Estimating Traffic Crash Counts Using Crowdsourced Data

Volpe Center
Erika Sudderth, PhD
Dan Flynn, PhD
Tanner Robart
Michelle Gilmore

BTS
Pat Hu
Ed Strocko

2018-06-14
Challenge: Tracking crashes in near real-time

• Crash data are typically available for certain crashes, after several months

• EDT (Electronic Data Transfer) of police accident reports available nightly for nine states

• Waze incident data available where user reported, all 50 states, every 2 minutes

• Waze and EDT could provide near-real time, granular estimates of crashes to inform safety policy and operations
Objectives

- Use crowdsourced data insights to improve transportation safety

Questions

- Can we integrate DOT data resources at large scales?
- Do Waze data support vision of a rapid crash indicator?
Analysis: Challenges and Solutions

Challenges
- Waze and EDT coordinates do not all align with FHWA road network
- How do we associate Waze events and EDT reports?
- Need to define zeros (time and places with no accidents)

Solutions
- Spatial aggregation of data to hexagonal grids (1-mile area)
- Match Waze to EDT on user-selected buffers in space and time
- Define zeros as grid cells and time periods with 1 or more non-accident Waze events but no EDT reports
Model Performance (April-Sept 2017 in MD)

Model estimates highly accurate overall; miss some precise patterns
Results – what have we learned?

Can we integrate DOT data resources at large scales?

- **YES** – Our data integration and analysis pipeline can support rapid crash estimates (when/where Waze signal present)
- Successfully integrated transportation data that are not originally intended to track traffic safety

Do Waze data support rapid crash indicator?

- **YES** – With Waze signal, models produce good overall estimates for MD (next test performance for other EDT states)
- Foundation for tool for rapid tracking of traffic safety trajectories
Next Steps

• Model testing and re-training for 4-5 EDT states

• Partnerships with state or local DOTs to identify use cases

• Cross-state Waze data assessment & dashboard

• Applications of segment-based models

Potential Applications

Rapid crash trend monitoring tool
• Flag anomalies
• Short-term intervention assessment
• Cross-state comparisons
• Effectiveness models

• Incident Duration
• Clearance Times
• Secondary Crashes
Additional Slides
Random Forests

- Machine learning approach which minimizes overfitting
- Trained models on 70% of data using EDT reports as our labeled “ground-truth”
- Tested model performance using 30% of data to compare estimated EDT crashes with observed EDT crashes
- Rigorously trained and tested data feature combinations (50+ models)
- Best crash estimation models minimize False Positives and False Negatives

Image credit: https://medium.com/@williamkoehrsen/random-forest-simple-explanation-377895a60d2d
SDI Waze Data Pipeline Development

**Waze**
- JSON files of events every 2 min
- Ingested in data lake

**Secure Data Commons**
- Curated to S3 and Redshift
- Derived to monthly, gridded data
- Combined with EDT and auxiliary data

**ATA Platform**
- Connection to SDC S3
- Estimation of EDT-level crash events using random forests
- Output model results to ATA S3 and local

**Output**
- Tabular and graphical outputs (ArcMap, Tableau)

**Amazon Web Services platform**
- Redshift database
- S3 buckets
- RStudio + Jupyter
- GitHub integration

**Amazon Web Services platform**
- S3 buckets
- RStudio + Jupyter
- GitHub integration
- Athena
Six months of geolocated Waze data for Maryland (April - September, 2017)
SDI Waze Data Pipeline Development

0. Waze Data Ingestion and Curation

1. Query, Clip, Reduce

2. Space-Time Match

- Waze API
- Lambda function
- Amazon S3
- Amazon EC2
- Amazon Redshift

Plotting original and clipped MD
SDI Waze Data Pipeline Development

3. Grid and Urban Area Overlay
   - Urban Areas
   - Hexagonal grid tessellations

4. Grid Aggregation

5. Weather Overlay
   - Raster weather reflectivity
SDI Waze Data Pipeline Development

6. Modeling

- Adding:
  - FARS
  - HPMS road class
  - AADT
  - LEHD

ATA + Local

7. Visualization and Reporting
Divide data into training and testing subsets

- Training data: Select **70%** of observations (random by rows, whole days, or whole weeks)
- Test data: Remaining **30%** of observations

**Training**: fit model parameters with a large set of known EDT crashes, associated Waze events and other predictors

**Testing**: apply fitted model parameters to a new set of Waze events and other predictors to generate estimated EDT crashes

Compare estimated EDT crashes to observed EDT crashes in the test data set to evaluate model performance
Variable Importance: Waze Accidents (April-Sept)

Mean decrease in Gini impurity:

- Variable is useful in separating a node of mixed classes (both 0 and 1 EDT crashes, in our case) into two nodes with pure classes (all 0 or all 1 EDT crashes).

- Across all nodes in all the trees, how much does this variable decrease node impurities, averaged over all trees?
Waze Data: Jams and Crash Sequence Analysis

Potential Applications

- Incident Duration
- Clearance Times
- Secondary Crashes

April. 2017 MD

<table>
<thead>
<tr>
<th>Type</th>
<th>UniqueCounts</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACCIDENT</td>
<td>15,139</td>
</tr>
<tr>
<td>HAZARD</td>
<td>242,787</td>
</tr>
<tr>
<td>JAM</td>
<td>180,347</td>
</tr>
<tr>
<td>ROAD_CLOSED</td>
<td>1,130</td>
</tr>
</tbody>
</table>

Type Legend:
- ACCIDENT
- HAZARD
- JAM
- ROAD_CLOSED