ASSOCIAZIONE ITALIANA per l'INGEGNERIA del TRAFFICO e dei TRASPORTI
AIIT
Giordano Editore

Classification of Spatial Anomalies in Bus Trajectories: A Machine Learning Approach

Fabio Borghetti[1*], Niccolò Fiore[2], Silvio Cabrini[2], Stefano Rossi[2]

[1] *Design Department, Mobility and Transport Laboratory, Politecnico di Milano, Via Candiani 10, 20158 Milan, Italy*
[2] *Autoguidovie S.p.A. - Via M.F. Quintiliano, 18 – 20138 Milano*

Abstract

Public transportation, essential in urban and suburban settings for its environmental and operational efficiencies, is becoming increasingly sophisticated with Automated Vehicle Monitoring (AVM) technologies. These systems capture and process a wide array of data in real-time, facilitating unprecedented service monitoring and management. This research uses vast GPS (Global Positioning System) datasets from bus routes to analyze spatial anomalies in bus trajectories. Using modern AVM systems, the study applies Machine Learning techniques, specifically Random Forest and Extreme Gradient Boosting (XGBoost), to classify trips (defined as individual journeys from a specific origin to a specific destination at a given time) based on detected anomalies such as route deviations and unexpected interruptions. This classification enables targeted corrective actions, enhancing service reliability. Public transport operators in Italy and Europe receive remuneration based on various factors, including kilometers covered, and face penalties for not meeting service quality conditions. The Machine Learning tool developed to detect and classify spatial anomalies offers significant advantages in cost management, compliance with service obligations, and operational efficiency. The study outlines a methodical approach involving feature engineering on GPS points, aggregation into trip datasets, preliminary categorization of anomalies, and detailed multi-class classification using advanced algorithms. Focusing on a practical application, the study evaluates GPS data from Autoguidovie, a local operator in northern Italy. The highest-performing model, Extreme Gradient Boosting, demonstrated a classification accuracy of 88.72%. In the case study, the model identified 9,511 trips affected by spatial anomalies out of 27,832 trips analyzed, generating nearly 4 million GPS points. This result enabled the appropriate management of anomalies through proper analysis and corrective actions.

Keywords: Mobility, Transportation, Public Transport Management, AVM, Systems, Machine Learning

1 Introduction

Integrating AVM (Automatic Vehicle Monitoring) systems is crucial in public transport management, particularly in densely populated areas. When combined with sophisticated data analysis techniques and Machine Learning algorithms, these systems create a potent synergy for enhancing service delivery and operational efficiency.

* Corresponding author: Fabio Borghetti (fabio.borghetti@polimi.it)

This study seeks to harness the power of GPS (Global Positioning System) data from trips provided by modern AVM systems in the bus sector to develop a Machine Learning model capable of handling trajectory spatial anomalies. The coding was performed using the R programming language to complete the work.

The primary goal is not only to detect but also to categorize these kinds of anomalies in bus route trajectories based on their manifestations to avoid them. With thousands of bus trips operated daily and millions of GPS data points generated, it is essential to continuously identify and quantify any anomaly to enhance service quality, thereby minimizing disruptions. However, merely knowing the number of trajectories affected by anomalies is insufficient; understanding the type and cause of each anomaly is crucial.

Many public transport operators in Italy and Europe receive remuneration based on various factors, including the number of kilometers covered. Additionally, there are specific penalties for failing to meet the minimum service quality conditions outlined in their contractual agreements (ART, 2019).

To effectively detect spatial anomalies in bus trajectories, it is crucial to have real data from AVM systems and theoretically programmed routes that buses should follow during service. This data is contained in the GTFS (General Transit Feed Specification) file, which acts as a pivotal repository of scheduled services, route descriptions, and programming timelines. GTFS is a global standard originally developed to facilitate the easier integration of transit data into Google Maps. Public transport operators widely adopt it worldwide to enhance the accessibility and interoperability of their data.

To tackle the identified challenge, this study employs a structured approach, composed of the following steps:

- Collect GPS data points gathered from bus trips, the GTFS file containing scheduled service information, and the bus depots' positions. Then, feature engineering will be performed on the GPS dataset based on individual GPS points, resulting in an integrated dataset of GPS points with associated features.
- Aggregate GPS points into trajectories and perform feature engineering to create a trip-based dataset.
- Implement a Preliminary Classification Model, which utilizes outlier detection methods alongside AVM data. This enriches the dataset with a new classification representing an additional feature.
- Apply a Multi-class Classification Machine Learning Model to the trip's dataset. The model outputs trips labeled with spatial anomaly classes detected.
- Perform Corrective Actions. This step emphasizes applying the multi-class classification model results to implement practical actions that enhance service quality.

This research focuses on six specific trip anomalies observed over nearly a year within Autoguidovie, a private capital local public transport company serving various northern Italy regions with urban and suburban services. The most common critical issues that can compromise a trip’s integrity have been selected:

- *Deviation* involves localized GPS points deviating from the planned route, where the bus follows an alternate path. Two primary causes can lead to deviations: changes in the road network (like new one-way streets or construction) necessitating route adjustments, and human errors by drivers, possibly due to inadequate training or mistakes. Analyzing these occurrences can pinpoint recurring issues, requiring route planning or driver communication adjustments.

