A Hybrid Framework for Real-Time Validation of Public Transport AVL and APC Data Streams

Elias Albert ARIFOVICI[1]

ABSTRACT

The increasing adoption of telematics technologies in public transport systems has enabled the continuous collection of operational data through Automatic Vehicle Location (AVL) and Automated Passenger Counting (APC) systems. These data streams support real-time monitoring, service planning, and performance evaluation. However, raw operational data are frequently affected by positioning inaccuracies, communication delays, asynchronous updates, sensor malfunctions, and missing observations, reducing their reliability for operational decision-making. Consequently, the transformation of raw telemetry into trustworthy operational information remains a significant challenge for transport authorities and operators.

This paper proposes a hybrid validation framework for the real-time processing of AVL and APC data streams in urban public transport systems. The framework combines multiple complementary validation mechanisms, including GTFS-based map matching, geospatial geofencing, edge-based sensor event detection, and stateful heuristic validation logic. Unlike approaches that rely exclusively on spatial filtering or statistical anomaly detection, the proposed framework integrates spatial, temporal, and operational context to improve data consistency and reduce the impact of noise and hardware anomalies.

The framework was implemented and evaluated using operational data from a metropolitan public transport network. The proposed approach enables the identification of unreliable vehicle positions, exclusion of inactive vehicles located in depots, validation of stop-related events, and stabilization of vehicle tracking under conditions of GPS drift and asynchronous updates. The results demonstrate that combining multiple validation layers improves the robustness of operational analytics and provides a reliable foundation for real-time performance monitoring. The proposed methodology is independent of specific hardware vendors and can be integrated into existing GTFS-based transport analytics platforms.

Keywords: public transport analytics, AVL, APC, GTFS, data validation, map matching, geofencing, intelligent transportation systems, real-time monitoring.

[1] Corresponding author: arifovicielias21@stud.ase.ro, The Bucharest University of Economic Studies (ASE), Bucharest, Romania

1. INTRODUCTION

The digital transformation of urban mobility has fundamentally changed the way public transport systems are monitored and managed. Modern transport operators increasingly rely on telematics infrastructures capable of continuously collecting operational data from vehicles, onboard equipment, and passenger information systems. Among the most widely adopted technologies are Automatic Vehicle Location (AVL) systems, which provide real-time information about vehicle positions and movements, and Automated Passenger Counting (APC) systems, which measure passenger boarding, alighting, and occupancy levels [3], [12]. Together, these technologies generate large volumes of operational data that can support performance monitoring, service planning, resource allocation, and evidence-based decision-making.

The growing availability of operational data has accelerated the adoption of data-driven approaches in public transport management. Transport authorities increasingly use real-time information to evaluate service reliability, monitor fleet utilization, estimate passenger demand, and support strategic planning activities. Furthermore, AVL and APC data constitute the foundation of many passenger-facing applications, including real-time arrival predictions, disruption management systems, and journey planning platforms [1], [8].

Despite their potential, raw telematics data cannot be directly used for analytical purposes without adequate validation and processing mechanisms. In practical operating environments, AVL and APC streams are frequently affected by data quality issues. GPS positioning errors may generate inaccurate vehicle locations, especially in dense urban environments characterized by signal reflections and reduced satellite visibility. Communication delays and asynchronous updates may cause inconsistencies between data sources, while onboard sensors may occasionally exhibit anomalous behavior due to hardware failures or calibration issues. These limitations introduce uncertainty into operational datasets and may compromise the reliability of downstream performance indicators.

The challenge is particularly significant in urban transport systems where operational decisions increasingly depend on near real-time information. Incorrect vehicle positioning may lead to false route assignments, delayed transmissions may distort punctuality calculations, and faulty passenger counting sensors may generate inaccurate occupancy estimates. As a consequence, transport analytics systems require robust validation mechanisms capable of transforming heterogeneous and potentially unreliable data streams into coherent and trustworthy operational information.

Previous research has extensively investigated individual aspects of transport data quality, including map matching techniques, GPS error correction methods, sensor calibration procedures, and anomaly detection algorithms [2], [7], [10], [11]. However, many existing approaches address these challenges separately and focus on a single type of data source or validation method. In operational environments, transport data quality problems are often multidimensional, involving simultaneous spatial, temporal, and logical inconsistencies. Therefore, relying on a single validation technique may not be sufficient to ensure robust operational monitoring.

