

A Unified Approach to Real-Time Public Transport Data Processing

Juraj Lazúr¹[0009–0008–0750–4080], Jiří Hynek¹[0000–0002–7292–6094], and Tomáš Hruška¹[0000–0001–5454–6592]

Faculty of Information Technology, Brno University of Technology, Božetěchova 1/2,
612 66 Brno, Czech Republic
`{ilazur,hynek,hruska}@fit.vut.cz`

Abstract. The use of real operations data is essential for the planning and management of modern public transport systems. With the expansion of universal formats for describing the structure of public transport systems, such as GTFS or Transmodel, the use of these data has expanded far beyond the public transport domain. On the other hand, the effort to use these data encounters the problem of its processing, storage and integration with the structure of the transport system due to the volume and speed of data generation from real operations. These problems are even more evident in the case of further use of these data as inputs for machine learning, or data mining, where integration of data from different systems into a single model is necessary. The purpose of this paper was to design a method by the which big data from real operations could be integrated with the changing structure of the transport system so that this data could be stored long term without loss of granularity, or entropy value. As a result, we proposed a data model with big data transformation algorithm, whose functionality has been verified in testing over the public transport system of the second largest city in the Czech Republic.

Keywords: Public Transport · Big Data Processing · Big Data Visualisation · GTFS

1 Introduction

Public transport has been undergoing a gradual digitisation process for more than 50 years [1]. On the side of transport companies and authorities, there are two interlinked areas. The first involves the planning of operations, while the second provides tools for real-time traffic management [2]. It is the data from real operations that serves as the necessary basis for development planning and problem solving in the system [3]. As a result, there is a closed loop of continuous improvement of the system based on its behaviour in the real environment.

The importance of processing and analysing data from the real operation has increased significantly in the last decades [4, 5]. This trend is partly due to the growth of urban agglomerations and the associated increase in the number of

passengers and the size of the area to be operated. This is also linked to the expansion of transport infrastructure, such as dedicated lanes, the funding of which is strictly based on data analysis. On the other hand, the de-carbonisation of public transport and the use of alternative powered vehicles is putting pressure on the efficiency of planning the use of individual resources [6, 7].

While in the past, the analysis of data from public transport systems was based on targeted data collection [8], current tools or algorithms use mainly standardized formats (e.g. GTFS) for describing transport systems as their input [9]. The use of these formats is partly based on the widespread use of these formats and the associated availability of data in a unified format [11], but also on the potential for replication of useful analyses over systems described by the same format [12]. In addition, the availability of these formats is used in areas other than the optimization of public transport systems [11].

On the other hand, the internal data models of the published tools, or algorithms, are different from each other in contrast to the unified inputs. This inconsistency then causes duplication of tools with the same functionalities. But more importantly, it makes difficult to compare different transport systems with each other [9]. The method of obtaining the necessary amount of raw real operations data also varies. In many cases, the studies themselves must be preceded by the data collection and storage [11, 12]. Some simplification is to retrieve this data via various APIs, but even in this case further annotation of the data is necessary [9]. In addition, when collecting real operations data, it is crucial to address how to process, compress and store this big data, due to its quantity and speed of generation.

One of the systems over which it has been necessary to address these issues is the Brno public transport system. As a part of the expansion of open datasets, the Data Department of the Municipality of Brno has obtained access to real operations data of this transport system. In order to extract new knowledge from these data, an analytical tool has been designed. Due to the amount of data that had to be stored, the use of an appropriate compression method seemed necessary. Although efficient compression methods exist not only in the field of Big Data [10], the requirement to integrate the real operation data with the transport system structure necessitated the design of a transformation algorithm, that would also be able to reduce the real operation data. The proposed solution, combines the transport system structure with real operations data. While the proposed data model allows to store the changing structure of the public transport system, the real operations data integration and compression requirements are solved by the proposed transformation algorithm. Thus the proposed solution would be able to store both the expected and at the same time the actual state of the transport system over a longer time horizon.