- *Interrupted Trip* occurs when a continuous GPS signal -tracking the bus route- suddenly stops, leaving a trip segment uncovered. The key is to determine whether these interruptions are due to communication failures between the onboard device and the control center or if the service physically stopped, leaving a part of the programmed route uncovered.
- *Signal Gaps* refer to temporary absences of GPS data along the planned route — for instance, when short portions of the trajectory are missing between otherwise continuous GPS points. Unlike Interrupted Trips, where tracking stops completely from a certain point onward, Signal Gaps indicate brief and localized signal losses. These gaps are useful to identify areas with weak GPS coverage or potential device malfunctions, without implying a complete trip interruption.
- *Depot:* this happens when a bus transmits typical service signals while heading to the depot instead of the end of its route. It is important to identify whether this is due to the system's failure to recognize the end of a trip or other factors since such movements should be filtered out, as they do not affect the actual service.
- *Double Run:* this involves the AVM system failing to recognize that a bus has completed a trip and is beginning another, possibly on the same or a different planned route.
- *Aborted Trip:* this special case of an Interrupted Trip occurs when the bus does not start its intended route but transmits a few GPS points before ceasing transmission entirely.

The goal is to support public transport operators in optimizing costs, avoiding contractual penalties, improving operational efficiency, and ensuring data quality for regulatory compliance. This work demonstrates how data-driven methods can contribute to smarter, more reliable public transport systems.

2 State of the Art

The increasing availability of GPS data through Automatic Vehicle Monitoring (AVM) systems has opened new possibilities for analyzing public transport operations. These data-rich environments have spurred scientific research to enhance transit performance, service planning, and passenger experience. Artificial Intelligence and Machine Learning techniques are central in extracting actionable insights from such data, particularly in areas like arrival time prediction, congestion monitoring, map-matching, and driver behavior analysis.

Recent contributions have focused on time reliability and passenger-oriented metrics, proposing methodologies that leverage AVM data to evaluate service regularity and punctuality (Barabino et al., 2017a; Barabino et al., 2017b). Other studies have investigated safety issues along routes and at bus stops, developing models to estimate crash risk or classify safety levels, even in contexts lacking historical crash data (Barabino et al., 2024; Ye et al., 2016). While these approaches do not directly address spatial trajectories, they highlight the importance of operational data interpretation centered on passenger experience. Among other approaches, detecting anomalous trajectories has gained growing attention as a critical tool for improving operational reliability.

Building on this background, the present study introduces a novel approach that shifts the focus from anomaly detection to the multi-class classification of route deviations. This distinction allows for a more nuanced understanding of trajectory anomalies' nature and causes, offering operators actionable feedback. The following sections summarize the

most relevant scientific contributions and outline the Machine Learning techniques selected for this work.

2.1 Scientific Research

The exploration of GPS data within public transportation research has highlighted several critical areas: i) time interval prediction (Kumar et al., 2017), ii) road congestion analysis (Kong et al., 2018), iii) map matching (Raymond and Imamichi, 2016), iv) driver behavior (Singh et al., 2024), and v) the identification of anomalous trajectories (Zou et al., 2023) as shown in Figure 1.

Figure 1: Scientific research scheme.

These areas benefit significantly from advanced computational methods, demonstrating the utility of GPS data across different facets of transportation management.

Regarding the specific case of anomaly detection, various methods exist in the literature, ranging from outlier detection to Machine Learning and Deep Learning models, to determine whether an anomaly is present in the trajectory.

The novel aspect concerning the state of the art is the shift from mere anomaly detection to the detailed classification of route deviations, highlighting the need for models that address the specific causes behind these anomalies. This shift is crucial for tailoring solutions that enhance the effectiveness of targeting corrective actions to improve the service.

2.2 Machine Learning algorithms

This study employs two powerful Machine Learning algorithms to classify bus trajectory anomalies: Random Forest (Breiman, 2001) and Extreme Gradient Boosting (Chen and Guestrin, 2016). Both algorithms are renowned for their efficacy in multi-class classification challenges.

The Random Forest algorithm is an ensemble Machine Learning model that uses multiple decision trees for classification tasks. Decision trees work by recursively splitting the dataset based on feature values, with each internal node representing a decision and each leaf representing a class label. While individual decision trees can suffer from overfitting due to high sensitivity to training data, Random Forest addresses this limitation through ensemble learning. Random Forest builds multiple decision trees using bootstrapping - creating random subsets of data with replacement, each containing random rows and features. Each tree is trained independently on its respective subset, capturing different feature correlations. For classification, the final prediction is

determined by majority voting among all trees (aggregation). This "bagging" process (bootstrapping + aggregation) significantly reduces overfitting and improves model robustness. Key parameters include the number of trees and the number of features each tree considers.

Extreme Gradient Boosting (XGBoost) is an advanced implementation of Gradient Boosting designed for speed and performance. Unlike Random Forest's parallel approach, XGBoost builds decision trees sequentially, where each tree corrects the errors of previous ones through "boosting". The process begins with an initial decision tree that predicts output classes. The algorithm then computes residuals (between actual and predicted values) and builds subsequent trees to predict these residuals rather than the original classes. Predictions are combined using a learning rate parameter that controls each new model's contribution. This iterative process continues until performance stops improving. XGBoost enhances standard Gradient Boosting through regularization techniques (L1 and L2) that prevent overfitting by penalizing model complexity, and parallel processing capabilities for improved efficiency and scalability. Critical parameters include the number of trees, learning rate, and maximum tree depth, which control model complexity and learning dynamics.

