DRISI

CALTRANS DIVISION OF RESEARCH, INNOVATION AND SYSTEM INFORMATION

TRANSFORMING IDEAS INTO SOLUTIONS

Research Notes

Develop evaluation criteria for ML/AI (machine learning/artificial intelligence)-generated or third-party traffic data

To develop a framework for assessing the quality of traffic data broadly and relative to inquiries related to vehicle trajectory, speed, and volume from third-party vendors.

Traffic Operations

November 2025

Project Title:
Develop evaluation criteria for ML/AI-generated or third-party traffic data

Task Number: 4472

Start Date: November 1, 2024

Completion Date: October 31, 2026

Task Manager:

Abdullah Faiyaz
Transportation Engineer, Electrical
Abdullah.Faiyaz@dot.ca.gov

WHAT IS THE NEED?

Traffic operations data is instrumental in providing valuable insights for traffic management, urban planning, and transportation optimization.

The common traffic data types include volume (vehicle counts and concentration in a segment), speed (segment average), flow (vehicles passing a point or through a segment), travel time (between two points), signal timing, parking, cyclist counts, pedestrian counts (at specific locations), incidents (accidents, construction, etc. at specific locations), etc.

Currently, the California Department of Transportation (Caltrans) lacks the mechanism or resources to sustainably acquire the required traffic operations data. With the advances in Connected and Autonomous Vehicles, Smart Mobile Phone Applications, Computer Vision, and Machine Learning technologies, many potential third-party traffic data vendors could meet Caltrans' needs. However, Caltrans does not have a standard approach to validate and evaluate the third-party data to make intelligent business decisions.

This research can be justified based on one recent event as an example:

When the I-10 freeway fire damage caused the closure of a segment within the Los Angeles (LA) city, due to a lack of relevant and timely data, Caltrans was not able to conduct a solid evaluation of the system-wide traffic impact due to this closure.

To address Caltrans' data, the above-listed shortcomings/

Caltrans

DRISI provides solutions and knowledge that improves California's transportation system.

recent events, and to collect complete and comprehensive traffic data, this research aims to evaluate and compare the third-party sourced data with the ground truth data (acquired from Performance Measurement System (PeMS)), census, and Weigh-in-motion (WIM) stations, and cameras). The third-party data could come from connected vehicles, cell phone Global Positioning System (GPS), satellite images, etc. For data evaluation, multiple criteria could be used, for example, the cost (licensing fees), accuracy, relevance/recency, time-resolution (5-mins, hourly, daily, etc.), availability (in the required locations), the type of data (speed, volume, classification, etc.), and ease of programmatic access (API, etc.)

WHAT ARE WE DOING?

The framework will be constructed in a three-step process. First, we will identify the metrics relevant to assessing the quality of a dataset. These go beyond quantities such as mean, standard deviations, and other moments, which are less useful for spatially and temporally varying data, and instead center on how valuable the dataset is when used as input to some estimation or decision-making process. For example, a dataset that allows for accurate analysis of an unknown metric (say, speed and volume data at a particular location) is of high quality relative to this metric. Second, we will develop and/or identify statistical methods for computing the metrics in practice. These methods will leverage the technique of cross-validation to compute aggregate metrics such as average imputing error (i.e., remove one data point, apply an imputing technique, and use the remaining data to estimate the value of the first data point, and compute the estimation error). The methods will also leverage sensitivity analysis, influence functions, and laws of traffic flow (such as consistency) to quantify how easily a particular dataset can estimate a particular metric of interest. Third, we will implement the statistical methods in Python using existing packages and customized code.

WHAT IS OUR GOAL?

The goal of this project is to provide a framework for assessing the quality of traffic data broadly and relative to specific queries. The framework will apply to vehicle trajectory data and other compiled data, including speed and volume information from third-party vendors such as Google Maps and INRIX. It will also apply to data generated by WIM stations and image data from stationary cameras. The framework will provide a systematic approach to comparing data sets, tracking the decline in data quality over time and space, and assessing the severity of any missing or corrupted data.

WHAT IS THE BENEFIT?

The research would lead to a general-purpose criterion and process for validating and evaluating any third-party data holistically.

The product of the research will enable Caltrans to identify suitable third-party traffic data to complement the organically collected data using existing infrastructure. The product will also enable Caltrans to select the most cost-effective and sustainable third-party data source.

Traffic data is one of the essential data products that support key functions of Caltrans, including, but not limited to, traffic incident management, permitting, maintenance, asset management, planning, design, and construction divisions. By supporting the functions of the above functional arm of Caltrans, this research supports all goals of Caltrans.

WHAT IS THE PROGRESS TO DATE?

The contract was executed on July 2, 2025, and the kickoff meeting was held on August 4, 2025. Since then, a comprehensive literature review has been completed, fulfilling the primary objective for Quarter 1. In addition, a preliminary analysis of traffic data was conducted using a small study area in Sacramento. This analysis involved evaluating the

temporal and spatial characteristics of the data and assessing their consistency with widely accepted physical principles in the transportation literature. Further details, including findings from the literature review and results from the preliminary traffic data evaluation, were provided in report format.