Palacký University Olomouc

Transactions on Transport Sciences
Peer-Reviewed Open Access Journal
SI SCSP conference/2025 DOI: 10.5507/tots.2025.006
journal homepage: www.tots.upol.cz

Approaches to the Data Governance in Transportation Domain – Data Quality Rules Design and an Example of Traffic Data Inspection

MARTIN LANGR, PAVEL HRUBEŠ

Czech Technical University in Prague, Faculty of Transportation Sciences, Department of Transport Tekematics, Konviktská 20, 110 00 Prague, Czech Republic

ABSTRACT: The paper focuses on the practical application of procedures for the design and creation of a complex system of data quality rules in transportation domain. The aim of the research was to verify the feasibility of implementing all relevant practices of the data governance concept for the specific needs of transportation data and to present possible ways of quality evaluation of these data. The paper describes the proposed requirements, tools and methodology for developing a complex approach to data governance and data quality evaluations. The functionality and benefits of the proposed procedures and their real implementation have been verified by applying the procedures on real data from selected profiles of the D11 motorway in the Czech Republic from the year 2023. The discussed results demonstrate the functionality of the proposed solution and allow for future large-scale application of the verified procedure across the full geographic coverage of the data. It is also possible to apply the procedures and rules to other data sources or data from other types of roads including urban traffic.

KEYWORDS: data governance, data quality, data quality rules, traffic data quality check

1. INTRODUCTION

In the previous two years, we have been working on research and application of data governance approaches in the transport domain, specifically on methods and procedures for data quality assessment, within the framework of a scientific project supported by The Technology Agency of the Czech Republic named "Data Quality Tools for Ensuring System Reliability of Transport Information Centers". As part of previous published work (Hrubeš, 2024), an extensive search of available publications confirmed that this topic is very relevant; practically, data quality is necessarily addressed in any transport data processing, but these practices are often only presented in general terms and the approaches and tools used are not sufficiently described or presented.

Thus, it was necessary to look for procedures, methods and tools to approach traffic data and its quality evaluation. In the course of the project, we have described in detail the methodological procedures (Mlynářová, 2023), which include exploring different dimensions of data quality such as validity, consistency, completeness, uniqueness, timeliness, and accuracy.

Methodologically, we have followed the general rules and procedures of the data governance approach and for application we have used different types of traffic data provided by the Road and Motorway Directorate. Primarily, the data were traffic volume and speed data. Due to the higher diversity and greater coverage of traffic data, the range of sources was extended to include datasets focusing on other specific aspects of traffic. These sources included data from truck rest areas, weather stations and vehicle weight in motion systems. For all these datasets, a baseline analysis was performed to provide insight into their structure, characteristics and the needs of data quality examination approaches. Building on these works, we studied the time dependence characteristics of the time series of specific measured and observed variables, where we explored the possibility of designing statistical models (Generalized Additive Model) to detect time series anomalies. One published paper (Purkrábková, 2024) describes the use of the model in the context of the whole data quality control procedure. A second paper (Purkrábková, 2025) describes the design and validation of a stand-alone model and is currently under review for future publication.

A comprehensive overview of the analyzed references is thus part of the results of the previous steps of the whole research. In this paper, we build on these steps and focus on the final integration of the described methods into the definition of traffic data quality control rules with implementation in the Accurity Quality application and demonstration of their direct application to data from the Czech D11 motorway recorded during 2023.

The following research objectives were formulated for this paper.

- RO1: based on previous results, to appropriately formulate a data quality rule design methodology with a focus on traffic data (including the formulation of data quality rule requirements).
- RO2: to develop and validate a suitable tool for effective design, realization, implementation and management of data quality rules.
- RO3: implement the created data quality rules in the Accurity Quality application environment and verify their functionality.
- RO4: test the functionality of the implemented rules and perform a quality assessment of a separate dataset.

2. METHODS AND DATA

The overall process of the implementation of the quality control rules for the specific area of traffic data, which uses the knowledge and results of previous activities, consisted of 3 main steps:

- formulation of requirements for traffic data quality rules and the process of their development;
- the design and realization of the data quality rules and their implementation;
- application of the implemented rules to a comprehensive traffic data set of the selected motorway.

The first key step was the formulation of Data Quality Rules Requirements. This was a general theoretical activity, but based on clearly described conditions. Based on the analysis of the specifics of the different data sources, sub-aspects requiring control by separate rules were formulated separately. These rules were then organized into groups based on the described data governance procedures, with the described links and logical sequence. This is due to the following possible phasing of rule creation and implementation.

The second step consisted in the physical design and implementation of all data quality rules. Here, it was necessary to respect both the formulated requirements and the variety of rule types and their expected inputs, as well as the function requirements, structure and parameters of the Accurity Quality application that was used to implementation the rules. The main aspects of this activity include the phasing and sequencing of the sub-steps of the development from the perspective of data check, as well as the phasing and sequencing of the sub-steps from the perspective of rule development. Related to this is the need to design a unified naming convention for all objects and rules to be created and to make efficient use of available functions, such as sharing parts of the rule design definitions, etc. For this activity, it was necessary to create a suitable tool that would allow all the necessary rules not only to be designed and created, but also to be managed and effectively implemented in the application. Each created rule was first tested for its function and then a procedure was created for its automated and bulk creation. After the creation of coherent groups of rules, they were imported and implemented in bulk.

