intervals given by the model but also aggregated information available from historical data is used.

The first step of model definition is the creation of a pooled dataset containing information from all sectors and lanes under consideration. It is helpful (but not necessary) to impute missing values for both the dependent variable as well as all independent variables before fitting the regression model.

The coefficients of the model are estimated using historical data. The predicted number of vehicles on a given sector and lane given the model on newly measured data is finally evaluated in order to make a decision if the measured data are likely to be correct.

The idea is to use statistical prediction intervals as well as historical information on the expected number of vehicles on the corresponding lane given day time, month, sector among other characteristics. If both the prediction interval given the model does not contain the observed number of vehicles and the measured number of vehicles is outside the range of the 2.5% and 97.5% quantile of the number of vehicles given historical information the model suspects a detector failure or abnormal road condition. The input data may also depend on the mean travel time between sectors.

This time lag may be estimated and included in the model. However given the limited data at hand (no travel times available) this information can not be used for modelling purposes. Nevertheless the model can be extended in order to apply further variables to detect malfunctions independently of sector properties since these properties are already indirectly used in the underlying regression model.

Another point worth mentioning is that even though traffic count data is modelled an ordinary least square regression (OLS) is used. While it might be beneficial to use a regression based on poisson or negative binomial distribution assumptions the main task is the identification if detectors are working (an not a highly accurate prediction of values). In addition the OLS approach has some benefits. First of, OLS-regression is available in virtually any statistical software which is good for later implementation purposes but it is also (considerably) faster than for example using a negative-binomial regression model. This could be useful especially during the online application.

Due to the variability of the variables a data aggregation needs to be carried out since one minute data especially for low-traffic time period can vary quite much. Consequently the data has been aggregated to 5-minute time intervals.

In order to create a profound model a set of independent variables needs to be created. Further choices that have to be made include the specification of a regression procedure such as ordinary least square regression, poisson-or negative binomial regression procedure or any robust regression method.

Once the required choices have been made and a suitable model has been found it is required to fit the model to historical data. This results in a set of regression coefficients. A very important limitation with respect to the independent variables that are used in the regression model is that these variables need to be available at the same time interval as the data on number of vehicles that are modelled (especially for an online application). Whenever new data (both on the number of vehicles driving at a specific lane at a given location as well as all the required independent variables have been collected), it is possible to plug these new data into the specified model. Then the model predicts the number of vehicles at the given lane and sector as well as the corresponding confidence interval for each predicted value based on historical information.

After this step both the observed and collected values of number of vehicles for a given time interval at a specific lane on a given sector as well as the predicted number of vehicles for the exactly same time, lane and sector are available.

The next task is then to describe the proposed idea to judge weather the observed and measures values significantly different from the expected values in order to identify a high chance of detector failure. The idea is simple and is outlined below.

The main idea is based on two different assumptions. A detector failure or abnormal road condition is identified if both assumptions are violated. The assumptions are:

1) the observed (measured) number of vehicles must lie within its corresponding prediction interval given the results of the regression model.
2) the observed (measured) number of vehicles must lie within the range of the 2.5% and 97.5% percentile (normal range) of the distribution of number of vehicles for a given sector, lane, hour and hour of the day based on historical information.

Only if both assumptions are violated a measurement is flagged as possibly faulty. In other words, if the corresponding prediction interval for a newly measured data point does not include the measured number of vehicles and the observed number of vehicles lies outside the normal range based on historical data, there is a strong indication that the detector measuring data for the lane and sector under consideration is at fault and needs to be checked.

Furthermore also standardized residuals can be used as separate rule for data assessment. The key idea is that the standardized model-residuals given as difference between observed and predicted values for a given measurement follow a (symmetric) distribution around with mean 0 and standard deviation of 1. If measurements with large standardized residuals are observed the measurement is possibly suspicious and flagged.

In general both proposed procedures are very flexible because the basic model based on historical data can be refitted at any given time. For example the input data used when fitting the model may be changed in a way that more current data (that have been proven to be measured by correctly-working detectors) may be included while old data may be removed. This results in modified regression coefficients that are then used as input parameters for predicting measurements and testing purposes.

Example for traffic data assessment on the German motorway A 8

For the year 2010 one-minute interval data was provided for selected road detection sites along the A 8 motorway. Extensive work is required for the import and manipulation of the traffic data outside of the proposed software platform of QUATRA. Furthermore different codes and formats are used in German and Austrian data streams - therefore additional work is required for data transformation.

Thus only a few sectors were extracted which feature different numbers of lanes. Furthermore instead of 5 minute intervals for the online application within QUATRA the data was aggregated and tested upon 10-minute intervals.

For testing purposes the following model was applied:

$$qKFZ \sim tNetto + classVA + weekday + vPKW + sectorLane + hour$$

This means that the number of vehicles (*qKFZ*) in a 10-minute interval is explained by a linear combination of the following independent variables along with an intercept:

tNetto: this variable represents the mean net gap between vehicles measured in 1/10 seconds

classVA: indicator of traffic volume which is defined as:

1 - if *qKFZ* >= 1 and *qKFZ* <= 5

2 - if *qKFZ* >= 6 and *qKFZ* <= 10

3 - if *qKFZ* >= 11 and *qKFZ* <= 20

4 - if *qKFZ* >= 21 and *qKFZ* <= 50

5 - if *qKFZ* >= 51 and *qKFZ* <= 100

6 - if *qKFZ* >= 101 and *qKFZ* <= 150

7 - if *qKFZ* >= 151 and *qKFZ* <= 200

8 - if *qKFZ* >= 201

weekday: factor variable specifying weekday of measurement (Monday to Sunday)

vPKW: mean velocity of cars

sectorLane: factor variable specifying all possible combinations of variables SST.Nr (sectors) and DE.Kanal (lanes)

hour: hour of the day

Three different approaches were used for the detection of suspicious measurements and possible detector failures

a) Strategy based on prediction intervals and historical quantiles: In this case possible suspicious measurements are found if the observed value of variable *qKfz* is not within the prediction interval derived in the regression model and the observed value of the number of vehicles does not lie within the 2.5% and 97.5% quantiles of the historical empirical data distribution given a specic sector, lane and hour of the day

b) Strategy based on standardized residuals and historical quantiles: In this case possible suspicious measurements are identified if the absolute values of standardized residuals given of differences between predicted and observed values of the variable *qKfz* are larger than a given threshold and if the observed value of *qKFZ* does not lie within within the 2.5% and 97.5% quantiles of the historical empirical data distribution given a specific sector, lane and hour of the day.

c) Strategy based on robust measures of location and deviance: In this scenario possible measurement errors or suspicious values are highlighted if an outlier is detected based upon median values (as measurements of location) as well as values of the *Qn*-estimator (as a robust measurement of dispersion) given a specific sector, lane and hour of the day:

$$pnorm(x, me, f * Qn) > 1 - \frac{\alpha}{2}$$

with *pnorm()* being the probability function of the normal distribution

x being an observed value of variable *qKFZ*

me and *Qn* the location and scale estimate for a specific sector, lane and hour of the day based on historical information,

f being a arbitrarily chosen constant and as usual. In the special case a multiplication factor for *Qn* was added to gain more flexibility.

