DATA GOVERNANCE IN TRAFFIC DATA: ANOMALY DETECTION WITH GENERALIZED ADDITIVE MODELS

Z. Purkrábková[*], *M. Langr*[*], *P. Hrubeš*[*], *M. Brabec*[†]

Abstract: The primary objective of the presented research is to enhance an existing data quality control application by integrating advanced anomaly detection mechanisms based on generalized additive models. This approach targets time-series traffic data, where traditional methods may fall short in identifying complex, non-linear patterns of anomalies. In collaboration with Simplity s.r.o., we are extending their current data quality assessment tool to incorporate generalized additive models, providing a more robust and dynamic solution for monitoring and ensuring the reliability of traffic datasets. The integration of these models aims to improve the accuracy of anomaly detection, leading to more effective data management in transport systems and contributing to higher standards of data quality in the field of traffic informatics.

Key words: *traffic anomalies, data quality, data application, generalized additive model, data governance*

Received: June 15, 2024
Revised and accepted: August 29, 2024

DOI: 10.14311/NNW.2024.34.011

1. Introduction

Quality public information services depend on precise data about the state and performance of the transportation system. A major challenge is the integration of data from multiple sources into coherent datasets, which is crucial for effective traffic management and reliable passenger information services. For systems that automatically provide passenger updates based on traffic monitoring inputs, data accuracy is vital. High-quality data is necessary for making sound decisions, and developing effective automated techniques is essential to prevent the dissemination of incorrect or misleading information.

In the realm of Intelligent Transportation Systems (ITS), the system environment has grown increasingly complex compared to earlier decades when specialized

[*]Zuzana Purkrábková; Martin Langr; Pavel Hrubeš; Department of Transport Telematics, Czech Technical University in Prague, Faculty of Transportation Sciences, Konviktská 20, CZ-110 00 Praha 1, Czech Republic, E-mail: purkrzuz@fd.cvut.cz, langr@k620.fd.cvut.cz, pavel.hrubes@cvut.cz

[†]Marek Brabec; Institute of Computer Science, The Czech Academy of Sciences, Pod Vodárenskou věží 271/2, CZ-182 07 Praha 8, Czech Republic, E-mail: mbrabec@cs.cas.cz

©CTU FTS 2024

technologies were more common. There is now a broader range of concepts such as Mobility as a Service (MaaS), emerging technologies, micromobility, IoT solutions, new communication methods, and advanced data processing techniques. As a result, the diversity in traffic data, both in structure and meaning, has greatly expanded. This variability is also reflected in the different ways data is processed across various spatial and temporal scenarios, which calls for better data linking, integration, and fusion within control systems and user applications.

With ongoing digitalization, there is a growing need to standardize processes and information. To meet this need, a framework known as data governance has been developed. Originally prominent in sectors with high trust requirements, like banking, data governance is a strategic approach that establishes roles, rules, procedures, and best practices to ensure the proper use, security, and quality of data. It provides guidance on setting up control mechanisms that promote an organized and effective use of data within organizations. The accompanying diagram (Fig. 1) outlines the key issues addressed by data governance.

Fig. 1 *Diagram illustrating the data governance procedure [1].*

Various components of metadata governance are highlighted in pink, while the elements related to reference or static data governance are shown in yellow. Red is used to represent remediation efforts and data cleansing activities. The blue block emphasizes key aspects of data understanding within the broader framework of data governance. Data quality governance components are differentiated in green. For this article, this green part is very important. This particular section of the data governance model details the procedures for developing and implementing data quality standards.

The data quality methodology formally classifies the standards into different quality dimensions, addressing both technical and substantive expectations for data quality. These dimensions include validity, consistency, completeness, uniqueness, timeliness, and accuracy. Tolerance thresholds are applied to evaluate the data's

compliance with these standards, ensuring adherence to the set quality benchmarks. This approach enables tracking of non-compliant data and observing its progress over time or across different datasets.

The paper introduces an application designed for assessing the quality of traffic data. It begins by describing the sources of the traffic data used in the study. Next, it details the methodological approach and the statistical model employed for detecting anomalies. The following section explains the data processing steps and the approach used in the application. Finally, the paper presents the results of the study and suggests potential avenues for future research.

2. State of the Art

The importance of data quality control is not a new concept. However, it remains under-explored in the scientific community when it comes to traffic data. The broader topics of data quality and management have been extensively discussed by various authors for example [2] or [3]. The quality of traffic data is very often addressed by authors in terms of one of the data quality assessment parameters such as completeness in accident data [4] or in data from detectors [5], consistency in accident data [6], timeliness in obtaining vehicle localization data [7], validity in term of autonomous systems [8] etc.

Unfortunately, no comprehensive research has been found that specifically addresses traffic data in general or data quality processing in the transportation field. The research [9] examines very similar data sources but focuses solely on data processing without considering data quality aspects. The authors [10] addressed related data in their study, focusing on Czech data and its adequate quality for assessing traffic flow categorization and modeling vehicle emissions estimation. The authors [11] work with spatial and temporal traffic data in traffic prediction; however, they only mention the large amount of diverse data needed for autonomous systems and do not further discuss its quality.

