**Multidisciplinary Journal of Research in Engineering and Technology**

ISSN: 2348-6953

Volume 13 Issue 2 (2026) | Pages: 254-259

# Bridging the Gap: A Data-Driven Approach Data-Driven Solutions for Public Transportation Challenges in India

Kalpana S. Dhende[1], P. A. Kamble[2], S. P. Kulkarni[3], S. V. Bhartal[4], A. R. Chandure [5], S. V. Changulpai [6]

[123456]MCA, MES IMCC, Pune, Maharashtra, India

[1]ksd.imcc@mespune.in, [2]theprasadanilkamble@gmail.com, [3]shantanukulkarni365@gmail.com, [4]sanskruti.bhartal@gmail.com, [5]rarnavchandurerox@gmail.com, [6]saurabhchangulpai2305@gmail.com

**Peer Review Information**

***Type**: Article*
***Received:** 28 March 2026*
***Revised:** 13 April 2026*
***Accepted:** 19 May 2026*
***Published:** 01 June 2026*

**Abstract**

This research investigates the operational inefficiencies in India's public transportation system through a case study of Pune Mahanagar Parivahan Mahamandal Limited (PMPML). Rapid urbanization has placed significant pressure on existing transport infrastructure, leading to issues such as overcrowding, unreliable services, and declining operational efficiency. This study utilizes data from PMPML's annual reports along with geospatial analysis of bus stop distributions to evaluate key performance indicators, including fleet utilization, service reliability, passenger demand, and complaint trends. The analysis reveals a critical operational paradox where mechanical reliability has improved, yet service performance has deteriorated, as evidenced by increasing service cancellations and passenger complaints. By integrating statistical and geospatial techniques, the study identifies inefficiencies in fleet deployment, demand–supply mismatches, and uneven spatial distribution of services. To address these challenges, the paper proposes a data-driven framework leveraging data science, artificial intelligence, and machine learning techniques, including demand forecasting models, route optimization algorithms, and real-time scheduling systems. These approaches enable dynamic decision-making, improved resource allocation, and enhanced service reliability. The findings demonstrate that the observed inefficiencies are not primarily due to infrastructure limitations but rather the absence of intelligent, data-driven management systems. The study emphasizes the need for a paradigm shift toward predictive and adaptive transport systems to improve efficiency, passenger satisfaction, and sustainability in urban public transportation.

**Keywords:** Public Transportation, Data-Driven Transportation, Geospatial Analysis, Route Optimization, Demand Forecasting.

**How to Cite This Article**

Dhende, K. Kamble, P. Kulkarni, S. Bhartal, S. Chandure, A. Changulpai, S. (2026). Bridging the Gap: A Data-Driven Approach Data-Driven Solutions for Public Transportation Challenges in India. *Multidisciplinary Journal of Research in Engineering and Technology*13(2), 254–259.

## Introduction

Public transportation serves as the backbone of urban mobility and plays a crucial role in fostering economic growth and social inclusivity, particularly in rapidly urbanizing nations such as India. With the continuous rise in urban population, public transport systems are under increasing pressure to meet growing travel demands. The expansion of urban centers has led to a significant imbalance between transportation demand and available infrastructure, resulting in overcrowding, reduced reliability, slow transit speeds, poor coordination, and compromised passenger experience.[1]

A major factor contributing to these inefficiencies is the limited integration of modern software systems and data-driven technologies in the planning and operation of public transportation.[1] The absence of real-time data analytics, predictive modeling, and intelligent scheduling mechanisms restricts the ability of transport authorities to optimize resource utilization and respond dynamically to changing demand patterns. This technological gap not only affects commuter experience but also has broader implications for environmental sustainability and economic productivity.[2] Challenges such as inefficient fleet utilization, service cancellations, increasing operational costs, and rising emissions further exacerbate the problem.

