Risk Analysis, Vol. 37, No. 8, 2017 DOI: 10.1111/risa.12785

Analysis of Traffic Crashes Involving Pedestrians Using Big Data: Investigation of Contributing Factors and Identification of Hotspots

Kun Xie,[1,*] Kaan Ozbay,[1] Abdullah Kurkcu,[1] and Hong Yang[2]

This study aims to explore the potential of using big data in advancing the pedestrian risk analysis including the investigation of contributing factors and the hotspot identification. Massive amounts of data of Manhattan from a variety of sources were collected, integrated, and processed, including taxi trips, subway turnstile counts, traffic volumes, road network, land use, sociodemographic, and social media data. The whole study area was uniformly split into grid cells as the basic geographical units of analysis. The cell-structured framework makes it easy to incorporate rich and diversified data into risk analysis. The cost of each crash, weighted by injury severity, was assigned to the cells based on the relative distance to the crash site using a kernel density function. A tobit model was developed to relate grid-cell-specific contributing factors to crash costs that are left-censored at zero. The potential for safety improvement (PSI) that could be obtained by using the actual crash cost minus the cost of "similar" sites estimated by the tobit model was used as a measure to identify and rank pedestrian crash hotspots. The proposed hotspot identification method takes into account two important factors that are generally ignored, i.e., injury severity and effects of exposure indicators. Big data, on the one hand, enable more precise estimation of the effects of risk factors by providing richer data for modeling, and on the other hand, enable large-scale hotspot identification with higher resolution than conventional methods based on census tracts or traffic analysis zones.

KEY WORDS: Big data; grid cell analysis; pedestrian risk

1. INTRODUCTION

In the last few decades, a variety of quantitative methods have been used to explore safety-related data and to provide inferences to essential tasks of safety management such as investigation of risk factors and hotspot identification. In the era of "Big Data," with increase in volume, variety, and acquisition rate of urban data, safety researchers face challenges that can also be turned into great opportunities. The advance in urban big data manifests itself in two ways: (1) massive amounts of data regarding traffic crashes, traffic volumes, road networks, land use, sociodemographic features, and weather have been digitalized and are available for larger urban areas rather than limited regions as before; and (2) emerging data sources such as GPS-equipped taxis, traffic cameras, electronic toll collection facilities, automatic vehicle identification detectors, transit counter turnstiles, cellular telephones, and social media can be leveraged to extract extremely detailed information for decision making. With urban big

[1]Department of Civil and Urban Engineering, Center for Urban Science and Progress, CitySMART Laboratory, New York University, Brooklyn, NY, USA.
[2]Department of Modeling, Simulation & Visualization Engineering, Old Dominion University, Norfolk, VA, USA.
*Address correspondence to Dr. Kun Xie, Department of Civil and Urban Engineering, Center for Urban Science and Progress, CitySMART Laboratory, New York University, Brooklyn, NY 11201, USA; tel: +1 646 997 0547; kun.xie@nyu.edu.

data, there is a potential to gain newer and deeper insights into traffic risk analysis. For example, with richer data for modeling, the effects of risk factors can be estimated more precisely by including into safety models additional explanatory variables that used to be unobservable. Another example is that with massive digitalized and geocoded data available, hotspot identification can be implemented in a larger scale (digitalized data can be obtained for larger urban areas) with higher resolution (when very large amounts of data are provided, statistical patterns of observations aggregated by smaller zones become robust) than conventional methods.

Pedestrians are prone to higher risk of injuries and fatalities when involved in traffic crashes compared with vehicle occupants. In 2013, 66,000 pedestrians were injured and 4,735 were killed by traffic crashes in the United States, accounting for about 3% and 14% of the total roadway injuries and fatalities, respectively.[1] In urban areas of big cities, pedestrian safety is even a more serious concern. Take New York City as an example, where pedestrians constituted approximately 33% of all severe injuries from 2005 to 2009 and 52% of all traffic fatalities from 2004 to 2008.[2] To address pedestrian safety issues, the investigation of contributing factors to pedestrian crashes is of great importance to transportation agencies. Statistical models have been widely used to capture the relationship between pedestrian crash occurrence and site-specific contributing factors. In addition, it is essential to identify hotspots prone to high risk of pedestrian crashes for further examination. Accurate identification of these hotspots can result in efficient allocation of government resources obligated to countermeasure development given time and budget constraints.

The objective of this study is to explore the potential of using big data in advancing the pedestrian safety analysis, including the investigation of contributing factors and hotspot identification. Manhattan, which is the most densely populated urban area of New York City, is used as a case study. Manhattan has four times as many pedestrians killed or severely injured per mile of street compared to the other four boroughs of New York City.[2] The New York City mayor launched the Vision Zero Action Plan in 2014, which has emphasis on pedestrian safety.[3] New York City's open data policy makes data from various government agencies available to the public and this enables in-depth data-driven analyses of pedestrian safety. The first reason for using the term "Big Data" in this study is that massive amounts of data from multiple sources were collected, integrated, and processed. It is worth mentioning that taxi trip data and subway turnstile usage data that were rarely used for safety modeling were also obtained and processed. A program that is designed to take advantage of the advances in parallel data processing was designed to process a large amount of taxi data in a Hadoop-based platform.[4] The second reason is that we are interested in investigating the pedestrian safety patterns of the whole study area instead of focusing on selected samples as in most previous studies on safety modeling such as Hess *et al.*[5] and Xie *et al.*[6] The entire study area was uniformly split into numerous grid cells, which differ by a wide variety of attributes. Grid cells were used as the basic geographical units to capture crash, transportation, land use, sociodemographic features, and social media data, and subsequently were used for model development.

2. LITERATURE REVIEW

2.1. Statistical Modeling

Table I summarizes previous studies on pedestrian crash models. Most previous studies use crash frequencies to indicate the pedestrian hazard and focus on modeling pedestrian crash frequencies. In the early practice, linear regression models were used to capture the relationship between pedestrian crash frequencies and contributing factors.[7,9,12,13] Poisson-based models such as the negative binomial (Poisson-Gamma) models[10,11,15–17] and Poisson-lognormal[18–20] models are proven to outperform linear regression models in accommodating the nonnegative, discrete, and overdispersed features of crash frequencies.[21] To account for the spatial autocorrelation of pedestrian crash data, simultaneous autoregressive models[8] and conditional autoregressive models[19,20] have been developed. Poisson-based models can also be extended by incorporating random parameters[17] to account for unobserved heterogeneity and be integrated with multivariate response models[19] to address correlation among different crash types. Other than studies on pedestrian crash frequency models, Brüde and Larsson[7] estimate the crash rate (crash count per million passing pedestrians) using a linear regression model and Hess *et al.*[5] model the pedestrian crash presence (0 for sites without pedestrian crashes and 1 for sites with pedestrian crashes) using logistic

regressions. Pedestrian safety indicators in the previous studies such as crash frequency, crash rate, and crash presence cannot reflect the injury severity levels of different crashes. Crash cost, differing by injury severity, can be a better safety measure for pedestrians. However, previous studies that focus on modeling crash cost are rare. Crash cost is used to measure pedestrian safety in this study. It is not appropriate to use linear regression models and Poisson-based models to estimate the crash cost, since crash cost is continuous and nonnegative. A tobit model is developed for the crash cost in this study. Details on the tobit model are available in Subsection 4.1.

2.2. Contributing Factors

Contributing factors to pedestrian crashes have been investigated in the literature. The most important and intuitive ones are traffic exposure indicators such as pedestrian volume,[7,8,15,16] vehicle volume,[5,7–10,12,15,16] and vehicle miles traveled (VMT).[18–20] Explanatory variables that can represent the scales of road networks are also commonly used, such as intersection number/density[11,17,18] and road length/density.[11,14,17,19,20] In addition, traffic control and design features[5,7,10,14,15] are found to have significant impacts on pedestrian safety. Shankar *et al.*[10] find that corridors with two-way center turn lanes and smaller signal spacing are prone to have higher risk of pedestrian crashes. Miranda-Moreno *et al.*[15] find four-leg intersections exhibit a higher pedestrian hazard than three-leg intersections after controlling for other variables. Public transit features have been investigated as well. Bus/subway stop number is found to be positively associated with pedestrian crashes.[16,17,19,20] Hess *et al.*[5] affirm that increase in bus ridership would lead to higher pedestrian crash risk. Wier *et al.*[13] state that areas with higher public transit accessibility are likely to have more pedestrian crashes. Furthermore, land-use patterns[12,13,16,17] can be related with the occurrence of pedestrian crashes. Wang and Kockelman[19] find that areas with mixed land-use patterns are associated with higher pedestrian crash frequencies. Demographic features, including population,[8,11–13,16,17,20] age composition,[8,11,13,20] and race composition,[12,17,18] and economic features, including employment/unemployment,[8,12,13] income,[14,20] and population below poverty level,[13,20] are found to have influences on pedestrian safety. In this study, in addition to the traditional data used in the previous safety studies such as vehicle volumes, road network, land use, demographic, and economic features, emerging data sets including taxi trips, subway turnstile counts, and social media are also used for safety modeling. The main objective of incorporating these new data sets is to understand the effect of ever increasing data generated in large and highly connected and densely monitored urban areas. It is difficult to collect the pedestrian volume over the whole study area, so we use surrogate measures such as taxi trip, subway ridership, bus stop density, and the number of tweets to reflect pedestrian exposure. More details on data are presented in Section 3.