The third step consisted of testing and verifying the functionality of the entire data quality control process. A complete section of motorway was selected for testing. For this section, all the rules related to the affected road profiles, carriageways and lanes were implemented. The rules were sequentially applied to data from all months of the year 2023. Thus, it was possible to analyze not only the functionality of all the rules themselves, but also the data quality results and their variation over the year. The application of the data quality rules therefore consisted of the following procedure:

- selecting a section of motorway suitable for testing the rules
- preparation of data for the selected motorway for each month of the year 2023
- implementation of the sequence of necessary data quality rules
- implementation of the rules in the application
- running all rules separately for each month
- evaluation and representation of the results of the rules

The evaluation of the results of the rules mainly concerns the evolution of the results of the top aggregated rules for the whole motorway section and its individual profiles. Furthermore, a more detailed view of the reasons for positive and negative aggregation results has been developed, especially in terms of data completeness, the number of undesirable "null" values in the data and the number of detected anomalies in the intensity and speed values of the traffic flow.

2.1 Data and Application

For the successful implementation of the research, the available data enabling their analysis and following control as well as the application itself for the implementation of data quality rules and the application of data governance approaches were essential.

2.1.1 Data used in the research

In the context of the project, different data sources were used to measure different aspects related to transport. All data were provided for research purposes by the Road and Motorway Directorate for the period of years 2021 and 2022. Specifically, data from Automatic Traffic Counters (ASD), Floating Car Data (FCD), Truck Rest Areas (ODP), Weather Stations (MET) and Weigh-in-Motion (WIM) systems were analyzed. For all these datasets, a basic analysis was carried out, describing the structure and content of the data as well as the potential and applicability for quality research. Based on these data, a methodology for designing data quality control rules was developed.

Then, due to the scale of the data and the content of all the characteristic parameters, the ASD data source was selected and further used for building and learning a GAM model for anomaly detection in the data. At the same time, this data was used for the purpose of testing the functionality of the proposed procedures and rules.

A separate ASD dataset from 2023 was used to implement and apply the developed quality rules.

The data labeled as ASD comes from a network of strategic detectors of various technologies/types installed on the transportation infrastructure throughout the Czech Republic. These detectors continuously detect traffic flow parameters in a specific communication profile. Most detectors use a pair of induction loops in each lane of the monitored profile, or they may utilize non-intrusive technologies such as microwave detectors.

The ASD data contains a total of 14 individual parameters in its structure. For quality check, the most important identifiers of each site are the detector location loopcounterID - lane, counterID - carriageway and their relation to the driving profiles (geoportalID - not included in the primary dataset – has been added). Then the time interval parameter and the value parameters sumvehiclecounts (traffic volume) and avgvehiclespeed (traffic speed).

2.1.2 Accurity Quality application

Accurity Quality application was used for the design, implementation and subsequent testing and evaluation of data quality in the research. It is a commercial product of the project's main research partner, the company Simplity. A separate instance of this application was used, available to members of the research team and linked to the source data database. The application allows creating and searching all rules and related objects in an intuitive web-based environment. The creation of rules follows a general procedure and consists of a sequential creation of partial steps starting from the data structures, through the quality rules themselves to the different levels of aggregated measurements.

3. RESULTS

The research results are presented in 3 parts. First, the formulated requirements for data quality rules are described, which also form the methodology of their creation. Subsequently, the design of a validated procedure for the physical creation and implementation of data quality rules is described. The third part consists of the results of the data quality rules applied to the 2023 data of the selected motorway.

3.1 Requirements for the quality rules development process

The requirements for the process of creating data quality rules that came out of the detailed traffic data analysis are present-

ed in Figure 1. The individual procedures are presented in the form of a description of the desired rule, so that the principle of each rule is clear. Overall, the quality control process can be divided into several parts that build on each other.

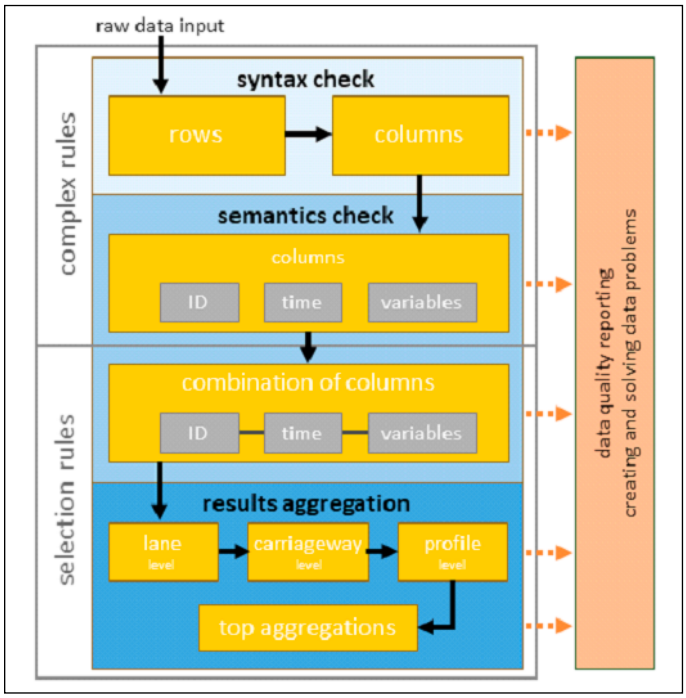

Figure 1. Logical Inspection Sequence for ASD Source Data (source: authors)

The primary check is the syntax check of the data source itself, followed by several other steps of checking the semantics of the data items. This is done first comprehensively for the whole dataset and then selectively, e.g. for individual specific traffic detectors.