To verify the proposed solution, this concept was implemented and tested over the Brno public transport system. As a result, the proposed data model based on GTFS format with transformation algorithm was able to reduce the raw data volume by 75% on average while maintaining the same granularity and provides a data model for analytical tools. The generality of the resulting

data model should also make easier to re-use this concept for another transport systems. Also, the proposed solution could further simplify the subsequent use of data from real operations in analyses, machine learning, or data mining. These benefits should bring efficiency gains, cost reductions and, most importantly, increased passenger satisfaction.

2 Transport Systems Structure Modeling

The complexity of public transport systems is reflected in the internal data models of the tools that operate them. The basic structure of transport systems consists of physical infrastructure, such as stops, and a timetable that determines from where, when and to where each connection will be made. Thus these data models have to deal both with inputs coming from different sources [13] and with continuous changes in transport systems. Modularity thus becomes an important feature of these models, resulting from the fact that some elements, such as stop position, change less frequently than, for example, line schedules [13]. Linking these data models, which implement these requirements differently, not only in the context of public transport system integration thus becomes a complex problem. Furthermore, these systems work with a very similar set of entities, but the ambiguous ontology causes some entities to represent different things in different systems [14]. These problems have been largely eliminated with the introduction of standardised open formats for describing the structure of transport systems.

These formats, such as GTFS¹, shown in the Figure 1, or Transmodel², were originally developed both to share information between transport systems [18] and to simplify the development of various third-party applications [15], such as trip planning application. The use of these formats as inputs for various analyses comes later, while the use of these formats is related both to the widespread availability of data in these formats by transport authorities [15] and to the unified ontology, as the different parts of the formats represent the same elements in different systems. Together with real operation data, they have become the most common basis for advanced transport systems analysis.

The majority of studies is generally based on two data sources. The first one is the structure of the transport system and the second one is the data from real operation [17]. While the structure is mostly available from standardized formats, the extraction of real operation data is often based on collecting data during the progress of the study from various APIs [16]. A certain simplification that more recent studies have been working with is standardisation of formats for sharing data from real operation. Formats such as GTFS Realtime or SIRI, contains real-time information on the current state of the system in standardized form. However, the integration as well as the real-time processing of these two data sources still needs to be addressed separately in almost every analysis. An exception is the Transmodel family sub-format, OpRa³. This standard should

¹ <https://gtfs.org/>

² <https://www.transmodel-cen.eu/>

³ <https://www.opra-cen.eu/>

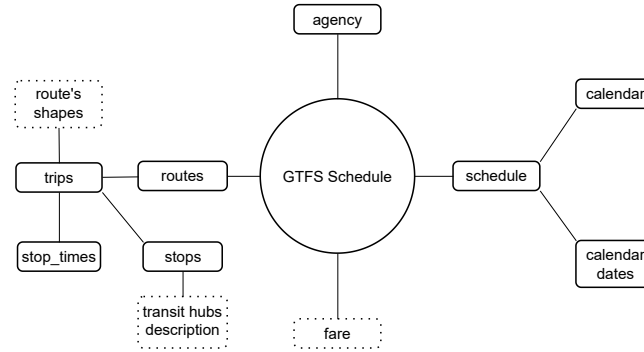


Fig. 1. A simplified GTFS structure diagram for describing the structure of a particular public transport system at a particular moment in time. Each element represents a single file. Elements with a dotted border are optional in the standard. By combining these data, it is thus possible to get an overview of when, when and which trip is served in the system.

primarily be used for sharing and enable storing raw information about the behaviour of the transport system, in order to study and optimise it. However, this format is still under development.

Expansion of open formats for the description of public transport systems has provided a new and highly efficient entry point for a wide variety of studies [16] whose purpose has long been beyond public transport [11]. On the other hand, there still remains the problem of collecting real operations data over a longer time interval. Despite efforts at unification of analyses input in the form of GTFS Realtime or OpRa standards, transformation, processing and storing the real operations data still represents a repeated part of the studies. In addition, the characteristics of the real operations big data create the need to address the volume as well as the velocity of the creation of this data.