To investigate these challenges in a real-world context, this study focuses on the Pune Mahanagar Parivahan Mahamandal Limited (PMPML), the primary public bus transport provider in Pune. Using data extracted from PMPML's annual reports along with geospatial analysis of bus stop distributions, this research examines key operational indicators such as fleet utilization, service reliability, passenger demand, and complaint trends. By integrating statistical analysis with geospatial techniques, the study aims to identify systemic inefficiencies and uncover spatial patterns that influence service performance.

The objective of this research is to demonstrate how a data-driven approach can provide actionable insights for improving public transportation systems. By analyzing real operational data from PMPML, this paper highlights the limitations of traditional management practices and emphasizes the need for adopting data science, artificial intelligence, and machine learning techniques for optimized route planning, demand prediction, and service management. The findings of this study aim to contribute toward the development of more efficient, reliable, and sustainable urban transportation systems.

## Literature Review

Public transportation systems in India face numerous structural and operational challenges that significantly impact efficiency and service quality. Overcrowding remains a persistent issue, with buses and trains frequently operating beyond their designed capacity, leading to passenger discomfort and safety concerns [3]. Additionally, unreliable services—characterized by inconsistent schedules, delays, and poor intermodal coordination—continue to undermine commuter trust and system effectiveness[3], [4]. Financial constraints further exacerbate these problems, as low fare structures limit revenue generation, restricting maintenance, infrastructure upgrades, and modernization efforts [3].

Infrastructure-related limitations also play a critical role in reducing system efficiency. Congested and narrow road networks, along with poor surface conditions, hinder smooth transit operations [4]. The lack of effective multimodal integration increases waiting times and results in inefficient transfers between different transport modes [3]. Accessibility challenges remain prominent, particularly for rural populations and individuals with disabilities, limiting equitable access to public transportation [3]. Furthermore, insufficient adoption of modern technologies—such as automated tracking systems, real-time monitoring, and route optimization tools—restricts the ability of transport systems to improve operational performance [4].

In contrast, global public transportation systems demonstrate the transformative potential of technological integration. Advanced digital solutions have significantly improved operational efficiency through optimized route planning, scheduling, and resource allocation [5]. Passenger experience has also been enhanced through real-time information systems, mobile ticketing, and integrated payment platforms [6]. Moreover, predictive maintenance enabled by artificial intelligence (AI) and Internet of Things (IoT) technologies allows early detection of system failures, reducing downtime and improving reliability [5], [7]. Cities such as Singapore, Stockholm, Helsinki, and Seoul exemplify successful large-scale implementation of intelligent transportation systems and integrated digital infrastructure [6].

The application of data science, artificial intelligence, and machine learning has further expanded the capabilities of modern transportation systems. Machine learning models are widely used for demand forecasting and route optimization, enabling transit authorities to dynamically adjust routes and schedules based on predicted passenger demand [8]. Predictive maintenance systems leverage historical and sensor data to proactively schedule repairs and minimize service disruptions [9]. Additionally, AI-driven traffic management systems optimize signal timings to improve traffic flow, while intelligent passenger information systems provide real-time updates and personalized travel recommendations [10]. Case studies in developing regions demonstrate the effectiveness of models such as Random Forest (RF) and Long Short-Term Memory (LSTM) networks in addressing complex transportation challenges [8], [9].

Despite these advancements, there remains a significant gap in the application of data-driven approaches within Indian public transportation systems, particularly in integrating operational data with spatial analysis for decision-making. Existing studies largely focus on theoretical models or isolated implementations, with limited emphasis on real-world system-level analysis. This research addresses this gap by utilizing PMPML operational data [11] in conjunction with geospatial analysis [12] to demonstrate how data-driven methodologies can be applied to identify inefficiencies and improve urban public transportation systems.

