

International Journal Research Publication Analysis

Page: 01-10

ANOMALY DETECTION IN URBAN MOBILITY DATA FOR IDENTIFYING UNUSUAL PASSENGER DEMAND PATTERNS

***Golla M. Moulika, G. N. R. Prasad**

Department of MCA (Masters in Computer Applications) Chaitanya Bharathi Institute of Technology (CBIT), Gandipet, Hyderabad, Telangana, India.

Article Received: 05 September 2026

Article Revised: 25 September 2026

Published on: 15 October 2026

***Corresponding Author: Golla M. Moulika**

Department of MCA (Masters in Computer Applications) Chaitanya Bharathi Institute of Technology (CBIT), Gandipet, Hyderabad, Telangana, India.

DOI: <https://doi-doi.org/101555/ijrpa.9233>

ABSTRACT

Urban public transportation systems generate large volumes of mobility data covering passenger demand, routes, transport type, time, weather, and occupancy. This paper reports a data-driven anomaly detection framework, applied to a 10,000-record demonstration dataset spanning 20 bus and metro routes over six months, that identifies unusual passenger-demand observations without requiring labelled data. The pipeline covers preprocessing, exploratory analysis, feature engineering, Isolation Forest as the primary detector, and DBSCAN as a density-based comparison. On the demonstration dataset, Isolation Forest (contamination = 0.05) flags 500 of 10,000 observations (5.00%) as anomalous, while DBSCAN (eps = 0.5, min_samples = 10) labels 198 observations (1.98%) as noise, with 102 observations — 51.5% of DBSCAN's noise points — flagged by both methods. Anomalies concentrate on bus routes (7.26% anomaly rate versus 2.78% for metro), peak hours (8.60% versus 3.48% off-peak), and weekends (Sunday 8.55%, Saturday 6.16%), and align closely with the dataset's existing crowd classification (51.6% of “High” crowd-level records are flagged, against 3.05% of “Low” records). The framework gives transportation planners a reproducible, statistically grounded view of unexpected demand variation to support capacity planning and resource allocation.

INDEX TERMS: Urban Mobility, Public Transportation, Anomaly Detection, Isolation Forest, Dbscan, Passenger Demand, Occupancy Rate, Data Analytics.

1. INTRODUCTION

Urban transportation systems generate large quantities of structured and semi-structured data through daily passenger activity. Information such as passenger counts, routes, transport types, time of travel, weather conditions, and vehicle capacity provides valuable insight into transportation demand. As data-driven transportation management grows, analyzing this information has become important for improving operational planning and resource utilization.

Traditional transportation analytics often focuses on descriptive statistics or passenger-flow prediction. While prediction can estimate expected demand, it does not necessarily identify observations that differ substantially from normal behaviour. For example, a route that generally carries 70–90 passengers during a given period may suddenly record far more; such an observation may represent an unusual demand pattern requiring further investigation.

Anomaly detection provides an alternative by identifying observations that deviate from the underlying patterns in the data without requiring every observation to be manually labelled as normal or abnormal — useful because real-world transportation datasets rarely contain confirmed anomaly labels. Recent research has applied Isolation Forest to shared mobility while incorporating contextual information such as weather and transit availability, and other studies have investigated abnormal passenger flow in urban rail systems.

This paper applies Isolation Forest as the primary anomaly detector and DBSCAN as a comparative density-based technique to a real 10,000-record urban transit dataset, moving from simply predicting crowd levels to identifying observations that deviate from normal demand patterns and examining the operating conditions under which they occur.

2. RELATED WORK

Passenger-flow analysis has been extensively studied in intelligent transportation systems, but traditional work has focused on forecasting demand and recurring patterns rather than abnormal behaviour, which need not follow historical trends.

Wang et al. investigated early warning of burst passenger flow using abnormal out-flow detection followed by flow estimation, demonstrating the link between abnormal behaviour and subsequent surges [3]. A 2025 study combined convolutional and recurrent neural networks to detect abnormal passenger flow during emergencies in urban rail transit [4].