In the following approach the general way is described how the test is conducted:

1) the data of the first eight months of 2010 of the section on the motorway A 8 were prepared as input for the first regression model

2) the number of vehicles for each site is predicted given the measured independent variables from all data points available

3) each of the remaining months is treated as newly measured data (September - December)

4) given the model-coefficients obtained in step 2) the number of vehicles for the data-set is predicted

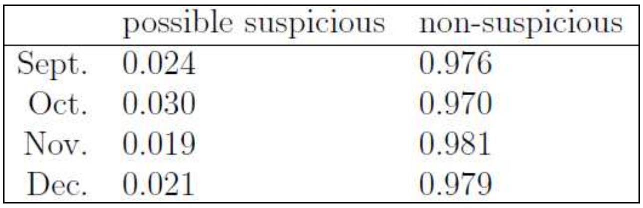
5) a set of measurements and diagnostics is calculated given three different approaches as described above (prediction intervals and historical quantiles, standardized residuals and historical quantiles and robust measures of location and deviance)

6) the data is updated by adding data for an additional month and new prediction intervals are calculated, consequently different estimates for regression coefficients are defined

Using this approach possible erroneous detector measurements can be identified.

Based on the preliminary results from the calculations for each of the four times that the model has been refitted with additional data, the coefficients of almost all independent variables were significant and also quite stable. The following tables 8-10 include rates of possibly suspicious values or measurement errors of the three proposed strategies that have been calculated for month September to December.

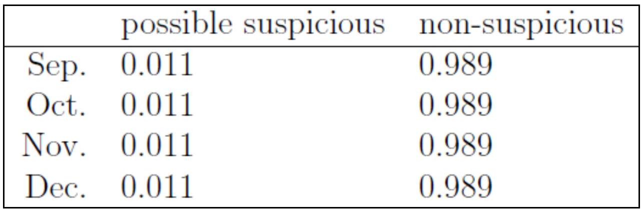
Table 8: Share of measurements classified as suspicious and non-suspicious (strategy based on prediction intervals and historical quantiles)



	possible suspicious	non-suspicious
Sept.	0.024	0.976
Oct.	0.030	0.970
Nov.	0.019	0.981
Dec.	0.021	0.979

Source: nast consulting, Technical University of Vienna

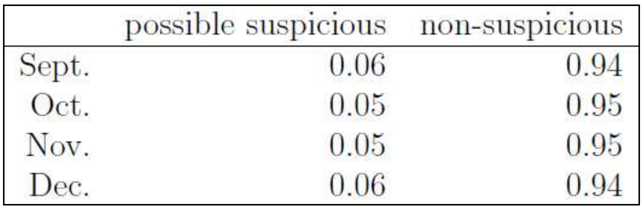
Table 9: Share of measurements classified as suspicious and non-suspicious (strategy based on standardized residuals and historical quantiles)



	possible suspicious	non-suspicious
Sep.	0.011	0.989
Oct.	0.011	0.989
Nov.	0.011	0.989
Dec.	0.011	0.989

Source: nast consulting, Technical University of Vienna

Table 10: Share of measurements classified as suspicious and non-suspicious (strategy based on using robust techniques)



	possible suspicious	non-suspicious
Sept.	0.06	0.94
Oct.	0.05	0.95
Nov.	0.05	0.95
Dec.	0.06	0.94

Source: nast consulting, Technical University of Vienna

The results from the above tables show that the strategies based on a combination of either standardized model-residuals or prediction intervals together with quantiles of historical data for the dependent variable have classified less observations as possibly erroneous data than the third strategy that is solely based on robust measurements of location and scale. Furthermore it can be seen that the strategy based on standardized residuals in table 10 leads to the lowest identification rates of possible measurement errors. However, from this information alone it is not possible to state whether any of these methods is superior to another. On a general note it should be stated that it is indeed quite difficult to compare different strategies since in the data a variable stating observed (and established) measurement errors are missing.

4.2.2 Data Imputation

The results from an analysis of the available traffic data are directly linked to the quality and completeness of the examined data set. For this reason, erroneous measured data values and also missing values must therefore be imputed. Surely, values that arise from malfunctioning detectors or communication failures as well as missing values in the data should not simple be removed (which is of course a possible procedure), but should be imputed with reasonable estimates.

Several imputation strategies might be used, each with different computational and implementation characteristics, advantages and disadvantages. Some methods can even take possible spatial dependencies into account.

Several question during the imputation process arise. Can a missing reading in a station be inferred from previous observations of that particular station? Can the missing value be inferred from available values at other nearby stations? Should only co-occurring measurements be considered or can measurements from the immediate past be incorporated? Should estimation occur evaluating elements in a data set strictly once or possibly multiple times?

The answers to all these questions determine and influence the imputation methods that should and can be used along with other user requirements such as for example the required computational performance of the imputation procedure.

One of the simplest imputation approaches consists of repeating the previously measured value whenever a missing or suspicious value is identified. This method is very easy to implement, computationally fast but cannot be trusted for long periods of missing values. Furthermore, this procedure gives worst estimates if a detector reports faulty values systematically and constantly or if values are constantly missing starting from a certain time stamp. However, this method can be improved easily by defining a time-window over temporal measurements, estimating a specific mean (arithmetic mean or median, for example) of a quantity of interest and replacing the missing values with that estimate. This approach can be used for both offline and online evaluation. For online imputation, one can use the mean of previous values to impute an incoming missing or erroneous value, and offline just the mean of historic data may be used. All in all, these methods seems to be too oversimplified to provide reasonable estimates of erroneous or missing values.

