Datasheet for 'Toronto Traffic Speed Analysis Dataset'*

Andy Jiang

December 2, 2024

Provides access to traffic speed data collected in Toronto's designated safety zones between 2017 and 2024. This dataset enables analysis of urban mobility, safety trends, and the relationship between traffic volume and speed percentiles. This datasheet enhances the reproducibility and understanding of research conducted using this dataset.

Motivation

1. *For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.*
   - The dataset was created to monitor traffic speeds, enforce road safety, and provide insights into traffic behavior in designated Safety Zones in Toronto. It allows for analyzing trends and assessing traffic management strategies.
2. *Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?*
   - The dataset was created by the City of Toronto as part of the Safety Zone Watch Your Speed Program (WYSP) (Dumas 2024).
3. *Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.*
   - The dataset is funded by the City of Toronto as part of its traffic safety initiatives.
4. *Any other comments?*
   - N/A

Composition

*Code and data are available at: https://github.com/AndyYanxunJiang/toronto-vehicles-speed-analysis.

1. *What do the instances that comprise the dataset represent (for example, documents, photos, people, countries)? Are there multiple types of instances (for example, movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.*

    - Each row represents aggregated monthly traffic data collected from radar-equipped Safety Zone speed signs, including speed percentiles, vehicle counts in speed bins, and total traffic volumes.

2. *How many instances are there in total (of each type, if appropriate)?*

    - The dataset contains 31,990 observations (monthly data from 2017–2024).

3. *Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (for example, geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (for example, to cover a more diverse range of instances, because instances were withheld or unavailable).*

    - It is a sample limited to traffic monitored in specific Safety Zones across Toronto.

4. *What data does each instance consist of? "Raw" data (for example, unprocessed text or images) or features? In either case, please provide a description.*

    - Each instance includes aggregated speed and volume data, such as 5th, 50th, and 95th percentile speeds, and counts of vehicles in predefined speed ranges.

5. *Is there a label or target associated with each instance? If so, please provide a description.*

    - No labels or targets, but each instance is timestamped (monthly).

6. *Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (for example, because it was unavailable). This does not include intentionally removed information, but might include, for example, redacted text.*

    - Some speed bins may have zero counts, reflecting low or no traffic in specific ranges during certain months.

7. *Are relationships between individual instances made explicit (for example, users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.*

    - No explicit relationships; each observation is independent and time-indexed.

8. *Are there recommended data splits (for example, training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.*

- No.

9. *Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.*

   - Potential behavioral biases exist due to the visibility of speed signs, which may cause drivers to slow down.

10. *Is the dataset self-contained, or does it link to or otherwise rely on external resources (for example, websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (that is, including the external resources as they existed at the time the dataset was created); c) are there any restrictions (for example, licenses, fees) associated with any of the external resources that might apply to a dataset consumer? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.*

    - The dataset is self-contained.

11. *Does the dataset contain data that might be considered confidential (for example, data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.*

    - No.

12. *Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.*

    - No.

13. *Does the dataset identify any sub-populations (for example, by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.*

    - No.

14. *Is it possible to identify individuals (that is, one or more natural persons), either directly or indirectly (that is, in combination with other data) from the dataset? If so, please describe how.*

    - No.

15. *Does the dataset contain data that might be considered sensitive in any way (for example, data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.*

    - No.