Isolation Forest [1] was introduced as an anomaly-detection method that isolates observations directly rather than first modelling normal data, and is designed to be computationally efficient on large datasets. DBSCAN [2] provides a density-based alternative in which points

in insufficiently dense regions are treated as noise; DBSCAN-based techniques have also been applied to public-transport smart-card data to analyze travel patterns [6], [7]. This paper combines both methods with contextual analysis of passenger demand in a single, interpretable framework.

3. PROBLEM STATEMENT AND OBJECTIVES

Public transportation demand varies with time, route, weather, day type, and transport mode. Although prediction models can estimate expected demand, unusual observations may not be adequately identified. The problem addressed is:

How can unsupervised machine learning techniques identify unusual passenger-demand patterns in urban public transportation data, and what contextual factors are associated with these anomalies?

The objectives are to: (i) preprocess and analyze the transportation dataset; (ii) identify the factors influencing demand; (iii) detect unusual observations using Isolation Forest; (iv) apply DBSCAN for density-based comparison; (v) compare the observations flagged by both approaches; (vi) analyze anomalies by route, time, weather, transport type, and occupancy; and (vii) visualize the results to support capacity planning and resource allocation.

4. DATASET AND PREPROCESSING

4.1 Dataset Description

The study uses dataset.csv, 10,000 observations across 13 attributes covering 20 routes (Route-101 to Route-120), 2 transport types (4,956 bus and 5,044 metro records), 10 stops, 4 weather conditions, and dates from 1 January to 30 June 2026 (Table 1).

Table 1: Attributes of the Urban Transit Dataset.

Attribute	Description
Date, Time, Day	Temporal identifiers for each observation
Route, Stop, Transport_Type	Route/stop identifiers; Bus or Metro
Weather, Temperature_C	Environmental conditions at time of travel
Passengers, Capacity	Recorded ridership and maximum vehicle capacity
Occupancy_Rate	Passengers as a percentage of capacity
Peak_Hour, Crowd_Level	Peak-period flag and existing crowd classification

4.2 Data Quality and Preprocessing

Quality checks found no missing values and no fully duplicated rows across the 10,000 records. Sixteen records (0.16%) show an occupancy rate above 100%, indicating vehicles recorded as carrying more passengers than nominal capacity; these are retained, since over-capacity loading is itself a demand signal rather than a data error. Preprocessing converts Time into an Hour field, encodes Day, Peak_Hour and Transport_Type, and recomputes Occupancy_Rate as

$$\text{Occupancy Rate} = (\text{Passengers} \div \text{Capacity}) \times 100$$

For example, 100 passengers against a capacity of 200 gives an occupancy rate of 50%. Bus records average 77.8 passengers ($\sigma = 43.1$) against a 200-passenger nominal capacity, while metro records average 179.1 passengers ($\sigma = 44.5$) against a 500-passenger capacity — confirming the two modes operate at different scales, which motivates including Capacity alongside Passengers as a model feature.

5. PROPOSED METHODOLOGY

The workflow has eight stages: data preprocessing, exploratory data analysis, feature engineering, Isolation Forest detection, DBSCAN comparison, comparative analysis, contextual analysis, and visualization (Fig. 1, Section 7).

5.1 Exploratory Data Analysis

EDA examines the passenger-count distribution, hourly and route-wise demand, weather effects, and the relationship between occupancy and available capacity, establishing what “normal” demand looks like before anomaly detection is applied (Figs. 2–3).

5.2 Feature Engineering

Raw fields are converted into the numeric feature vector used by both models:

$$\mathbf{X} = [\text{Passengers}, \text{Capacity}, \text{Occupancy Rate}, \text{Temperature}, \text{Hour}]$$

Day of week, Peak_Hour, Route and Transport_Type are retained separately for the contextual analysis in Section 9.