2.3. Units of Analysis

Regarding the units of analysis, a portion of previous studies use transportation facilities, including intersections[7,9,15,16] and road segments,[5,10] while others use geographical units for zone-level modeling. There are a variety of geographical units used as analysis zones such as block groups,[18] census tracts,[8,12–14,17,18] traffic analysis zones (TAZs),[11,18] Thiessen polygons based on census tracts,[19] and ZIP areas.[20] Census blocks are the smallest geographic units used by the U.S. Census Bureau. Block groups are composed of census blocks and then assembled into census tracts. Both block groups and census tracts can be easily connected to the demographic and economic features from census data and thus are widely used as units of analysis. TAZs, which are usually collection of census blocks,[22] are delineated by state departments of transportation or metropolitan planning organizations for tabulating transportation-related census data.[18] To be consistent with the zoning system used in transportation planning, it is advantageous to use TAZs for macroscopic safety analysis. Abdel-Aty *et al.*[18] give a detailed discussion on the application of block groups, census tracts, and TAZs for transportation safety planning. However, the boundaries of block groups, census tracts, and TAZs generally coincide with major arterials that could be high-crash locations. Crashes occurring on those boundaries are arbitrarily assigned to adjacent zones in most cases and this would lead to biased inferences. To address this issue, Wang and Kockelman[19] build Thiessen polygons based on the centroid of census tracts and use them for model development. Kim *et al.*[23] and Gladhill and Monsere[24] propose to use uniformly sized grid cells as units of analysis. Using grid cells allows inclusion of crashes without giving special consideration to crashes on the boundaries. The size

of grid cells, which is much smaller than Thiessen polygons, can be helpful in capturing contributing factors more precisely and can provide higher resolution for hotspot identification. In this study, grid cells of Manhattan are used as the geographic units of analysis. The cell-structured framework makes it easy to accommodate diversified data sets. The size of samples used for model development is also much larger than those employed in the literature and it enables high-risk location (hotspot) identification at a higher resolution with enhanced accuracy.

2.4. Hotspot Identification Methods

The naïve methods that simply rely on the raw crash observations such as crash frequencies[25] and crash rates[26] are among the early practice of hotspot identification. A well-known limitation of the naïve methods is the regression-to-the-mean (RTM) issue, especially in cases when data are only available for a short term (e.g., two years or less). Since crashes are rare and random events, sites flagged as hotspots due to high crash frequencies in one period can experience lower crash frequencies subsequently even when no treatment is implemented.[27,28] To address the RTM issue, the empirical Bayes (EB) approach[29–31] and the full Bayes (FB)approach[32,33] that are developed based on safety performance functions have been widely used. The RTM issue can be addressed by using EB-/FB-adjusted crash frequency as a safety measure for ranking. Another safety measure usually used in combination with the EB and FB approaches is the potential for safety improvement (PSI),[34,35] which can properly account for the safety effects of traffic volume and other exposure indicators. Detailed A detailed introduction to PSI is presented in Subsection 4.3. However, most studies on EB and FB approaches neglect the injury severity of crashes. Only a few researchers proposed to incorporate crash severity into risk measures.[34,36]

Spatial analysis techniques such as the local spatial autocorrelation method[37–39] and the kernel density estimation method[38–40] have been used recently in hotspot identification. The local spatial autocorrelation method uses the similarity between one observation and its neighboring observations (local Morgan's I index[41]) to measure the crash concentration. The kernal density estimation method spreads the risk of each crash based on the assumption that crash occurrence is attributed to the spatial interaction existing between neighboring sites. In the previous studies, local spatial autocorrelation and kernel density estimation methods are based on nonparametric estimation and effects of exposure indicators cannot be accounted for. As mentioned above, we use crash cost that can reflect the injury severity levels of different crashes to indicate the pedestrian crash risk and the grid cells as analysis units. The kernel density function is used to distribute the cost of each crash to its neighboring cells. The crash cost of each cell is correlated with cell-specific features using the tobit model. The safety effects of exposure indicators can be accounted for by using PSI estimated from the tobit model to rank hotspots.

3. DATA PREPARATION

The map of Manhattan was uniformly split into a total of 6,204 grid cells with size of 300×300 feet^2, which are used as the units of analysis. The width of a standard block (264 feet) in Manhattan is close to 300 feet and the length of it (900 feet) is divisible by 300 feet.[42] Using cells with lengths of 300 feet can capture location-specific features more precisely and provide street-by-street resolution for risk analysis. Crash, transportation, land use, sociodemographic features, and social media data were captured for each cell using spatial analysis tools in ArcGIS.[43] Advantages of using grid cells as units of analysis over the traditional methods that are based on facilities (intersections and road segments) include: (1) there is no need to decide whether crashes are intersection related or road segment related, which can be a complicated process;[44] (2) there is no need to conduct road segmentation (e.g., splitting roadways at each intersections, removing dangle points); and (3) it is less convenient to incorporate land-use features, taxi trips, and subway ridership into modeling.

We obtained five-year crash record data (2008–2012) from the New York State Department of Transportation[3] (NYSDOT). The crash aggregations over five years have less natural variation and the RTM effect can be relieved. A total of 6,192 crashes in which pedestrians were involved were identified. According to their injury severity, crashes were categorized into five types: no injury (13.49%), possible injury (49.42%), nonincapacitating injury (27.68%), incapacitating injury (9.12%), and fatality (0.29%). The annual pedestrian crash frequencies by injury severity during the study period are presented in Fig. 1. No increase/decrease tendency is observed in crash frequencies from year to year.

[3]Source: http://www.dmv.ny.gov/stats.htm.

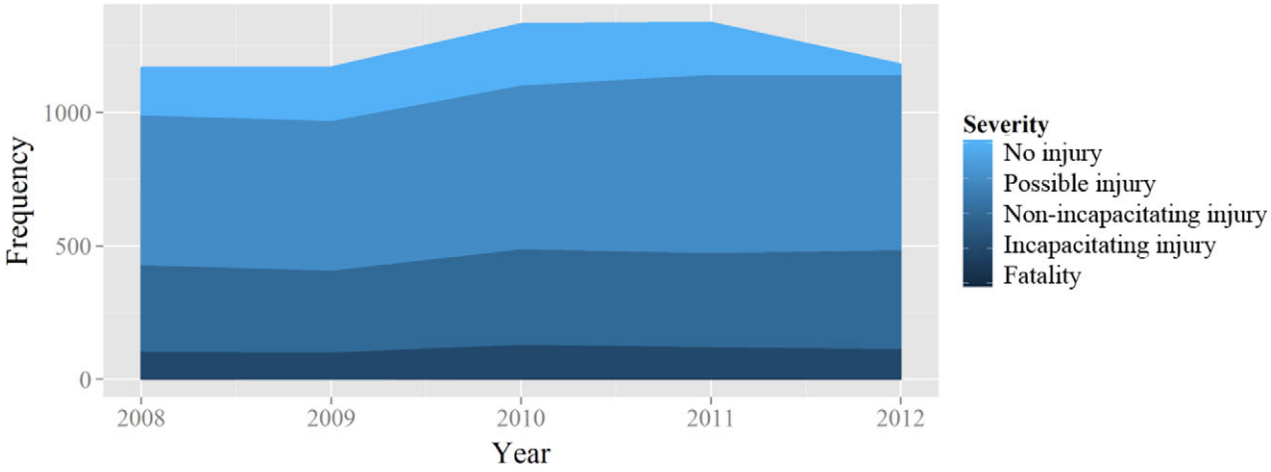

Fig. 1. Annual pedestrian crash frequencies by injury severity (2008–2012).

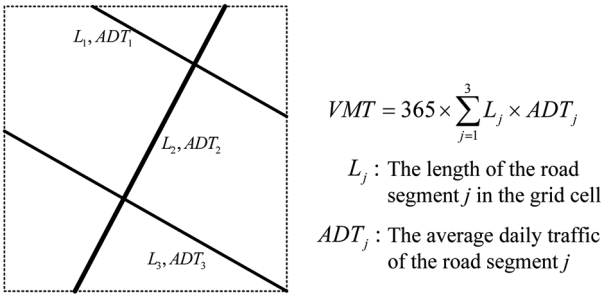

Fig. 2. Demonstration of computing VMT for a grid cell.