While both algorithms were tested, XGBoost consistently outperformed Random Forest regarding accuracy and interpretability of the classification results and was therefore selected for the final model.

3 Methodology

3.1 Overview

The developed methodology aims to process extensive GPS data from bus AVM systems. This Machine Learning approach focuses on individual bus trips: it classifies each trip as either a "Correct Trip" or, if anomalies are present, into one of six distinct anomalous classes. Each class has unique characteristics and operational implications that can be understood and managed appropriately. The possible classes, developed from extensive operational data analysis and on-field experience, are the following:

- Correct Trip: it represents an ideal trip where the bus follows the planned route and stops at designated terminals as expected.
- Deviation: identified when the bus diverges from its planned route, taking an alternate path before rejoining the intended course.
- Interrupted Trip: a trip that starts normally but suddenly stops transmitting data, indicating an unfinished journey.
- Depot: it represents trips including GPS points leading to or from a bus depot, which should not be considered part of the trip.
- Signal Gaps: when short trajectory portions are missing between otherwise continuous GPS points. This may be due to GPS disruptions.
- Double Run: the case where the AVM system mistakenly records two sequential trips under one trip identifier.
- An Aborted Trip occurs when a bus does not start at all its intended route despite being at the departure point.

To compare the data from actual trajectories with the programmed route that buses are supposed to follow, the General Transit Feed Specification (GTFS) file can be used, which acts as a pivotal repository of scheduled services, route descriptions, and programming timelines. GTFS is a global standard originally developed to facilitate the easier integration of transit data into Google Maps. Public transport operators widely adopt it worldwide to enhance the accessibility and interoperability of their data.

As can be seen in the diagram in Figure 2, the methodology consists of the following steps:

1. Collect GPS data points gathered from the buses' AVM system, the GTFS file, and the position of the bus depots. Then, feature engineering will be performed on the GPS dataset based on individual GPS points, resulting in an integrated dataset of GPS points with associated features.
2. Aggregate GPS points into trajectories to create a trip-based dataset.
3. Apply feature engineering to the trip dataset and build a preliminary classification model that integrates outlier detection techniques with AVM data. The resulting enriched dataset should include a binary anomaly flag (anomalous vs. non-anomalous trips) as a new feature.
4. Apply a Multi-Class Classification Machine Learning model to the enhanced trip-based dataset. The model outputs trips labeled with spatial anomaly classes detected. This outcome will be used to undertake corrective actions on the service.

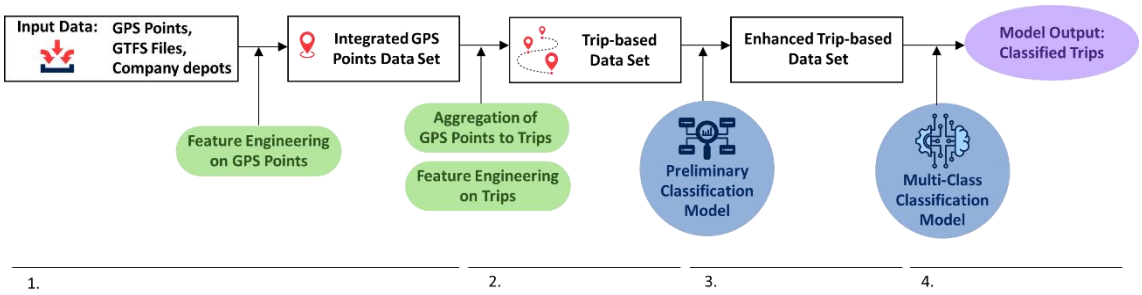

Figure 2: Methodology scheme.

3.2 Features on the trip-based data set

Let a trip T be defined as an ordered sequence of GPS observations $T = \{p_1, p_2, \dots, p_{N_{trip}}\}$, where each point $p_i = (lat_i, lon_i, t_i)$ represents the position and timestamp of the vehicle at index i. The trip starts at the origin p_1 and ends at destination $p_{N_{trip}}$, occurring at a specific time interval.

Let $R = \{r_1, r_2, \dots, r_{N_R}\}$ be the programmed route, i.e., the sequence of geographical coordinates that define the expected path of the trip, as provided by the GTFS feed. Each point $r_j = (lat_j, lon_j)$ belongs to the cartographic representation of the planned service.

Each feature is computed based on these GPS points and is designed to capture the trip's spatial, temporal, and operational characteristics, supporting the identification of anomalies by Machine Learning models.