Complex rules are those rules that provide comprehensive control of all data for a given source. Such rules are applied across the board, regardless of the different detector parameters, locations or interconnections.

Selection rules are rules that are applied separately to logically selected parts of the data. In the case of the ASD data source under investigation, these were mainly for individual measurement locations in the data, i.e. separately for unique identifiers in the data, *loopcounterID* or *counterID* columns.

Syntax check consists of those rules that check the content of individual data items in terms of their form, i.e. the expected assumptions on the type and number of characters, the structure of the character string, etc. The analysis of traffic data shows that syntax checking is applied only in the context of complex rules.

Semantics check consists of rules that in any way check the relevance or correctness of data in terms of its content. Semantics check follows procedurally after syntax checking, and from this perspective it can be assumed that the data already conforms to the expected formal requirements.

The overall procedure of applying the individual steps can be illustrated in the ASD data inspection diagram. The individual steps build on each other. This means, among other things, that no step can be implemented without implementing the previous step. Some steps may contain multiple parallel checks (e.g., column combinations) in a real procedure.

3.1.1 Rules for checking the acceptability of row processing

These rules check that each data record (row in the database) matches the expected structure of the data source. The input is a complete text string containing all the partial entries of the data record, i.e. all the data of the record are inserted in one single column. The data of a particular row are separated by semicolons ";" or commas ",", for example, and it is on the basis of the number of separating characters that the rule checks the number of columns in each data record. Records with a different number of columns are discarded from the next checking process.

The data quality tolerance in this case is the strictest possible (1 record) - i.e. the rule will not pass if there is only one wrong record in the whole data source.

3.1.2 Rules for syntactic checks

Rules for the syntax checking of a given source's data check the formalities. This check includes both an overall check of the data structure (number of columns) and a check of individual columns according to their data type.

An important principle of syntax checking in the whole data quality control process is that all records that are identified as incorrect are excluded from all further checking. This is based on the logic that if an entire data record or a single data record item does not conform to the requirements, the entire record loses credibility. It is not possible to deal correctly with checking the content relevance of the data. This principle implies in particular the requirement for a suitable definition of the data structures of the downstream rules.

3.1.3 Rules for checking individual columns by value type

The rules for syntactic checking of individual columns are applied for each data column separately. The input for applying the rules is data that has passed the previous check, i.e. does not contain records that did not pass the previous check. Before the necessary data structures are created, the data remain in the form of a text chain, but these are split into individual columns. Separate rules are also defined for these columns based on their types. This includes checking whether the content of a given column is, for example, text, an identifier of a given character range or a numeric value.

The data quality tolerance in this case is the strictest possible (1 record) - i.e. the rule will not pass if there is only one incorrect record in the whole data source.

3.1.4 Rules for semantic checking

The data semantics checking rules of a given transport source check the content relevance of the stored values. This part of the check is significantly more extensive than the syntax check. Also, at this stage of the check, erroneous entries are no longer excluded from further checking steps. Rules that "fail" based on defined requirements can be quantified and used for data quality reporting or also incorporated into subsequent aggregate measurements.

Semantic checking rules are applied both for complex checking of the entire range of a given data type across the entire data source, as well as for selective checks of specific detectors, etc.

The input to these rules is all the data in the data source from which those that did not satisfy the conditions of the previous syntax checking rules have been discarded. Further, when defining data structures, in practice, the data in each column are finally assigned the corresponding data types.

The data analyses gradually revealed the need to use several partial types of rules. These have a common principle resulting from the type of semantic meaning of the data under inspection. These are the following rule types:

- Checking according to a codebook consists in the existence of a codebook, i.e. a finite list of values admissible for the given parameter that is checked by this rule. Primarily, this type of rules is applied to columns in the form of an identifier or time information.

- Checking specific values allows you to create rules for specific data values (e.g. NULL), based on a fixed condition. Primarily this type of rules is applied to columns with continuous variables.
- Anomaly detection contains rules that do not have fixed parameters and limits set for data checking, but use a statistical model that takes into account the temporal behavior of the variables. Primarily, this type of rules is again applied to columns with continuous variables.

3.1.5 Aggregated rules

Aggregate checks or aggregate measurements allow you to combine 2 or more rules and calculate a new "aggregate" result based on their results (i.e. whether or not the rules passed). It is practical to structure aggregated rules in multiple levels in a meaningful way. This is partly because the results of already aggregated rules from lower levels can enter the aggregations at higher levels.