Traffic data mining is featured in several studies. Research [12] utilizes extensive datasets from Italian highways to perform cluster analysis aimed at estimating the Annual Average Daily Traffic. The authors [13] discuss the use of automated sensor data (mainly travel time and traffic density estimation) in India for predicting traffic conditions in Indian traffic conditions. They use k-nearest neighbor and artificial neural network for prediction. Data mining over data from the Czech Republic was the focus of [14] research, where they ran cluster analysis over big data. However, this research did not use traffic data.

When examining anomaly detection, we encounter several studies that have already been conducted. The study [15] utilized Kernel Density Estimation for detecting anomalies in the density-flow relationship. This approach was tested on data from an English motorway. Research [16] deals with outlier detection approaches in urban traffic analysis in case of flow outliers and also trajectory outliers. The authors [17] use a computational data science approach to detect anomalies in traffic data from traffic detectors for autonomous vehicles. They highlight that combining data science with advanced artificial intelligence techniques provides a higher level of anomaly detection. As a result, it becomes possible to reduce congestion and traffic incidents.

The research [18] deals with surveillance-related research on anomaly detection in public places, mainly on the road. The authors analyze more vision-guided anomaly detection techniques and also describe gaps in the available datasets and anomaly detection capability.

The use of the generalized additive model (GAM) framework in traffic data analysis is not uncommon. For instance, [19] apply GAM to assess the safety of connected vehicles in a pilot project in Wyoming. Their study compares the generalized additive model with generalized linear and nonlinear models, demonstrating that GAM provides better insights into crash patterns along corridors. Similarly, paper [20] uses the generalized additive model to study the varying levels of autonomous vehicle integration into traffic. Thanks to GAM, the authors were able to effectively model the macroscopic fundamental diagram with different levels of autonomous vehicle involvement.

A study by [21] deals with somewhat different data, focusing on meteorological information and emission measurements during the COVID-19 lockdown in Beijing. This research employs the generalized additive model to distinguish the effects of lockdowns from the impact of weather on concentrations of nitrogen dioxide and fine particulate matter in the city. Another emission-related study by [22] uses the GAM model to investigate the causes of spatial heterogeneity in traffic emissions. The authors leverage GAM's ability to characterize the functional relationships between vehicle emissions and urban features relevant to city planning.

It is evident that all areas of study, data quality control, anomaly detection, and the application of generalized additive models, have already found use in transportation research. However, there is a notable gap in the literature when it comes to combining these approaches. Few studies focus on the integration of data quality assessment with advanced anomaly detection methods like GAM in traffic data analysis. This research aims to bridge that gap, offering a novel approach that leverages both data quality control and generalized additive models-based anomaly detection to enhance the reliability and insightfulness of traffic data analysis.

3. Data

The main goal of this research is to extend the quality control application with the possibility of control mechanisms using generalized additive models to detect anomalies in time series traffic data. The initiative focuses on improving data quality within the broader context of data governance. The dataset in question includes information from the main road network in the Czech Republic, covering the years 2021 and 2022. These datasets represent detailed collections of data in their original format, obtained by the road and motorway directorate (RSD) using various technologies and systems. The primary data sources are automatic traffic counters (ASD) and floating car data (FCD).

ASD data is gathered from a network of strategically placed detectors, utilizing various technologies, installed across the Czech Republic's transportation infrastructure. These detectors continuously track traffic flow parameters at specific points on the road. Most detectors operate with pairs of induction loops in each lane, while some use non-intrusive technologies like microwave sensors. The data includes identifiers for road profiles and lanes, along with key traffic parameters,

such as traffic volume and speed, measured at specific time intervals for different vehicle categories. This is supplemented by the total vehicle count and their average speed. The time intervals (either 5 or 60 minutes) and the number of vehicle categories vary depending on the detector technology. Some older detectors do not capture speed information. ASD data is not collected in real-time but is typically updated once a month, meaning it is not available online and is primarily used for RSD's internal purposes. Currently, it is not provided as open data.

FCD data, on the other hand, provides information on traffic conditions for individual road sections (TMC segments), based on data from floating vehicles. The source comprises data from over 100,000 vehicles in the Czech Republic, at least 75% of which are passenger cars. FCD data includes traffic flow speed parameters and derived qualitative metrics, such as travel time, delay, traffic quality, and congestion detection for each segment. However, FCD does not offer quantitative data on total traffic volume. This data is produced in real-time for all road segments, in line with the latest TMC location tables. The current real-time data (excluding historical datasets) is accessible through the Czech Transport Data Register portal [23]. A more detailed description of both data sources can be found in earlier works [24].

Other data is also available, such as weighing data while driving or parking at rest areas. These data will not be considered in the paper.

4. Methodology

The data quality control methodology has clearly defined sequential steps that must be followed. The steps are also described on the workflow (Fig. 2), where the ones that are affected by the following description are highlighted.