Since we are dealing with time-related data (e.g.. time series) it is possible to assume that characteristics observed at a particular point in time are closely related to an immediately preceding time interval. The time series of interest (e.g. amount of vehicles measured at a given lane in a specific sector per minute) should first be made stationary. After that step, a filter (for e.g. by using a moving average) may smoothen the time series. Supposing that a value is auto-correlated with the previous values, an autoregressive model can be estimated. This leads to an ARIMA model[b]. The disadvantage of such time series models is that important covariates cannot be easily taken into account and also that historic information that may very well be available cannot be included directly.

The following figure shows the partial autocorrelation of the amount of vehicles (Kfz) at a specified sector on the Austrian motorway A 12. As expected the partial autocorrelation is rather high for small time lags and decreases with increasing length of the time lags. Therefore, assuming an ARIMA process is realistic and imputation of missing or suspicious data values using a time series model will likely lead to plausible imputation results.

[b] cf. BOX / JENKINS (1970)

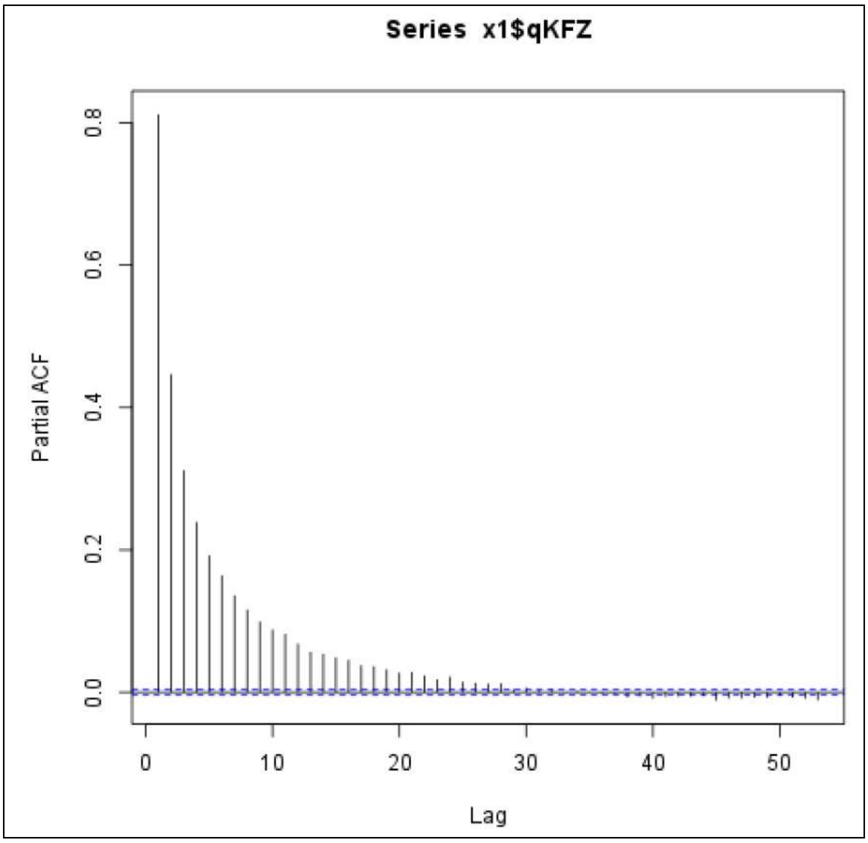


Figure 4: Partial autocorrelation of the amount of KFZ

Source: nast consulting, Technical University of Vienna

Univariate imputation methods as described before become less useful as the length of measurement failures or missing values at a single station increases. In addition, the variance of the variable of interest is decreased by imputing, for example, using a very basic method such as mean-imputation.

Regression methods can better deal with longer periods of failure because more information is used to impute the missing data. In general, valid information of neighbouring lanes may be used and information in other sectors as well as other covariates in the data that is used. It is of course also possible to use historic information using a regression based imputation approach. However, using only information from neighbourhood lanes, imputation is of no success when both lanes do either have no data or erroneous data.

Regression consists of finding a functional description of an observed response variable to a set of independent predictor variables. While regression plays also a role for auto-regressive models, covariates should be used that will help to improve model-based predictions that will be used as replacements for missing values. Methods with a certain random component can be used for multiple imputation. However, the main interest is not to multiple impute values but to construct one complete data set.

To optimally build the models, the whole data set have to be restructured. Additional columns are built for the amount and speed of vehicles on the other lane at same sector and the same point in time. For using the information of previous or following sectors we create variables for the amount and speed of vehicles on the previous or following sector (on the same lane).

Additionally, the time lag between two neighbouring sectors, so that for lane 1 and 2 of sector a the corresponding lane of sector b is used with a time lag of −1 minute. For lane 1 and 2 of sector b the corresponding lane of sector a is used with a time lag of +1 minute. In addition the constraint is considered that, for example, speed of a vehicle will only be imputed if the amount of KFZ is non-zero.

When using a regression model for imputation all covariates should be non-missing. This would decrease the number of possible covariates tremendously. The selected approach starts with an "optimal" and full model with the information of all necessary covariates in it and imputation of all missing values of the target variable with no co-missingness in the covariates. After that in each step a variable is deleted that include missing values, preferable those variables with the highest number of missing values.

Using the model fit and the corresponding regression coefficients, the missing values (with no missing values in the covariates) in vPkw are predicted using that information. Finally, the missing values in vPkw are then replaced by the exponential values of the predictions. After the full model a series of reduced models is applied and missing values are predicted. It is not always necessary that all models are used. In most cases a smaller number of models are sufficient to impute all missing values in vPkw.

4.2.3 Conclusion

Methods for detection of bad detectors from their outputs, and imputation of missing data from historic data have been developed. Traditional existing methods are often quite oversimplified, either by looking or modelling univariate information or by using very plain models where only the neighbourhood detector is used to impute missing values or to identify malfunctions of detectors.

However, there is much more information included in how detectors behave over time, and that information of one variable is dependent on other covariates. The proposed algorithm performs better than traditional methods and methods used in the past, because historical information on all variables is used that are related with the variable to impute or to detect malfunction.

In particular, two methods for the detection of erroneous data were presented. First, robust Mahalanobis distances are used to determine observations that are far away from the main bulk of the data. Secondly, a general regression model is used to identify observations that are unusual far away from their predictions. In combination with a simple second rule based on univariate outlyingness, this gives realistic estimates which observations are suspicious or abnormal. The general model does have - as the name suggests - a broad application area and it is not limited to the currently used data. Data from more than two sectors or more than two lanes can be considered.

