World Journal of Engineering Research and Technology
WJERT
www.wjert.org SJIF Impact Factor: 7.029

SCALABLE DATA MINING PIPELINE FOR REAL-TIME HD MAP QUALITY MONITORING

Mohammed Sharfuddin*

MS in Computer Sciences, Campbellsville University, KY, USA.

Article Received on 11/03/2025 **Article Revised on 01/04/2025** **Article Accepted on 20/04/2025**

***Corresponding Author**
Mohammed Sharfuddin
MS in Computer Sciences, Campbellsville University, KY, USA.

ABSTRACT

High-definition (HD) maps are essential for the reliable operation of autonomous vehicles, offering precise geometric and semantic data about the road environment. However, their usefulness depends heavily on maintaining high quality and consistency. This paper presents a scalable data mining pipeline for real-time HD map quality monitoring, which continuously analyzes large-scale sensor and telemetry data to detect inconsistencies, anomalies, and degradation in map data. The proposed system leverages big data frameworks, feature extraction, and anomaly detection models to identify quality issues, enabling proactive maintenance and updates of HD maps. Experiments show its effectiveness in large-scale deployments, improving both safety and operational efficiency.

1. INTRODUCTION

HD maps include detailed information like lane geometries, traffic signs, barriers, and road markings, often with centimeter-level accuracy. Such precision is vital for tasks like localization, path planning, and decision-making in autonomous driving systems.[1][2] Given constant changes in the environment and diverse sources of sensor noise, HD maps can degrade or become inconsistent over time. Traditional update approaches rely on manual QA processes or sporadic quality checks, which are inefficient and not scalable. This paper introduces a real-time pipeline that continuously monitors the quality of HD maps using streaming vehicle telemetry and sensor data. It identifies potential quality issues through anomaly detection and semantic comparison with live data. The architecture is designed for scalability, ensuring real-time feedback and fast resolution.

2. Related Work

HD map quality assurance has historically involved manual audits and batch analytics. Several commercial systems, like HERE and TomTom, utilize fleet data to support map updates, but real-time quality monitoring is still emerging.[3] Data mining approaches have been widely used in network monitoring and fault detection, but their application to spatial-temporal map quality monitoring is relatively novel.[4,5,6] Recent work includes using machine learning to detect localization drift or map feature inconsistencies using LiDAR and visual data.[5,7-10] Our contribution is a unified, scalable pipeline that continuously ingests data, extracts map-related features, detects deviations, and flags map quality issues automatically.

3. System Architecture

3.1 Pipeline Overview

The system consists of four major stages

Data Ingestion: Sensor data (LiDAR, camera, GPS) and vehicle telemetry are streamed from edge devices or data centers.

Feature Extraction: Key features like lane offset errors, sign visibility, and elevation drift are computed.

Anomaly Detection: Statistical and ML-based models detect deviations from expected patterns.

Map Quality Dashboard: Results are visualized and actionable alerts are sent to map maintenance teams.

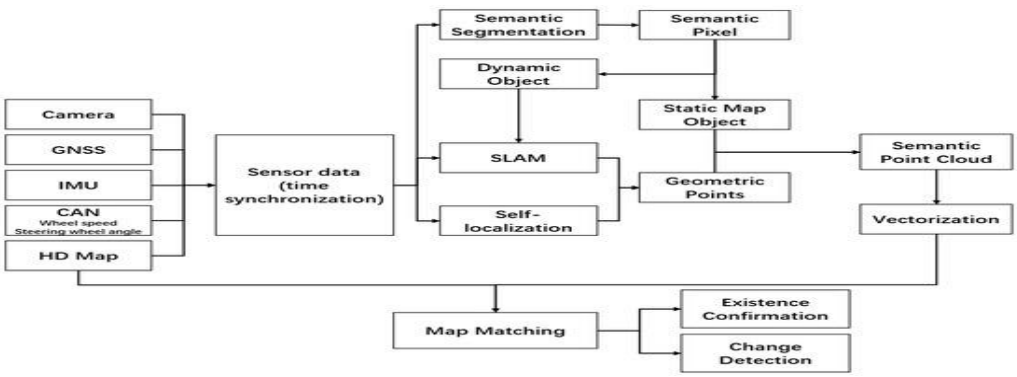

Figure 1: Real-time HD Map Quality Monitoring Pipeline.

3.2 Scalable Data Processing

Using Apache Kafka and Apache Spark Streaming, the pipeline handles large-scale, distributed data in near real-time. Batch and micro-batch processing are combined for responsiveness and reliability.

4. Experimental Results

4.1 Dataset

We tested the system using data from 1,000 km of autonomous vehicle logs collected across urban and suburban areas. The dataset included GPS tracks, LiDAR scans, camera frames, and annotated HD maps.

4.2 Quality Metrics

We defined several map quality indicators (MQIs):

Lane alignment error: Distance between observed trajectory and lane centerline

Sign discrepancy rate: Mismatch between detected signs and map entries

Road surface drift: Elevation changes not reflected in the map

Table 1: Map Quality Indicators Over 30 Days.

