PI-0381 - Big Data Deployment Requirements for Transportation Use Cases

Requested by
Ping Qiu, Traffic Operations
Author: Abdullah Faiyaz, DRISI

March 3, 2025

The Caltrans Division of Research, Innovation and System Information (DRISI) receives and evaluates numerous research problem statements for funding every year. DRISI conducts Preliminary Investigations on these problem statements to better scope and prioritize the proposed research in light of existing credible work on the topics nationally and internationally. Online and print sources for Preliminary Investigations include the National Cooperative Highway Research Program (NCHRP) and other Transportation Research Board (TRB) programs, the American Association of State Highway and Transportation Officials (AASHTO), the research and practices of other transportation agencies, and related academic and industry research. The views and conclusions in cited works, while generally peer reviewed or published by authoritative sources, may not be accepted without qualification by all experts in the field. The contents of this document reflect the views of the authors, who are responsible for the facts and accuracy of the data presented herein. The contents do not necessarily reflect the official views or policies of the California Department of Transportation, the State of California, or the Federal Highway Administration. This document does not constitute a standard, specification, or regulation. No part of this publication should be construed as an endorsement for a commercial product, manufacturer, contractor, or consultant. Any trade names or photos of commercial products appearing in this publication are for clarity only.

Table of Contents

Executive Summary

Background

The increasing availability of big data has significantly impacted transportation research and operations, supporting a wide range of use cases such as managed lane performance assessment, route estimation, and traffic demand modeling. However, with the rapid advancement of sensing technologies, network communications, and artificial intelligence (AI), transportation agencies face significant challenges in effectively utilizing big data for planning, monitoring, and management.

One of the major challenges is data quality, particularly concerning velocity, variety, and veracity. Speed measurements from probe vehicles or smartphones may be biased due to low penetration rates, while GPS-based route inference can be inaccurate due to low sampling rates and GPS limitations. Similarly, travel demand estimation through data fusion from diverse sources—such as travel surveys, probe traces, and traffic sensors—may introduce errors due to varying sampling methods and demographic biases. Additionally, AI-powered video and LiDAR sensors offer new opportunities for intersection performance assessment but require updated intelligent transportation system (ITS) architectures for data management and communication. The rise of Generative AI (GenAI) further complicates data reliability, particularly in assessing social media-sourced information.

Given these challenges, further research is needed to establish practical deployment requirements for big data that enhance transportation applications. These requirements will define essential traffic measurement inputs and specify minimum technical criteria related to sampling rate, penetration rate, latency, data fusion, and aggregation methods to ensure accurate and reliable traffic data analytics.

Currently, there is a lack of clearly defined standards available to validate and manage data sources. As a result, there is a critical gap in the transportation industry's ability to ensure high accuracy big data solutions.

This research is needed to address that gap. It aims to provide transportation agencies with foundational knowledge and standards that are needed to assess the suitability and quality of emerging data sources, improve the operational effectiveness of data pipelines, and build systems that can evolve with the emergence of new technologies. It is necessary for transportation agencies to understand the impact of big data on the future of mobility, ensuring accuracy and usability while maintaining the safety of drivers.

Summary of Findings

"Assessing veracity of big data: An in-depth evaluation process from the comparison of Mobile phone traces and groundtruth data in traffic monitoring"

This study highlights the role of data quality in the context of Big Data. An in-depth analysis is done comparing Cellphone Big Data to ground truth data, utilizing traffic related data collected along a major Italian trunk road. The results of this analysis reveal a need for a thorough evaluation of Big Data when it is widely integrated due to the accuracy of Cellphone Big Data being sensitive to factors such as traffic speed, mobile phone network characteristics, and vehicle occupancy rates. Additionally, the paper delves into the implications of these findings for traffic monitoring and management applications that rely on Cellphone Big Data. It discusses the potential biases and inaccuracies that can arise if these sensitivities are not properly accounted for, suggesting the necessity of developing robust

methodologies for data validation and error correction. The study concludes by emphasizing the importance of understanding the limitations and strengths of Cellphone Big Data to ensure its reliable and effective use in transportation planning and real-time traffic management systems.

Gaps in Findings

The result of this study is based on monitoring & analysis of an Italian trunk road, so it is not certain that similar results would be duplicated on California roads and highways. This could also have a direct impact on the quality of the data. Additionally, there is no incorporation of artificial intelligence in the context of this study either. This would mean that in addition to potentially inaccurate results due to inability to recreate the setting, we also wouldn't be able to fully get an idea of the impact of these new technologies.

Next Steps

To bridge these gaps, future research should focus on establishing some sort of standardized framework for further analysis and improvement of Big Data sources. Additionally, it would also be of great use to understand the use of Cellphone Big Data in various real-world traffic settings within California. It would also be of best interest to evaluate current data pipelines and transportation infrastructures to see where improvements could be built upon.

Detailed Findings

Background

With transportation agencies increasingly relying on big data to support their application, a question arises, how reliable is the data? When looking at recent evaluations of mobile phone-based traffic data, the vulnerability in the data quality is one of the main highlights. Mobile phone-based traffic data is just one example, but it helps us deeply understand why further research is necessary.

For example, in the case study comparing mobile traffic data to ground-truth data along a major Italian trunk road, significant fluctuations in the accuracy of the data were revealed. They had found that external factors such as traffic speed and network connectivity introduced major inconsistencies that affected the usability of the data. Due to how common this type of data source is, this study alone raises significant concern for further research for transportation agencies around the world.

Although the study focuses specifically on mobile-phone data, the findings directly align with the broader concerns for other forms of data capture such as AI-powered videos or LIDAR. With the new age of generative AI, these are new technologies that need extensive research to further instill confidence in the accuracy of the generated content. Additionally, the study focuses on the Italian highway system, so there could be geographical concerns that could lead to different results when doing the same study using California's diverse roadway environments and traffic conditions. By testing within our own environments and using new technologies, we will be able to further understand and enhance our data pipeline and assess the suitability of big data technologies.

Research is needed to provide a standardized framework to help fill a critical gap within the transportation sector's ability to assess big data technologies. In order to address some of these limitations, future research needs to focus on developing standardized methods to validate and deploy big data while considering the various transportation contexts that exist in real-world settings. Simulating real-world traffic conditions via the California roadway system would allow for improved data collecting techniques. Establishing this framework will enable transportation agencies to create and improve data pipelines to ensure that they would be able to accurately and efficiently operate with evolving needs.

Related Research and Resources

"Assessing veracity of big data: An in-depth evaluation process from the comparison of Mobile phone traces and ground truth data in traffic monitoring", Journal of Transport Geography, Volume 118, June 2024, 103930, ISSN 0966-6923, Alessandro Nalin, Valeria Vignali, Claudio Lantieri, Denis Cappellari, Bruno Zamengo, Andrea Simone, https://doi.org/10.1016/j.jtrangeo.2024.103930. (https://www.sciencedirect.com/science/article/pii/S096669232400139X)