Road Traffic Crashes Derived from Crowdsourced Reports from Ma3Route

Overview

The datasets contain the time and location of road traffic crashes in Kenya (primarily Nairobi); crash information is derived from crowdsourced reports from @Ma3Route. Ma3Route is a mobile/web/SMS platform that crowdsources transport data and provides users with information on traffic, road traffic crash (RTC), matatu directions and driving reports. Users post RTC or traffic information to Ma3Route, where Ma3Route then publishes the post on Twitter. Tweets from @Ma3Route were queried using the Twitter API (tweets were no longer queried once Twitter rebranded to X).

The following process is used to transform tweets into crashes:

1. Tweets are coded as to whether they report a crash
2. If a tweet reports a crash, the crash location is geocoded relying on the text of the tweet relying on mentions of landmarks and roads
3. Geolocated crash reports are clustered into individual crashes; crash reports within 500 meters and 4 hours are grouped into individual crashes

This process is implemented in two ways:

1. First, we manually code and geolocate crashes from July 2017 through July 2018.
2. Second, we implement an algorithm that automates coding and geocoding crashes from August 2012 through July 2023.

The following paper provides additional details on how the datasets were created. Note that the algorithm has some error in determining the whether a tweet reports a crash and in geolocated the location of the crash; please refer to the paper for details on accuracy:

> Milusheva S, Marty R, Bedoya G, Williams S, Resor E, et al. (2021) "Applying machine learning and geolocation techniques to social media data (Twitter) to develop a resource for urban planning." PLOS ONE 16(2): e0244317. https://doi.org/10.1371/journal.pone.0244317

Datasets

Two datasets are provided:

- **ma3route_crashes_truthcode:** Includes crashes from July 2017 through July 2018; whether a tweet reports a crash and the geolocation of the crash are manually coded.
- **ma3route_crashes_algorithmcode:** Includes crashes from August 2012 through July 2023; whether a tweet reports a crash and the geolocation of the crash are determined by the described in Milusheva et al. (2021)

Variables

Both datasets provide the following variables

Variable	Description	Variable Type
crash_id	Unique crash ID	Numeric
crash_datetime	Date/time of the crash (using date/time of the first tweet that reported the crash)	Datetime
crash_date	Date of the crash (using date of the first tweet that reported the crash)	Date
latitude	Latitude of crash	Numeric
longitude	Longitude of crash	Numeric
n_crash_reports	Number of tweets that reported crash	Numeric
contains_fatality_words	Whether the tweet contains one of the words: 'dead', 'died', 'body', 'killed', or 'fatal'	Numeric • 1 = Yes • 0 = N0
contains_pedestrian_words	Whether the tweet contains the word: 'pedestrian'	Numeric • 1 = Yes • 0 = N0
contains_matatu_words	Whether the tweet contains the word: 'matatu'	Numeric • 1 = Yes • 0 = N0
contains_motorcycle_words	Whether the tweet contains one of the words: 'boda', 'motorcycle', or 'motor cycle'	Numeric • 1 = Yes • 0 = N0