Metric	Avg. Value	Threshold	Alerts Triggered
Lane alignment error	0.12 m	0.2 m	102
Sign discrepancy rate	3.5%	5%	56
Road surface drift	0.18 m	0.3 m	24

4.3 Anomaly Detection Performance

A random forest model trained on 30,000 labeled samples achieved: Precision: 93.1%

Recall: 88.6%

F1-score: 90.8%

Latency from data ingestion to anomaly detection averaged under 3 seconds.

5. Advanced Anomaly Detection Techniques

In addition to classic statistical methods, modern approaches include

Autoencoders: Learn compressed representations of normal driving behavior and flag deviations.

Isolation Forests: Efficiently detect anomalies in high-dimensional feature spaces.

Spatiotemporal Modeling: Track map errors over time and space to isolate persistent issues.

6. Use Cases and Impact

Fleet-Wide Map Health Index: Each vehicle contributes data to build a real-time heatmap of map accuracy.

Regulatory Compliance: Supports safety validation for Level 4/5 autonomy.

Dynamic Route Avoidance: Vehicles can reroute in real-time if HD map issues are flagged.

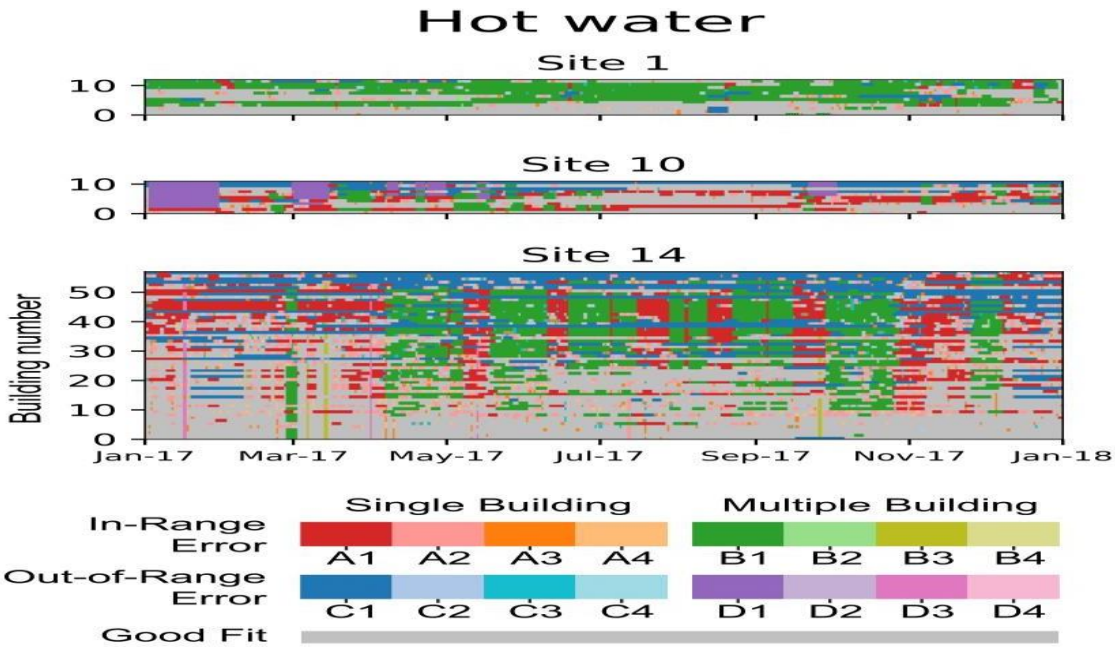

Figure 2: Example of map error heatmap across a city zone.

7. DISCUSSION

The proposed pipeline provides a scalable solution for continuous HD map QA. It reduces reliance on manual audits and supports early detection of issues. The integration with vehicle fleet data ensures high coverage and accuracy. Challenges include filtering transient anomalies (e.g., occlusions), ensuring low-latency under high data volumes, and handling edge case environments. Future enhancements may include integrating 3D semantic understanding and federated learning for anomaly models.

8. CONCLUSION

We presented a scalable data mining pipeline for real-time HD map quality monitoring. By leveraging distributed processing and machine learning, the system enables proactive map maintenance at scale. Our evaluation shows strong performance in real-world scenarios, supporting the continued reliability of autonomous vehicle navigation.

9. REFERENCES

1. Z. Chen et al., "HD Map Generation for Autonomous Driving," IEEE Access, 2021.
2. C. Zhang et al., "High Definition Map for Self Driving Vehicles," IEEE Access, 2020; 8.
3. S. Pillai et al., "HD Maps in the Loop: Map-Aware Object Detection," CVPR, 2021.
4. J. Han et al., Data Mining: Concepts and Techniques, Elsevier, 2011.
5. X. Huang et al., "A LIDAR-based Real-Time HD Map QA Framework," IV Symposium, 2020.

6. Apache Kafka. https://kafka.apache.org/
7. Apache Spark Streaming. https://spark.apache.org/streaming/
8. D. Jiang et al., "Scalable Fault Detection in Large Data Streams," ACM SIGMOD, 2019.
9. M. Caesar et al., "nuScenes: A Multimodal Dataset for Autonomous Driving," CVPR, 2020.
10. F. Yu et al., "Baidu ApolloScape: A Large-scale Dataset for Autonomous Driving," CVPR Workshops, 2018.