- **Spatial dynamics:**
  - Median Consecutive Point Distance

$$\overline{d_{trip}} = median\left(\{d_i = dist(p_i, p_{i+1})\}_{i=1}^{N_{trip}-1}\right) [m] \quad (1)$$

- Maximum Consecutive Point Distance

$$d_{max,trip} = max\left(\{d_i = dist(p_i, p_{i+1})\}_{i=1}^{N_{trip}-1}\right)\ [m] \quad (2)$$

- Median Distance from Programmed Route

Given a programmed route R, let $s_i = dist(p_i, R)$, then:

$$\bar{s} = median\left(\{s_i\}_{i=1}^{N_{trip}}\right)\ [m] \quad (3)$$

- Maximum Distance from Programmed Route

$$s_{max} = max\left(\{s_i\}_{i=1}^{N_{trip}}\right)\ [m] \quad (4)$$

- **Temporal dynamics:**
  - Median Consecutive Point Time Difference

$$\overline{t_{trip}} = median\left(\{t_{i+1} - t_i\}_{i=1}^{N_{trip}-1}\right)\ [s] \quad (5)$$

  - Maximum Consecutive Point Time Difference

$$t_{max,trip} = max\left(\{t_{i+1} - t_i\}_{i=1}^{N_{trip}-1}\right)\ [s] \quad (6)$$

- **Operational integrity:**
  - Density of GPS Points

Let N_{trip} be the number of GPS points describing a trip, and let L_{trip} be the cumulative distance of the real GPS trajectory, computed as:

$$L_{trip} = \sum_{i=i}^{N_{trip}-1} dist(p_i, p_{i+1})\ [km] \quad (8)$$

Then, the Density of GPS Points is defined as:

$$\delta = \frac{N_{trip}}{L_{trip}} * 100\ [km^{-1}] \quad (7)$$

  - Out of Route Percentage

Let $N_{trip,Out\ of\ Route}$ be the number of GPS points marked as Out of *Route* by the AVM system, i.e., when the bus is more than 50 meters away from the planned route:

$$P_{\%,Out\ of\ Route} = \frac{N_{trip,Out\ of\ Route}}{N_{trip}}\ [\%] \quad (9)$$

- **Route compliance:**
  - Detected vs. Programmed Distance Ratio

Let $L_{theoretical}$ be the programmed distance of the route:

$$l_{\%,trip} = \frac{L_{trip}}{L_{theoretical}}\ [\%] \quad (10)$$

  - Proximity to Nearest Depot

Let $D_{depot,i}$ be the distance from point p_i to the nearest depot:

$$prox_{depot} = \min_{1 \le i \le N_{trip}} (D_{depot,i})\ [m] \quad (11)$$

  - Proximity to Starting Terminal

Let $D_{st,i}$ be the distance from point p_i to the starting terminal:

$$prox_{st} = \min_{1 \le i \le N_{trip}} (D_{st,i})\ [m] \quad (12)$$

- Proximity to Arrival Terminal

  Let $D_{at,i}$ be the distance from point p_i to the arrival terminal:

$$prox_{at} = \min_{1 \le i \le N_{trip}} (D_{at,i}) \; [m] \tag{13}$$

Maximum Speed $v_{\max}$

Let v_i be the average speed of point p_i, computed as

$$v_i = \frac{dist(p_i p_{i-1})}{t_i - t_{i-1}} \; [ms^{-1}] \tag{14}$$

Then, the Maximum Speed is defined as:

$$v_{\max} = \max_{1 \le i \le N_{trip}} v_i \; [ms^{-1}] \tag{15}$$

3.3 Preliminary Classification Model

A preliminary classification model with features that refer to entire trips has been developed. This model inputs the trip-based data set with numeric features calculated and selected in the previous section. The first step of the model involves performing a binary classification of trips into "Correct Trip" and "Anomalous Trip", based on outlier detection applied to the features characterizing each trip.

The second step integrates the binary output with information from the AVM system regarding the presence of points interpreted as Out of Route. The final output is a classification into four classes, given by the combination of the two steps, and the incorporation of this classification into the trip-based data set as an additional feature characterizing the trips. The scheme of the Preliminary Classification Model is depicted in Figure 3.

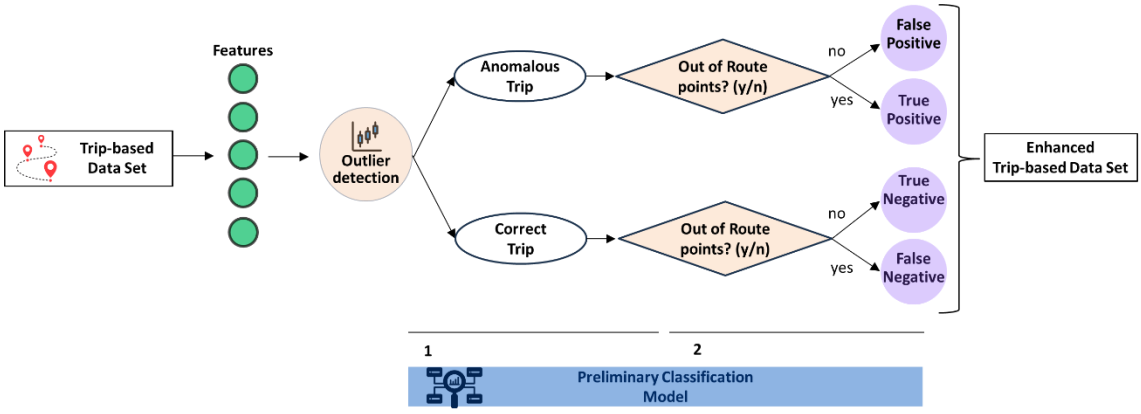

Figure 3: Preliminary Classification scheme.

