Datasheet for Red-Light Camera Fines Dataset*

Analyzing Traffic Enforcement and Compliance Patterns

Qisheng Yu

December 1, 2024

This datasheet details the composition, collection process, and intended uses of the red-light camera fines dataset. It is designed to support research on traffic enforcement strategies and compliance behavior. The dataset includes fines issued across multiple regions and years, with additional enforcement timing data. This document provides insights into the methodology, limitations, and potential applications.

Motivation

- The dataset was created to enable analysis of regional and temporal trends in traffic compliance through fines issued at red-light camera intersections. It aims to address gaps in enforcement-specific datasets and support evidence-based policymaking.
- The dataset was sourced from Toronto's municipal open data portal and compiled by the Red-Light Camera program for public transparency and research.
- This dataset development was not externally funded.
- The dataset fills a critical gap by offering detailed information specific to red-light camera enforcement and regional trends, which are not commonly available in public datasets.

Composition

- Instances represent annual counts of traffic fines issued at red-light cameras, enforcement start dates, and regional identifiers.
- There are 4,508 instances in the long-format dataset, covering 25 wards over 18 years (2007–2024).
- The dataset is nearly exhaustive for Toronto's red-light enforcement programs; however, it excludes locations without cameras or alternate enforcement measures.

*Code and data are available at: https://github.com/qishengyu/red-camera-charge.

- Each instance consists of variables like fines, year, ward number, and enforcement status.
- The target variable is the annual fine count.
- No missing data exists in the cleaned dataset; missing values in raw data were addressed during preprocessing.
- Relationships are captured implicitly via shared regional identifiers and years, which enable trend analysis.
- The dataset is split into 80% training and 20% testing for modeling purposes.
- No errors were identified post-cleaning, though real-world data may still contain minor inaccuracies.
- The dataset is self-contained and does not rely on external resources.
- The dataset does not include confidential data.
- There is no offensive or distressing content in the dataset.
- The dataset includes regional subpopulations identified by ward numbers but lacks individual-level demographic details.
- The dataset does not identify individuals.
- The dataset is not sensitive and complies with open data licensing.
- Additional notes: The dataset was transformed into a long format for modeling purposes.

Collection Process

- Data were collected directly via red-light camera systems, processed for public release by Toronto's traffic enforcement department.
- Cameras and automated detection systems recorded violations; the data were curated via software pipelines.
- The data represent a deterministic census of locations under monitoring, ensuring broad coverage.
- Data collection was automated with oversight from municipal agencies; no compensation applies.
- Data span 2007–2024, matching the period of enforcement program operation.
- Ethical reviews were conducted at the municipal level before public release.
- Data were obtained from the municipal database, ensuring compliance with public use guidelines.
- No direct notification or consent was required as the data pertain to violations, not individuals.

- Consent mechanisms are not applicable due to anonymized data.
- No explicit retention limits are set, but updates may include newer enforcement data.

Preprocessing/Cleaning/Labeling

- Data cleaning involved date standardization, removal of duplicate rows, and transformation to long format for analysis.
- Both raw and cleaned datasets are stored for reproducibility.
- Software and R packages used include `dplyr`, `tidyr`, and `arrow`.

Uses

- The dataset has been used to model the impact of enforcement start time on regional fine trends.
- A repository linking related papers and models is available on the GitHub project page.
- The dataset could support studies on traffic compliance, accident prevention, or urban mobility.
- Data limitations must be acknowledged, such as its focus on fines without accident data.
- Tasks involving individual-level analyses or demographic profiling are not supported.

Distribution

- The dataset is publicly available via Toronto's open data portal.
- It is distributed as CSV and Parquet files on GitHub.
- Updates follow Toronto's open data schedule.
- Licensing is governed by the city's open data terms.

Maintenance

- The dataset is maintained by the Toronto municipal open data team.
- Contact: `open.data@toronto.ca`.
- Updates will include new years of data, accessible via GitHub or the open data portal.