3 Proposed Solution

Goal of our solution is to improve the availability of data describing the real behaviour of transport systems. It connects the public transport systems structure and real operations data. Designed in this way, the solution enables to store the real behaviour over a longer time horizon. The proposed solution can be divided into two main parts. The first part is the proposed data model, based on the GTFS format, which enables to store transport system's expected structure. The second part is an transformation algorithm, which enables compression and integration of real operations data into the proposed data model. By integrating real operations data, the actual behaviour of the transport system is stored in the proposed data model. The structure of the solution itself is then shown in the Figure 2.

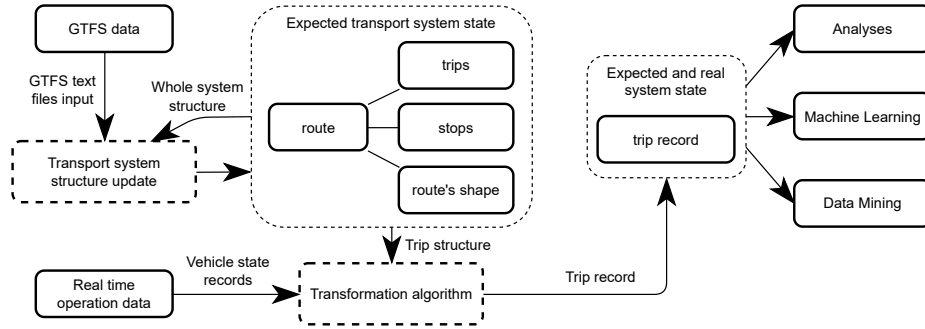


Fig. 2. The model of usage. The expected state of the system is periodically updated based on the current GTFS data. Afterwards, the global expected system state is stored for each day. The real operation data is integrated and stored into the data model using a proposed transformation algorithm. Trip records, which include the expected as well as the actual state of the system, serve as the basis for subsequent use of the data.

The proposed data model uses GTFS format data as its input. This is due to the fact that the GTFS format provides an unambiguous ontology. Also, it is provided by a high number of transport authorities and is fully adapted for automated processing. The basic element of the data model is a route. Each route consists of a list of stops, the definition of the route's shape over a road or rail network and trips. A trip is the realization of a specific route on a specific time. By connecting these basic elements, the proposed model produces the expected system structure for the current day. The proposed data model is then daily updated with the current GTFS data. As a result, a list of trip records that should have been made is stored for each day. This not only stores the expected structure of the system, but it is possible to track individual changes in the structure by comparing the stored states.

The proposed transformation algorithm connects the data from the real operation with the trip records. While the structure of the transport system is described by available standardised formats, the real operation data have generally different record structures, and come from various sources. In our design, we divided the real operation data into two groups. The first group includes the data pertaining to an entire trip record or group of trip records. Such data are, for example, the records of cancellation of a trip, replacement of a trip, or maintenance of an entire line. This data can easily be assigned as an additional attribute to a trip record. In contrast, the second group consists of data that changes during the course of the trip, such as vehicle delays or occupancy. By simply aggregation the second data group into a single value, there would be a significant loss of granularity. However, storing all the raw data would be unsustainable in the long term, both in terms of speed of creation and quantity.

Our solution to these conflicting requirements is the proposed transformation and compression algorithm shown in Figure 3.



Fig. 3. Three main steps of the transformation algorithm for integrating and compression of real operation big data into the proposed data model. Organizing and storing raw real operations data based on the trips to which that data belongs allows compression, but also preserves all of its external relationships.

The essence of this proposed algorithm is the use of the geographical component of the real operations data and the assumption that the data values changes only at certain route points. The first step of the algorithm is to group the input records from the real operation according to the individual trips. The second step is to assign each record, based on its geographical part, to the corresponding logical segment of the trip. The third step is the compression, which is graphically illustrated in Figure 4.