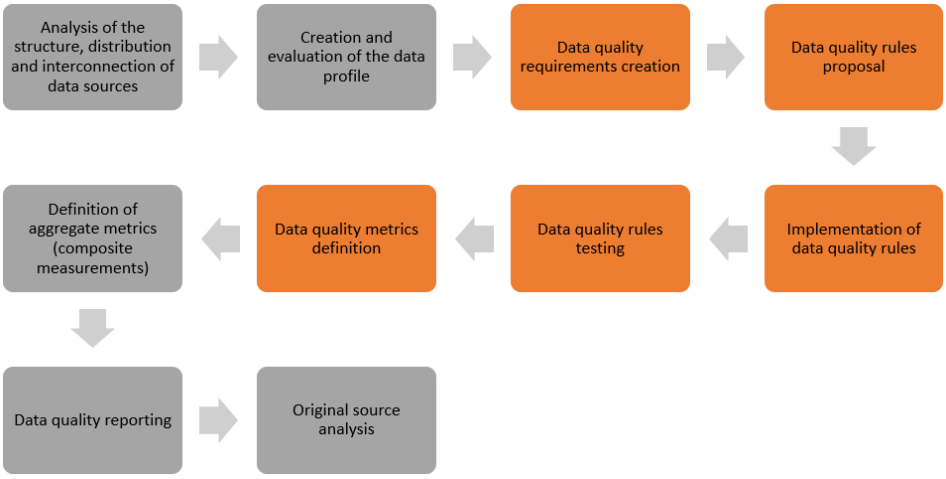

Fig. 2 *Workflow diagram of data governance process, steps related in the application shown in orange (Source: authors).*

The initial step in the process involves analyzing the structure of the data, which is where the six dimensions of data quality are applied to evaluate the raw dataset. At this stage, it is crucial to address foundational data quality issues by removing duplicates, invalid entries, and meaningless values. By thoroughly cleaning and preparing the data from the outset, we ensure that the inputs to subsequent models are reliable and suitable for analysis. This foundational step is essential for achieving a deeper level of data quality and enhancing the accuracy and credibility of the overall analysis. This step is discussed in greater detail in the article [24].

The next phase leverages the knowledge of the dataset to develop fundamental data validation rules, addressing both syntactic and semantic aspects. These rules are implemented within the web application, as the datasets have become sufficiently large and complex to necessitate automated validation processes. Following the initial validation, more advanced data quality rules are created, with cross-implementation ensuring a comprehensive approach. The outcome of this process is the identification of data points requiring further investigation, alongside alerts for datasets with a high percentage of failed validations. This systematic approach enhances the ability to maintain data quality at scale, ensuring reliability and actionable insights.

The presented objective of this work is to design an overall data quality control procedure and to provide a detailed specification of each individual step. This includes the development of an internal methodology for the design, creation, and management of quality rules and their implementation into the utilized application. Methodologically, these activities must be fully aligned with data governance practices, the actual structure and characteristics of the data used, their technological sources (detectors), and the functional capabilities of the application in use. A key activity is the design of a model for data anomaly detection and its integration into the application.

These steps directly support the primary objective of the application, which is anomaly detection. This is facilitated by a specialized, designed tool integrated directly into the web application. The tool employs advanced algorithms to identify deviations from expected patterns, allowing for real-time detection of anomalies within the data.

The statistical approach in the presented methodology is grounded in the Generalized Additive Model (GAM) framework [25], [26]. This framework supports modular modeling, enabling the model to consist of several easily interpretable components. For instance, the model can include a flexible, nonparametric decomposition of a traffic characteristic time series into components such as long-term (inter-annual) trends, annual periodic patterns, and weekly periodic patterns. However, various other configurations can be used to tailor the model to meet specific practical needs in traffic data quality control. In our case, we model the data at an hourly resolution, addressing the issue of incompatible time resolutions in the floating vehicle and counter data, which is analogous to the Modifiable Areal Unit Problem (MAUP) [27].

The potential of GAM modeling is demonstrated through a specific model formulation, where the explanatory variables can be easily adjusted to implement data quality procedures, such as checking consistency between consecutive sensors

or between a sensor and floating vehicle data.

$$Y_t = \mu + \sum_s \alpha_s I(day\ t\ is\ in\ year\ s) + s_{sea}(day\_within\_year(t)) \\ + s_{HT}(hour\_within\_week(t)) + \varepsilon_t,$$

where Y_t is a traffic characteristic variable from a given location at time t, μ is an unknown location parameter to be estimated from empirical data. α_s is an effect of calendar year s (we have S years of data). For identifiability reasons common in ANOVA-like (sub-)models [28], we use the so-called treatment contrast. *I(.)* is an indicator function (the value of 1 if its argument is true and 0 otherwise). s_{sea} is an unknown smooth periodic function (i.e. a "functional" parameter) to be estimated from the data. s_{HT} is an unknown smooth periodic function to be estimated from data. ε_t is a random (identically, independently distributed across time t) variable with working gaussianity assumption, namely $\varepsilon_t\ N(0, \sigma^2)$.

5. Results

As stated above, the basic analysis of the data source from the perspective of data quality dimensions was the subject of previous activities and related publications. This made it possible to design a detailed procedure for further activities, verify their functionality, and effectively implement them into the available data governance tool.