In the research design, the aggregated rules logic for the ASD resource was designed into a total of 4 levels. These levels correspond to the logic of a real detector arrangement and thus allow a clear and understandable interpretation. These levels are as follows:

- Level 1 – lane
- Level 2 – carriageway (direction of travel)
- Level 3 – road profile (both directions)
- Level 4 – top and other specific aggregations (geographical, technological, etc.)

3.2 Creation and implementation of data quality rules and other objects

The design and development of the individual rules themselves proved to be a key task for their implementation and application. It was desirable to design the whole process to be sustainable and as clear and efficient as possible. This is mainly due to the large number of rules and other related objects, but also to the relationships between them and their applicability to the overall quality assessment. Thus, the final design of the rule generation method includes an unambiguous structure and sequence resulting from the functions of the application used, as well as an unambiguous naming system for all rules.

Each rule is defined by a number of parameters that need to be clearly described in a given structure. Furthermore, the structure of the different types of rules is also specified, including the links between them. The condition for the implementation of a rule is thus the existence of their required inputs. The general structure of the rule types is shown in Figure 2. It shows that other aggregated measurements (AGR) can be used for aggregated measurements and furthermore Business Rules (BR) or Data Integrity Rules (DIR), for the implementation of which Basic Measurements (BM) or Data Integrity Analyses (DIA) are necessary. At the beginning of all defined rules is then the existence of Data Quality Views (DQV), which determine the selection of the range of data to be checked. Furthermore, Global Parameters (GP) are useful for defining rules.

During the problem solving process, the complexity of this task became more evident due to the total expected scope of all rules for each evaluated data source. It is not realistic to create and implement such a large number of rules manually (via the web interface). Automated preparation and then mass import of all rules, i.e. the ability to prepare and create rules outside the control application, proved to be a suitable way. In view of the number of rules, it was necessary to find and use a tool that would allow the bulk preparation of rules for their subsequent import. Given the needs, a spreadsheet processor was pragmatically chosen to enable this:

- efficient and clear automated rule design,
- direct relation of instances of individual object types via their name and other parameters
- minimizing errors when defining a large number of objects,
- preparing definitions for import into the web application

The selected tool allows to define individual objects in a structured way and to provide bindings in the rule structure. The benefit of using this tool is to minimize the number of manually entered parameters. The main links include the names of parameters (columns), the names of existing rules, data sources or global parameters, as well as the conditions of the rules themselves formulated in the form of SQL code.

Overall, automation of the creation of a comprehensive set of rules is achieved by using this tool. This ensures the efficiency, consistency and correctness of the created rules.

To achieve a functional and efficient way of creating and implementing rules, a well-designed naming system for all objects created is crucial. Three main aspects had to be taken into account: the uniqueness of the names within the whole ecosystem of quality rules (the names themselves can be considered as unique identifiers); the possibility of efficient search in the rules; the name is a carrier of information about the type and purpose of the rule. Thus, a prescribed structure has been created for rule names, using established abbreviations for rule types and data sources, as well as parameter names and specific identifiers.

3.3 Application of quality rules

A section of the D11 motorway was selected for verification of the implementation of the data quality rules. This motorway currently runs for a total length of 113 kilometers from Prague eastwards around Hradec Králové and then northwards towards Poland, where it ends for the moment at Jaroměř. There are currently 17 profiles with ASD detectors on the road, which contain a total of 68 individual monitored locations in the traffic lanes. The profiles in the first half of this motorway were used for testing purposes. An example of the location of the profiles, which are also the subject of the

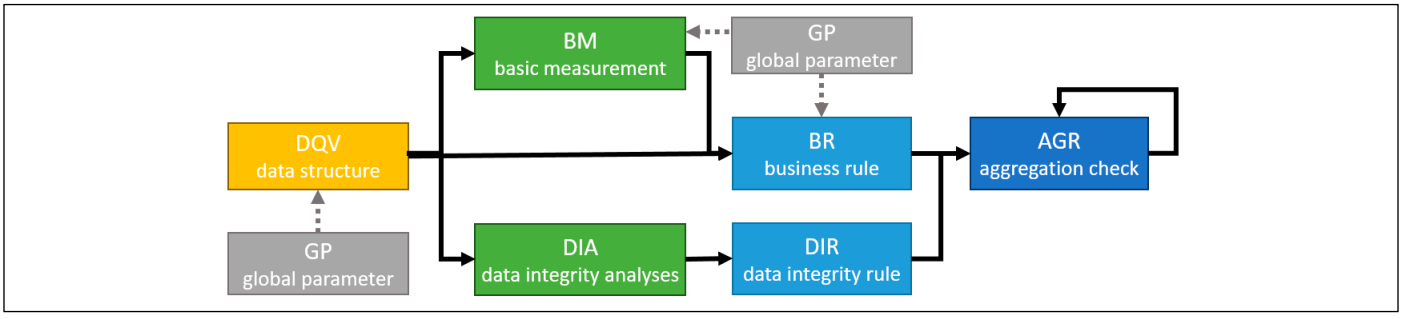

Figure 2. Relationship of functions and objects used for creating data quality rules (source: authors)