3.4 Labeling

Preparing a labeled dataset for training and testing was necessary before advancing to the construction and application of Machine Learning models. To this end, data from a full operational day, a typical school-day Monday, of Autoguidovie was selected.

Autoguidovie is a private capital local public transport operator active across several regions in northern Italy.

The dataset, extracted from the AVM system, included one row per detected GPS point, totaling 211,860 data points across 1,179 bus trips. Each trip was manually labeled with one of the seven predefined anomaly classes (including Correct Trips). This labeling phase was essential to enable the supervised training, testing, and evaluation of the Machine Learning algorithms.

A strong class imbalance was initially present since many trips were labeled Correct. An under-sampling procedure was applied to prevent the model from learning in a biased or ineffective way: the number of Correct Trips was reduced to match the volume of the less frequent classes, resulting in a balanced dataset across all classes. Although this balancing step is not detailed in the results, it was fundamental to ensure reliable model performance.

3.5 Multi-class Classification Model

Two powerful Machine Learning algorithms were applied: Random Forest and XGBoost. Since XGBoost consistently outperformed Random Forest regarding prediction accuracy and overall reliability, only the results from XGBoost are presented. The model development process included the following steps:

1. 10-Fold Cross-Validation: the dataset was divided into ten parts to iteratively train and validate the model, ensuring robustness and generalization to unseen data.
2. Evaluation using classification metrics: the model was also assessed through standard classification metrics (e.g., sensitivity, specificity, balanced accuracy, etc.).
3. Parameter tuning: model hyperparameters were optimized iteratively to maximize accuracy and improve predictive performance.
4. Feature sensitivity analysis: important features influencing the model's predictions were identified and analyzed.

The Extreme Gradient Boosting (XGBoost) model yielded highly satisfactory results. The best-performing configuration achieved a mean accuracy of 88.72%, using 150 rounds, a maximum tree depth of 3, a learning rate (eta) of 0.7, and full subsampling, as shown in Table 1. The confusion matrix in Figure 4 illustrates robust performance across all classes, with particularly high accuracy for “Correct Trip” and “Interrupted Trip”.

The Evaluation metrics in Table 2, such as recall, precision, specificity, and balanced accuracy, further validate the model’s effectiveness: recall exceeds 0.88 for five out of the seven classes, and balanced accuracy remains consistently high, peaking at 95.9% for “Interrupted Trip”. Moreover, feature importance analysis reveals that the most influential variables were the Detected vs. Programmed Distance Ratio, the Preliminary Classification output, the Maximum Distance from Programmed Route, the Out of Route percentage, and the Proximity to the depot.

These features proved crucial in accurately distinguishing among the various anomaly classes. Overall, the model demonstrates excellent classification capabilities and offers strong potential for operational deployment.

Table 1: Parameters maximizing accuracy - Extreme Gradient Boosting.

Model	*Parameters*	*Values*	*Mean Accuracy*
XGBoost	nrounds	150	0.8872
	max depth	3	
	eta	0.7	
	subsample	1	

		REFERENCE						
		Signal gaps	Depot	Deviation	Double run	Aborted trip	Interrupted trip	Correct trip
PREDICTION	Signal gaps	29	0	0	0	0	2	2
	Depot	0	28	3	2	1	0	0
	Deviation	1	2	57	2	0	1	4
	Double run	0	3	1	48	0	0	0
	Aborted trip	0	0	0	0	15	1	0
	Interrupted trip	10	0	1	0	1	95	0
	Correct trip	2	1	4	0	0	0	74

Figure 4: Confusion matrix - Extreme Gradient Boosting.

Table 2: Evaluation metrics - Extreme Gradient Boosting.

	Signal gaps	*Depot*	*Deviation*	*Double run*	*Aborted trip*	*Interrupted trip*	*Correct trip*
Recall	0.690	0.824	0.864	0.923	0.882	0.960	0.925
Specificity	0.989	0.983	0.969	0.988	0.997	0.959	0.977
Precision	0.879	0.824	0.851	0.923	0.938	0.888	0.914
Neg Pred Value	0.964	0.983	0.972	0.988	0.995	0.986	0.981
Prevalence	0.108	0.087	0.169	0.133	0.044	0.254	0.205
Detection Rate	0.074	0.072	0.146	0.123	0.038	0.244	0.190
Detection Prevalence	0.085	0.087	0.172	0.133	0.041	0.274	0.208
Balanced Accuracy	0.839	0.903	0.916	0.956	0.940	0.959	0.951

3.6 Corrective actions

Besides identifying spatial anomalies in trajectories, the methodology's ultimate purpose is to use this information at an operational level to improve the service and ensure that each anomaly is appropriately addressed whenever possible.

This can be achieved by classifying anomalous trajectories on a sufficiently large sample of service days or weeks. Many anomalies may be found to repeat across days, persist particularly on a certain programmed route, or be associated with a specific bus or driver. Practical examples will be shown in the case study.

4 Case study: Autoguidovie

The classification methodology for spatial anomalies in bus trajectories was applied to real-world data from Autoguidovie, a public transport operator active in Italy in

Lombardia, Piemonte, and Veneto regions. The analysis focused on the service areas of Milano Sud-Est, Cremona, and Monza e Brianza, where the AVM systems are uniform, ensuring consistent data quality for accurate model application.