The procedure includes the following steps:

- Analysis of data structure and attributes, along with the preparation of auxiliary variables and definitions
  - Definition of data structures
  - Definition of global variables
- Verification of the structure and content of individual attributes (columns) in the data source
  - Data integrity check-comparing data content against defined reference tables
  - Data anomaly detection: identifying data entries that fall outside the range of values established by the statistical model
- Defining and executing data quality rules: evaluating data quality
  - Specific rules for different parameters and data sets
    - Utilize performed data integrity checks and anomaly detection
    - Additional independently created rules (e.g., data completeness or specific value content, etc.)
  - Aggregation of data quality rules across multiple levels

  - Selection of data quality rules corresponding to similar types, detectors, and locations
  - Assignment of weights to individual rules
  - Aggregation of rules based on their mutual associations and assigned weights

- Overall data quality assessment according to aggregations at the lane level, lane segment, roadway profile, and also, for example, by technology type, route, or region

The individual steps are implemented in SQL within an application used by the authors as part of a research project. It is an online application hosted on servers running the PostgreSQL system. This approach handles data at the server level, minimizing the load on the application itself. Additionally, all procedures and guidelines for creating the necessary SQL code to establish data structures, rules, and associated variables are documented in the project's internal methodology, making them broadly applicable. This also enables further work, modifications, expansions, or application to other data sources.

To efficiently create all rules and related structures and parameters, it was essential to develop a dedicated methodology for their description. This approach helps reduce the complexity caused by the increasing number of rules. Each generally created rule consists of multiple components that must be uniquely identifiable. Many of these rules are then applied to hundreds of specific detectors, significantly increasing the total number of rules. The proposed approach was also chosen to facilitate the search and filtering of individual rules.

The label for each rule always begins with an abbreviation defining the specific step (e.g., data structure, integrity rule, anomaly detection, business rule, etc.), followed by the dataset designation (ASD or FCD) and an indication of the relevant column. An example label might be *DIR-ASD-13-starttime-data-ciselnik-right*, where "DIR" denotes a data integrity rule, "ASD" indicates strategic detectors, "13" refers to the thirteenth column, and a brief description or name follows.

Along with developing a unique methodology for labeling and describing rules, a procedure and tool were designed for automatically generating names and SQL code to enable the bulk and automated creation of all data rules and their subsequent import into the application.

The first step is defining the data structures, which involves selecting only those parameters (columns) or data records that will be subject to checks in subsequent steps. The goal is to optimize data handling by avoiding unnecessary loading and processing of data that the given rule does not pertain to.

Given the growing number of basic rules, it was essential to define so-called global parameters. These parameters function as predefined variables used in the creation of data quality control rules, enabling efficient and repeated use of pre-prepared code segments when defining rules. An example of a defined global variable within the application is shown in Fig. 3. This variable is then applied within the data structure definition. By using global variables and applying negation, it's also possible to use a rule that detects specific data errors to define a data structure that excludes such erroneous data.

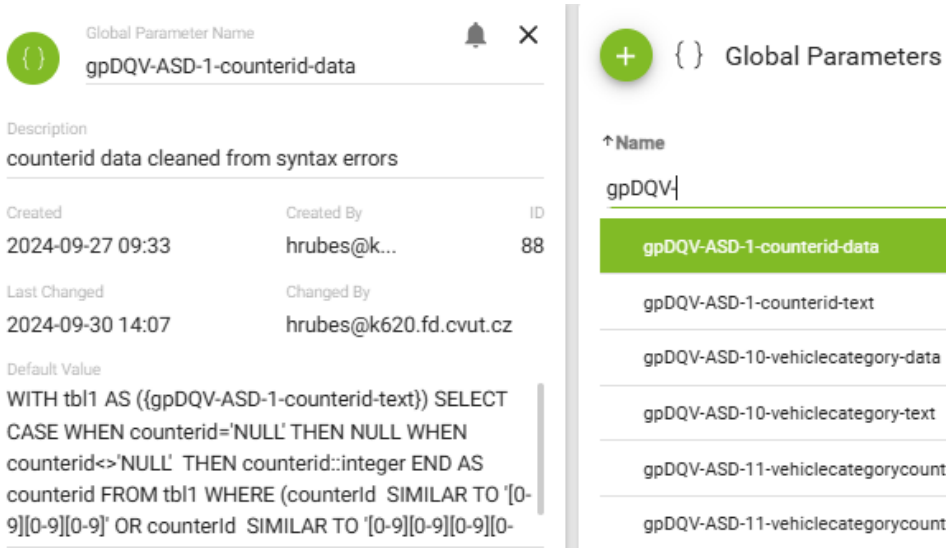

Fig. 3 *Definition of global parameters [Source: authors].*

These global variables were subsequently used to define control rules (known as "provide") that focus on validating the data structure. Typically, this involves ensuring that the data does not contain nonsensical values from a syntactic perspective and that the columns in the datasets are correctly assigned, among other things. This configuration is illustrated in the following Fig. 4.