6. ISOLATION FOREST AND DBSCAN

6.1 Isolation Forest

Isolation Forest isolates each observation by recursive random partitioning; anomalous points, being sparse or extreme, are isolated in fewer partitions than normal points [1]. This study uses 200 trees with contamination set to 0.05 (an assumed anomaly rate of 5%), applied

directly to X . Each observation receives an anomaly score and a label: 1 for normal, -1 for anomalous.

6.2 DBSCAN

DBSCAN groups standardized observations into dense regions defined by a neighbourhood radius (eps) and a minimum point count (min_samples); points outside every dense region are labelled noise [2]. A k-distance plot ($k = 5$) on the standardized features indicated a working radius near 0.5; with eps = 0.5 and min_samples = 10, DBSCAN converges to two clusters, corresponding closely to the bus-scale and metro-scale occupancy bands identified in Section 4.2.

7. SYSTEM ARCHITECTURE

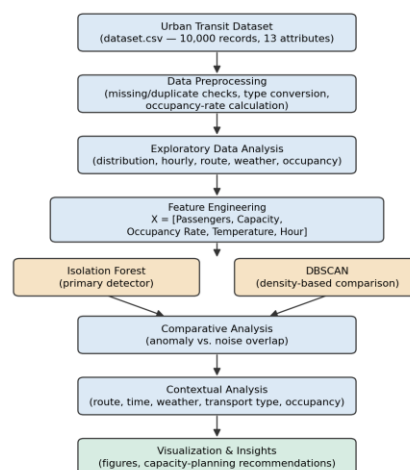


Fig. 1. Architecture of the anomaly-detection workflow, from raw dataset to visualization and insights.

8. IMPLEMENTATION FRAMEWORK

Table 2: Technology Stack.

Technology	Role in the project
Python 3	Data processing and machine learning
Pandas, NumPy	Cleaning, aggregation and numerical operations
Scikit-learn	IsolationForest and DBSCAN implementations
Matplotlib	Static visualization of results
CSV	Dataset storage (dataset.csv, 10,000 rows)

9. EXPERIMENTAL RESULTS AND ANALYSIS

The results below come from running Isolation Forest and DBSCAN on the full 10,000-record dataset.

9.1 Overall Detection Counts

Table 3: Isolation Forest vs. DBSCAN — Detection Summary.

Metric	Isolation Forest	DBSCAN
Total observations	10,000	10,000
Anomalies / noise points	500	198
Percentage of dataset	5.00%	1.98%
Parameters used	contamination = 0.05	eps = 0.5, min_samples = 10
Clusters found	—	2

Of the 198 DBSCAN noise points, 102 (51.5%) are also flagged by Isolation Forest, and these 102 form 20.4% of Isolation Forest's 500 anomalies. The two methods therefore agree on roughly half of DBSCAN's noise set while Isolation Forest, with its higher target rate, additionally flags a further set of borderline observations that DBSCAN's density criterion treats as part of the bus- or metro-scale cluster. Agreement between the two is treated as a statistical signal of unusualness, not as confirmation of a real-world incident.

9.2 Route-Wise Distribution

Anomalies are not concentrated on a single route: the five routes with the most flagged observations are Route-120 (33), Route-118 (31), Route-117 (30), Route-112 (30) and Route-104 (29), out of 20 routes averaging 25 anomalies each — a spread consistent with generally distributed, rather than route-specific, demand variation (Fig. 4).

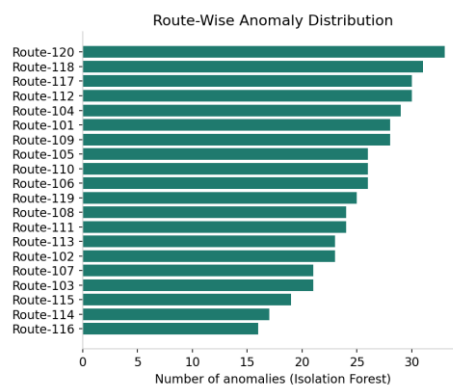


Fig. 4. Route-wise anomaly counts (Isolation Forest).