**Dataset & Preprocessing**

Data Preprocessing and Feature Engineering Techniques:

- Cleaning and validating geospatial data by removing missing, duplicate, and invalid latitude–longitude records.
- Standardizing coordinate systems (e.g., WGS84) for consistent spatial analysis.
- Transforming tabular data into geospatial formats (GeoPandas) for mapping and spatial operations.
- Structuring and sequencing bus routes to reconstruct actual travel paths.
- Detecting and removing spatial outliers to improve data accuracy.
- Computing distance-based features such as inter-stop distance and total route length (Haversine formula).
- Applying clustering techniques (e.g., DBSCAN, K-Means) to identify high-density regions and transit hubs.
- Modelling the transport network as a graph (stops as nodes, routes as edges) to extract connectivity metrics.
- Engineering route-level features such as number of stops, route overlap, and service frequency indicators.

**Proposed Model / Methodology**

This study adopts a data-driven analytical framework by integrating PMPML operational data with geospatial analysis of bus routes and stops to identify inefficiencies in urban public transportation systems.

*Data Sources*

The analysis is based on two primary datasets:

- **PMPML Annual Report Data:**
  Provides operational metrics such as fleet size, vehicle utilization, number of operated schedules, passenger count, cancellations, revenue, and passenger complaints.
- **PMPML Bus Route and Stop Dataset:**
  Contains detailed geospatial and structural information including route names, stop sequences, and latitude–longitude coordinates of bus stops.

Additional supporting data sources such as government open data platforms (e.g., www.kaggle.com, data.opencity.in) and transit APIs are referenced for contextual understanding and validation.

*Dataset and Annual Report Description*

The PMPML Annual Report (2023–2024 and 2024–2025) provides comprehensive operational data, including fleet statistics, service performance metrics, passenger trends, revenue details, and maintenance indicators, enabling a comparative analysis of system efficiency over time

The PMPML Bus Route Dataset consists of multiple structured tables capturing route and stop-level information, as summarized below in TABLE 1

*Data Preprocessing*

To ensure data quality and consistency, the following preprocessing steps are performed:

- Cleaning missing, duplicate, and inconsistent records
- Standardizing formats for numerical and categorical variables
- Converting latitude–longitude data into geospatial objects using GeoPandas

*Statistical Analysis of PMPML Data*

The operational dataset is analyzed to extract key performance indicators:

- Fleet utilization and vehicle deployment efficiency
- Demand–supply gap between sanctioned and operated schedules
- Service reliability through cancelled kilometres
- Passenger trends and complaint analysis

This analysis identifies system-level inefficiencies in transport operations.

*Geospatial Analysis of Bus Routes and Stops*

Geospatial techniques are applied to analyze spatial patterns:

- Mapping bus stops and routes using GIS tools
- Generating heatmaps to identify high-density service regions
- Applying clustering algorithms (e.g., DBSCAN) to detect transit hubs
- Identifying under-served and over-served areas

This step helps determine where inefficiencies occur geographically.4

**TABLE 1:** PMPML Bus Route Dataset Features

| **Table** | **Rows** | Columns |
|---|---|---|
| Main | 9 | 2 |
| 376 Rout name, Stage & LL | 34089 | 24 |
| BRTS UNIQUE STOP NAME | 1116 | 11 |
| Non BRTS UNIQUE STOP NAME | 4318 | 8 |
| Route Description | 1030 | 9 |
| Short Rout name Stage | 1428 | 9 |

**Results & Performance Metrics**

*Fleet Utilization and Operational Efficiency*

The analysis of PMPML data reveals a decline in operational efficiency between 2023–24 and 2024–25. as shown in Figure 1. The total number of vehicles held per day decreased from 2066 to 1933, while the average number of vehicles on road declined from 1658 to 1558 . Additionally, fleet utilization reduced from 74.44% to 71.03%.

*Figure SEQ Figure * ARABIC 1: Fleet Utilization and Operational Efficiency*

This indicates that despite maintaining a relatively stable fleet size, the effective deployment of vehicles has decreased. The reduction suggests inefficiencies in fleet scheduling and allocation, potentially due to the absence of real-time monitoring and optimization systems.

*Demand–Supply Gap in Service Operations*

A significant discrepancy was observed between scheduled and operated services. While the total number of sanctioned schedules decreased marginally (1823 to 1784), the actual operated schedules dropped more sharply (1607 to 1481). as shown in Figure 2.