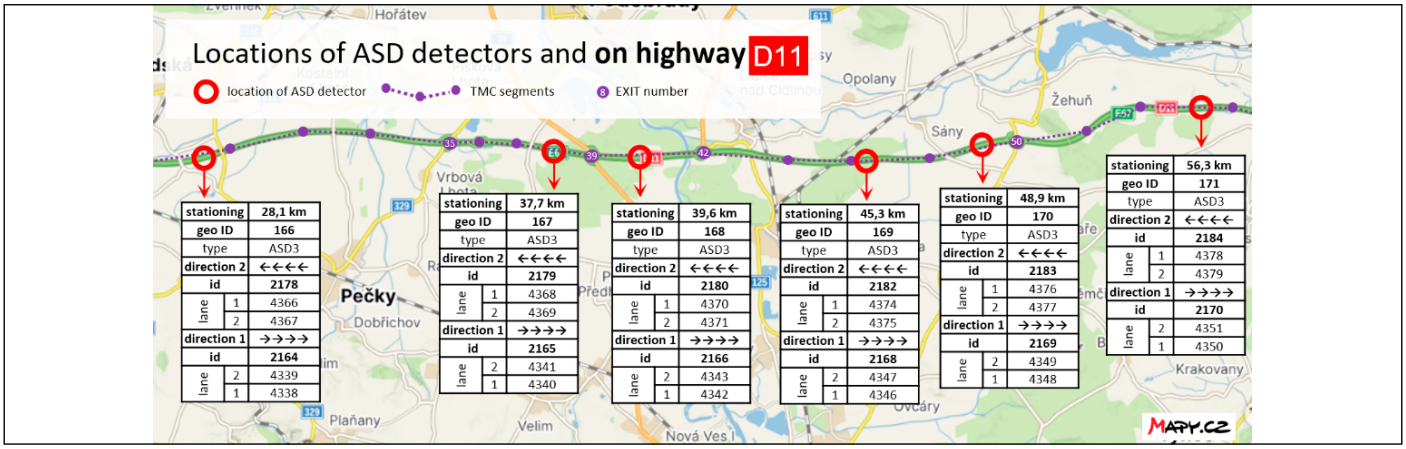

Figure 3. Example of locations of ASD detectors (source: authors)

presented results, and the carriageway and lane identification, is shown in Figure 3.

Then the data were prepared for the selected locations. This involved the formal preparation of the raw source data into 12 separate tables for each month of 2023. The data was then entered into a database that was accessed by the application.

Then, according to the procedures described above, the design, implementation and import of all the basic and downstream aggregation rules for the D11 motorway was done.

The design and implementation of rules in the Accurity Quality application for this motorway part created a total of 321 Aggregate Checks, 80 Data Integrity Analyses, and 200 Data Quality Rules. This makes a total of 601 rules and a number of other source objects. Using these rules, a data quality analysis was performed for each month of 2023 in sequence.

The complex results of this analysis are shown in the graphs in Figure 4, implemented by the Accurity Quality application. The graphs show the yearly evolution of the percentage success rate for the top aggregation rule of the whole motorway (first row) and similar rules for partial selected motorway profiles.

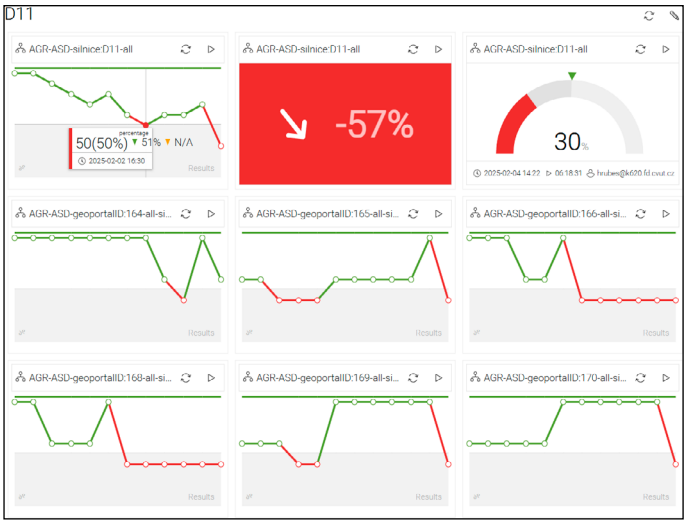

Figure 4. Top aggregate rules results – view for D11 motorway in 2023

We can see that in January and February the overall data quality was 100 percent according to the set parameters of the rules, then in the following months it decreases gradually up to 50 percent in August, then increases slightly up to 20 percent in November but falls to 30 percent in December. The second and third graphs show the trend and the final value of the last measurement, namely December 2023. The remaining 6 graphs show the evolution of the data quality of the individual motorway profiles.

The overall data quality scores are the output of the aggregated rules. Their positive or negative result depends on the partial results of the input rules. Thus, the summaries of these results were also evaluated to verify the functionality. With these, the overall results can be further interpreted and analyzed. For the selection view of the individual detectors, the results concerning the parameters of the number of vehicles (*sumvehiclecount*) and the average speed of the vehicles (*avgvehiclespeed*) were evaluated. Specifically, the evaluation of completeness (existence of time series); the number of "null" values (there is real – numerical content within the time series); and the number of values detected as anomalies (outliers from the expected values).