Traffic volume and road network data were obtained from NYSDOT,[4] and based on those data sets, VMT was computed for each grid cell. Fig. 2 presents an example of computing VMT for one grid cell. Roadways were split by the boundary of each grid cell using spatial tools in ArcGIS, so that the length of each road segment within the cell could be obtained. VMT could be obtained by computing the sum of the products of road lengths and average daily traffic of road segments.

The truck flow ratio was estimated based on the outcomes of the best practice model (BPM) developed by the New York Metropolitan Transportation Council[5] (NYMTC). More details on the truck flow estimation using the BPM are presented in our previous study.[45] The geographic information system (GIS) data of bus and subway stations were obtained from the Metropolitan Transportation Authority[6] (MTA). Additionally, the ridership for each subway station was computed using the turnstile data provided by MTA.[7] The GIS data of sidewalks and bike paths were obtained from the New York City Department of City Planning5F[8] (NYCDCP) and New York City Department of Transportation (NYCDOT),[9] respectively.

NYCDCP provides detailed information about land use.[10] The land use was classified, into four main categories, including commercial, residential, mixed, and park. A Visual Basic for Applications (VBA) program was developed in ArcGIS to compute the areas by zoning category for each grid cell, and sequentially, the ratio for each zoning category was calculated.

The sociodemographic data based on the 2011 census survey were retrieved from the U.S. Census Bureau.[11] The sociodemographic data are composed of demographic features (e.g., total population, population under 18, and population over 65), economic features (e.g., employment and median income), and housing features (e.g., median value and household average size). It should be noted that sociodemographic data were organized by census tracts, which were larger than the grid cells. The

[4]Source: https://gis.ny.gov/gisdata.
[5]Source: http://www.nymtc.org/project/bpm/bpmindex.html.
[6]Source: http://web.mta.info/developers/download.html.
[7]Source: http://web.mta.info/developers/turnstile.html.
[8]Source: http://www.nyc.gov/html/dcp/html/bytes/dwnsidewalk.shtml.
[9]Source: http://www.nyc.gov/html/dot/html/about/datafeeds.shtml#bikes.
[10]Source: http://www.nyc.gov/html/dcp/html/bytes/dwn_pluto_mappluto.shtml.
[11]Source: http://factfinder.census.gov.

Table I. Previous Studies on Pedestrian Crash Models

Study	Response Variable	Data Set	Location	Methodology	Key Explanatory Variables
Brüde and Larsson[7]	Crash rate	Intersections ($N = 285$)	Sweden	Linear regression model	Pedestrian volume, vehicle volume, and intersection type (signalized, unsignalized, and roundabouts).
LaScala *et al.*[8]	Crash frequency	Census tracts ($N = 149$)	San Francisco, California	Simultaneous autoregressive model	Vehicle volume, population density, age composition of the local population, unemployment, gender, and education.
Lyon and Persaud[9]	Crash frequency	Intersections ($N = 1{,}069$)	Toronto, Canada	Linear regression model	Pedestrian volume and vehicle volume.
Shankar *et al.*[10]	Crash frequency	Corridors ($N = 440$)	Washington State	Negative binomial model and zero-inflated Poisson model	Vehicle volume, traffic signal spacing, presence of center-turn lane, and illumination.
Hess *et al.*[5]	Crash presence (1 for yes and 0 for no)	Highways and urban arterials ($N = 181$)	Washington State	Logistic regression model	Vehicle volume, the number of traffic lanes, transit stop usage, and retail location size.
Ladrón de Guevara[11]	Crash frequency	Traffic analysis zones ($N = 859$)	Tucson, Arizona	Negative binomial model	Intersection density, percentage of miles of principal arterial, percentage of miles of minor arterials, percentage of miles of urban collectors, population density, population under 17, number of employees.
Loukaitou-Sideris *et al.*[12]	Crash frequency	Census tracts ($N = 860$)	Los Angeles, California	Linear regression model	Vehicle volume, land use, population density, employment density, and race.
Wier *et al.*[13]	Crash frequency	Census tracts ($N = 176$)	San Francisco, California	Linear regression model (log-transformed)	Traffic volume, arterial streets without public transit, land use, employment, resident population, population below poverty level, and the proportion over 65.
Cottrill and Thakuriah[14]	Crash frequency	Census tracts ($N = 886$)	Chicago, Illinois	Poisson model (corrected for underreported crashes)	Road length, suitability for walking, transit availability, crime rates, income, and presence of children.
Miranda-Moreno *et al.* (2011) [15]	Crash frequency	Intersections ($N = 519$)	Montreal, Canada	Negative binomial model	Pedestrian volume, vehicle volume, intersection configuration (four-leg and three-leg).
Pulugurtha and Sambhara[16]	Crash frequency	Intersections ($N = 176$)	City of Charlotte, North Carolina	Negative binomial model	Pedestrian volume, vehicle volume, bus stop number, land use, population.
Ukkusuri *et al.*[17]	Crash frequency	Census tracts ($N = 2{,}216$)	New York City	Negative binomial model with random parameters	Intersection number, road length, bus stop number, subway station number, land use, population, and race.
Abdel-Aty *et al.*[18]	Crash frequency	Census tracts ($N = 457$), block groups ($N = 1{,}338$), and traffic analysis zones ($N = 1{,}479$)	Florida	Poisson-lognormal model	VMT, intersection number, the number of workers commuting by public transportation, the workers commuting by walking, and the proportion of minority population.
Wang and Kockelman[19]	Crash frequency	Thiessen polygons based on census tracts ($N = 218$)	Austin, Texas	Poisson-lognormal with multivariate conditional autoregressive effects	VMT, bus stop density, sidewalk density, network intensity, land-use entropy, and population density.
Lee *et al.*[20]	Crash frequency	ZIP areas ($N = 983$)	Florida	Poisson-lognormal with conditional autoregressive effects	VMT, proportion of high-speed roads, density of rail and bus stops, density of hotels, motels, and guest houses, density of ferry terminals, density of K–12 schools, population, proportion of children, proportion of people working at home, proportion of households without available vehicle, proportion of households below poverty level, and median household income.

Table II. Average Comprehensive Cost by Injury Severity

Severity	Unit Cost ($)
Fatality	4,538,000
Incapacitating injury	230,000
Nonincapacitating injury	58,700
Possible injury	28,000
No injury	2,500

aforementioned VBA program was used to capture the area of each census tract for each grid cell. Based on the assumption that all the sociodemographic features were distributed evenly within each census tract, the cell-based features were aggregated after being weighted by census tract areas.

3.1. Spatial Processing

It is assumed that the crashes are not only caused by the risk factors of the cells they are located at but also attributed to the risk factors of neighboring cells. For example, travel demand in the central areas can induce the traffic in the surrounding areas and thus increase the crash risks in both the central areas and the surrounding areas. Therefore, it is essential to spread the hazard of each crash to its surrounding areas. The crash hazard can be measured by crash cost, which varies among crashes with different injury severities. The unit cost of crashes adopted was obtained from the National Safety Council,[46] as shown in Table II.

The kernel density tool in ArcGIS 9.3[43] was employed to spread the cost of each crash spatially with the highest value at the crash site and tapering to zero at the search radius. Raster cells (10×10 feet^2) with values indicating location-specific crash costs were generated using a quartic polynomial as the kernel density function. The crash cost assigned to each raster cell can be expressed as:

$$RC(s) = \sum_{i=1}^{n} \rho \left[1 - \left(\frac{d_{is}}{r} \right)^2 \right]^2 C_i, \qquad (1)$$

where $RC(s)$ is the crash cost assigned to the raster cell s, C_i is the cost of crash i, d_{is} is the distance from the location s to the crash i, and r is the search radius (or bandwidth). The quartic term $\rho[1 - (\frac{d_{is}}{r})^2]^2$ represents the proportion of cost distributed from crash i to raster cell s, where ρ is a constant scaling factor to ensure $\sum_s \rho[1 - (\frac{d_{is}}{r})^2]^2 = 1$. There is still not a well-established quantitative way to determine the search radius r. The average spacing between north–south avenues in Manhattan is about 790 feet. We assume the influence radius of crashes should be greater than the average spacing of avenues. The final selection of search radius is 1,000 feet in this study.[12] The crash cost of each grid cell defined was obtained by aggregating the raster values.

As mentioned, ridership of each subway station was computed from the MTA subway turnstile data. However, these point values cannot properly represent the spread of passengers over the space. The kernel density function was also used to predict the spatial distribution of passengers after leaving the subway stations. C_i in Equation (1) was replaced by the ridership of each station when computing the passenger density. Since the bus ridership is not available, we use the density of bus stops as a surrogate measure for bus ridership. Similarly, the kernel density function was used to compute the bus stop density with C_i in Equation (1) equal to 1. Fig. 3 presents the spatial distribution of crash cost, subway ridership, and bus stop density in Manhattan.