*Figure SEQ Figure * ARABIC 2: Demand-Supply Gap in Service Operations*

This widening gap indicates that a growing number of planned services are not being executed. Such inefficiencies highlight limitations in operational planning and the inability to dynamically adapt to real-time constraints such as traffic, demand fluctuations, or resource availability.

*Increase in Service Cancellations*

Cancelled kilometers increased significantly from 1.30 crore to 1.74 crore, with daily cancelled kilometers rising from 35,587 to 47,831. Can be Observed in Figure 3.

*Figure SEQ Figure * ARABIC 3: Increase in Service Cancellations*

Interestingly, this increase occurred despite a reduction in breakdown rates (1.51 to 1.22 per 10,000 km) and accidents. This suggests that cancellations are primarily due to planning inefficiencies rather than mechanical failures.

*Customer Satisfaction Indicators*

Passenger complaints increased sharply from 14,842 to 26,359 , indicating a significant deterioration in perceived service quality. as Figure 4 Shows, This rise reflects inefficiencies in service delivery, delays, and potential mismatches between passenger expectations and actual system performance.

*Figure SEQ Figure * ARABIC 4: Customer Satisfaction Indicators*

*Energy Transition and Sustainability Trends*

The dataset shows an increase in electric bus utilization, with effective kilometers rising from 8,512 to 30,114 . Additionally, energy efficiency improved from 1.10 to 1.36 km per unit. Can be observed in Figure 5.

*Figure SEQ Figure * ARABIC 5: Energy Transition and Sustainability Trends*

However, the contribution of electric buses remains minimal compared to total system operations, indicating that the transition toward sustainable mobility is still in its early stages.

*Operational Paradox*

A critical finding of this study is the presence of an operational paradox: while mechanical breakdowns decreased from 7,705 to 6,145, as seen in Figure 6.

*Figure SEQ Figure * ARABIC 6: Breakdowns vs Cancellations*

service cancellations increased significantly. This indicates that inefficiencies are not driven by vehicle reliability but by planning and management shortcomings.

Despite reduced breakdowns, cancellations increased, indicating planning inefficiencies rather than mechanical failures.

*Geospatial Insights*

The geospatial analysis of PMPML bus stops and routes Figure 7 highlights significant spatial imbalances in service distribution. Heatmap and clustering techniques reveal a high concentration of bus stops and routes in central urban regions, indicating over-served areas with route redundancy. In contrast, peripheral and suburban regions exhibit sparse stop density, reflecting under-served areas with limited accessibility to public transport services.

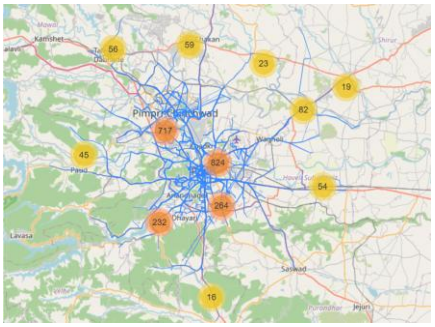


*Figure SEQ Figure * ARABIC 7: Geospatial Distribution of Bus Stops and Routes*

This uneven spatial distribution suggests inefficient route planning, This can be Seen in Figure 8 where resources are concentrated in already well-connected zones while emerging and low-density areas remain underserved. Such imbalances contribute to reduced overall system efficiency, lower ridership in peripheral regions, and increased operational strain in central corridors.

*Figure SEQ Figure * ARABIC 8: Heatmap Distribution of Service Availability*

The integrated analysis of all the Insights, the Geospatial Analysis and Heatmap reveals that over-served regions lead to low fleet utilization due to service redundancy, while under-served regions result in increased passenger complaints. Additionally, the mismatch between service distribution and demand contributes to rising service cancellations. These findings indicate that system inefficiencies are primarily driven by imbalanced service allocation and lack of data-driven planning.