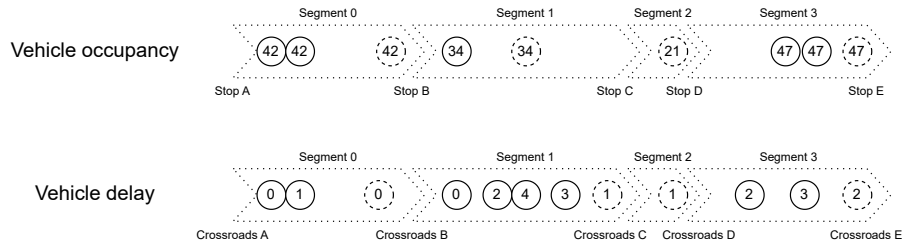


Fig. 4. Only one record is stored for each segment, indicated by a dotted line. The records are sorted from oldest to newest. For clearly given logical sections, such as vehicle occupancy, the assignment of the most recent record is unambiguous. However, for example vehicle delays, the reduction is dependent on the chosen length of the logical segments. Thus, in some cases, information may be lost. An example is segment 1, where the sharp increase and decrease in delay is lost. However, an overall delay increase of delay value 1 over segment 0 is recorded.

Record reduction is based on the assumption that data values from real operations change only at certain route points. Therefore, the route is divided

into logical segments, with only one value assigned to each segment and stored. While in some cases, such as vehicle occupancy, the determination of these points is unambiguous, for example in the case of actual vehicle delay, the location of the points depends on the desired level of granularity. This is due to the fact that the delay can vary even at a single route point in the case of traffic jam. Our proposed algorithm determines the logical length of the segments based on the route's shape. By merging the segments, it is then possible to obtain a higher compression rate at the cost of losing granularity, or vice versa.

The structure of a transportation system without real operation data cannot track the dynamics of the system behavior, while without knowledge of the structure, modeling the relationships between real operation data is difficult. This relationship becomes even deeper when we need to store real time data over a longer time horizon. Our proposed data model attempts to reflect these requirements in an attempt to achieve the best possible compression to granularity ratio.

4 Results and Evaluation

To validate our proposed concept, we implemented and tested the whole model on the Brno public transport system in cooperation with the Data Department of the Municipality of Brno. The chosen transport system represents the second largest city in the Czech Republic. It provides sufficient complexity, as well as the availability of suitable data sources. Testing the implemented model focused on two areas and included an average of 50 lines and 6500 trips per day during the 30 days period. The first area was the ability of the proposed solution to store the structure of the transport system and its changes. The second area of testing and evaluation was calculating the transformation algorithm compression rate.

The structure of the monitored Brno public transport system consists of 11 tram lines, 13 trolleybus lines and 37 bus lines. For a greater diversification of the results, 28 railway lines of the integrated system of the South Moravian Region were also included in the testing. The source of input data was a regularly updated GTFS dataset and a database of real operations data, which contained raw records for 1 previous day each time. A significant problem was caused by the absence of route's shapes in the input GTFS data, which we solved by interpolating these data using our own routing algorithm.

The implementation itself consisted of a full-stack application for processing and basic aggregation and visualization of the stored records. The NoSQL database MongoDB was used for data storage, while the server part, which implements the proposed data model and transformation algorithm, uses NodeJS technology and the Express framework. The user interface was implemented using the React library. The entire implementation that was used in the testing is then available in the public repository⁴. The live version of the system in the test run is then available on the Brno city website⁵.

⁴ <https://github.com/Jorgen98/BPTSAT-Public>

⁵ <https://kod.brno.cz/bptsat/>

The first tested area was to verify the ability of the proposed solution to store the structure of the transport system. The stored data was compared with the reference data of the transport operator. The comparison then showed that the proposed solution is able to store the structure with an accuracy of 93%, while being able to deal with a wide set of anomalies. On the other hand, the bottlenecks were some inaccuracies in the input GTFS data, as well as the inability of the system to deal with specific changes in stop identifiers and the duplication of some trips in the case of replacement services.