To address this limitation, this paper proposes a hybrid validation framework for real-time AVL and APC data streams. The framework integrates multiple complementary validation mechanisms, including GTFS-based map matching, geofencing of operational facilities, edge-based event detection, and stateful validation logic. By combining these techniques, the framework aims to reduce the impact of GPS drift, asynchronous updates, sensor anomalies, and operational ambiguities while preserving the ability to process data in near real time.

The proposed approach was developed and evaluated within the context of a metropolitan public transport network operating under a GTFS-based data environment. Rather than focusing on a particular software implementation, the paper emphasizes the methodological principles required to improve the reliability of operational transport data and to support more robust performance analytics.

The main contributions of this research are the following:

• the design of a hybrid validation framework combining spatial, temporal, and operational validation layers;
• the introduction of a stateful validation approach for stabilizing vehicle tracking under noisy operating conditions;
• the integration of geospatial and sensor-based validation mechanisms into a unified processing workflow;
• the demonstration of the framework's applicability within a real-world public transport environment.

The remainder of the paper is structured as follows. Section 2 reviews related work on AVL systems, APC technologies, and transport data quality. Section 3 presents the proposed validation framework and its architecture. Section 4 describes the validation algorithms and processing mechanisms. Section 5 discusses the case study and practical implementation results. Finally, Section 6 summarizes the conclusions and outlines future research directions.

2. RELATED WORK

AVL Data Quality and Map Matching

Automatic Vehicle Location (AVL) systems have become a fundamental component of modern public transport operations, providing continuous information regarding vehicle positions, movement patterns, and service execution. The widespread deployment of GPS-enabled tracking devices has significantly improved the ability of transport operators to monitor fleets and evaluate operational performance. However, the reliability of AVL-based analytics depends directly on the quality of the collected positioning data.

One of the most frequently reported challenges in AVL systems is the presence of positioning errors caused by satellite signal degradation, multipath effects, communication delays, and reduced visibility in dense urban environments. These phenomena may generate spatial noise, commonly referred to as GPS drift, which can lead to unrealistic vehicle trajectories, incorrect route assignments, and inaccurate travel time estimation [7].

To address these limitations, numerous studies have investigated map matching techniques, which aim to align noisy GPS observations with a reference transport network. One of the most comprehensive reviews of map-matching algorithms was presented by Quddus et al. [10], highlighting the importance of spatial context in improving positioning accuracy. Existing approaches range from simple geometric matching methods to probabilistic and hidden Markov model-based techniques designed for complex road networks.

Within public transport environments, map matching plays a particularly important role because vehicle trajectories are constrained by predefined routes. GTFS-based route geometries offer an additional source of contextual information that can be used to validate vehicle positions and estimate route progress. Nevertheless, most existing approaches primarily focus on correcting positional inaccuracies and less frequently consider the integration of map matching with other validation mechanisms capable of addressing temporal inconsistencies and sensor-related anomalies.

Consequently, although map matching significantly improves AVL reliability, it alone cannot fully address the broader challenges associated with operational transport data streams.

Reliability Challenges in APC Systems

Automated Passenger Counting (APC) systems represent another important source of operational intelligence in public transport. These systems estimate passenger boarding, alighting, and onboard occupancy levels using technologies such as infrared sensors, stereoscopic cameras, pressure-sensitive devices, or hybrid sensor configurations.

APC data are increasingly used to support service planning, capacity allocation, and passenger demand analysis. By providing information on vehicle occupancy and stop-level demand patterns, APC systems enable transport authorities to move beyond supply-oriented performance indicators and incorporate passenger-centered perspectives into decision-making processes.

Despite these advantages, APC systems are also affected by several reliability challenges. Counting accuracy may decrease under crowded conditions, during simultaneous boarding and alighting events, or as a result of hardware calibration issues. Sensor malfunctions may generate incomplete observations, duplicated records, or persistently repeated states that do not accurately reflect operational reality.

Previous research has primarily focused on improving counting accuracy through sensor calibration, computer vision techniques, and statistical correction models. However, relatively limited attention has been given to the validation of APC event streams within integrated operational analytics platforms. In practice, passenger counting information is rarely analysed in isolation and is often combined with vehicle location data and timetable information. This creates a need for validation mechanisms capable of identifying anomalous sensor behaviour while preserving real-time processing capabilities.