**Comparative Analysis With Baseline Models**

*Limitations of Traditional Methods*

- Reliance on static route planning and fixed schedules without adapting to real-time demand.
- Inability to respond dynamically to traffic conditions and operational constraints.
- Inefficient fleet utilization and widening demand–supply gaps observed in PMPML data.
- Lack of integration between operational data and spatial distribution of services.
- Redundant services in over-served regions and inadequate coverage in under-served areas.
- Absence of predictive capabilities for demand forecasting and maintenance planning.
- Reduced service reliability leading to increased cancellations and passenger dissatisfaction.

*Advantages of Data-Driven Solution*

- Enables dynamic decision-making using historical and real-time data.
- Supports demand forecasting to align service supply with passenger demand.
- Optimizes route planning and scheduling to reduce redundancy and improve coverage.
- Integrates geospatial analysis to identify over-served and under-served regions.
- Improves fleet utilization and reduces service cancellations.

- Enables predictive maintenance to minimize breakdowns and disruptions.
- Enhances passenger experience through real-time information and adaptive services.
- Provides a scalable and efficient framework for modern public transportation systems.

**Conclusion & Future Scope**

*Conclusion*

- PMPML analysis reveals declining fleet utilization and operational efficiency.
- Significant gap between sanctioned and operated services indicates execution inefficiencies.
- Increase in cancellations and complaints reflects declining service reliability and passenger satisfaction.
- Geospatial analysis identifies over-served central regions and under-served peripheral areas.
- Integrated analysis confirms inefficiencies are due to poor planning, not infrastructure limitations.
- Highlights the need for data-driven decision-making in public transportation systems.

*Future Scope*

- Implementation of machine learning models for demand forecasting and route optimization.
- Integration of real-time data from GPS, traffic APIs, and passenger applications.
- Development of adaptive scheduling systems for dynamic service management.
- Use of predictive maintenance models to reduce operational disruptions.
- Expansion of geospatial analysis with real-time and user behavior data.
- Application of AI-driven frameworks for smart and sustainable transportation systems.

**References**

1. Chandiramani, J., & Nayak, S. (2019). Big Data Analytics and Internet of Things for Urban Transportation: A Case of Pune City, Maharashtra, India. In *Big Data Analytics for Smart and Connected Cities* (pp. 244-277). IGI Global.
2. Ying, G. (2025). Machine Learning and Cloud-Enhanced Real-Time Distributed Systems for Intelligent Urban Services. *Journal of System and Information Sciences*, 1(1), 189-200.
3. Singh, P. (2024). The Crisis of Public Transport in India: Overwhelming Needs but Limited Resources. ResearchGate.
4. Sharma, S., & Singh, R. (2026). An Outlook of Challenges and Implications in Indian Transport Industry. International Journal of Engineering Research & Technology (IJERT), 15(3).
5. Magginas, V., et al. (2020). Artificial Intelligence, Transport and the Smart City: Definitions and Dimensions of a New Mobility Era. Sustainability, 12(7), 2789.
6. Allam, Z., et al. (2024). Artificial Intelligence-Enabled Metaverse for Sustainable Smart Cities: Technologies, Applications, Challenges, and Future Directions. Electronics, 13(24), 4874.
7. Jimenez, F. (2021). Intelligent Transportation Systems (ITS). MDPI.
8. Tozzo, V., et al. (2024). Machine Learning for public transportation demand prediction: A Systematic Literature Review. ResearchGate.
9. Rezaie, M. (2021). Machine Learning Applications in Surface Transportation Systems: A Systematic Review. ProQuest Dissertations.
10. Lopez, J., et al. (2024). Congestion Forecasting Using Machine Learning Techniques: A Systematic Review. Future Transportation, 5(3), 76.
11. Jay Vardhan Open City Pune Dataset 2024 Pune PMPML Annual Data - Dataset - CKAN
12. Viraj Kadam Kaggle Dataset 2021 Geospatial Analysis of Bus Routes