Table I gives an overview of the completeness and number of "null" values for each detector during the year. The results for both the number of vehicles and their speed parameters proved to be identical. For profile 166, the drop in completeness in August and the complete loss of data in the following 3 months are particularly significant. Other minor decreases in completeness are also evident. Their occurrence in a given month for several profiles is characteristic. Checking the number of "null" values for these detectors did not reveal any occurrence in the data, so this check is not behind the decrease in overall data quality in any of the profiles.

The results of the number of anomalies detected are shown in Table II. The anomalies, that is values outside the interval defined by the statistical model, are different and shown separately for the number of vehicles and their speeds. The anomalies are detected for the carriageways of the road. The records in the table, shown in orange and red, indicate those situations where the overall data quality was reduced, the threshold for evaluating the pass/fail quality control rules was set at 50 records. It is also useful to remind that the total number of records in a month, depending on the number of days, is 672 - 744 (1 record every hour).

4. DISCUSSION AND CONCLUSION

In relation to the defined research objectives, we discuss their achievement here. RO1 was fulfilled by the design and development of a rulemaking methodology. Its suitability was verified by their subsequent creation and application. Another objective RO2 was achieved by creating a rule generation and management tool in the form of a spreadsheet. In addition to bulk rule creation and import, the tool allows for the preparation of data for other applications or ways of quality control. Another RO3 objective was achieved by successfully imported rules into the Accurity Quality application and verified their

Traffic profiles, carriageways and lanes of the D11 highway			months January to December 2023																					
			completeness of transport variables												number of "null" values									
geoportalID	counterID	loopcounterID	1	2	3	4	5	6	7	8	9	10	11	12	1	2	3	4	5	6	7	8	9	10	11	12
164	2162	4334	100	100	100	100	100	100	100	100	100	100	100	100	0	0	0	0	0	0	0	0	0	0	0	0
		4335	100	100	100	100	100	100	100	100	100	100	100	100	0	0	0	0	0	0	0	0	0	0	0	0
	2176	4362	100	100	100	100	100	100	100	100	100	100	100	100	0	0	0	0	0	0	0	0	0	0	0	0
		4363	100	100	100	100	100	100	100	100	100	100	100	100	0	0	0	0	0	0	0	0	0	0	0	0
165	2163	4336	100	100	100	100	100	100	100	100	100	99	100	100	0	0	0	0	0	0	0	0	0	0	0	0
		4337	100	100	100	100	100	100	100	100	100	99	100	100	0	0	0	0	0	0	0	0	0	0	0	0
	2177	4364	100	100	100	100	100	100	100	100	100	99	100	100	0	0	0	0	0	0	0	0	0	0	0	0
		4365	100	100	100	100	100	100	100	100	100	99	100	100	0	0	0	0	0	0	0	0	0	0	0	0
166	2178	4366	100	100	100	100	100	100	100	75	0	0	0	100	0	0	0	0	0	0	0	0	0	0	0	0
		4367	100	100	100	100	100	100	100	75	0	0	0	100	0	0	0	0	0	0	0	0	0	0	0	0
168	2166	4342	100	100	100	100	100	97	100	98	92	100	100	100	0	0	0	0	0	0	0	0	0	0	0	0
		4343	100	100	100	100	100	97	100	98	92	100	100	100	0	0	0	0	0	0	0	0	0	0	0	0
	2180	4370	100	100	100	100	100	97	100	98	92	100	100	100	0	0	0	0	0	0	0	0	0	0	0	0
		4371	100	100	100	100	100	97	100	98	92	100	100	100	0	0	0	0	0	0	0	0	0	0	0	0
169	2168	4346	74	100	100	100	100	99	100	100	99	100	100	100	0	0	0	0	0	0	0	0	0	0	0	0
		4347	74	100	100	100	100	99	100	100	99	100	100	100	0	0	0	0	0	0	0	0	0	0	0	0
	2182	4374	74	100	100	100	100	99	100	100	99	100	100	100	0	0	0	0	0	0	0	0	0	0	0	0
		4375	74	100	100	100	100	99	100	100	99	100	100	100	0	0	0	0	0	0	0	0	0	0	0	0
170	2169	4348	100	100	100	100	100	99	100	100	99	100	100	100	0	0	0	0	0	0	0	0	0	0	0	0
		4349	100	100	100	100	100	99	100	100	99	100	100	100	0	0	0	0	0	0	0	0	0	0	0	0
	2183	4376	100	100	100	100	100	99	100	100	99	100	100	100	0	0	0	0	0	0	0	0	0	0	0	0
		4377	100	100	100	100	100	99	100	100	99	100	100	100	0	0	0	0	0	0	0	0	0	0	0	0
171	2170	4350	100	100	100	100	100	100	100	100	100	100	100	100	0	0	0	0	0	0	0	0	0	0	0	0
		4351	100	100	100	100	100	100	100	100	100	100	100	100	0	0	0	0	0	0	0	0	0	0	0	0
	2184	7378	100	100	100	100	100	100	100	100	100	100	100	100	0	0	0	0	0	0	0	0	0	0	0	0
		7379	100	100	100	100	100	100	100	100	100	100	100	100	0	0	0	0	0	0	0	0	0	0	0	0