For operational monitoring systems, the challenge is therefore not only the accuracy of passenger counts but also the reliability of the events generated by APC sensors and their consistency with the broader operational context.

Data Fusion and Real-Time Transport Analytics

Recent advances in intelligent transportation systems have increasingly promoted the integration of heterogeneous operational data sources. Rather than analysing AVL, APC, and schedule data separately, modern transport analytics platforms seek to combine multiple information streams in order to obtain a more comprehensive representation of system performance.

Data fusion approaches have demonstrated significant benefits for public transport monitoring, including improved service reliability assessment, enhanced passenger demand analysis, and more accurate operational decision support [4], [5], [6]. GTFS datasets have become a particularly important integration layer because they provide a standardized representation of routes, stops, schedules, and service structures that can be linked with real-time operational observations [1], [8].

Several studies have proposed architectures that combine AVL data with GTFS schedules to estimate delays, identify missed trips, or evaluate punctuality. Other approaches integrate passenger counting information to analyse vehicle occupancy and demand distribution. However, most existing systems focus primarily on performance measurement and operational reporting, while data validation is often treated as a preliminary preprocessing step rather than as a core analytical component.

In real-world operational environments, data quality issues frequently occur simultaneously across multiple sources. GPS drift, asynchronous updates, missing observations, and sensor anomalies may interact and propagate through the analytical pipeline, affecting downstream indicators. As a result, the reliability of performance metrics depends not only on the availability of data but also on the robustness of the validation mechanisms applied before analysis.

This observation reveals an important research gap. While individual solutions exist for map matching, sensor validation, and data integration, fewer studies propose unified frameworks capable of combining spatial, temporal, and operational validation mechanisms within a single real-time processing workflow. The present research addresses this gap by introducing a hybrid validation framework that integrates map matching, geofencing, edge-based event detection, and stateful validation logic into a coherent methodology for improving the reliability of AVL and APC data streams.

3. FRAMEWORK ARCHITECTURE

The proposed framework is designed to improve the reliability of operational public transport data by integrating multiple validation mechanisms into a unified processing workflow. Rather than relying on a single filtering technique, the framework combines spatial, temporal, and operational validation layers capable of addressing the most common sources of uncertainty found in AVL and APC data streams.

The architecture follows a three-layer design consisting of data acquisition, validation, and operational output generation. The objective is to transform heterogeneous and potentially unreliable telemetry streams into validated operational information suitable for real-time analytics and decision support applications.

Data Sources

The framework integrates three complementary categories of data sources.

- GTFS Static Data

GTFS Static provides the theoretical representation of the transport network, including route geometries, stop locations, timetables, service calendars, and trip definitions. Within the framework, GTFS data act as the reference layer against which real-time observations are validated.

- AVL Data Streams

Automatic Vehicle Location (AVL) feeds provide continuously updated information regarding vehicle positions, timestamps, route identifiers, and operational status. These streams constitute the primary source of real-time operational information but are also affected by positioning errors, delayed updates, and communication disruptions.

- APC Data Streams

Automated Passenger Counting (APC) feeds provide information related to passenger occupancy, boarding and alighting activity, and door sensor states. APC data introduce an additional operational context that can be used to validate stop-related events and assess service utilization.

Validation Layer

The core contribution of the framework is represented by a multi-stage validation layer composed of four complementary mechanisms.

- Map Matching

Vehicle positions are projected onto GTFS route geometries to reduce the impact of GPS drift and improve route consistency.

- Geofencing

Spatial filtering mechanisms identify vehicles located inside depots or operational facilities and prevent inactive vehicles from being included in service-level analytics.

- Edge Detection

Sensor event validation is based on state transitions rather than persistent logical states, allowing the framework to detect meaningful operational events while reducing the impact of sensor anomalies.

Stateful Validation Logic**:** A state-aware tracking mechanism preserves temporal continuity between successive observations and prevents unrealistic operational transitions caused by noisy or incomplete data.

Operational Outputs

After passing through the validation layer, the framework generates a set of validated operational entities that can be directly used for performance monitoring and analytics.

These outputs include:

- validated vehicle positions;
- validated stop events;
- validated occupancy measurements;
- validated operational states;
- route progress estimates.

The resulting information forms a reliable foundation for downstream applications such as punctuality analysis, fleet monitoring, service reliability assessment, and operational dashboards.