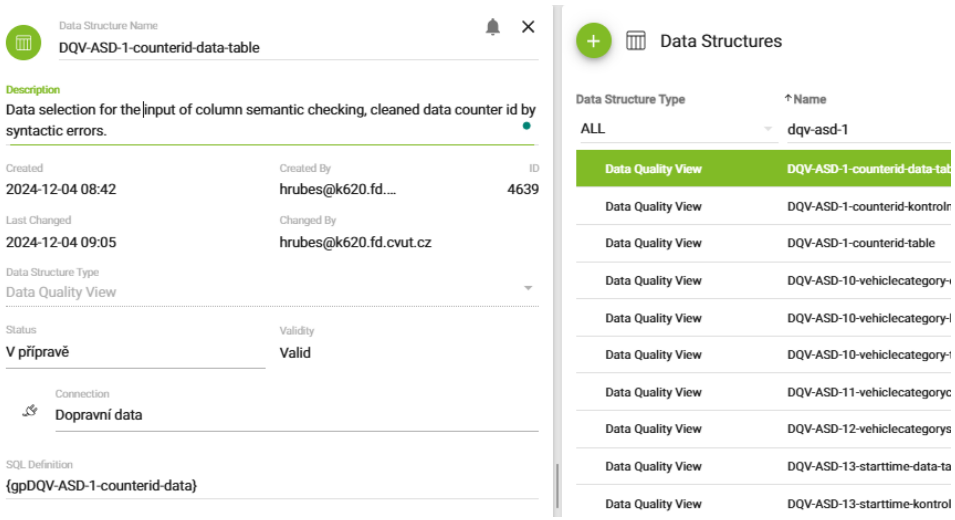

Fig. 4 *Using of global parameters in data structure definition [Source: authors].*

The next step is integrity analysis, during which all data is compared with predefined reference tables. The result is a cross-analysis of these two datasets,

determining matching and differing values and quantifying their percentage representation, as demonstrated in Fig. 5. This step is crucial for identifying inconsistent values, as data that do not meet integrity criteria will not be processed further. Therefore, it is important to quantify the percentage of the data that will be excluded from further analysis through this process. The integrity analysis is primarily intended to check the syntax of the data, ensuring that the contained information is formally correct, takes on specific expected values, or has logically corresponding relationships among different identifiers.

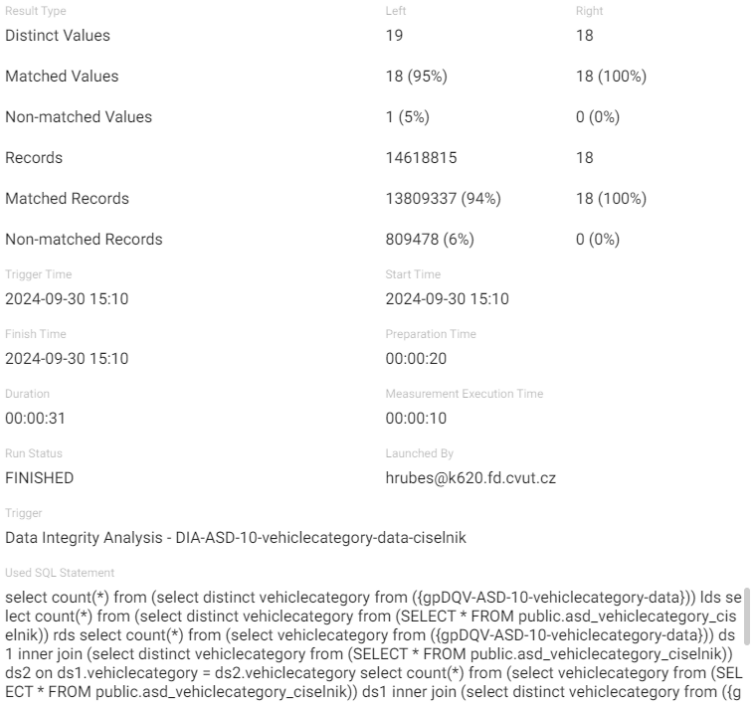

Fig. 5 *Creation of data integrity analysis [Source: authors].*

For possible semantic data checks, a statistical model has been created and implemented for anomaly detection in the data, as described in the previous methodological section. The output of this model consists of three vectors representing the average value, lower tolerance limit, and upper tolerance limit, corresponding to the time intervals of the available data source. If the value contained in the data falls outside the range established by the model, it is recorded as a data anomaly. Visualization of this process is provided in Fig. 6. If it is found that the measured value exceeds the defined tolerances, further analysis of this data is necessary, as it may not represent a genuine anomaly indicating an error in the data or the detector's function, but rather a real traffic condition. The next step involves defining the data quality rules themselves. The basic type is known as business rules, as illustrated in Fig. 7. The principle of these rules is to determine how many values from the defined data structure meet the specified rule and how many do not. Other