3.2. Big Taxi Data

Three-year New York City taxi data from 2010 to 2012 were obtained from the New York City Taxi & Limousine Commission[13] (NYCTL). The generated taxi trips are approximately 175 million per year and 525 million in total. The pick-up and drop-off locations of each taxi trip are provided in the data set. The total number of pick-ups and drop-offs in each cell can be used as one of the surrogate measures for pedestrian exposure in the safety models. However, it is time consuming and challenging to assign 525 million taxi trips (each include both pick-up and drop-off coordinates) to 6,204 grid cells. Therefore, we designed a MapReduce program to process the massive taxi data set. MapReduce is a programming model for expressing distributed and parallel computations on large-scale data processing.[47] In this study, the MapReduce program is composed of a Mapper that performs counting and sorting and a Reducer that performs a summation operation. More specifically, the Mapper generated a key-value pair for each taxi pick-up/drop-off, with key

[12]For your reference, the search radius used in the study by Anderson (2009) is 200 m (656 feet).

[13]Source: http://www.nyc.gov/html/tlc.

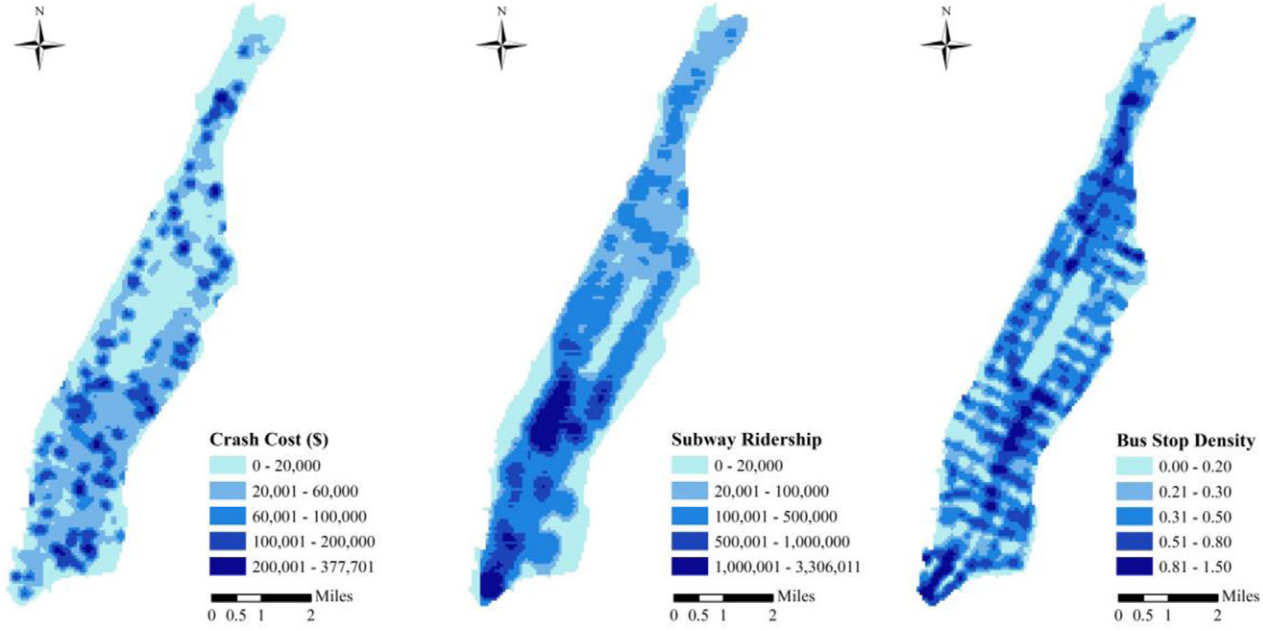

Fig. 3. Crash cost, subway ridership, and bus stop density at grid-cell level in Manhattan.

corresponding to the grid cell ID and value equal to one. Then output from the Mapper was sorted by grid cell ID and sent to the Reducer, where taxi pick-ups/drop-offs were aggregated according to the grid cell ID.

R-tree proposed by Guttman in 1984[48] is a dynamic index structure for spatial searching. The basic idea of R-tree is to group spatial objects with minimum bounding rectangles and organize those bounding rectangles in a tree structure. When a query is conducted, only the objects within the bounding rectangles intersected with the query are checked. Thus, most of the objects in the R-tree do not need be read during a query. The R-tree indexing approach was integrated in the Mapper to expedite taxi data processing. It helped to reduce the computation time tremendously. The MapReduce program was designed via an open-source implementation Hadoop and was operated in computing clusters provided by the Amazon Web Service (AWS).[49] Annual taxi trips (counting both pick-ups and drop-offs) for grid cells were obtained for the years 2010, 2011, and 2012. It was found that the year-to-year variation of total taxi trips is quite small (within 5% difference). The average annual taxi trips were computed and used as a surrogate measure for pedestrian volume in the crash cost models.

3.3. Social Media Data

Social media data have the potential to be used as information providers in transportation research. A recent study by Kurkcu *et al.*[50] presents the application of social media data in incident management. In this study, social media data are used to extract potential indicators of pedestrian exposure. Gnip[14] is a social media application programming interface (API) aggregation company that allows users to collect data from various social media APIs simultaneously. The "Historical Power Track" tool of Gnip, which delivers 100% of publicly available tweet messages from Twitter since 2006, was used to gather geo-tagged tweets. It is worth mentioning that geo-tagged tweets are not available prior to 2011 for Twitter's compliance reasons on Gnip. A bounding box was used to filter the geo-tagged tweets with the rule that the tweets' geolocations should be fully contained within the defined region. The bounding box for this study was defined by −40° 41′ 51″ N, −74° 1′ 39″ W, and 40° 52′ 38″ N, −73° 54′ 11″, which contains the whole study area. This filtering job can be performed by making a HTTP POST request to the Gnip's API interface. The period of tweets, filtering rules, and some additional meta-data have to be included in the POST request.

[14]Source: https://gnip.com/.

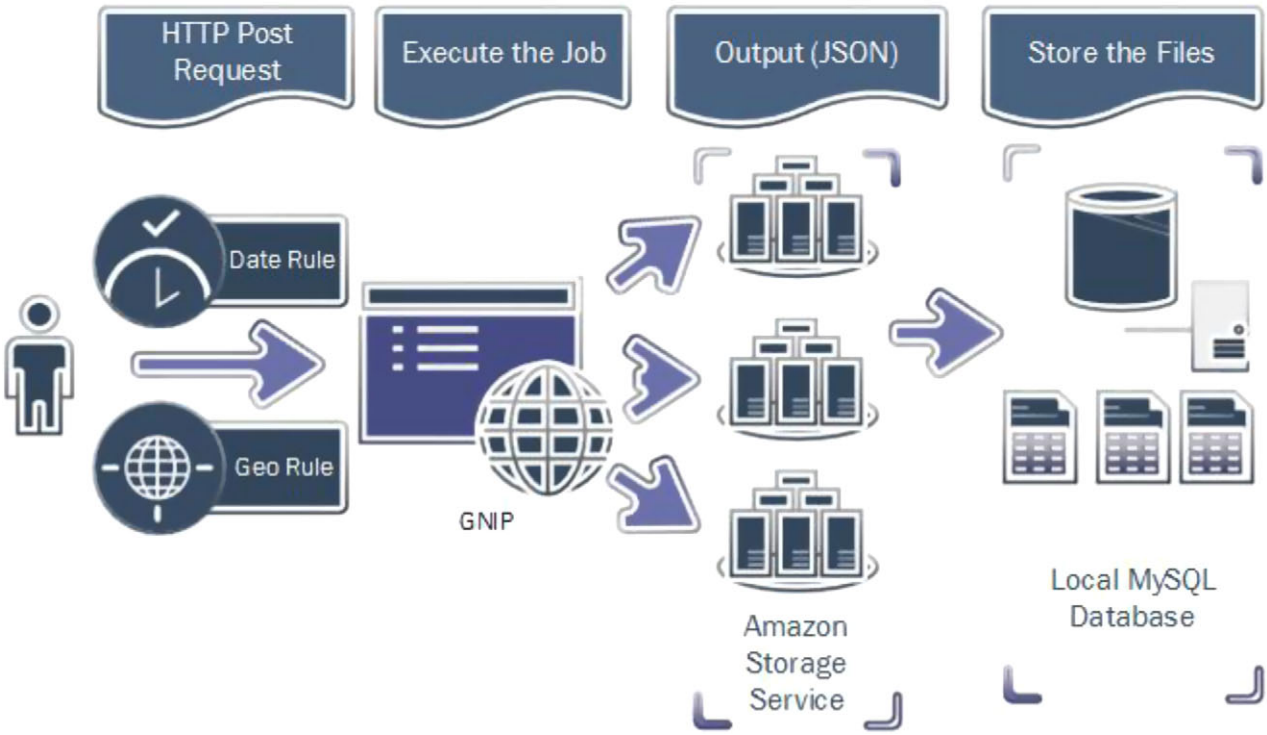

Fig. 4. Twitter data acquisition procedure.