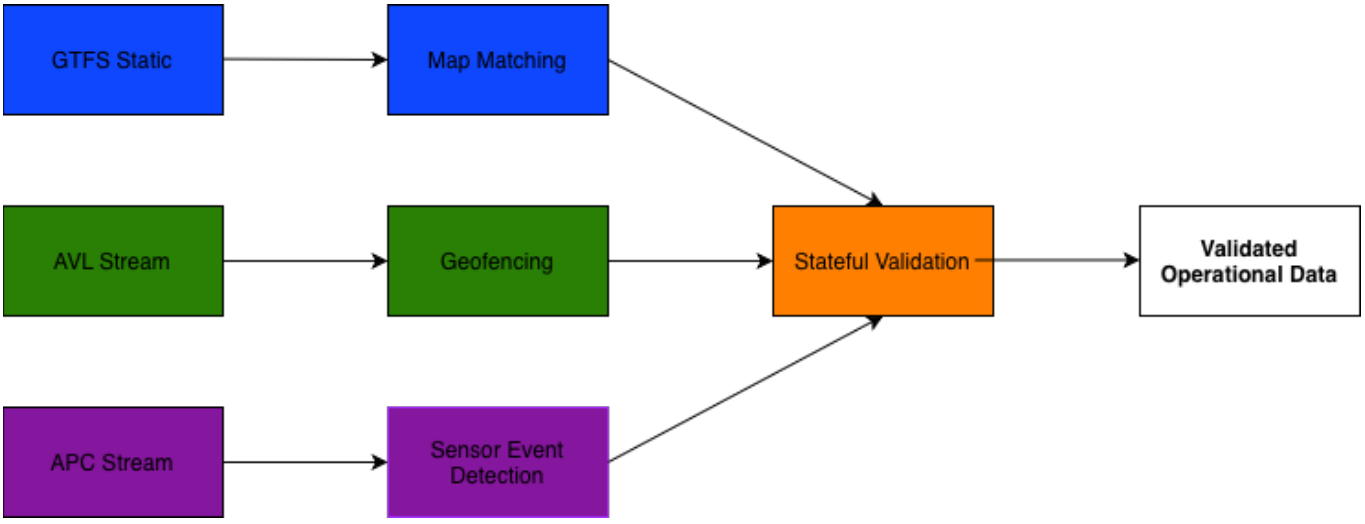

Figure 1. Proposed hybrid validation framework integrating GTFS-based map matching, geofencing, sensor event detection, and stateful validation for real-time AVL and APC data streams.

Table 1. Validation Layers and Their Operational Roles:

Validation Layer	Main Problem	Input	Output
Map Matching	GPS Drift	AVL + GTFS	Validated Position
Geofencing	Depot Vehicles	AVL	Operational Status
Sensor Event Detection	Sensor Faults	APC	Validated Events
Stateful Validation	Temporal Inconsistency	All Sources	Validated Operational Data

4. HYBRID VALIDATION FRAMEWORK

Operational public transport data streams are frequently affected by spatial inaccuracies, asynchronous updates, and hardware anomalies. The proposed framework addresses these challenges through a sequence of complementary validation mechanisms that progressively increase the reliability of incoming observations.

Unlike traditional approaches that focus exclusively on spatial filtering or anomaly detection, the proposed methodology combines spatial validation, operational context, and temporal continuity in order to obtain more robust analytical outputs.

Map Matching Validation

GPS Drift as a Source of Positional Uncertainty. GPS positioning systems operating in urban environments are affected by multipath effects, signal reflections, temporary satellite visibility loss, and communication delays. As a consequence, AVL records may place vehicles outside their actual route geometry, generate unrealistic jumps between consecutive observations, or falsely indicate reverse movement.

Direct use of raw coordinates can therefore compromise route assignment, stop detection, and performance analysis.

Projection onto GTFS Route Geometry: To mitigate these issues, the framework projects each incoming AVL observation onto the corresponding GTFS route geometry.

For every vehicle position, the algorithm searches for the nearest point located on the predefined route shape. Rather than evaluating the entire geometry for every update, a constrained search

window is used around the last validated position. This approach reduces computational complexity while preserving trajectory continuity.

The validation process consists of three consecutive steps:

1. Identification of a local search window around the previously validated route position.
2. Determination of the nearest geometry point using a distance minimization procedure.
3. Estimation of vehicle progress along the route.