Data from April 15 to May 20, 2024, were analyzed using daily GPS tracks and weekly GTFS files. This period and dataset were chosen to validate the model's effectiveness in a controlled and homogeneous environment.

The application of the Extreme Gradient Boosting algorithm on the set of trips operated during the period under study leads to the classification displayed in the graph in Figure 5. Fortunately, most trips are identified as "Correct Trips". "Deviation" and "Interrupted Trip" classes represent 20.2% of anomalies. Anomalies constitute 34.2% of the trips.

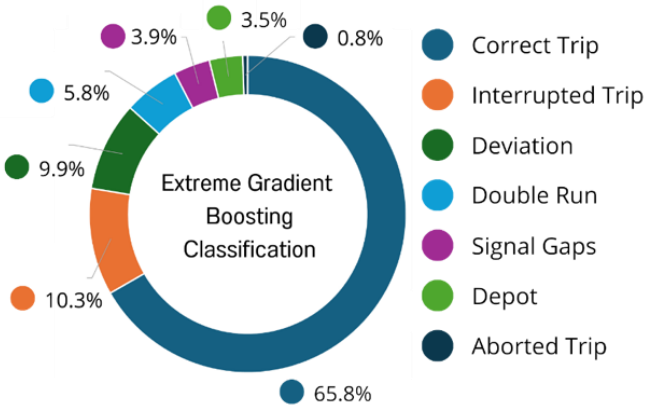

Figure 5: Donut Chart - Output of Extreme Gradient Boosting classification.

4.1 Corrective Actions

This section shows the practical implications of applying this model to a public transportation company. The great ability to analyze and classify thousands of trips in a matter of minutes allows us to observe patterns of repeated anomalies and errors, and analyzing the data by hand or with traditional tools would be very time-consuming and resource-intensive. Some cases have been analyzed based on the classification performed by the Extreme Gradient Boosting model, which has proven to be the most accurate. For example, it is possible to identify which routes are most affected by each type of anomaly. Additionally, identifying if there are specific vehicles (identified by their serial number) that, regardless of the route, recurrently exhibit the same problems, is achievable. Lastly, reviewing drivers' performance is also possible, as their behavior can generate certain anomalies. Thus, this classification model can monitor the service from a scheduling perspective (necessitating appropriate network redesign), find defects involving certain vehicles, and supervise drivers' actions.

A corrective action is shown in Figure 6. In this case, drivers were not following the programmed route (the black line), performing a detour, indicated by the stream of GPS points in red in Figure 6, generating trips classified as "Deviation" by the model. The drivers made this type of detour because a change was made to the traffic circulation: the original route from the scheduled plan has become a one-way street in the opposite direction. In this instance, Autoguidovie changed the route design associated with this type of error. Once this corrective action was taken, as seen in Figure 6, the programmed route and the actual route identified by the GPS trajectory began to coincide in this segment, no longer generating anomalies.

Assigning an economic value to this adjustment is also possible, with some approximations. The company is remunerated based on the kilometers covered. When the scheduled route was not adjusted, this detour increased by 450 meters for each trip. Over a year, 5229 trips are made on this route (data from 2024). Assuming a standard remuneration for public transport operators in a suburban environment (between 1.5 €/km and 2.5 €/km), the detour corresponds to a total unearned compensation ranging between approximately €3500 and €5800 per year. Given the magnitude of the detour, this is a modest amount but considering that a huge amount of these mistakes on important routes can happen, they can cause significant damage. Therefore, identifying and correcting these mistakes is critical.

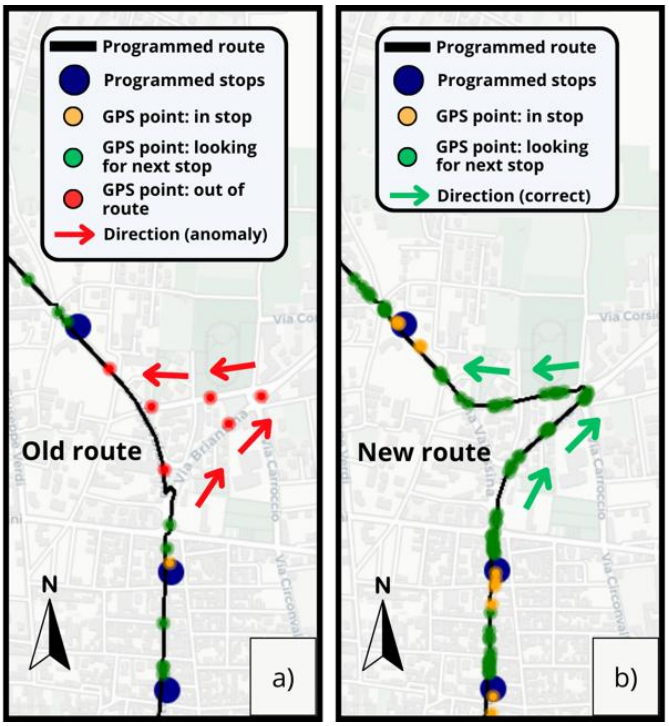

Figure 6: a): Trip affected by "Deviation" anomaly; b): Effect of route adjustment.