To impute missing values or erroneous data points, certain models have been established that take all useful information into account. This allows much better imputations as traditional methods like hot-deck, k-nearest neighbour or mean imputation.

5 Development of the software system

Within work package 4 "Software Development" a software system platform is to be created along with the corresponding service for data processing and data quality analysis. Based on the requirements of the different tools that need to be developed (online-freeway-tool and urban-offline-tool) the platform will need to be able to process data from different streams and formats during online and offline applications. Furthermore all corresponding layout and operating requirements need to be addressed for manual and automated processing and analysis of traffic data.

As part of the software concept already existing software modules from the software products LOTRAN-DQ (TRANSVER GMBH) will be used and upgraded to cater for the needs of the QUATRA objectives.

5.1 Software Concept

5.1.1 Functionality

Rather than inventing a complete new software tool the functions, data interfaces and visualisation tools of the existing software LOTRAN-DQ are taken into account for the software tool development. The following functionalities are provided within LOTRAN-DQ:

- Automatic sequence control that starts the import of data and the calculation of quality indicators at defined times (e. g. each night, each hour or each minute)
- Import of infrastructure data (e. g. number of lanes, ID and location of detectors) as basis for the import of traffic data and the calculation and visualization of quality indicators
- Import of traffic data from a data archive
- Management of parameters for calculation of quality indicators (e. g. thresholds)
- Calculation of quality indicators for defined intervals (e.g. 1 hour, 1 minute)
- Saving of calculated quality indicators at a data base
- Selection of infrastructure (e.g. stretches, detectors) and period for visualization of quality indicators
- Visualization of infrastructure
- Visualisation of quality indicators in tables and diagrams
- Configuration of diagrams (e.g. labels, scaling)
- Print of selected tables and diagrams for selected infrastructure
- Export of selected tables (as csv-files) and diagrams (as images) for selected infrastructure
- Help

During several workshops with road authorities in Austria and Germany the following new functionalities have been identified, cost estimated and prioritised[c] as part of the QUATRA software development process:

- Calculation and Visualization of new quality indicators:
  - New statistical indicators for freeway and urban areas
  - New indicator for identification of hanging detectors (delivering the same plausible dataset for a defined period)
  - Logical interconnection of different indicators for plausibility checks
  - Automatic selection of periods without missing data for calculation of “global indicators” (accounting)
  - Automatic interpretation of “global indicators” (accounting)
- Specific calculation parameters for different types of infrastructure (e.g. main road, ramp)
- Online calculation of quality indicators (for traffic state estimation and control purposes)
- Marking of hours with implausible data (that this data will not be used for statistical purposes)
- Interface to bug tracking system to get information about detectors and known errors and to generate tickets if errors have been identified by indicators
- Interface to road works management system to get information about closed lanes etc. (reason for no counting)
- Display of detector information (e.g. type, manufacturer, model)
- Entry of flags and comments (e.g. for detectors with known problems)
- List of indicators that can be filtered and sorted
- Visualization of problems (indicators > thresholds) per detector and interval (coloured matrix)
- Point out changes compared to the last period
- Generation of monthly quality reports for management purposes
- Comparison of detector date with data sources, if available (e.g. toll data, floating car data, mobile phone data)

[c] Not all identified new functionality can be realized in this project

5.1.2 System Architecture

To provide the listed existing and proposed functionalities of the QUATRA system the following components of LOTRAN-DQ are used in order to supply a stable and already tested system:

- LOTRAN-DQ Server
- LOTRAN-DQ Graphical User Interface (Client)
- LOTRAN-DQ Database (for results)

The LOTRAN-DQ server is being started by the automatic sequence control at defined times. It imports the infrastructure data from an infrastructure file (XML-format) by using the infrastructure data interface and the traffic data from a traffic data archive by using the traffic data interface. The LOTRAN-DQ DB interface reads the calculation parameters from the results database. Thereafter the data quality indicators are calculated and written in the results database by using the LOTRAN-DQ DB interface.

The results database provides data to the graphical user interface based upon the selections made for infrastructure, period and indicators. The visualization is provided in tables and diagrams, which can be printed by using the print manager or exported in different formats by using the export manager. The infrastructure visualization uses the infrastructure file and its respective interface. The graphical interface furthermore provides the possibility to set the global calculation parameters in the parameter manager. The help module assists the users with the handling of the software.