Route Progress Estimation**:** Once the optimal projection point has been identified, the corresponding geometry index is transformed into a normalized route progress indicator.

The route progress of a vehicle can be expressed as a normalized indicator representing the position of the validated point along the route geometry:

$$Progress = \frac{i}{N}$$

where:

- i is the index of the validated point on the route shape;
- N is the total number of points describing the route geometry.

The resulting value is bounded within the interval [0,1], where 0 represents the beginning of the route and 1 corresponds to the route terminus. This normalized representation facilitates route progress estimation, vehicle comparison, and service monitoring across different routes and vehicle types.

This indicator provides a linear representation of vehicle advancement and enables:

- comparison between vehicles operating on the same route;
- delay estimation;
- headway monitoring;
- route-level performance analysis.

By constraining vehicle positions to valid route geometries, the framework substantially reduces the impact of GPS drift and improves the stability of operational tracking.

Depot Geofencing

Identification of Non-Operational Vehicles: A common source of analytical errors originates from vehicles located inside depots, garages, or operational facilities. Although these vehicles may continue transmitting valid AVL coordinates, they are not actively participating in passenger service.

If such records are processed without additional validation, they may generate false trip assignments, distort fleet utilization indicators, and inflate the number of active vehicles.

Point-in-Polygon Validation**:** To address this problem, the framework employs a geofencing mechanism based on depot polygons extracted from geospatial datasets. Geofencing-based approaches have been increasingly adopted in public transport analytics to identify operational route segments, distinguish active from inactive vehicles, and improve the reliability of AVL-based performance monitoring [9].

Each incoming AVL position is evaluated through a Point-in-Polygon test. If the reported coordinate lies inside a predefined depot area, the vehicle is classified as inactive and excluded from operational analytics.

Formally, for a vehicle position P(x,y), the validation condition is defined as:

$P \in$ Depot Area

Vehicles satisfying this condition are temporarily removed from route assignment and performance calculations.

Operational Benefits**:** The geofencing layer provides several practical advantages:

- elimination of false trip assignments;
- improved fleet utilization estimates;
- reduction of tracking noise;
- exclusion of maintenance and reserve vehicles from operational indicators.

As a result, only vehicles actively participating in commercial service are considered by downstream analytical processes.

Edge Detection for Sensor Validation

Limitations of Persistent Sensor States: Operational analytics frequently rely on logical sensor information such as door states and passenger counting events. However, hardware anomalies may cause sensors to remain indefinitely in a single state despite changes in actual operational conditions.

For example, a faulty door sensor may continuously report an OPEN status during an entire trip. If interpreted directly, this behaviour could generate false stop events and incorrect occupancy calculations.

Transition-Based Event Detection: Instead of analysing the current sensor value, the proposed framework focuses on detecting state transitions.

The methodology evaluates changes between consecutive observations:

- 0 → 1 = event detected
- 1 → 1 = no new event
- 1 → 0 = event completed

This approach allows the framework to identify meaningful operational events while filtering out sensor states that remain unchanged for prolonged periods.

Operational Impact: Edge detection improves the reliability of stop-event validation and reduces the influence of hardware anomalies on downstream analytics.

Combined with map matching and geospatial validation, this mechanism provides an additional layer of confidence for interpreting APC-generated events.

Stateful Validation Logic

Motivation for Stateful Validation: Spatial validation techniques such as map matching and geofencing significantly improve the quality of AVL observations. Similarly, edge detection reduces the impact of sensor anomalies by identifying meaningful operational events. However, these mechanisms evaluate observations individually and do not fully consider the temporal continuity of vehicle behaviour.

In real-world operations, public transport vehicles follow predictable operational patterns. Vehicles do not instantly switch between unrelated operational states, nor do they repeatedly change direction or route assignment without a corresponding operational context. Nevertheless, noisy AVL streams may generate precisely such unrealistic transitions.

For example, GPS drift may temporarily place a vehicle behind its previously validated position, suggesting reverse movement along the route. Likewise, delayed updates may create the impression that a vehicle has suddenly advanced a significant distance without traversing intermediate segments. When interpreted independently, such observations may destabilize vehicle tracking and lead to incorrect operational conclusions.

To address these limitations, the proposed framework introduces a stateful validation mechanism that evaluates each observation within the context of previously validated states.