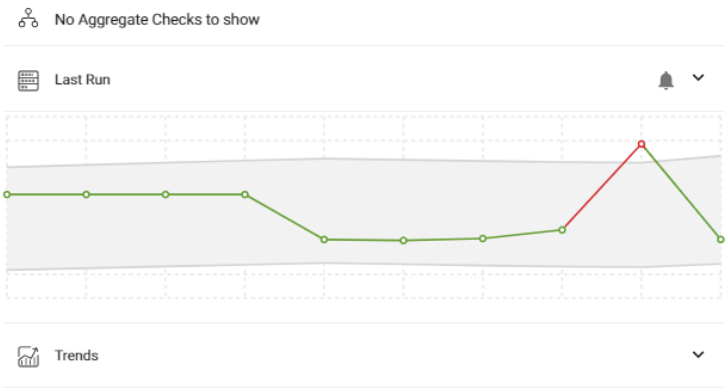

Fig. 6 *Anomaly detection result display in the application [Source: authors].*

specific rules build on the previous integrity checks or anomaly detections. Within all quality rules, establishing criteria for the tolerability of samples is crucial. Here, it is determined what percentage or number of samples can be considered acceptable for successfully passing the check and designating the sample as valid. These criteria must be precisely set according to the specific requirements of the dataset, as shown in Fig. 8. With appropriate settings of these criteria, it becomes possible to analyze trends in the results of checks when repeated, as well as whether the rule was satisfied or not, which factors into the subsequent steps of aggregating quality rules. The summary of all created quality rules forms a very broad base of data regarding data quality. However, each rule by itself only provides information about a specific parameter or a particular detector. To evaluate data quality as a whole, the final phase is necessary, known as the aggregation of rules (aggregate checks). The principle of this step is to logically group the results of individual quality rules so that data quality can be assessed at multiple levels. Given the nature of ASD data (and similarly FCD), there is a need to group checks primarily based on their location and type of control. Therefore, it is possible to aggregate the results of completeness and accuracy checks of individual traffic parameters (intensity and speed) within a single lane and further aggregate these for the entire lane segment (in one direction) and for the entire profile (in both directions) of the roadway. If needed, aggregation can also be performed based on cohesive traffic routes or the type of detector technology used. This step is closely tied to the interpretation of measured data quality and will thus be more aligned with the actual implementation of the complete database of data.

6. Conclusion and Future Work

The article presents the procedure and methods for addressing the specific steps in evaluating traffic data, focusing on both their formal correctness and content accuracy. For these purposes, a data validation application was successfully utilized, which was tailored to meet the needs of this traffic data.