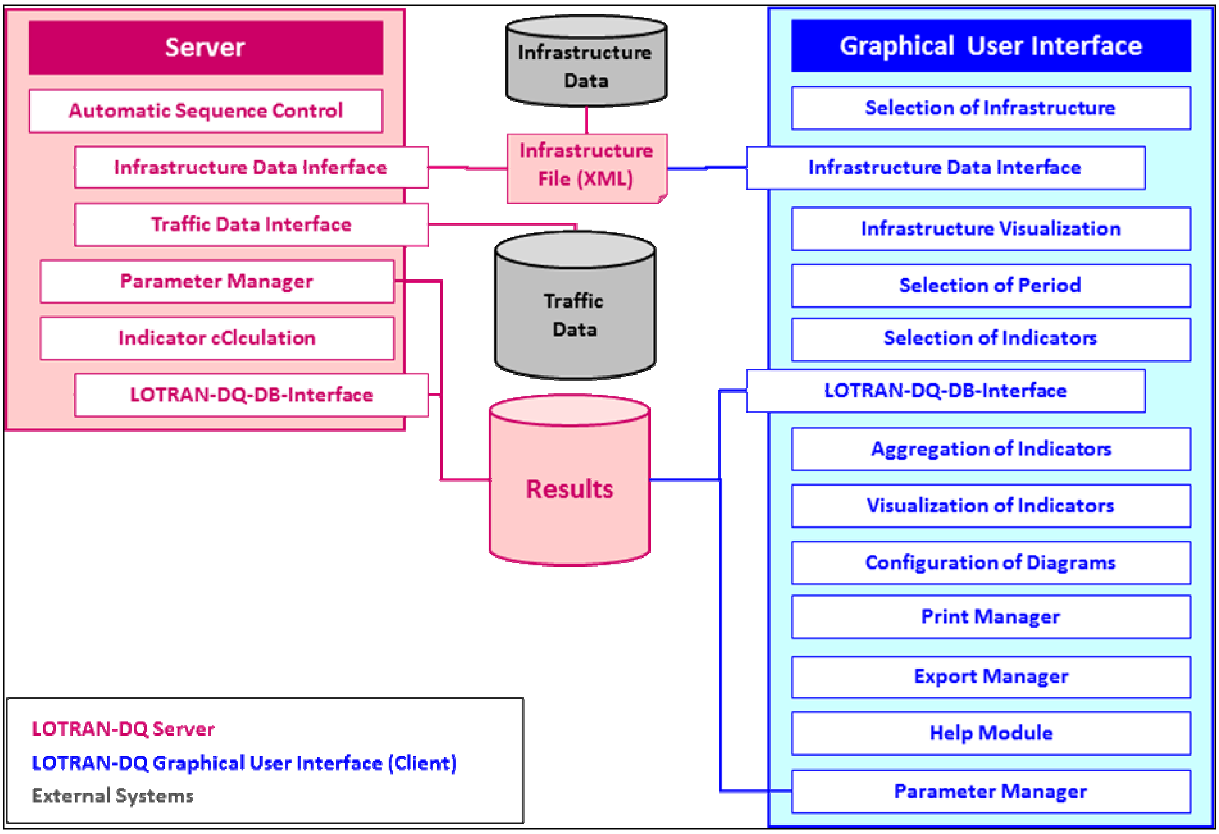


Figure 5: System Architecture

Source: TRANSVER

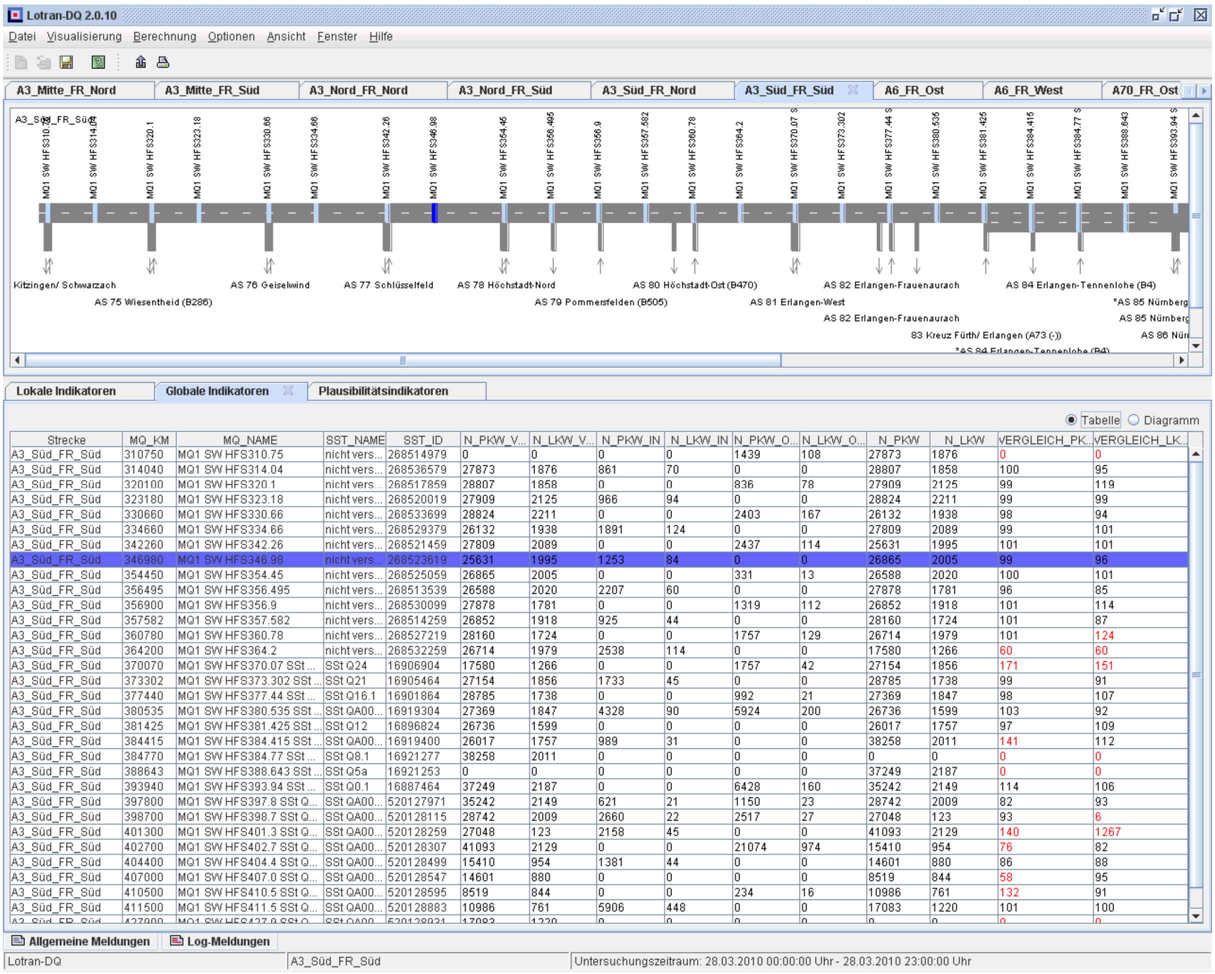
5.1.3 Visualization (Graphical User Interface - GUI)

The proposed visualisation consists of one register for each section that provides the following components:

- Visualization of infrastructure (number of lanes, entries and exits, measurement cross sections)
- One register for each of the three types of indicators (local, global, plausibility)
- Provision table view in figure 6 and diagram view in figure 7

If a data set (at the table) or a detector (at the visualization of the infrastructure) is selected the corresponding detector or data set will be highlighted in blue.

If an indicator value is outside of the visualization thresholds then this value/detector will be highlighted in red at the table and diagram. The visualization thresholds can be individually set for each client (while the calculation thresholds are global).