State Estimation Model**:** The operational state of a vehicle is determined using information from the current observation, previous validated observations, movement characteristics, and sensor-derived events. This relationship can be represented conceptually as:

$$State_t = f(Position_t, Position_{t-1}, Speed_t, Events_t)$$

where:

- $Position_t$ represents the current validated position;
- $Position_{t-1}$ represents the previously validated position;
- $Speed_t$ denotes the current vehicle speed;
- $Events_t$ represents sensor-derived operational events.

This formulation highlights the stateful nature of the proposed framework, where operational decisions are based not only on current observations but also on temporal continuity and contextual information derived from previous system states.

Vehicle Operational States: The framework models vehicle behaviour through a finite set of operational states representing the most common situations encountered during service execution.

The following states are defined:

- **Moving** – the vehicle is progressing along a route and actively performing service;
- **Stopped** – the vehicle is temporarily stationary, typically at a stop or traffic signal;
- **Terminal Waiting** – the vehicle is located near a route terminal and waiting to begin a new trip;
- **Depot** – the vehicle is positioned within an operational facility and is not considered active;
- **Uncertain** – insufficient information is available to determine a reliable operational state.

Each vehicle is continuously associated with one of these states, which is updated whenever new observations become available.

State Transition Validation**:** Rather than accepting all incoming observations at face value, the framework evaluates whether the implied transition between two consecutive states is operationally plausible.

Examples of valid transitions include:

- Moving → Stopped
- Stopped → Moving
- Moving → Terminal Waiting
- Terminal Waiting → Moving
- Depot → Moving

In contrast, certain transitions are considered unlikely and require additional validation before being accepted. Examples include:

- Moving → Depot
- Terminal Waiting → Depot
- Moving → Uncertain

Such transitions may indicate temporary GPS anomalies, communication interruptions, or inconsistent route assignments rather than genuine operational changes.

By enforcing logical transition rules, the framework prevents isolated anomalous observations from immediately altering the estimated operational state of the vehicle.

Temporal Consistency and Fail-Safe Tracking**:** A key feature of the proposed approach is the preservation of temporal consistency. The framework maintains a short operational memory containing previously validated positions, route assignments, movement direction, and state history.

When a new observation conflicts with the recent operational history, the framework does not immediately discard it. Instead, the observation is evaluated against multiple criteria, including:

- consistency with route geometry;
- movement direction;
- recent speed profile;
- terminal proximity;
- depot status;
- sensor-derived events.

Only when sufficient evidence supports the new interpretation is the operational state updated.

This fail-safe strategy reduces oscillations caused by GPS jitter and prevents repeated state switching under unstable operating conditions.

Benefits of Stateful Validation**:** The introduction of stateful validation provides several advantages over purely spatial or event-based approaches.

First, it improves trajectory stability by preserving continuity between consecutive observations. Second, it reduces the impact of temporary positioning errors that would otherwise generate unrealistic route progress estimates. Third, it enhances the reliability of trip assignment and stop-event detection by incorporating operational context into the validation process.

Most importantly, stateful validation transforms the interpretation of transport telemetry from a collection of independent observations into a coherent representation of vehicle behaviour over time. This additional contextual layer increases the robustness of operational analytics and provides a more reliable foundation for real-time monitoring applications.

5. CASE STUDY AND FRAMEWORK EVALUATION

Operational Environment

The proposed framework was implemented and evaluated using operational data collected from a metropolitan public transport network. The evaluation environment combines GTFS Static schedules, real-time AVL vehicle location streams, APC passenger counting data, and geospatial information describing operational facilities.

The analysed data represent a typical urban transport environment characterized by high vehicle density, frequent service updates, heterogeneous hardware infrastructure, and varying communication quality. Such conditions provide a suitable context for evaluating the robustness of data validation mechanisms under realistic operating circumstances.

The objective of the evaluation is not to assess a particular transport operator, but to demonstrate how different validation layers contribute to improving the consistency and reliability of operational data streams.

GPS Drift and Route Position Validation

One of the most common anomalies observed in AVL streams is GPS drift. In several situations, vehicles were temporarily reported outside their actual route geometry due to signal reflections and positioning inaccuracies.

Without validation, these observations generated unrealistic route progress estimates and occasional false indications of reverse movement. After applying GTFS-based map matching, vehicle trajectories remained aligned with the planned route geometry and route progress estimation became significantly more stable.