After creating the request, the job should be accepted with user credentials to retrieve historical data. When it is completed, Gnip will deliver a data uniform resource locator (URL) endpoint that contains a list of file URLs that can be downloaded simultaneously. The generated compressed JavaScript object notation data files are hosted at AWS's Simple Storage Service (S3) and they are available for 15 days. Requested jobs on Gnip may end up delivering millions of tweets that require large amounts of storage space. Therefore, it generates up to six files for each hour of the requested time period. On average, 50,364 files are generated for each year of data. A Python code is developed to send the POST requests, accept the job, retrieve the list of URLs provided by Gnip, and download them in parallel. Fig. 4 shows the Twitter data acquisition procedure. The final data set is stored in a local MySQL database, which contains about 7 million geo-tagged tweets posted in defined bounding boxes between September 1, 2011 and October 1, 2015. For the study horizon, a total of 948,238 geo-located Twitter messages was extracted. The annual number of tweets for each grid cell was computed.

The descriptions and descriptive statistics of crash, transportation, land use, sociodemographic, and social media data are listed in Table III.

4. METHODOLOGY

4.1. Tobit Model

Crash cost that accounts for both crash frequency and severity is used as the response variable in model development. The tobit model (also referred to as a censored regression model) first proposed by Tobin[51] can accommodate left-censored dependent variables. The tobit model assumes that there is a latent variable Y_i^*, which can be regressed by explanatory variables. The dependent variable Y_i is equal to Y_i^* when Y_i^* is positive and is observed to be zero when Y_i^* is less than or equal to zero. Crash cost with a value of zero can be regarded as left-censored because the corresponding latent variable is ensured to be less than or equal to zero, although its real value cannot be measured. If the censoring effect is not considered, for example, using a linear model to replicate the cost distribution, negative estimates for cost can be generated, which

Table III. Descriptions and Descriptive Statistics of Key Variables ($N = 6{,}204$ grid cells)

Variable	Description	Mean	*SD*
Crash			
Crash cost	Annual average cost of pedestrian crashes after spatial processing (10^3 \$) (136 zeros)	42.64	45.22
Transportation			
VMT	Annual vehicle miles traveled (10^6 veh. mile)	900.72	1,479.12
Truck ratio	The average ratio of truck flow to total flow	0.04	0.05
Subway ridership	Annual subway ridership after spatial processing (10^3)	245.76	392.20
Bus stop density	Number of bus stops after spatial processing	0.36	0.23
Sidewalk	Total length of sidewalks (mile)	0.07	0.07
Bike path	Total length of bike paths (mile)	0.02	0.03
Taxi trip	Average of annual taxi pick-ups and drop-offs (10^3)	48.16	75.57
Land use			
Commercial ratio	The ratio of commercial zone area to the whole area	0.29	0.40
Residential ratio	The ratio of residential zone area to the whole area	0.50	0.44
Mixed ratio	The ratio of mixed zone area to the whole area	0.06	0.22
Park ratio	The ratio of park area to the whole area	0.14	0.31
Sociodemographic			
Population	Total population	241.83	151.22
Population under 14	The population under 14 years	30.13	24.44
Population over 65	The population 65 years and over	32.10	25.70
Male	The population of males	113.86	70.73
Female	The population of females	127.95	82.21
White	The white population	116.28	116.01
Black	The black population	31.30	51.60
Asian	The Asian population	26.57	40.02
Hispanic	The Hispanic population	61.72	91.03
Median age	Median age of population	1.58	0.99
Median income	Median income per household (10^3 \$)	3.26	2.71
Employed	Number of the employed	129.51	87.47
Unemployed	Number of the unemployed	11.77	10.37
Social media			
Tweet number	Average number of tweets per year	114.90	194.78

is unrealistic, whereas the tobit model can account for the censoring effect and restrict the outputs to be nonnegative. Tobit models have been applied in transportation safety research to model the crash rates.[52,53] The tobit model can be described as:

$$Y_i^* = \beta \mathbf{X}_i + \varepsilon_i,$$
$$Y_i = \begin{cases} Y_i^* & if\ Y_i^* > 0 \\ 0 & if\ Y_i^* \le 0 \end{cases}, \quad (2)$$

where Y_i is the dependent variable (crash cost) for site i ($i = 1, 2, ..., n$, n is the number of observations), Y_i^* is the latent variable, $\mathbf{X}_i$ is a vector of explanatory variables (transportation, land use, and demo-economic features), β is a vector of coefficients to be estimated, and ε_i is the error term, which follows a Gaussian distribution with mean zero and variance σ^2. The log-likelihood function for the tobit model is:

$$\ln L = \sum_{Y_i>0} \ln\left[\mathrm{Prob}\left(Y_i^* = Y_i\right)\right] + \sum_{Y_i=0} \ln\left[\mathrm{Prob}\left(Y_i^* < 0\right)\right]$$
$$= \sum_{Y_i>0} \ln\left[\phi\left(\frac{Y_i - \beta \mathbf{X}_i}{\sigma}\right)\sigma^{-1}\right] + \sum_{Y_i=0} \ln\left[1 - \Phi\left(\frac{\beta \mathbf{X}_i}{\sigma}\right)\right], \quad (3)$$

where $\phi(.)$ and $\Phi(.)$ are the probability density function and cumulative density function of the standard normal distribution, respectively. Tobit models are calibrated by maximizing the log-likelihood given in Equation (3).

In Equation (2), since ε_i is normally distributed with mean zero, the expectation of latent variable $E(Y_i^*)$ is $\beta \mathbf{X}_i$, and the marginal effects of $\mathbf{X}_i$ on the latent variable are β. The censoring effects have to be considered to obtain the expectation of the

dependent variable $E(Y_i)$ and it is given by:[54]

$$\begin{aligned} E(Y_i) &= \Pr ob(Y_i = 0) \times E[Y_i | Y_i = 0] \\ &\quad + \Pr ob(Y_i > 0) \times E[Y_i | Y_i > 0] \\ &= \Pr ob\left(Y_i^* \le 0\right) \times 0 + \Pr ob\left(Y_i^* > 0\right) \\ &\quad \times E[Y_i^* | Y_i^* > 0] \\ &= \Phi\left(\frac{\beta \mathbf{X}_i}{\sigma}\right) \times \int_0^{\infty} Y_i^* \phi\left(\frac{Y_i^* - \beta \mathbf{X}_i}{\sigma}\right) \sigma^{-1} dY_i^* \\ &= \phi\left(\frac{\beta \mathbf{X}_i}{\sigma}\right) \sigma + \Phi\left(\frac{\beta \mathbf{X}_i}{\sigma}\right) \beta \mathbf{X}_i. \end{aligned} \tag{4}$$

Equation (4) is used to estimate the expected average crash cost in the section of hotspot identification. The marginal effects of $\mathbf{X}_i$ on the dependent variable can be obtained by:

$$\begin{aligned} \frac{\partial E(Y_i)}{\partial \mathbf{X}_i} &= \frac{\partial\left[\phi\left(\frac{\beta \mathbf{X}_i}{\sigma}\right)\sigma\right]}{\partial \mathbf{X}_i} + \frac{\partial\left[\Phi\left(\frac{\beta \mathbf{X}_i}{\sigma}\right)\beta \mathbf{X}_i\right]}{\partial \mathbf{X}_i} \\ &= \frac{\partial\left[\phi\left(\frac{\beta \mathbf{X}_i}{\sigma}\right)\sigma\right]}{\partial \mathbf{X}_i} + \frac{\partial\left[\Phi\left(\frac{\beta \mathbf{X}_i}{\sigma}\right)\right]}{\partial \mathbf{X}_i} \times \beta \mathbf{X}_i \\ &\quad + \Phi\left(\frac{\beta \mathbf{X}_i}{\sigma}\right) \times \frac{\partial \beta \mathbf{X}_i}{\partial \mathbf{X}_i} \\ &= -\frac{\beta \mathbf{X}_i}{\sigma} \times \phi\left(\frac{\beta \mathbf{X}_i}{\sigma}\right) \times \frac{\beta}{\sigma} \times \sigma \\ &\quad + \phi\left(\frac{\beta \mathbf{X}_i}{\sigma}\right) \times \frac{\beta}{\sigma} \times \beta \mathbf{X}_i + \Phi\left(\frac{\beta \mathbf{X}_i}{\sigma}\right) \beta \\ &= \Phi\left(\frac{\beta \mathbf{X}_i}{\sigma}\right) \beta. \end{aligned} \tag{5}$$

According to Equation (5), the marginal effects of tobit models can be regarded as the estimated coefficients β times the expected proportion of uncensored observations $\Phi(\frac{\beta \mathbf{X}_i}{\sigma})$. Equation (5) is used for variable interpretation in the following section.