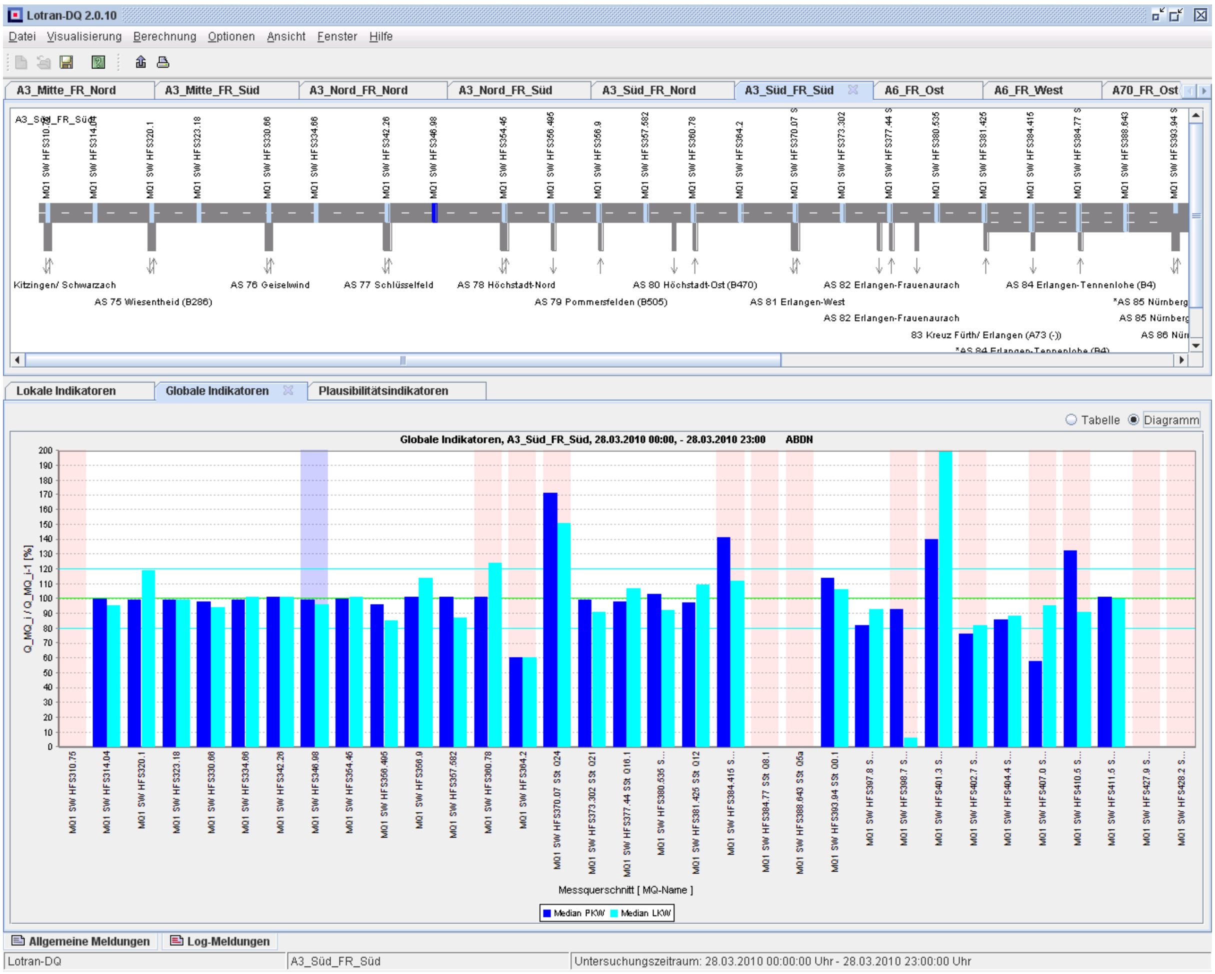


Strecke	MQ_KM	MQ_NAME	SST_NAME	SST_ID	N_PKW_V...	N_LKW_V...	N_PKW_IN	N_LKW_IN	N_PKW_O...	N_LKW_O...	N_PKW	N_LKW	VERGLEICH_PK...	VERGLEICH_LK...
A3_Süd_FR_Süd	310750	MQ1 SW HFS310.75	nicht vers...	268514979	0	0	0	0	1439	108	27873	1876	0	0
A3_Süd_FR_Süd	314040	MQ1 SW HFS314.04	nicht vers...	268536579	27873	1876	861	70	0	0	28807	1858	100	95
A3_Süd_FR_Süd	320100	MQ1 SW HFS320.1	nicht vers...	268517859	28807	1858	0	0	836	78	27909	2125	99	119
A3_Süd_FR_Süd	323180	MQ1 SW HFS323.18	nicht vers...	268520019	27909	2125	966	94	0	0	28824	2211	99	99
A3_Süd_FR_Süd	330660	MQ1 SW HFS330.66	nicht vers...	268533699	28824	2211	0	0	2403	167	26132	1938	98	94
A3_Süd_FR_Süd	334660	MQ1 SW HFS334.66	nicht vers...	268529379	26132	1938	1891	124	0	0	27809	2089	99	101
A3_Süd_FR_Süd	342260	MQ1 SW HFS342.26	nicht vers...	268521459	27809	2089	0	0	2437	114	25631	1995	101	101
A3_Süd_FR_Süd	346980	MQ1 SW HFS346.98	nicht vers...	268523619	25631	1995	1253	84	0	0	26865	2005	99	96
A3_Süd_FR_Süd	354450	MQ1 SW HFS354.45	nicht vers...	268525059	26865	2005	0	0	331	13	26588	2020	100	101
A3_Süd_FR_Süd	356495	MQ1 SW HFS356.495	nicht vers...	268513539	26588	2020	2207	60	0	0	27878	1781	96	85
A3_Süd_FR_Süd	356900	MQ1 SW HFS356.9	nicht vers...	268530099	27878	1781	0	0	1319	112	26852	1918	101	114
A3_Süd_FR_Süd	357582	MQ1 SW HFS357.582	nicht vers...	268514259	26852	1918	925	44	0	0	28160	1724	101	87
A3_Süd_FR_Süd	360780	MQ1 SW HFS360.78	nicht vers...	268527219	28160	1724	0	0	1757	129	26714	1979	101	124
A3_Süd_FR_Süd	364200	MQ1 SW HFS364.2	nicht vers...	268532259	26714	1979	2538	114	0	0	17580	1266	60	60
A3_Süd_FR_Süd	370070	MQ1 SW HFS370.07 SSt ...	SSt Q24	16906904	17580	1266	0	0	1757	42	27154	1856	171	151
A3_Süd_FR_Süd	373302	MQ1 SW HFS373.302 SSt ...	SSt Q21	16905464	27154	1856	1733	45	0	0	28785	1738	99	91
A3_Süd_FR_Süd	377440	MQ1 SW HFS377.44 SSt ...	SSt Q16.1	16901864	28785	1738	0	0	992	21	27369	1847	98	107
A3_Süd_FR_Süd	380535	MQ1 SW HFS380.535 SSt ...	SSt QA00...	16919304	27369	1847	4328	90	5924	200	26736	1599	103	92
A3_Süd_FR_Süd	381425	MQ1 SW HFS381.425 SSt ...	SSt Q12	16896824	26736	1599	0	0	0	0	26017	1757	97	109
A3_Süd_FR_Süd	384415	MQ1 SW HFS384.415 SSt ...	SSt QA00...	16919400	26017	1757	989	31	0	0	38258	2011	141	112
A3_Süd_FR_Süd	384770	MQ1 SW HFS384.77 SSt ...	SSt Q8.1	16921277	38258	2011	0	0	0	0	0	0	0	0
A3_Süd_FR_Süd	388643	MQ1 SW HFS388.643 SSt ...	SSt Q5a	16921253	0	0	0	0	0	0	37249	2187	0	0
A3_Süd_FR_Süd	393940	MQ1 SW HFS393.94 SSt ...	SSt Q0.1	16887464	37249	2187	0	0	6428	160	35242	2149	114	106
A3_Süd_FR_Süd	397800	MQ1 SW HFS397.8 SSt Q...	SSt QA00...	520127971	35242	2149	621	21	1150	23	28742	2009	82	93
A3_Süd_FR_Süd	398700	MQ1 SW HFS398.7 SSt Q...	SSt QA00...	520128115	28742	2009	2660	22	2517	27	27048	123	93	6
A3_Süd_FR_Süd	401300	MQ1 SW HFS401.3 SSt Q...	SSt QA00...	520128259	27048	123	2158	45	0	0	41093	2129	140	1267
A3_Süd_FR_Süd	402700	MQ1 SW HFS402.7 SSt Q...	SSt QA00...	520128307	41093	2129	0	0	21074	974	15410	954	76	82
A3_Süd_FR_Süd	404400	MQ1 SW HFS404.4 SSt Q...	SSt QA00...	520128499	15410	954	1381	44	0	0	14601	880	86	88
A3_Süd_FR_Süd	407000	MQ1 SW HFS407.0 SSt Q...	SSt QA00...	520128547	14601	880	0	0	0	0	8519	844	58	95
A3_Süd_FR_Süd	410500	MQ1 SW HFS410.5 SSt Q...	SSt QA00...	520128595	8519	844	0	0	234	16	10986	761	132	91
A3_Süd_FR_Süd	411500	MQ1 SW HFS411.5 SSt Q...	SSt QA00...	520128883	10986	761	5906	448	0	0	17083	1220	101	100
A3_Süd_FR_Süd	427000	MQ1 SW HFS427.0 SSt Q...	SSt QA00...	520128931	17083	1220	0	0	0	0	0	0	0	0