9.3 Transport Type

Bus records show a substantially higher anomaly rate (360 of 4,956 records, 7.26%) than metro records (140 of 5,044, 2.78%). Buses appear more exposed to irregular loading — plausibly due to smaller nominal capacity and greater sensitivity to local events — while metro demand is comparatively steady.

9.4 Time and Day Effects

Anomalies are more frequent during peak hours (255 of 2,966 peak records, 8.60%) than off-peak (245 of 7,034, 3.48%). By hour of day, anomalies cluster around 20:00 (52), 09:00 (44), 07:00 (40) and 08:00 (34) — the morning and evening commute windows — alongside a smaller late-night group at 00:00 (38). By day of week, Sunday (123 anomalies, 8.55% of Sunday records) and Saturday (88, 6.16%) show markedly higher anomaly rates than weekdays (3.15%–4.75%), suggesting weekend travel is less predictable than the routine weekday commute (Fig. 5).

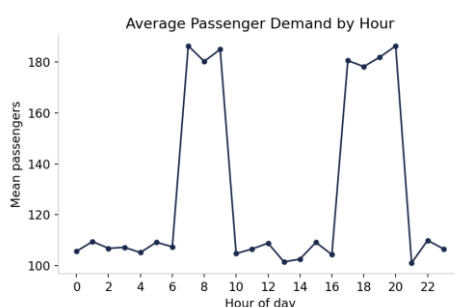


Fig. 5. Average passenger demand by hour of day.

9.5 Weather

Anomaly rates are fairly uniform across weather conditions — Rainy 5.30%, Clear 5.22%, Cloudy 4.58%, Hot 4.57% — indicating that in this dataset weather is a secondary factor compared with time-of-day and transport type (Fig. 6).

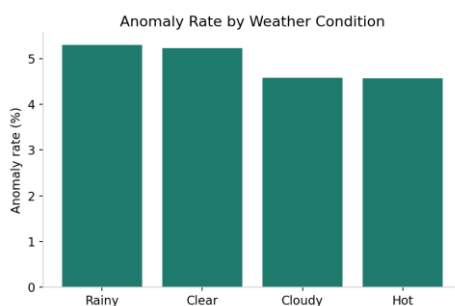


Fig. 6. Anomaly rate by weather condition.

9.6 Occupancy and Crowd-Level Validation

Flagged observations run higher on occupancy than normal ones: a mean of 43.40% and median of 49.95% for anomalies, against a mean of 37.02% and median of 34.20% for normal observations, with anomaly occupancy spanning 2.5%–111.5% (both very low and very high occupancy can be unusual). This is corroborated by the dataset's independent Crowd_Level field, which the model never sees directly: 51.63% of “High” crowd-level records are flagged as anomalous, against 7.95% of “Medium” and only 3.05% of “Low” records — a strong external check that the statistical anomalies correspond to genuinely unusual crowding (Figs. 7–8).

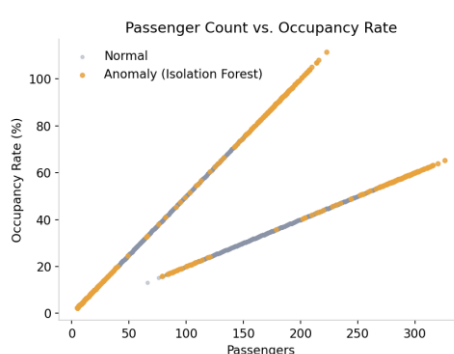


Fig. 7. Passenger count vs. occupancy rate, normal vs. Isolation Forest anomalies.