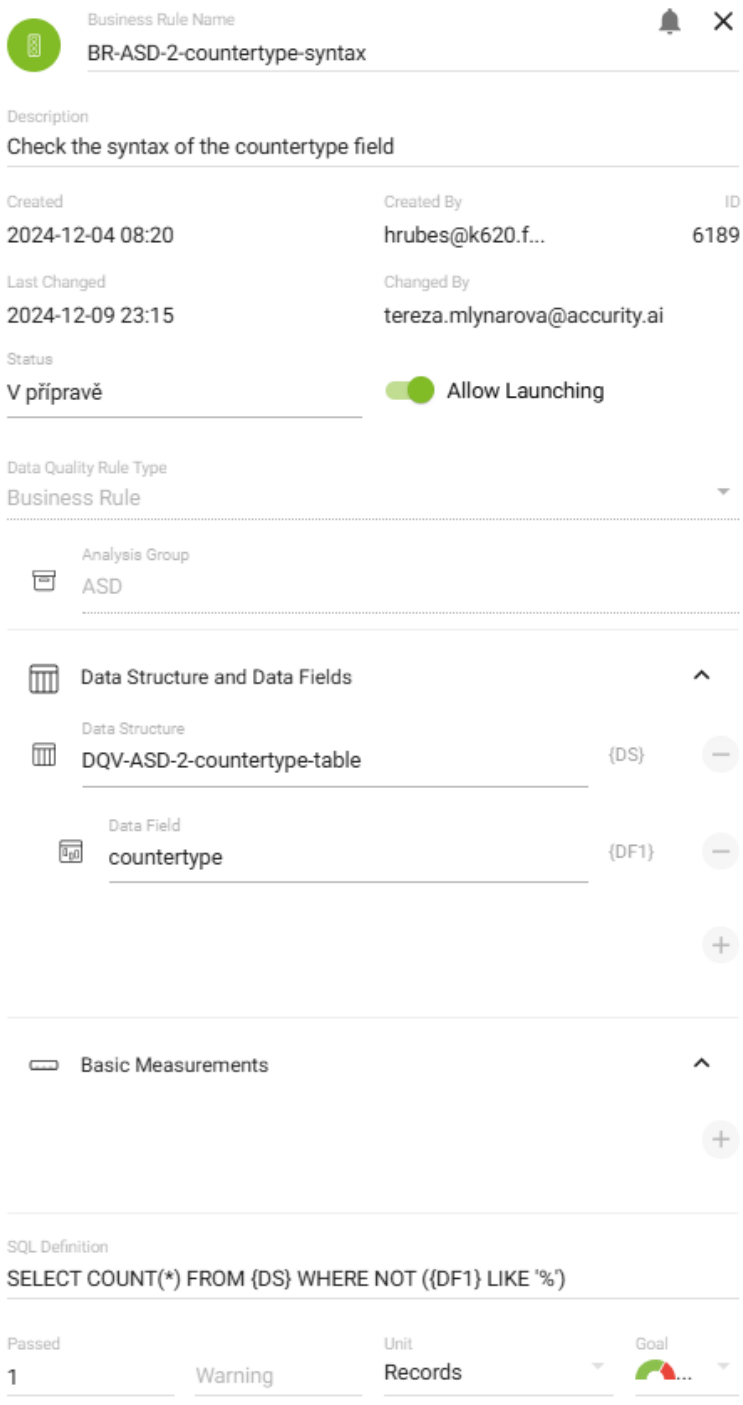

Fig. 7 *Definition of Bussiness rule [Source: authors].*

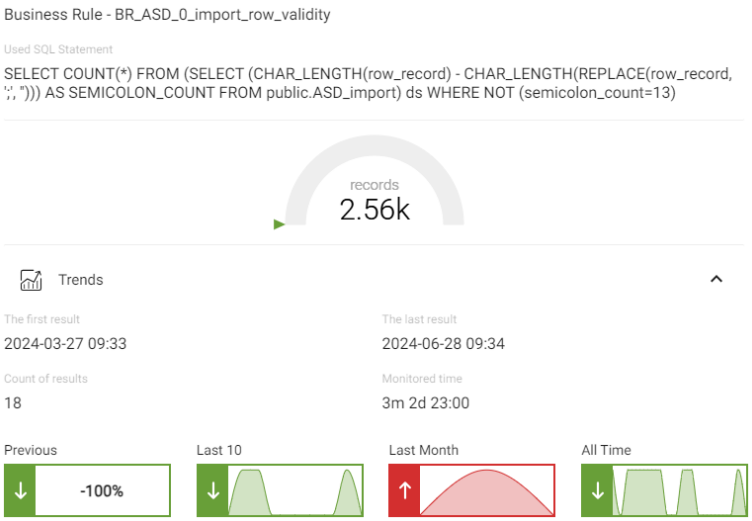

Fig. 8 *Calculation of Bussiness rule – trends [Source: authors].*

A significant element that has been designed and implemented is the detection of data anomalies using a statistical Generalized Additive Model. This model enables content validation of the measured data based on the long-term characteristics of traffic in a given location and time. This is a very important part of the entire process, as it is often challenging to determine whether atypical values in a time series of measured traffic parameters are erroneous or correspond to an extraordinary (but real) traffic situation.

It has been demonstrated that all individual steps of data validation are functional and can be automated using information tools. This has led to the establishment of a complete ecosystem for data quality control. The next task will be to expand the database of all defined rules to encompass data from all available detectors and to apply the logic and sequence of executing all rules over time. This will enable the interpretation of the actual data quality and its development over time in accordance with the data governance approach.

The advantage of the proposed method lies in its versatility, allowing application to a variety of datasets beyond just highway networks. Additional datasets managed by the Directorate of Roads and Highways were analyzed, including data from truck parking areas and floating vehicles. The approach is generally based on three key elements: an identifier, a timestamp, and a primary parameter (such as intensity or speed), which are commonly present in most datasets.

It is also important to mention the limitations of the presented research. First, the reliance on generalized additive models assumes that long-term traffic characteristics are well-represented in the available data. However, this may not fully account for scenarios with sparse or irregular data collection. Additionally, while the anomaly detection component enhances data quality assurance, its efficacy is contingent on the comprehensiveness of the defined rules. Furthermore, if the detected anomalies are not discussed with the data managers, who possess domain-

specific knowledge of the datasets, these anomalies risk being interpreted without the necessary contextual understanding. Addressing these limitations in future research will be essential for further improving the generalizability and scalability of the approach.

To compare accuracy, it is possible to use, for example, data from floating vehicles, which describe well the real behaviour of the traffic flow. Another possible data source for comparison is traffic events, where information is stored that traffic was unrestricted and unusual. The accuracy of the proposed model could be further validated through complementary non-statistical approaches. For instance, integrating machine-learning techniques, such as unsupervised clustering or neural networks, could offer a comparative perspective on the effectiveness of the generalized additive models. Additionally, expert review and manual anomaly verification could provide valuable insights into the contextual appropriateness of detected anomalies. These alternative methods could supplement the statistical approach and provide a more holistic evaluation of the model's performance.

Incorporating anomaly detection enhances the application's ability to monitor and maintain high data quality standards, ensuring that irregularities are promptly flagged for further analysis and resolution.

Acknowledgement

The project (CK04000189/ Data quality tools for ensuring system reliability of transport information centers) is co-financed with the state support of the Technology Agency of the Czech Republic within the Transport 2020+ Programme.

The authors acknowledge the partner, the Road and Motorway Directorate, namely Ing. Filip Týc, for providing the data and for cooperation in the analysis and discussions.

References

[1] MLYNÁŘOVÁ T., HRUBEŠ P. a kolektiv. Metodika přístupu k datové kvalitě. Research Report. Praha, 2023.

[2] BRONSELAER A. Data Quality Management: An Overview of Methods and Challenges. In: *Flexible Query Answering Systems.* FQAS 2021. Lecture Notes in Computer Science, 12871, *Springer, Cham*, 2021.

[3] SALIH F.I., ISMAIL S.A., HAMED M.M., MOHD YUSOP O., AZMI A., MOHD AZMI N.F. Data Quality Issues in Big Data: A Review. In: *Recent Trends in Data Science and Soft Computing.* IRICT 2018. Advances in Intelligent Systems and Computing, *Springer, Cham.*, 2019, 843.

[4] MWALE M. et al. Estimation of the completeness of road traffic mortality data in Zambia using a three source capture recapture method. Accident Analysis & Prevention, 2023, 186, 107048.