Lotran-DQ | A3_Süd_FR_Süd | Untersuchungszeitraum: 28.03.2010 00:00:00 Uhr - 28.03.2010 23:00:00 Uhr

Figure 6: GUI – Visualization of infrastructure and table with global indicators

Source: TRANSVER



○ Tabelle ● Diagramm

Lotran-DQ | A3_Süd_FR_Süd | Untersuchungszeitraum: 28.03.2010 00:00:00 Uhr - 28.03.2010 23:00:00 Uhr

Figure 7: GUI – Visualization of infrastructure and diagram with global indicators

Source: TRANSVER

6 Networking with road authorities

In August and September 2012 intensive networking was carried out with Austrian and German road authorities. Main objective of the workshops with representatives were discussions of requirements of the authorities and possible solutions that could be solved with QUATRA. In general there are deficiencies in relation to available data especially during use cases were online data is required. These deficiencies mainly arise out of non technical error sources during the detection process (e.g. wrong installation of sensors).

According to the authorities not only total traffic volumes show erroneous data rather also single vehicle categories. If these category based deficiencies could be identified and reported that would be a clear benefit upon existing approaches.

Furthermore potential use cases were identified as follows:

Use case 1: reconstruction of the traffic situation for traffic control purposes: at present implausible data is not used for the traffic control algorithms on Austrian motorways. An online procedure would be required to analyse the data and create substitute data (mainly traffic volumes and average vehicle speed) in case of error detection. The validated data could then be forwarded to the traffic control algorithms. Furthermore a virtual detection site could be inserted in the user interface that presents the imputated data for a specified section. The QUATRA tool could be installed at sub-master stations.

Use case 2: labelling of erroneous data for traffic statistics: currently traffic statistics are based on offline data that needs to be manually assessed and checked prior to the statistics preparation. The plausibility check is carried out after the traffic data has been collected and saved. An automated labelling during QUATRA analyses would ease the estimation process because these labelled data entries could simply be excluded for statistics. The labelling would mainly be relevant for 8+1 data detection sites (in a second stage also for 2+0 data detection sites)

Use case 3: monitoring of operations and sensor availability: data quality is essential for motorway operations. There are already systems in place that identify and record technical failures of sensors and detectors. The evaluation of the transferred content of the traffic data (e.g. accuracy) is not being done at the moment. Once a day would be sufficient for operations purposes. A possible solution should be able to generate automatic reports of malfunctions and wrong contents because experts („Second Level“) are needed for data quality assessment (experts need to go through the data of all installations and sensors manually). Furthermore the expert system should decide by itself if the operator („First Level“) should be informed through report generation.

In general traffic data from detectors in road work sections need to be validated because they often provide inaccurate measuring. Systems should allow an interface to road work databases in order to label those time periods when road works occurred.

For validation purposes reference data from the Austrian toll system could be used.

The authorities also stated that a user friendliness and automated processing of the data would be very important in terms of acceptance of the product.