The results indicate that route-constrained positioning provides an effective mechanism for reducing the operational impact of spatial noise while preserving near real-time processing capabilities.

Depot Detection and Fleet Filtering

Another frequently observed situation involved vehicles located inside depots while continuing to transmit valid AVL coordinates.

When processed without additional filtering, these records could be interpreted as active vehicles and incorrectly included in fleet utilization calculations. The introduction of depot geofencing enabled the automatic identification of inactive vehicles through Point-in-Polygon validation.

As a result, non-operational vehicles were excluded from route assignment procedures and real-time operational indicators reflected only vehicles actively participating in passenger service.

Sensor Anomalies and Event Validation

The APC data stream occasionally exhibited persistent sensor states, particularly in relation to door-status indicators. In some cases, door sensors remained continuously in an OPEN state for extended periods despite normal vehicle movement.

The edge detection mechanism successfully distinguished actual operational events from static sensor states by focusing on state transitions rather than current values.

This approach reduced the likelihood of generating false stop events and improved the consistency of occupancy-related observations.

Table 2. Examples of Operational Anomalies Addressed by the Proposed Framework

Anomaly Type	Validation Layer	Expected Impact
GPS drift causing route deviation	Map Matching	Improved route alignment and progress estimation
Vehicle located inside depot	Geofencing	Exclusion from active fleet analytics
Persistent OPEN door state	Sensor Event Detection	Elimination of false stop events
Repeated sensor state without transition	Sensor Event Detection	Improved event reliability
Sudden reverse position jump	Stateful Validation	Preservation of tracking continuity
Inconsistent operational state transition	Stateful Validation	Increased stability of vehicle tracking

Benefits of the Hybrid Validation Approach

The evaluation demonstrates that individual validation techniques address different categories of data quality issues.

Map matching primarily improves spatial consistency, geofencing eliminates inactive operational entities, edge detection enhances event reliability, and stateful validation preserves temporal continuity.

When combined, these mechanisms create a complementary validation workflow capable of mitigating multiple sources of uncertainty simultaneously.

The proposed framework therefore provides a more reliable representation of operational reality than approaches based on a single validation strategy.

6. DISCUSSION

The results suggest that operational data quality should not be treated as a purely spatial problem. Although GPS correction remains an important component of transport analytics, many operational inconsistencies originate from interactions between positioning errors, asynchronous updates, and sensor anomalies.

The proposed framework addresses this challenge through a layered validation strategy that combines spatial, temporal, and operational context. Unlike traditional approaches that rely exclusively on map matching or statistical filtering, the methodology evaluates observations within a broader operational framework.

An important advantage of the proposed approach is its independence from specific hardware vendors and proprietary platforms. The framework relies on commonly available data sources, including GTFS feeds, AVL streams, APC observations, and geospatial datasets, making it applicable across different public transport environments.

Nevertheless, several limitations should be acknowledged. The validation logic relies on heuristic rules that may require calibration for different operational contexts. Furthermore, the framework does not currently incorporate machine learning techniques that could potentially improve anomaly detection under highly complex operating conditions.

Future research may investigate the integration of predictive models, confidence scoring mechanisms, and adaptive validation thresholds capable of automatically adjusting to changing operational environments.

7. CONCLUSIONS

The growing availability of operational data generated by modern public transport systems has created new opportunities for data-driven decision-making and real-time service monitoring. However, the practical value of AVL and APC data streams depends directly on their reliability. In real-world operating environments, telemetry data are frequently affected by GPS drift, communication delays, asynchronous updates, missing observations, and sensor anomalies. These issues can significantly reduce the accuracy of operational indicators and compromise analytical results if not adequately addressed.

This paper proposed a hybrid validation framework designed to improve the reliability of real-time public transport data streams. The framework integrates multiple complementary validation mechanisms, including GTFS-based map matching, depot geofencing, edge-based sensor event detection, and stateful validation logic. Unlike approaches that rely on a single validation technique, the proposed methodology combines spatial, temporal, and operational context to provide a more robust representation of vehicle behaviour and service execution.

The evaluation demonstrated that different validation layers address distinct categories of data quality problems. Map matching improves spatial consistency by reducing the impact of GPS drift, geofencing prevents inactive vehicles from influencing operational indicators, edge detection increases the reliability of sensor-derived events, and stateful validation preserves continuity between successive observations. Together, these mechanisms contribute to the generation of more reliable operational information suitable for real-time analytics and decision-support applications.