[5] HAMAD K., QUIROGA C. Assessment of Quality and Completeness of Archived ITS Sensor Data: TransGuide Case Study, 2014.

[6] ZHOU Z. et al. Investigating the uniqueness of crash injury severity in freeway tunnels: A comparative study in Guizhou, China, *Journal of safety research*, 2021, 77, pp. 105–113.

[7] GRZENDA M., KWASIBORSKA K., ZAREMBA T. Hybrid short term prediction to address limited timeliness of public transport data streams. *Neurocomputing*, 2020, 391, pp. 305–317.

[8] LOKAJ Z., SROTY M., VANIS M., BROZ J., MLADA M. C-ITS SIM as a tool for V2X communication and its validity assessment. In: 2021 *Smart City Symposium Prague*, (SCSP), IEEE, 2021, pp. 1–5.

[9] RŮŽIČKA J., TICHÝ T., HAJČIAROVÁ E. Big Data Application for Urban Transport Solutions, In: *Smart City Symposium Prague* (SCSP), Prague, Czech Republic, 2022, pp. 1–7.

[10] BENEŠ V., SVÍTEK M., MICHALÍKOVÁ A., MELICHERČÍK M. Situation model of the transport, transport emissions and meteorological conditions, *Neural Network World*, 2024, 34, 1, pp. 27–36, doi: 10.14311/nnw.2024.34.002.

[11] DIVYA J., CHANDRASEKAR A. DRGNN-dilated recurrent graph neural network framework incorporating spatial and temporal features signifying social relationships in IOT network based traffic prediction. *Neural Network World*, 33, 6, 2023, doi: 10.14311/nnw.2023.33.026.

[12] GECCHELE G., ROSSI R., GASTALDI M., CAPRINI A. Data mining methods for traffic monitoring data analysis: A case study. *Procedia-Social And Behavioral Sciences*, 2011, 20 pp. 455–464.

[13] RAJ J., BAHULEYAN H., VANAJAKSHI L. Application of data mining techniques for traffic density estimation and prediction. *Transportation Research Procedia*, 2016, 17, pp. 321–330.

[14] UGLICKICH E., NAGY I. 3D local crime type models based on crime hotspot detection. *Neural Network World*, 2024, 34, 1, pp. 89–110, doi: 10.14311/nnw.2024.34.006.

[15] KALAIR K., CONNAUGHTON C. Anomaly detection and classification in traffic flow data from fluctuations in the flow–density relationship. Transportation Research Part C: Emerging Technologies, 2021, 127, 103178.

[16] DJENOURI Y., et al. A survey on urban traffic anomalies detection algorithms. IEEE Access, 2019, 7, pp. 12192–12205.

[17] RAIYN J. Detection of road traffic anomalies based on computational data science. Discover Internet of Things, 2022, 2, 6.

[18] KUMARAN S.K., DOGRA P.D., ROY P.P. Anomaly detection in road traffic using visual surveillance: A survey. ACM Computing Surveys (CSUR), 2020, 53, 6, pp. 1–26, doi: 10.48550/arXiv.1901.08292.

[19] KHODA BAKHSHI A., AHMED M. M. Real-time crash prediction for a long low-traffic volume corridor using corrected-impurity importance and semi-parametric generalized additive model. Journal of Transportation Safety & Security, 2021, 14, 7, pp. 1165–1200.

[20] LU Q., TETTAMANTI T., HÖRCHER D., VARGA I. The impact of autonomous vehicles on urban traffic network capacity: an experimental analysis by microscopic traffic simulation. *Transportation Letters*, 2019, 12, 8, pp. 540–549.

[21] Hua J., Zhang Y., de Foy B., Shang J., Schauer J.J., Mei X., Sulaymon I.D., Han T. Quantitative estimation of meteorological impacts and the COVID-19 lockdown reductions on NO2 and PM2. 5 over the Beijing area using Generalized Additive Models (GAM). *Journal of Environmental Management*, 2021, 291, doi: 10.1016/j.jenvman.2021.112676.

[22] Wen, Yifan, WEN Y., WU R., ZHOU Z., ZHANG S., YANG S., WALLINGTON T.J., SHEN W., TAN Q., DENG Y., WU Y. A data-driven method of traffic emissions mapping with land use random forest models. Applied Energy, 2022, 305, doi: 10.1016/j.apenergy.2021.117916.

[23] National Transport Information Register of the Czech Republic. *Directorate of Road and Motorway*, 2023, Retrieved from https://registr.dopravniinfo.cz/cs/

[24] HRUBEŠ P., LANGR M., PURKRÁBKOVÁ Z. Review of data governance approaches in the field of transportation domain. In: 2024 *Smart City Symposium Prague*, IEEE, New York: IEEE Press, 2024.

[25] PEKÁR S., BRABEC M. Modern analysis of biological data. 3rd part. Nonlinear models in R (in Czech). Masarykova univerzita, 2019.

[26] HASTIE T.J., TIBSHIRANI R.J. Generalized Additive Models. Chapman & Hall/CRC, 1990.

[27] CRESSIE N. Change of support and the modifiable areal unit problem, 1996.

[28] GRAYBILL F. Theory and application of the linear model, 1976.