Using Big Data to Identify Hotspots of Pedestrian Crashes in Manhattan

Presented by Prof. Kaan Ozbay
November 19th, 2014

Kaan Ozbay, Kun Xie and Hong Yang
Center for Urban Science + Progress (CUSP)
Department of Civil & Urban Engineering (CUE)
New York University
Introduction

- Pedestrian Safety Situation in Manhattan (2005~2012)
  - A total of 9664 pedestrian crashes occurred
  - About 9.4% of them (906) involved serious injuries and fatalities

- Importance of Identifying Hotspots of Pedestrian Crashes
  - Vision Zero Action Plan was launched in 2014, aimed at reducing the crash rate and relieving crash severity
  - Accurate identification of these hotspots can result in efficient allocation of government resources

- Two Important Factors in Hotspot Identification:
  - a) Different costs of crashes by severity
  - b) Effects of crash exposures such as traffic volume, road length, etc.
Descriptive Analysis

Pedestrian Crash Frequency by Severity (2005~2012)
Descriptive Analysis

- Pedestrian Crash Causes

- DRIVER INATTENTION
- FAILURE TO YIELD RIGHT OF WAY
- PEDESTRIAN ERROR OR CONFUSION
- BACKING UNSAFELY
- TRAFFIC CONTROL DEVICES DISREGARDED
- AGGRESSIVE DRIVING OR ROAD RAGE
- VIEW OBSTRUCTED OR LIMITED
- UNSAFE SPEED
“Big Data” Used

- A massive amount of data from a variety of sources were collected. The total size of datasets is over 100 GB.

**Crash**
- No injury
- Possible injury
- Fatality
- ...
  *(Source: NYSDOT)*

**Traffic**
- Traffic volume
- Taxi trips
- MTA turnstile
- ...
  *(Source: NYSDOT, TCL, MTA)*

**Land Use**
- Source:
  - Residential
  - Commercial
  - ...
  *(Source: NYCDCP)*

**Socioeconomic**
- Population
- Employment
- GDP
- ...
  *(Source: US Census Bureau)*
"Big Data" Used: Taxi Trip Data

- Taxi pick-up and drop-off data from 2008 to 2012. Size of dataset is over 100 GB
- Taxi trips concentrate on main corridors such as 5 Ave and 6 Ave.
“Big Data” Used: MTA Turnstile Data

- Refreshed weekly, available up until May 05, 2010
- Midtown and downtown have large passenger volumes
Grid Cells

- Basic geographical unites of analysis: grid cells (300×300 feet²)
- Traffic, land use, demographic and socioeconomic features were captured for each cell
Spread of Crash Cost

- Crash Cost by Severity

<table>
<thead>
<tr>
<th>Crash Type</th>
<th>Comprehensive Cost per Crash ($)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fatality</td>
<td>4,538,000</td>
</tr>
<tr>
<td>Incapacitating injury</td>
<td>230,000</td>
</tr>
<tr>
<td>Non-incapacitating injury</td>
<td>58,700</td>
</tr>
<tr>
<td>Possible injury</td>
<td>28,000</td>
</tr>
<tr>
<td>Property damage only</td>
<td>2,500</td>
</tr>
</tbody>
</table>

(Source: National Safety Council. All values were converted to 2012 dollars)

- 2-D Kernel Density Function

\[ \lambda(s) = \sum_{i=1}^{n} \frac{1}{\pi r^2} k\left(\frac{d_{is}}{r}\right) \]

\( \lambda(s) \): Density at location \( s \)
\( r \): Bandwidth (1000 feet is used here)
\( d_{is} \): Distance from location \( s \) to crash \( i \)
\( k(.) \): kernel function (Gaussian function is used here)
The potential for safety improvement (PSI) was used as a measure to rank crash hotspots.

- Base cost of “similar” sites can be estimated by the crash cost model.
- Effects of crash exposures can be accounted for.
Crash Cost Model

- Linear Model
  - Develop a linear relationship between dependent variable crash cost and independent variables such as taxi trips, truck ratio, population, etc.

\[ y_i = \beta x_i + \mu_i, \mu_i \sim N(0,\sigma^2) \]

- Weakness of linear model
  - Ignore the fact that crash cost is left-censored at zero.
  - Have the chance to give a negative prediction of the crash cost
Crash Cost Model

- Tobit Model
  - Appropriate for describing relationship between a non-negative dependent variables (crash cost) and independent variables.

\[
y_i = \begin{cases} 
  y_i^* & \text{if } y_i^* > 0 \\
  0 & \text{if } y_i^* \leq 0
\end{cases}
\]

\[y_i^* = \beta x_i + \mu_i, \quad \mu_i \sim N(0,\sigma^2)\]

- \(y_i\): Pedestrian crash cost per year ($)
- \(y_i^*\): Latent variables ($)
- \(x_i\): Independent variables
- \(\beta\): Coefficients of \(x_i\)
- \(\mu_i\): Error term
Modeling Results

- **Model Comparison: Tobit model vs Linear Model**

<table>
<thead>
<tr>
<th></th>
<th>Log-likelihood</th>
<th>AIC</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear model</td>
<td>-74373.18</td>
<td>148774.4</td>
<td>148868.60</td>
</tr>
<tr>
<td>Tobit model</td>
<td>-72883.64</td>
<td>145795.3</td>
<td>145889.50</td>
</tr>
</tbody>
</table>

- The Tobit model outperforms the linear model by presenting higher log-likelihood and lower AIC and BIC.

- **Results of the Tobit Model**

|                      | Estimate | Std. Error | z value | Pr(>|z|) |
|----------------------|----------|------------|---------|----------|
| Intercept            | -1.34E+04| 1.73E+03   | -7.745  | 9.55e-15 * |
| Vehicle mile traveled| 7.99E-04 | 3.65E-04   | 2.189   | 0.028603 * |
| Taxi trips (10^3)    | 3.15E+01 | 3.45E-00   | 3.027   | 0.002471 * |
| Subway passengers (10^3) | 1.77E+01 | 1.52E+01   | 10.942  | <2e-16 *   |
| Truck ratio          | 9.04E+03 | 3.23E+03   | 2.839   | <2e-16 *   |
| Bus stop density     | 17.052   | 17         | <2e-16  | *          |
| Length of sidewalks  | 4.081    | 4.48e-05   | 9.054   | <2e-16 *   |
| Total population     | 7.614    | 2.66e-14   | 2.634   | 0.008432 * |
| Ratio of population over 65 | 2.34E+01 | 5.094      | 3.51e-07 | *          |
| Unemployment         | 1.44E+04 | 2.84E+03   | 5.051   | 4.39e-07 * |
| Ratio of commercial areas | 7.97E+03 | 2.34E+03   | 3.403   | 0.000667 * |
| Ratio of residential areas | 8.37E+03 | 3.09E+03   | 2.708   | 0.006764 * |

One unit increase is expected to increase the annual crash cost by 31.5 $. One unit increase is expected to increase the annual crash cost by 17.7 $. *Indicate variables which are statistically significant
Hotspot Identification

Spot with the greatest improvement potential: Broadway (from 180th to 181st ST)

272,384 $ can be saved annually from this spot!
Comparisons of Hotspots Identified

- Identify top 300 hotspots: by crash frequency vs by PSI
- Only 40 hotspots (about 13.3%) are overlapped
- Hotspots identified by PSI tend to be on continuous regions
Thank You!

Kaan Ozbay, Kun Xie and Hong Yang
Center for Urban Science + Progress (CUSP)
Department of Civil & Urban Engineering (CUE)
New York University