4.2. Model Assessment

The coefficient of determination R^2 is usually used to measure the model goodness of fit.[55] In addition to R^2, criteria based on likelihood estimation such as the Akaike information criterion[56] (AIC) and the Bayesian information criterion[57] (BIC) are used. AIC introduces parameter number as a penalty term and can serve as a comprehensive measure of model fitting and model complexity. As an alternative to AIC, BIC combines parameter number and sample size into the penalty term. The AIC and BIC can be expressed as:

$$AIC = -2LL_{\max} + 2k, \tag{6}$$

$$BIC = -2LL_{\max} + k\ln(N), \tag{7}$$

where $LL_{\max}$ is the maximum of log-likelihood function (Equation (3)), k the parameter number, and N is the sample size. If the AIC/BIC difference is greater than 10, the model with a lower AIC and BIC should be favored.[58,59]

4.3. PSI

Crash hotspots are not simply the ones with the highest crash costs, but the ones that are less safe than "similar" sites as a result of site-specific deficiency. The PSI has been widely used as a measure to identify crash hotspots.[34,35] PSI can be defined as the actual crash cost minus the expected cost of "similar" sites that can be obtained from the crash cost models. The safety effects of exposure indicators (e.g., VMT) can be accounted for in the crash cost models and thus PSI can capture the portion of crash cost that is caused by unobserved site-specific risk factors. The sites with higher PSI are expected to have far less crash costs after the implementation of the countermeasures. PSI is given by:

$$PSI_i = Y_i - E(Y_i), \tag{8}$$

where PSI_i is the PSI for site i. $E(Y_i)$ represents the expected average crash cost for sites that are similar to site i and can be estimated using Equation (4).

5. MODELING RESULTS AND VARIABLE INTERPRETATION

After introducing the methodology in the previous section, this section presents the modeling results and the variable interpretation. The linear regression and tobit models proposed were developed to estimate the annual cost of pedestrian crashes. The two models have the same selection of explanatory variables so that effective model comparison can be performed. Twelve explanatory variables were included after diagnosing multicollinearity using variance inflation factors (VIF). A VIF greater than 5 indicates the existence of a multicollinearity problem.[60] As presented in Table IV, the VIF of each explanatory variable is less than 5, and thus no multicollinearity is detected using this test.

Table IV. Detection of Multicollinearity Using Variance Inflation Factors (VIF)

Variables	VIF
Transportation	
VMT	1.086
Truck ratio	1.264
Subway ridership	1.665
Bus stop density	1.294
Taxi trip	1.718
Land use	
Commercial ratio	3.701
Residential ratio	3.885
Mixed ratio	1.487
Sociodemographic	
Population	3.104
Ratio of population over 65	1.267
Unemployed	2.628
Social media	
Tweet number	1.419

Maximum likelihood method was used for model estimation. Marginal effects of the tobit model were estimated using Equation (5). Coefficient estimates and marginal effects of explanatory variables, as well as assessment measures are reported in Table V. Statistic indicator p-value was used to test the significance of explanatory variables. All the explanatory variables were regarded as statistically significant at the 95% level (p-values < 0.05) in the tobit model, whereas the variables VMT, residential ratio, and ratio of population over 65 in the linear regression model were found to be insignificant.

It is clear from Fig. 5 that the grid-cell-based pedestrian crash cost is not normally distributed and has a lower bound at 0 (i.e., left-censored at 0). So theoretically, the tobit model, which can account for the censoring effect, should accommodate the crash cost data better. According to R^2 in Table V, the tobit model could explain 26.2% ($R^2 = 0.262$) of the variance in the crash cost that is greater than that of the linear regression model. Additionally, the LL_{max} values indicate that the tobit model is more likely to fit the data compared with the linear regression model. Comprehensive measures including AIC and BIC also suggest that the tobit model has significantly better performance (AIC and BIC differences are greater than 10). Overall, all those statistics provide evidence that the tobit model is superior to the linear regression model by accommodating the censored data. If the censoring is ignored, it will lead to biased estimates and unreliable statistical inferences.

Table V. Results of the Linear Regression and Tobit Models

	Linear Regression Model				Tobit Model			
	Estimate	Std. Error	p-Value	Marginal Effect	Estimate	Std. Error	p-Value	Marginal Effect
Intercept	−6.806	1.526	<0.001		−13.605	1.728	<0.001	
Transportation								
VMT	0.681	0.349	0.051	0.681	0.807	0.365	0.027	0.789
Truck ratio	106.500	11.059	<0.001	106.500	106.808	11.274	<0.001	104.467
Subway ridership	0.016	0.002	<0.001	0.016	0.017	0.002	<0.001	0.017
Bus stop density	40.595	2.436	<0.001	40.595	42.907	2.495	<0.001	41.966
Taxi trip	0.019	0.008	0.022	0.019	0.017	0.009	0.049	0.017
Land use								
Commercial ratio	15.170	2.362	<0.001	15.170	19.161	2.488	<0.001	18.741
Residential ratio	4.136	2.198	0.060	4.136	8.299	2.337	<0.001	8.117
Mixed ratio	7.045	2.713	0.009	7.045	11.898	2.861	<0.001	11.637
Sociodemographic								
Population	0.044	0.006	<0.001	0.044	0.046	0.006	<0.001	0.045
Ratio of population over 65	13.162	7.198	0.068	13.162	20.156	7.402	0.006	19.714
Unemployed	0.405	0.077	<0.001	0.405	0.410	0.079	<0.001	0.401
Social media								
Tweet number	0.012	0.003	<0.001	0.012	0.012	0.003	<0.0001	0.012
Model assessment								
R^2		0.258				0.262		
LL_{max}		−31,518				−30,968		
AIC		63,065				61,965		
BIC		63,159				62,059		

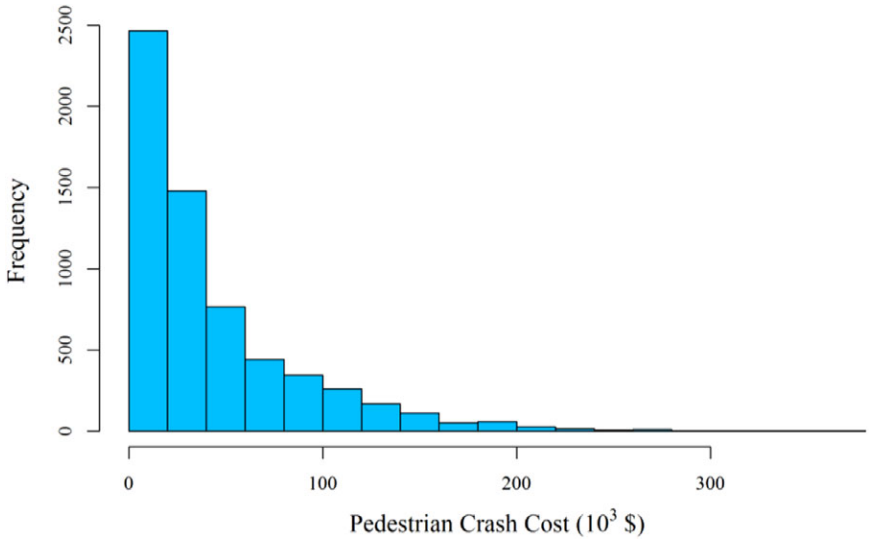

Fig. 5. Distribution of grid-cell-based pedestrian crash costs.

Due to its relatively better performance, the tobit model is used to interpret the effects of explanatory variables on pedestrian safety. VMT is found to be positively associated with crash cost and this finding is consistent with the previous studies.[18–20] Greater miles traveled by vehicles provide more opportunities for collisions with pedestrians. The marginal effect of VMT can be interpreted as: a one unit (10^6 vehicle mile) increase in VMT is predicted to raise the pedestrian crash cost by \$789 (0.789×10^3). Similarly, a 1% increase in truck ratio will lead to approximately \$1,045 ($1\% \times 104.467 \times 10^3$) more pedestrian crash cost. Intuitively, trucks can disturb the traffic flow and cause more severe crashes due to their heavy weight. Subway ridership and bus stop density, which are two pedestrian exposure indicators, are found to have positive impacts on the crash cost. Previous studies[16,17,19,20] show positive association between bus/subway stop number and pedestrian crashes, but the effect of subway ridership has not been investigated yet. According to the marginal effect of subway ridership, each increase of 1,000 subway ridership is accompanied with an increase in crash cost of \$17 in that region, keeping other variables constant. It should be noted that we are not claiming that the higher share of public transit leads to higher pedestrian crash cost, but regions with higher bus density or subway ridership have a higher number of pedestrians who are public transit users, and thus are associated with higher pedestrian crash cost.