16. *Any other comments?*
    - N/A

Collection process

1. *How was the data associated with each instance acquired? Was the data directly observable (for example, raw text, movie ratings), reported by subjects (for example, survey responses), or indirectly inferred/derived from other data (for example, part-of-speech tags, model-based guesses for age or language)? If the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.*
    - Data was collected using radar-equipped speed signs installed in Toronto Safety Zones.
2. *What mechanisms or procedures were used to collect the data (for example, hardware apparatuses or sensors, manual human curation, software programs, software APIs)? How were these mechanisms or procedures validated?*
    - Radar speed signs measure oncoming vehicle speeds with an accuracy of ± 1 km/h and aggregate the results monthly.
3. *If the dataset is a sample from a larger set, what was the sampling strategy (for example, deterministic, probabilistic with specific sampling probabilities)?*
    - Locations for monitoring were selected by the City of Toronto, focusing on Safety Zones.
4. *Who was involved in the data collection process (for example, students, crowdworkers, contractors) and how were they compensated (for example, how much were crowdworkers paid)?*
    - City of Toronto employees.
5. *Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (for example, recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.*
    - Data was collected monthly from 2017–2024.
6. *Were any ethical review processes conducted (for example, by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.*
    - No.
7. *Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (for example, websites)?*

- Data was collected automatically through radar speed signs.

8. *Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.*

    - Drivers passing by signs may be aware of data collection based on posted signage.

9. *Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.*

    - Not required; no personal information was collected.

10. *If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).*

    - N/A

11. *Has an analysis of the potential impact of the dataset and its use on data subjects (for example, a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.*

    - No.

12. *Any other comments?*

    - N/A

Preprocessing/cleaning/labeling

1. *Was any preprocessing/cleaning/labeling of the data done (for example, discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remaining questions in this section.*

    - Data was aggregated and cleaned to remove erroneous or incomplete observations.

2. *Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (for example, to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.*

    - N/A

3. *Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point.*

- N/A

4. *Any other comments?*
    - N/A

Uses

1. *Has the dataset been used for any tasks already? If so, please provide a description.*
    - Yes, for analyzing trends in traffic speed and volume, as detailed in this paper.
2. *Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other acess point.*
    - No known repository beyond the Open Data Toronto portal.
3. *What (other) tasks could the dataset be used for?*
    - Studying seasonal variations in traffic, evaluating traffic policy impacts, or identifying high-risk zones for accidents.
4. *Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a dataset consumer might need to know to avoid uses that could result in unfair treatment of individuals or groups (for example, stereotyping, quality of service issues) or other risks or harms (for example, legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?*
    - The dataset's focus on Safety Zones may limit its generalizability to other urban or rural areas.
5. *Are there tasks for which the dataset should not be used? If so, please provide a description.*
    - The dataset should not be used to generalize beyond the specific Safety Zones it covers.
6. *Any other comments?*
    - N/A

Distribution

1. *Will the dataset be distributed to third parties outside of the entity (for example, company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.*
    - Yes, through the Open Data Toronto portal.

2. *How will the dataset be distributed (for example, tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?*
    - As a CSV file or through an API provided by Open Data Toronto.
3. *When will the dataset be distributed?*
    - It is available now.
4. *Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/ or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.*
    - Yes, as per the Open Data Toronto terms of use.
5. *Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.*
    - No.
6. *Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.*
    - No.
7. *Any other comments?*
    - N/A

Maintenance

1. *Who will be supporting/hosting/maintaining the dataset?*
    - Open Data Toronto.
2. *How can the owner/curator/manager of the dataset be contacted (for example, email address)?*
    - Through the Open Data Toronto portal.
3. *Is there an erratum? If so, please provide a link or other access point.*
    - No.
4. *Will the dataset be updated (for example, to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to dataset consumers (for example, mailing list, GitHub)?*

- Yes, as new data is collected by the City of Toronto.

5. *If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (for example, were the individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.*

    - N/A

6. *Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to dataset consumers.*

    - Unknown, but likely updated periodically.

7. *If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to dataset consumers? If so, please provide a description.*

    - The dataset is available through the Open Data Toronto portal. Users who wish to augment or contribute to the dataset must contact the City of Toronto's Open Data team via their official website (https://open.toronto.ca). Contributions are typically validated by the data custodians to ensure consistency with municipal data standards.

8. *Any other comments?*

    - N/A

Reference

Dumas, Raphael. 2024. "WYSP Monthly Summary." City of Toronto, OpenData. https://open.toronto.ca/dataset/safety-zone-watch-your-speed-program-monthly-summary/.