Table 1. Completeness and "null" values – d11 motorway in 2023

Traffic profiles and carriageways of the D11 highway		months January to December 2023																							
		number of anomalies in sumvehiclecount values												number of anomalies in avgvehiclespeed values											
geoportalID	counterID	1	2	3	4	5	6	7	8	9	10	11	12	1	2	3	4	5	6	7	8	9	10	11	12
164	2162	11	2	26	42	31	21	29	8	79	566	21	360	15	1	4	11	9	8	13	19	16	110	28	47
	2176	5	0	1	22	28	19	31	13	81	562	13	307	13	2	33	31	39	22	17	22	42	108	30	46
165	2163	44	49	79	78	57	47	41	26	6	4	23	117	738	670	740	720	742	720	741	742	695	248	21	394
	2177	6	2	7	23	30	18	26	14	4	3	2	48	25	3	15	24	22	15	21	38	13	29	43	130
166	2164	14	6	31	59	55	26	136	351	0	0	0	110	8	2	9	6	8	2	298	555	0	0	0	60
	2178	0	0	0	0	0	0	0	0	0	0	0	0	11	1	5	10	28	8	342	482	0	0	0	91
168	2166	16	4	38	52	37	25	490	638	588	637	635	525	14	2	4	19	21	43	160	82	121	196	276	258
	2180	18	3	51	47	37	24	35	26	10	4	9	84	14	6	4	16	102	8	30	16	3	26	21	91
169	2168	11	4	49	73	56	27	30	12	17	12	13	144	536	672	737	719	531	0	0	0	1	0	0	341
	2182	4	0	5	20	25	17	8	5	1	0	2	42	542	672	733	714	611	9	0	0	0	0	0	449
170	2169	12	2	14	39	40	26	16	12	10	11	11	94	509	635	701	680	503	14	20	9	15	7	6	194
	2183	8	0	2	19	17	16	9	7	1	1	3	36	385	637	679	655	447	45	28	19	6	8	8	125
171	2170	18	9	4	12	126	429	447	427	418	442	473	501	28	14	13	19	28	76	77	97	109	92	145	135
	2184	6	1	5	22	24	12	12	8	3	0	8	47	174	458	244	226	199	11	9	29	8	6	22	113

Table 2. Anomalies in data – D11 motorway in 2023

runnability. The key objective RO4 was completed by applying all the rules to data from each month of 2023 for the selected D11 motorway in sequence. The outputs of the application show the evolution of the overall data quality of this motorway over time and allow the status and causes of reduced quality to be analyzed in the form of partial analyses.

According to the overall results, the data quality on the D11 motorway has a decreasing trend in 2023. This evaluation is consistent with the results of the individual profiles. These have different characteristics depending on their geographical location within the whole motorway. Some of the drops in quality can be interpreted by possible planned action along the route (e.g. profiles 166 and 168) for others further detailed analysis would be appropriate.

The analysis of the results of the completeness assessment shows possible reasons for the reduction of data quality in the form of partial or complete data failure. However, data completeness alone is not a key aspect of the data quality evaluation during 2023.

It shows that the detection of the number of anomalies in the data has a more significant impact on the overall quality. In this context, it needs to be further investigated whether the anomalies detected are real errors in the data, may reflect an imperfection in the model (model ageing), or are a standard traffic situation that the model is unable to detect.

The outputs of Table II show that the model confirm its quality even on data on which it was not trained, and it represents very well the evolution of the traffic variables speed and intensity. In most cases, a low number of anomalies are detected, which are due to common traffic anomalies. The results also change significantly during the year. This is demonstrated by a more detailed analysis of the data trends in selected months. The graphs in Figure 5 show an example for March 2023 in one direction of profile 169, where there is a high number of anomalies detected for the speed parameter.

The graph shows some typical situations. The flow of the number of vehicles corresponds to the periodicity and character of the model and only 5 values outside the model limits are detected. In the case of speed, on the other hand, there is a constant overstepping of the maximum of the model limit, which, however, is not realistically justified (average hourly