Consistent with the findings by previous studies,[12,13,16,17] land-use patterns are found to be related to the risk of pedestrian crashes. Modeling results indicate that the ratios of commercial, residential, and mixed areas have positive impacts on crash cost. Among all the land-use variables, the ratio of commercial area has the highest marginal effect on crash cost. A possible reason for this finding is that greater traffic attracted to commercial areas poses a greater risk of pedestrian crashes.

A number of studies indicate that the total population is positively associated with pedestrian crash occurrence[8,11,13,16,17,20] and this has been reconfirmed in this study. An increase of 1,000 in population is expected to promote the crash cost by \$45 (0.045×10^3). The regions with higher ratio of population over 65 tend to have higher pedestrian crash risk. A similar finding has been uncovered by Wier *et al.*[13] The elderly are more likely to be involved in a crash since they suffer from weak vision and hearing, and have longer perception–response time than others. In addition, the number of the unemployed is found to be positively related to pedestrian crash cost. Impacts of employment/unemployment have been discussed in previous studies.[8,12,13]

Regarding the social media data, the relationship between the number of tweets and the crash cost is found to be highly significant. It means the number of tweets can serve as a good indicator of pedestrian exposure. This finding shows the great

Fig. 6. Comparison of expected and observed annual pedestrian crash costs.

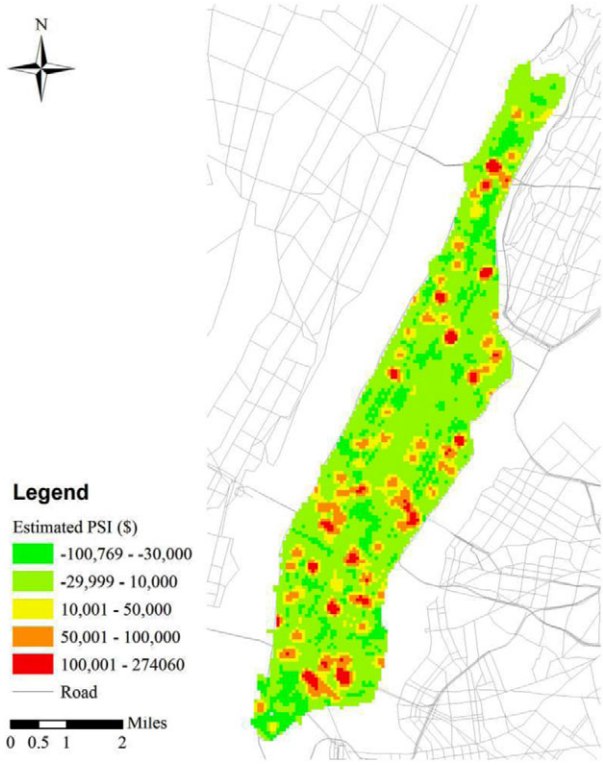

Fig. 7. Distribution of cell-based potential for safety improvement (PSI) in Manhattan.

potential of using social media data to extract helpful information for safety research.

6. HOTSPOT IDENTIFICATION

The expectation of annual pedestrian crash cost for each grid cell can be obtained by using Equation (4). It can be seen from Fig. 6 that an increase in the expected annual pedestrian crash cost is accompanied by an overall increase trend in the observed annual pedestrian crash costs. In Fig. 6, the grid cells that have higher observed crash costs than the expected ones are denoted with "x." These grid cells have positive PSI according to Equation (8) and can be flagged as hotspot candidates. Grid cells denoted with "o" can be regarded as relatively safer since their PSIs are less than or equal to zero. Generally, grid cells with higher observed crash costs are more likely to have positive PSIs.

PSI for each grid cell in Manhattan was computed using Equation (8). Fig. 7 shows the distribution of cell-based PSIs with green color indicating safe zones with low PSIs and red color indicating hazardous zones with high PSIs (color visible in on-line version). As demonstrated in Fig. 7, spatial clustering of high-risk zones can be observed (i.e., cells with red colors tend to be close to each other). An interesting finding is that numerous clusters of high-risk zones are located in the regions with access to the entrances/exits of tunnels/bridges, although these regions do not have the highest pedestrian volumes. One possible reason can be the disruptions in traffic flows caused by a large number of vehicles entering and leaving these tunnels/bridges. Also, a further check of the crash data shows that more severe crashes could be found in these regions. The grid cell with the highest PSI lies in Washington Heights, including the Broadway segment from 180th Street to 181th Street and the 180th Street segment from Broadway to Wadsworth Ave. Its PSI value implies that the pedestrian crash cost within the cell is approximately $274,060 higher than "similar" sites. Its high pedestrian crash cost can be attributed to risk factors not included in the model, such as poor traffic control device visibility, inadequate channelization, and sharp crossing angle. If countermeasures are implemented completely,

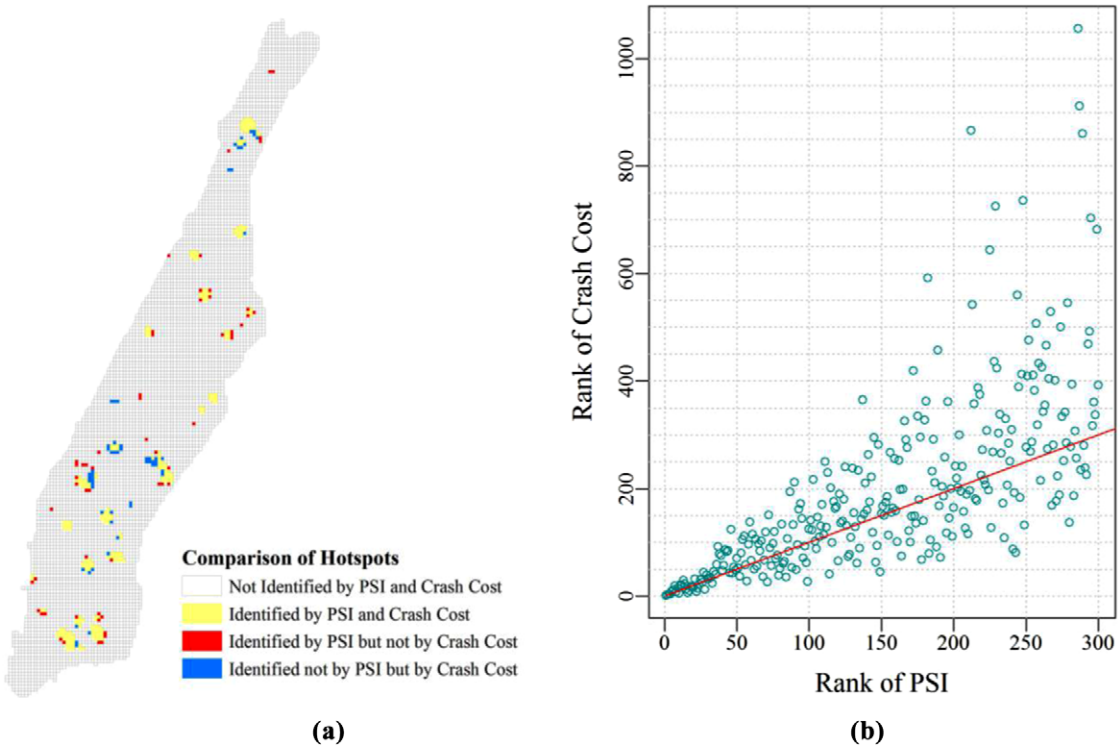

Fig. 8. Comparisons of hotspots identified by potential for safety improvement (PSI) and crash cost.

theoretically, $274,060 can be saved from pedestrian crashes within the cell each year.

Given time and budget constraints, only a portion of the grid cells should be selected as hotspots to implement treatments although most of them have the potential to be improved. In this study, we took the worst 300 cells (about 5%) as the hotspots to be examined. Comparisons were conducted between the hotspots selected by PSI (i.e., grid cells ranked top 300 by PSI) and those by crash cost (i.e., grid cells ranked top 300 by crash cost). Fig. 8(a) shows the difference between the hotspots identified by PSI and crash cost. In summary, 242 grid cells were identified as hotspots by both the PSI and crash cost, 58 grid cells were identified by PSI only, and another 58 were identified by crash cost only. Fig. 8(b) shows the difference between the ranks by PSI and the ranks by crash cost for the 300 grid cells with the highest PSIs. The *X*-axis represents the rank in decreasing order of the estimated PSI, and the *Y*-axis represents the rank in decreasing order of the observed crash cost. The spread of points around the red line indicates the difference in identifying the hotspots of pedestrian crashes. A tendency toward greater ranking difference between PSI and crash cost is observed as the rank of PSI increases. Additionally, it can be seen that a portion of the hotspots identified by PSI have high ranks of crash cost (over 600). The hotspot identification approach based on PSI has the potential to find sites with relatively low crash costs, which would otherwise be neglected by method based on crash cost.