Another example of corrective action is shown in Figure 7, where two cases of an Interrupted Trip are shown. In both cases, which are associated with the same driver, there is a similar and repeated behavior: instead of completing the trip (black route), the driver stops the trip early. From a certain point onward, there are no more GPS points.

In the analysis period, XGBoost identified several cases of Interrupted Trips, generated by a total of 335 drivers. These cases are often caused by the misbehavior of drivers who, perhaps, see that no one is on board or are late, stop the journey ahead of time, skipping the last stops. This is inappropriate, and having a tool to easily identify this type of anomaly can be very helpful in mitigating this phenomenon.

It was observed that 36% of the drivers have several associated anomalies, fewer than 5. Additionally, 87% of the drivers have some anomalies with less than 15. The driver with the highest number of associated anomalies had 31. Two examples associated with the driver are shown in Figure 7.

In this case, it was possible to communicate with the driver to convey the importance of carrying out the entire service, because the public agency that entrusted public transportation to the company in question demands that the trips be completed. Skipping

a large number of stops can result in users who had counted on the bus being left without the service that should have been there.

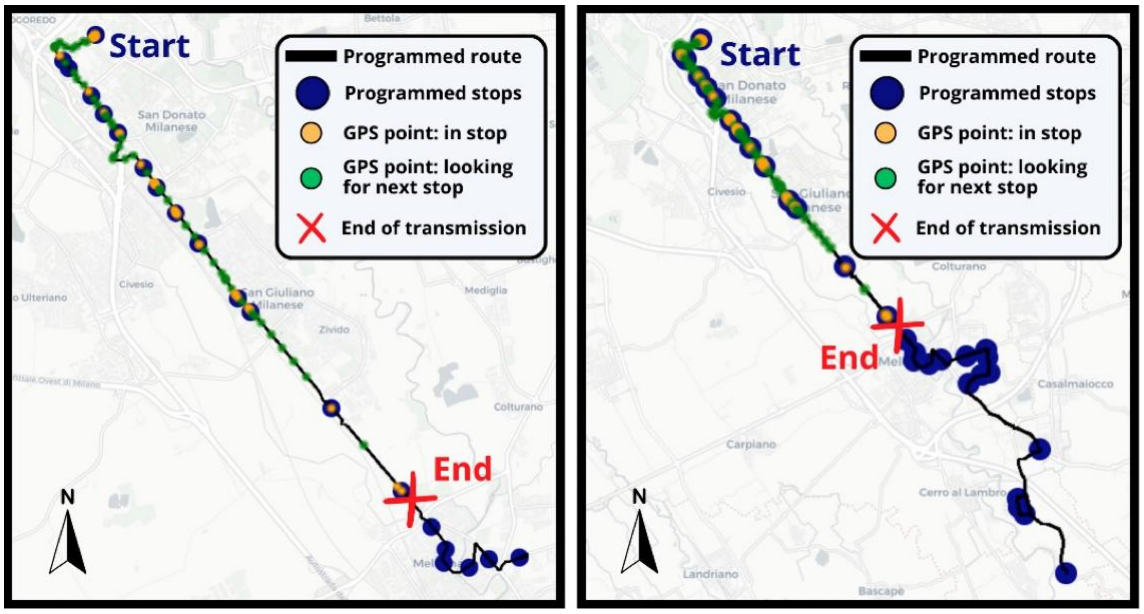

Figure 7: Two cases of "Interrupted Trip" anomaly associated with the same driver.

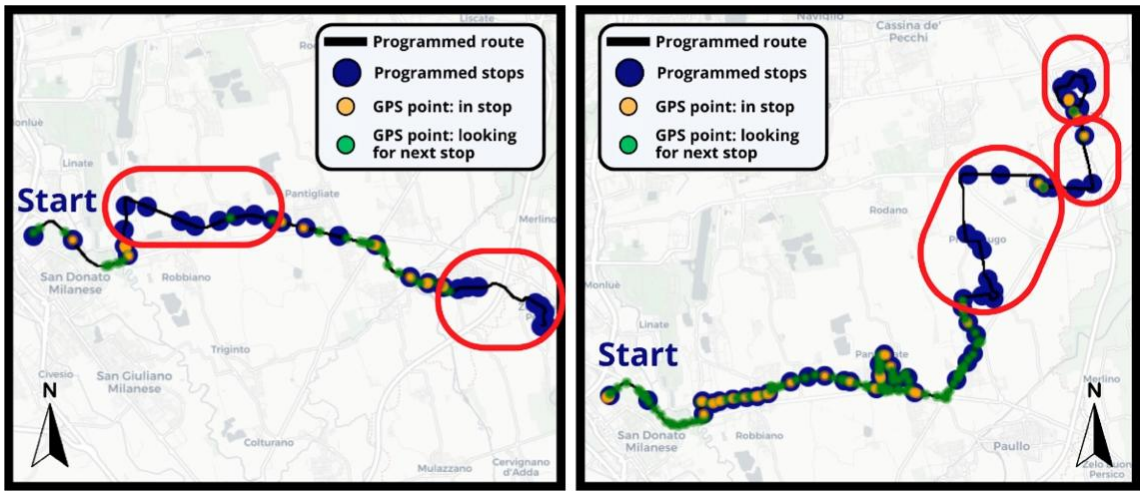

Figure 8: Two cases of "Signal gaps" anomaly associated with the same bus.