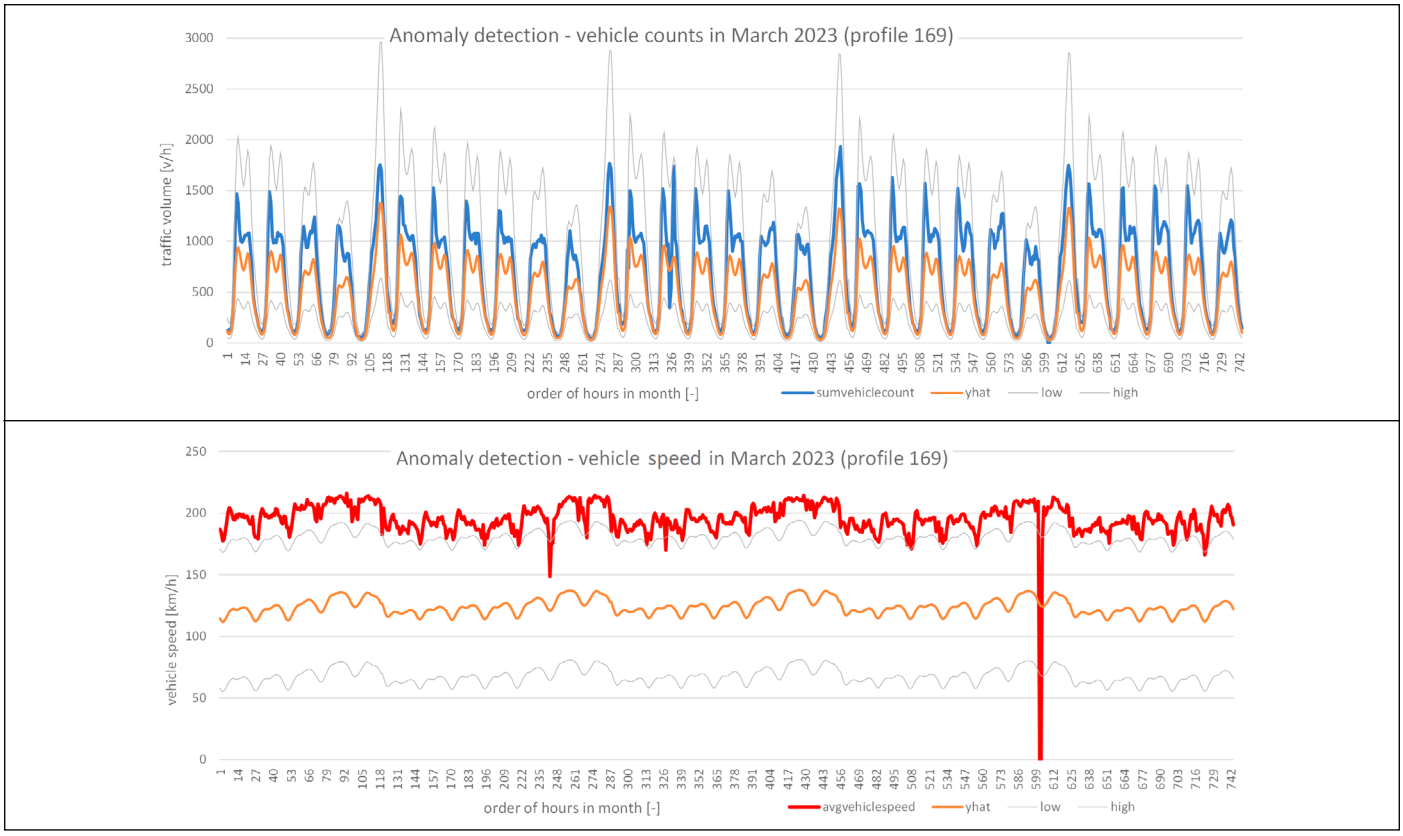

Figure 5. Anomaly check run for one profile direction 169 in March 2023

speeds exceeding 200 km/h). This is therefore not a model error or a standard traffic situation, but very probably a problem in the measured data. It is also noticeable (see Table II) that there has been a change in the data since June and this type of anomaly is no longer detected.

All of the reported results and evaluation were achieved based on the rules set by the expert view of the researchers. Therefore, it is also advisable to carefully discuss all the parameters with the operational staff of the data owner according to their real needs and experience to see if the rules we have introduced are set too strictly.

An actual quality evaluation demonstration was performed and presented for the one motorway. However, the rules in the tool were created for the entire geographic data coverage, so it is possible to extend the implementation to the entire Czech Republic and other data sources.

From the implementation experience, we see the need to make some modifications or additions that would further improve the global deployment. In particular, the following partial modifications:

- revision of the setting of threshold limits for different types of rules;
- implementing back-checking (e.g. of imported rules)
- discuss a ways of presenting the results of the quality assessment in Accurity Quality (according to the requirements of the data owner).

ACKNOWLEDGEMENTS

The research was realized within the research project (CK04000189/ Data quality tools for ensuring system reliability of transport information centers), which was financed with the state support of the Technology Agency of the Czech Republic within the Transport 2020+ program.

The authors acknowledge the partner, Road and Motorway Directorate of the Czech Republic, state-owned enterprise, namely Ing. Filip Týc, for providing the data and for cooperation in the analysis and discussions.

The authors also acknowledge their partner the company Simplity, namely Ing. Tereza Slováková for providing the Accurity Quality application and support in its use.

REFERENCES

Hrubeš P., Langr M., Purkrábková Z. (2024). Review of Data Governance Approaches in the Field of Transportation Domain, Smart City Symposium Prague – IEEE proceedings, Prague, Czech Republic, 2024, pp. 1-7, doi: 10.1109/SCSP61506.2024.10552682.

Mlynářová T., Hrubeš P. et al.(2023). Metodika přístupu k datové kvalitě, Research Report (in czech). Prague.

Purkrábková Z., Langr M., Hrubeš P., Brabec M. (2024). Data Governance in Traffic Data: Anomaly Detection with Generalized Additive Models, Neural Network World, 2024, pp. 203-218

Purkrábková Z., Langr M., Hrubeš P., Brabec M. (2025). Detecting anomalies in traffic data using a flexible semi-parametric model, European Transport Research Review 2025, to be publish