7 Sources

AUSTRIATECH, PÖYRY INFRA TRAFFIC GmbH, BAVARIAN ROAD ADMINISTRATION, TECHNICAL RESEARCH CENTRE OF FINLAND, WSP GROUP (2010): "QUANTIS - Quality Assessment and Assurance methodology for Traffic data and Information Services"

BALMBERGER, M. ET AL. (2000): „Überprüfung von Dauerzählstellen im Autobahnnetz der Autobahndirektion Nordbayern“, Straßenverkehrstechnik Heft 12/2000, p. 648ff

BICKEL, P.J. ET AL. (2007): "Measuring traffic", Statistical Science, Volume 22, p. 581-597

BOX, G.E.P. / JENKINS, G.M. (1970): "Time series analysis: forecasting and control", Holden-Day series in time series analysis, Source: http://books.google.at/books?id=5BVfnXaq03oC

BUSCH, F. ET AL. (2006): "Benchmarking for traffic data acquisition systems and traffic control systems", Scientific journal series research road building and traffic engineering, book 949, Bonn

CHAN, K.F. (2008): “Leading traffic data to a high quality standard”, DaVinci, 16th ITS World Congress Final Paper

CHEN, L. / MAY, A.D. (1987): “Traffic detector errors and diagnostics", Transportation Research Record No. 1132, Freeway Management and Operations. p. 82-93

CHEN, C. ET AL. (2003): “Detecting errors and imputing missing data for single-loop surveillance systems”, in: Transportation Research Record, Issue: 1855, p. 160-167

COIFMAN, B. / DHOORJATY, S. (2004): "Event data-based traffic detector validation test", Journal of Transportation Engineering, 130/3, p. 313-321

COIFMAN, B. / LEE, H. (2006): "A Single Loop Detector Diagnostic: Mode on-Time Test”, Applications of Advanced Technology in Transportation, Proceedings of the 9th International Conference, American Society of civil engineers, p. 623-628

COIFMAN, B. / LEE, H. (2011): "Diagnosing Chronic Errors in Freeway Loop Detectors from Existing Field Hardware“, UC Berkeley Transportation Library, Source: http://www.dot.ca.gov/new..._tid_0978_final_report.pdf

COREY, J. ET AL. (2011): "Detection and Correction of Inductive Loop Detector Sensitivity Errors Using Gaussian Mixture Models“, Conference: Transportation Research Board 90th Annual Meeting

DE WAAL, T. (2008): "An overview of statistical data editing", Discussion paper 08018, Statistics Netherland

FERNANDEZ-MOCTEZUMA, R.J. ET AL. (2009): „Toward improved and transparent imputation techniques for online trac data streams and archiving applications”, Proceedings of the 88th Annual Meeting of the Transportation Research Board (TRB)

FILZMOSER, P. ET AL. (2008): “Outlier identification in high dimensions”, CSDA, Volume 52, p. 1694-1711, p. 41-42

FORSCHUNGSGESELLSCHAFT FÜR STRASSEN- UND VERKEHRSWESEN (2005): "AK 3.5.20 - Reference notes for quality requirements and safety of local traffic data collection“, Source: http://www.fgsv-verlag.de/catalog/product_info.php?products_id=2219

FREUDENBERGER, P. (2001): "Analyses of detector loop data und development of a method for plausibility checks”, Diploma thesis on the chair of traffic engineering and control of Technical University of Munich (FGV-TUM)

HARDIN, J. / ROCKE, D. (2005): “The distribution of robust distances”, Journal of Computational and Graphical Statistics, Volume 14, Number 4, p. 928–946

HIGHWAYS AGENCY (2011): "MCH1529 - Guidance Notes for Assessment of Detector Technology/Systems"

HOOPS, M. (2002): "Methodology for quality management of aggregated data of a measurement system when operating traffic management systems", Dissertation on the chair of traffic engineering and control of Technical University of Munich (FGV-TUM)

KWON, J. ET AL. (2004): "Statistical methods for detecting spatial configuration errors in traffic surveillance sensors", Transportation Research Board, Issue: 1870, p. 124-132

LITTLE, R.J.A. / RUBIN, D.B. (1987): “Statistical Analysis with Missing Data”, Wiley, New York

MARONNA, R.A. ET AL. (2006): “Robust Statistics: Theory and Methods”, Wiley, New York

MARZ 99 - Bundesanstalt für Straßenwesen (1999): "Information sheet for the equipment of traffic control centres and sub centres"

MOMATEC GMBH (2012): "Traffic IQ - Pilot project Informationsqualität im Verkehrswesen"

NAST CONSULTING, TU WIEN - Institut für Wahrscheinlichtkeitstheorie (2008): "Machbarkeitsstudie Qualitätskontrolle für Verkehrsdetektoren im ASFINAG-Netz", on behalf of the ASFINAG

NIHAN, N.L. / WONG, M. (1995): "Improved error detection using prediction techniques and video imaging - final report“, Report No: WA-RD 386.1,TNW 95-01

NIHAN, N.L. / WONG, M. (2000): „Freeway traffic speed estimation using single loop outputs“, Department of Civil Engineering University of Washington, Source: http://faculty.washington.edu/yinhai/wangpublication_files/TRB_00_SP.pdf

NIHAN, N.L. ET AL. (1990): “Detector Data Validity”, Report No.: WA-RD 208.1

NIHAN, N.L. (1997): “Aid to Determining Freeway Metering Rates and Detecting Loop Errors”, Journal of Transportation Engineering Volume 123, p. 454-458, Washington State

TEMPL, M. / TODOROV, V. (2011): “Screening: methods and tools for unido indstat database”, Discussion paper 1, UNIDO

TEMPL, M. ET AL. (2011): “Iterative stepwise regression imputation using standard and robust methods”, Comput Stat Data Anal, Volume 55, p. 2793-2806

TEMPL, M. ET AL. (2012): “Exploring incomplete data using visualization techniques. Advances in Data Analysis and Classification”, Volume 6, p. 29-47, DOI: 10.1007/s11634-011-0102-y

TRANSVER (2008): "Quality control of traffic data in the federal highway network in Germany", LOTRAN-DQ

TUROCHY, R.E. / SMITH, B.L. (2000): “New procedure for detector data screening in traffic management systems”, Transportation Research Record, Issue Number: 1727

VANAJAKSHI, L.D. / RILETT, L.D. (2006): "System-Wide Data Quality Control of Inductance Loop Data using Nonlinear Optimization“, Journal of Computing in Civil Engineering, 2006/5, p. 187-196

VERSAVEL, J. (2007): "Traffic Data Collection: Quality Aspects of Video Detection”, Transportation Research Board 86th Annual Meeting, TRB 86th Annual Meeting Compendium of Papers, Source: http://trid.trb.org/view/2007/C/801361

WEIJERMARS, W.A.M. / VAN BERKUM, E.C. (2006): "Detection of Invalid Loop Detector Data in Urban Areas”, Transportation Research Record: Journal of the Transportation Research Board, Issue: 1945, p. 82-88