Lastly, we will show an example of how to deal with anomaly signal gaps. With cases like these, the potential issue of skipping scheduled stops becomes even clearer concerning those generated by Deviation anomalies, as they involve a greater number of programmed bus stops. If such occurrences happen frequently, they may breach the minimum quality conditions of the service contract between the operator and the public transport authority, potentially leading to penalties. This tool allows the monitoring of the situation, to ensure that such cases do not occur without serious and unavoidable reasons and to trace the underlying causes.

During the period considered for the analysis, the classification algorithm recognized a few cases of Signal Gaps. These were triggered by 190 buses. It was decided to analyze whether this type of anomaly is present randomly on all buses or whether there are specific buses where this anomaly is more frequent.

It was observed that throughout the period analyzed, out of 190 buses, 70% had several Signal Gaps type anomalies less than or equal to 5. Additionally, 93% presented a number less than 15 anomalies. The worst performance was by "Bus11", presenting 46 anomalies. Two examples of anomalies referred to Bus11 are shown in Figure 8. In such cases, it is appropriate to conduct an internal investigation, analyze the health of the bus transmission components, and perform the appropriate maintenance. Thus, with the classification of this type of case, the explained model is useful to monitor the status of the bus components involved in the transmission or collection of GPS data and to check if this is related to the degradation of the entire AVM equipment of the bus.

5 Conclusions

This study was conducted to develop and test a model that integrates the extensive GPS data from bus AVM systems with scheduled programming data to analyze bus trajectories. The best implemented model, Extreme Gradient Boosting, has demonstrated high accuracy and performance in the case study. The developed model has been made available to the Autoguidovie company to analyze service quality and to guide corrective actions. This automatic process is very useful for the company to improve service quality and to avoid contractual penalties or lower-than-expected kilometer compensation.

Some improvements can be made, such as enhancing the model's ability to analyze service areas with diverse AVM equipment, modifying anomaly classes (by merging or further differentiating them), extending the model to other transportation sectors, and integrating spatial anomalies with a punctuality analysis to explore potential correlations. Regardless of the possible improvements to the model, it has demonstrated how a Machine Learning approach to this type of data and in this sector can be extremely helpful in managing the vast amount of data available today, aiming for increasingly efficient mobility.

References

Autorità di Regolazione dei Trasporti (ART). Delibera n. 154/2019. Technical report, ART, 2019.

Barabino, B., Lai, C., Casari, C., Demontis, R., & Mozzoni, S. (2017a). Rethinking Transit Time Reliability by Integrating Automated Vehicle Location Data, Passenger Patterns, and Web Tools. IEEE Transactions on Intelligent Transportation Systems, 18(4), 756–766. https://doi.org/10.1109/TITS.2016.2585342

Barabino, B., di Francesco, M., & Mozzoni, S. (2017b). An Offline Framework for the Diagnosis of Time Reliability by Automatic Vehicle Location Data. IEEE Transactions on Intelligent Transportation Systems, 18(3), 583–594. https://doi.org/10.1109/TITS.2016.2581024

Barabino, B., Bonera, M., Maternini, G., Porcu, F., & Ventura, R. (2024). Refining a crash risk framework for urban bus safety assessment: Evidence from Sardinia (Italy). Reliability Engineering and System Safety, 245. https://doi.org/10.1016/j.ress.2024.110003

Ye, Z., Wang, C., Yu, Y., Shi, X., & Wang, W. (2016). Modeling level-of-safety for bus stops in China. Traffic Injury Prevention, 17(6), 656–661. https://doi.org/10.1080/15389588.2015.1133905

Kumar, B. A., Kumar, V., Vanajakshi, L., & Subramanian, S. C. (2017). Performance comparison of data driven and less data demanding techniques for bus travel time prediction. European Transport - Trasporti Europei, 65.

Kong, X., Song, X., Xia, F., Guo, H., Wang, J., & Tolba, A. (2018). LoTAD: long-term traffic anomaly detection based on crowdsourced bus trajectory data. World Wide Web, 21(3). https://doi.org/10.1007/s11280-017-0487-4

Raymond, R., & Imamichi, T. (2016). Bus trajectory identification by map-matching. Proceedings - International Conference on Pattern Recognition, 0. https://doi.org/10.1109/ICPR.2016.7899868

Singh, S. K., Anand, U., Patel, A., & Boro, D. (2024). Driving Behavior Analysis Using Deep Learning on GPS Data. Lecture Notes in Electrical Engineering, 1061 LNEE. https://doi.org/10.1007/978-981-99-4362-3_29

Zou, Q., Xiong, W., Wang, X., & Qin, F. (2023). Research on Real-Time Anomaly Detection Method of Bus Trajectory Based on Flink. Electronics (Switzerland), 12(18). https://doi.org/10.3390/electronics12183897

Breiman, L. (2001). Random forests. Machine Learning, 45(1). https://doi.org/10.1023/A:1010933404324

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 13-17-August-2016. https://doi.org/10.1145/2939672.2939785