The second tested area was testing the compression rate of the data, which was calculated by simply comparing the raw and processed data sizes. The results are shown in Table 1. The size of the processed data also includes the proportional part of the whole transport system structure.

Table 1. Results of measuring the average size of input and processed records belonging to a single trip. The resulting compression rate depends on the length of the route and the number of input records, whereby the number of records depends on the quality of the signal between the vehicle and the dispatcher centre within the RIS system.

	Trams	Trolleybuses	Buses	Trains
Avarage raw data record size [kB]	40.35	20.04	71.47	98.71
Avarage stored trip record size [kB]	12.68	10.12	11.63	16.41
Compression ratio	3.18	1.98	6.15	6.02

5 Discussion

The result of our work is the solution that can connect and store the expected and the actual state of transport systems over time. We were able to validate this concept during a test implementation over the Brno public transport system. This resulted in a functional data model, transformation algorithm that can reduce the input data by an average of 75% while keeping the same information and in the analytical application shown in Figure 5. The solution designed in this way can simplify the use of data from real operations, which should increase the efficiency of transport systems and the satisfaction of passengers. On the other hand, there still remains a small set of anomalies, e.g. duplication of trips during service replace, that the proposed model is not able to deal with yet.

Further development of the proposed solution could focus on its extension with other optional attributes of the GTFS format, such as fare. Also, the robustness of the proposed data model should be enhanced, e.g., by checking the input based on the interpolation of already stored records. In the case of the transformation algorithm, improving the determination of logical segments, e.g. by determining them in real time or based on a statistic computed from raw data, seems to be a very suitable area for further development.

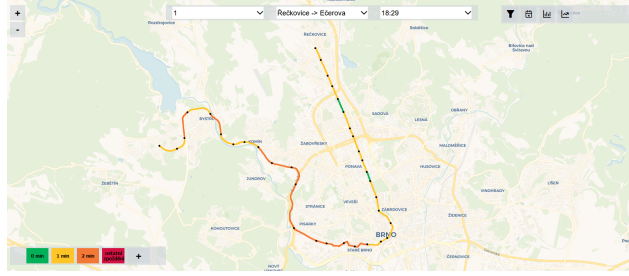


Fig. 5. The implemented user application for the analysis of trip delays over time. The average delay for the selected trip and time period are displayed. Delay categories, separated by colour, can be selected by the user. More detailed data can be obtained by clicking on the relevant part of the trip's shape.

6 Conclusion

The purpose of this paper was to find a way in which the structure and behaviour of public transport systems could be connected and stored. For this purpose, we proposed the data model for storing the structure of transport systems based on the GTFS format and a transformation algorithm that aims to integrate data from real operations into this model. We have validated our concept by implementing and testing the proposed solution on the Brno public transport system. As a result, the proposed data model and transformation algorithm should further simplify the access and use of the behaviour data of public transport systems for advanced analyses, the use of which goes beyond the field of public transport.

Acknowledgments This work was supported by project Smart information technology for a resilient society, FIT-S-23-8209, funded by Brno University of Technology.

References

1. Lampkin, B. a Wren, A. Computers in Transport Planning and Operation. Operational Research Quarterly (1970-1977). JSTOR. Sep 1972, Vol. 23, No. 3, pp. 404. <https://doi.org/10.2307/3007903>
2. P. Zito, G. Amato, S. Amoroso, and M. Berrittella, "The effect of Advanced Traveller Information Systems on public transport demand and its uncertainty," *Transportmetrica*, vol. 7, no. 1, pp. 31–43, Jan. 2011, <https://doi.org/10.1080/18128600903244727>
3. D. J. Symes, "Automatic vehicle monitoring: A tool for vehicle fleet operations," *IEEE Transactions on Vehicular Technology*, vol. 29, no. 2, pp. 235–237, May 1980, <https://doi.org/10.1109/t-vt.1980.23846>
4. M.-P. Pelletier, M. Trépanier, and C. Morency, "Smart card data use in public transit: A literature review," *Transportation Research Part C: Emerging Technologies*, vol. 19, no. 4, pp. 557–568, Aug. 2011, <https://doi.org/10.1016/j.trc.2010.12.003>