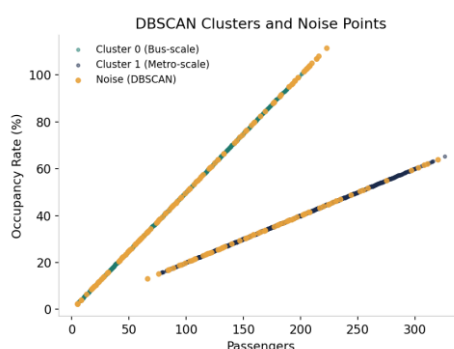


Fig. 8. DBSCAN clusters (bus-scale and metro-scale) and noise points.

10. DISCUSSION

The two detectors are complementary rather than redundant. Isolation Forest, tuned to a fixed 5% contamination rate, gives a controllable anomaly budget useful for operational alerting, while DBSCAN's density criterion is parameter-driven and yields a smaller, more conservative noise set that still captures roughly half of what Isolation Forest flags. The clearest, most consistent finding across both methods and the independent crowd-level field is that unusual demand is driven far more by when and on what mode a trip occurs (peak

hours, weekends, bus routes) than by weather. This is broadly consistent with prior transportation-anomaly studies that link abnormal flow to temporal and operational factors, while the crowd-level agreement (Section 9.6) provides a form of external validation not available in purely synthetic benchmarks.

11. APPLICATIONS AND ADVANTAGES

The framework supports capacity planning (flagging periods that may need additional service, especially bus routes at peak hours and weekends), resource allocation toward routes with recurring anomalies, and operational monitoring alongside existing crowd-prediction systems. It requires no manually labelled anomaly data, combines an established isolation-based method with a density-based cross-check, and scales to the 10,000-record dataset used here without difficulty.

12. LIMITATIONS AND FUTURE SCOPE

The framework flags statistical deviations; it does not itself establish a real-world cause, and results depend on the choice of contamination, `eps` and `min_samples`. The dataset is static, so genuine real-time alerting would require a streaming pipeline, and external factors such as public events, road closures or service disruptions are not captured. Future work could add streaming analytics, autoencoder or Transformer-based detectors, explainable anomaly scoring, and integration of external event data.

13. CONCLUSION

This paper presents and evaluates a data-driven framework for identifying unusual passenger-demand patterns in a real 10,000-record urban transit dataset. Isolation Forest and DBSCAN flag 5.00% and 1.98% of observations respectively, with 51.5% overlap on DBSCAN's noise set, and both align with the dataset's independent crowd classification. Anomalies concentrate on bus routes, peak hours and weekends rather than on weather, giving transportation planners a concrete, statistically grounded basis for capacity planning and resource allocation. Future work can extend the framework toward real-time processing, explainable anomaly detection, and external event integration.

14. REFERENCES

1. F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation Forest," in Proc. 8th IEEE Int. Conf. Data Mining, 2008, pp. 413–422.

2. M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, “A density-based algorithm for discovering clusters in large spatial databases with noise,” in Proc. 2nd Int. Conf. Knowledge Discovery and Data Mining, 1996.
3. H. Wang, L. Li, P. Pan, Y. Wang, and Y. Jin, “Early warning of burst passenger flow in public transportation system,” *Transportation Research Part C*, vol. 105, pp. 580–598, 2019.
4. L. Shao and L. Song, “Data-driven detection of passenger flow anomalies in urban rail transit systems during emergencies,” *Automation in Construction*, vol. 178, 106393, 2025.
5. “Towards Explainable Anomaly Detection in Shared Mobility Systems,” *IFAC-PapersOnLine*, vol. 59, no. 26, pp. 241–246, 2025.
6. N. Fulman, M. Marinov, and I. Benenson, “Investigating occasional travel patterns based on smartcard transactions,” *Transport Policy*, vol. 141, pp. 152–166, 2023.
7. “Increasing the precision of public transit user activity location detection from smart card data analysis via spatial–temporal DBSCAN,” *Data & Knowledge Engineering*, vol. 153, 102343, 2024.
8. Scikit-learn, “`sklearn.ensemble.IsolationForest` / `sklearn.cluster.DBSCAN`.” [Online]. Available: <https://scikit-learn.org>