An important characteristic of the proposed framework is its practical applicability. The methodology relies on widely adopted standards and data sources, including GTFS feeds, AVL telemetry, APC observations, and open geospatial datasets. As a result, the framework can be integrated into existing transport analytics platforms without requiring specialized hardware or proprietary infrastructure.

The research contributes to the growing body of work on transport data quality by demonstrating that reliable operational analytics require more than isolated filtering techniques. Effective validation must incorporate multiple sources of contextual information and evaluate observations within their operational environment. By combining several validation strategies into a unified workflow, the proposed framework provides a practical approach for transforming noisy telemetry streams into trustworthy operational intelligence.

Future work may extend the framework through the incorporation of machine learning techniques for anomaly detection, adaptive confidence scoring models, and predictive validation mechanisms capable of identifying potential inconsistencies before they propagate through operational analytics pipelines. Additional research may also investigate the applicability of the proposed methodology across different transport modes and metropolitan contexts.

Overall, the results indicate that hybrid validation approaches represent a promising direction for improving the quality, robustness, and operational value of real-time public transport data streams.

ACKNOWLEDGEMENT

The author declares no external funding and no conflict of interest.

ORCID

Elias-Albert Arifovici, https://orcid.org/0009-0005-2267-0069

REFERENCES

[1] A. Antrim and S. J. Barbeau, “The Many Uses of GTFS Data – Opening the Door to Transit and Multimodal Applications,” Center for Urban Transportation Research, University of South Florida, Tampa, FL, USA, 2013.

[2] C. Batini and M. Scannapieco, Data and Information Quality: Dimensions, Principles and Techniques. Cham, Switzerland: Springer International Publishing, 2016. doi: 10.1007/978-3-319-24106-7.

[3] J. E. Ehrlich, “Applications of Automatic Vehicle Location Systems Towards Improving Service Reliability and Operations Planning in London,” M.S. thesis, Massachusetts Institute of Technology, Cambridge, MA, USA, 2010.

[4] N.-E. E. Faouzi and L. A. Klein, “Data Fusion for ITS: Techniques and Research Needs,” Transportation Research Procedia, vol. 15, pp. 495–512, 2016. doi: 10.1016/j.trpro.2016.06.042.

[5] N.-E. E. Faouzi, H. Leung and A. Kurian, “Data fusion in intelligent transportation systems: Progress and challenges – A survey,” Information Fusion, vol. 12, no. 1, pp. 4–10, 2011. doi: 10.1016/j.inffus.2010.06.001.

[6] D. L. Hall and J. Llinas, “An introduction to multisensor data fusion,” Proceedings of the IEEE, vol. 85, no. 1, pp. 6–23, 1997. doi: 10.1109/5.554205.

[7] E. Mazloumi, G. Currie and G. Rose, “Using GPS Data to Gain Insight into Public Transport Travel Time Variability,” Journal of Transportation Engineering, vol. 136, no. 7, pp. 623–631, 2010. doi: 10.1061/(ASCE)TE.1943-5436.0000126.

[8] MobilityData, “General Transit Feed Specification (GTFS),” 2026. [Online]. Available: https://gtfs.org/documentation/overview/

[9] N. B. Pandikashala Ambalakkal, A. Drabicki and C. Antoniou, “Data-driven geofencing approach to identify bus route segments for travel time improvements using automatic vehicle location (AVL) data,” Transportation Research Procedia, vol. 95, pp. 976–983, 2026. doi: 10.1016/j.trpro.2026.02.123.

[10] M. A. Quddus, W. Y. Ochieng and R. B. Noland, “Current map-matching algorithms for transport applications: State-of-the-art and future research directions,” Transportation Research Part C: Emerging Technologies, vol. 15, no. 5, pp. 312–328, 2007. doi: 10.1016/j.trc.2007.05.002.

[11] S. Si, W. Xiong and X. Che, “Data Quality Analysis and Improvement: A Case Study of a Bus Transportation System,” Applied Sciences, vol. 13, no. 19, Art. no. 11020, 2023. doi: 10.3390/app131911020.

[12] V. R. Vuchic, Urban Transit: Operations, Planning and Economics. Hoboken, NJ, USA: John Wiley & Sons, 2005.