5. “Using Archived AVL-APC Data to Improve Transit Performance and Management,” Sep. 2006, <https://doi.org/10.17226/13907>
6. M. Gallet, T. Massier, and T. Hamacher, “Estimation of the energy demand of electric buses based on real-world data for large-scale public transport networks,” *Applied Energy*, vol. 230, pp. 344–356, Nov. 2018, <https://doi.org/10.1016/j.apenergy.2018.08.086>
7. J.-Q. Li, “Battery-electric transit bus developments and operations: A review,” *International Journal of Sustainable Transportation*, vol. 10, no. 3, pp. 157–169, Aug. 2014, <https://doi.org/10.1080/15568318.2013.872737>
8. P. van Egmond, P. Nijkamp, and G. Vindigni, “A comparative analysis of the performance of urban public transport systems in Europe,” *International Social Science Journal*, vol. 55, no. 176, pp. 235–247, Jun. 2003, <https://doi.org/10.1111/j.1468-2451.2003.05502005.x>
9. Z. Aemmer, A. Ranjbari, and D. MacKenzie, “Measurement and classification of transit delays using GTFS-RT data,” *Public Transport*, vol. 14, no. 2, pp. 263–285, Feb. 2022, <https://doi.org/10.1007/s12469-022-00291-7>
10. Y. Wiseman, K. Schwan, and P. Widener, “Efficient end to end data exchange using configurable compression,” *ACM SIGOPS Operating Systems Review*, vol. 39, no. 3, pp. 4–23, Jul. 2005, <https://doi.org/10.1145/1075395.1075396>
11. A. Nishino, A. Kodaka, M. Nakajima, and N. Kohtake, “A Model for Calculating the Spatial Coverage of Audible Disaster Warnings Using GTFS Realtime Data,” *Sustainability*, vol. 13, no. 23, p. 13471, Dec. 2021, <https://doi.org/10.3390/su132313471>
12. E. Chondrodima, H. Georgiou, N. Pelekis, and Y. Theodoridis, “Particle swarm optimization and RBF neural networks for public transport arrival time prediction using GTFS data,” *International Journal of Information Management Data Insights*, vol. 2, no. 2, p. 100086, Nov. 2022, <https://doi.org/10.1016/j.jjimei.2022.100086>
13. KIZOOM, N.; MILLER, P. A Transmodel based XML schema for the Google Transit Feed Specification with a GTFS/Transmodel comparison. Kizoom Ltd., London, 2008.
14. E. Ruckhaus, A. Anton-Bravo, M. Scrocca, and O. Corcho, “Applying the LOT Methodology to a Public Bus Transport Ontology aligned with Transmodel: Challenges and Results” *Semantic Web*, vol. 14, no. 4, pp. 639–657, Apr. 2023, <https://doi.org/10.3233/sw-210451>
15. ANTRIM, Aaron, et al. The many uses of GTFS data—opening the door to transit and multimodal applications. Location-Aware Information Systems Laboratory at the University of South Florida, 2013, 4
16. N. Wessel, J. Allen, and S. Farber, “Constructing a routable retrospective transit timetable from a real-time vehicle location feed and GTFS,” *Journal of Transport Geography*, vol. 62, pp. 92–97, Jun. 2017, <https://doi.org/10.1016/j.jtrangeo.2017.04.012>
17. N. Wessel and M. J. Widener, “Discovering the space–time dimensions of schedule padding and delay from GTFS and real-time transit data,” *Journal of Geographical Systems*, vol. 19, no. 1, pp. 93–107, Dec. 2016, <https://doi.org/10.1007/s10109-016-0244-8>
18. KNOWLES, Nick; MILLER, Peter; DRUMMOND, Paul. Transmodel and GTFS—Comparison and Convergence. Briefing Paper for the Public Transport Coordination Group (PTIC), Version, 2009, 4