7. SUMMARY AND CONCLUSIONS

This study explores the advantages of using big data in pedestrian risk analysis. A novel grid-cell-structured framework is proposed to investigate the effects of contributing factors to pedestrian crash cost and to identify the hotspots of pedestrian crashes. Manhattan, which is the most densely populated urban area of New York City, is used as a case study. Massive amounts of data from multiple sources such as taxi trip, subway turnstile, traffic volume, road network, land use, sociodemographic, and social media data were collected and used for modeling pedestrian crash cost. A parallel computation program was designed in a Hadoop-based platform to process a large amount of taxi data. It is

worth mentioning that the Twitter data were used to extract potential indicators of pedestrian exposure.

To investigate the overall safety patterns of pedestrians, the whole study area was uniformly split into grid cells as the basic geographical units of analysis. The cost of each crash, differing by injury severity, was assigned to the neighboring cells using a kernel density function. One advantage of using grid cells is that it enables inclusion of crashes without giving special consideration to crashes on the boundaries. Two cell-based crash cost models were developed for pedestrian crash cost. Statistic measures suggest that the proposed tobit model outperforms the linear regression model by accommodating the left-censored feature of crash cost. The tobit model was applied to investigate the effects of explanatory variables on pedestrian crash cost. Results show that VMT, truck ratio, subway ridership, bus stop density, taxi trip, ratio of commercial area, ratio of residential ratio, ratio of mixed area, population, ratio of population over 65, unemployment, and number of tweets are positively associated with crash cost.

This study further contributes to the literature by proposing a grid-cell-based hotspot identification approach. The PSI, which could be obtained by using the actual crash cost minus the cost of "similar" sites estimated by the crash cost model, is used as a measure to identify pedestrian crash hotspots. This approach takes into account two important factors that are generally ignored: (1) injury severity—use crash cost to indicate pedestrian crash hazard instead of crash frequency; and (2) effects of exposure indicators—use PSI to identify the hotspots. In addition, the grid-cell-based hotspot identification approach provides a pedestrian crash risk map of the whole study area with higher resolution than conventional methods based on census tracts or TAZs. Comparisons were conducted between the hotspots selected by PSI and those by crash cost. Note that 242 grid cells out of 300 were identified as hotspots by both PSI and crash cost. Furthermore, the hotspot ranks by PSI were compared with those by crash cost. Results show that PSI has the potential to find high-risk sites with relatively low crash costs, which would otherwise be neglected by methods based on crash cost. It should be noted that after identifying the hotspots, field visits and knowledge on the effectiveness of countermeasures gained through before–after safety studies are still needed for the development of countermeasures to improve the safety performance. The proposed methodology has potential transferability and can be implemented in less populated regions by adjusting the size of the grid cells and the bandwidth of kernel density functions for spatial processing.

The potential of harnessing big data to advance risk analysis is presented in this article. On the one hand, big data enable more precise estimation of the effects of risk factors by providing richer data for modeling. Explanatory variables rarely exploited in the literature, including taxi trips, subway ridership, and tweet number, are used to represent pedestrian exposure. Biased inferences would be obtained if pedestrian exposure is not accounted for properly. On the other hand, big data enable large-scale hotspot identification at a much higher resolution than conventional methods based on census tracts or TAZs. The crash, taxi trip, and Twitter data contain specific coordinate information, which makes it possible to explore traffic safety patterns at a street-by-street level. A high-resolution crash hotspot map of the whole study area was generated with the detailed hotspot ranking that can help road safety managers in prioritizing interventions at the citywide level. Overall, big data analytics has the potential to help government agencies gain deeper insights and support them in making better decisions on the allocation of resources for safety improvement.

This article aims to serve as a stepping stone for grid-cell-based risk analysis. Using a cell-structured framework to model the potential for risk reduction is first published in this journal. The cell-structured framework has the potential to incorporate richer and more diversified data sets into safety modeling. Future study is needed to evaluate the effectiveness of the proposed hotspot identification method and compare it with the traditional ones. Other than the taxi GPS data and social media data used in this study, additional crowdsourced data such as mobile devices, in-car sensors, and surveillance cameras can be used for proactive safety management. Those emerging data sets not only provide location-specific but also time-specific information. The spatiotemporal relationship between crash occurrence and its contributing factors can be established in real time or near real time. Time-dependent hotspots can be identified and used to support the patrol routes and frequency of police cars on a daily or even hourly basis. Furthermore, big data could provide more information for real-time crash risk assessment, and as connected vehicle technologies continue to advance, it will be possible to take active actions (e.g., notifying drivers, lower speed limits) to prevent the occurrence of crashes before they actually do. In

addition, the proposed methodology has the chance to be applied in other fields such as health, public security, and environment. For example, the regions with excessive certain disease types, crimes, and natural hazard could be identified.

ACKNOWLEDGMENTS

The work was partially funded by the CityS-MART laboratory of the UrbanITS center at the Tandon School of Engineering, and the Center for Urban Science and Progress (CUSP) at New York University (NYU). The authors would like to thank the New York State Department of Transportation, the New York City Department of Transportation, the New York Metropolitan Transportation Council, the Metropolitan Transportation Authority, and the New York City Department of City Planning for providing data for the study. The contents of this article reflect views of the authors who are responsible for the facts and accuracy of the data presented herein. The contents of the article do not necessarily reflect the official views or policies of the agencies.

REFERENCES

1. NHTSA. 2014. NHTSA. Traffic Safety Facts 2012. Washington, DC: U.S. Department of Transportation, National Highway Traffic Safety Administration, 2014.
2. Viola R, Roe M, Shin H. The New York City Pedestrian Safety Study and Action Plan. New York: New York City Department of Transportation, 2010.
3. Government NYC. Vision zero action plan 2014. In Series Vision Zero Action Plan 2014. 2014. New York City Government. Zero Action Plan 2014. Available at: http://www.nyc.gov/html/visionzero/pdf/nyc-vision-zero-action-plan.pdf.
4. White T. Hadoop: The Definitive Guide, 3rd ed. Sebastopol, CA: O'Reilly Media, Inc., 2012.
5. Hess PM, Moudon AV, Matlick JM. Pedestrian safety and transit corridors. Journal of Public Transportation, 2004; 7(2):5.
6. Xie K, Wang X, Huang H, Chen X. Corridor-level signalized intersection safety analysis in Shanghai, China using Bayesian hierarchical models. Accident Analysis and Prevention, 2013; 50:25–33.
7. Brüde U, Larsson J. Models for predicting accidents at junctions where pedestrians and cyclists are involved. How well do they fit? Accident Analysis & Prevention, 1993; 25(5):499–509.
8. LaScala EA, Gerber D, Gruenewald PJ. Demographic and environmental correlates of pedestrian injury collisions: A spatial analysis. Accident Analysis & Prevention, 2000; 32(5):651–658.
9. Lyon C, Persaud B. Pedestrian collision prediction models for urban intersections. Transportation Research Record: Journal of the Transportation Research Board, 2002; (1818):102–107.
10. Shankar VN, Ulfarsson GF, Pendyala RM, Nebergall MB. Modeling crashes involving pedestrians and motorized traffic. Safety Science, 2003; 41(7):627–640.
11. Ladrón de Guevara F, Washington S, Oh J. Forecasting crashes at the planning level: Simultaneous negative binomial crash model applied in Tucson, Arizona. Transportation Research Record: Journal of the Transportation Research Board, 2004; (1897):191–199.
12. Loukaitou-Sideris A, Liggett R, Sung H-G. Death on the crosswalk: A study of pedestrian-automobile collisions in Los Angeles. Journal of Planning Education and Research, 2007; 26(3):338–351.
13. Wier M, Weintraub J, Humphreys EH, Bhatia R. An area-level model of vehicle-pedestrian injury collisions with implications for land use and transportation planning. Accident Analysis & Prevention, 2009; 41(1):137–145.
14. Cottrill CD, Thakuriah PV. Evaluating pedestrian crashes in areas with high low-income or minority populations. Accident Analysis & Prevention, 2010; 42(6):1718–1728.
15. Miranda-Moreno LF, Morency P, El-Geneidy AM. The link between built environment, pedestrian activity and pedestrian—vehicle collision occurrence at signalized intersections. Accident Analysis & Prevention, 2011; 43(5):1624–1634.
16. Pulugurtha SS, Sambhara VR. Pedestrian crash estimation models for signalized intersections. Accident Analysis & Prevention, 2011; 43(1):439–446.
17. Ukkusuri S, Hasan S, Aziz H. Random parameter model used to explain effects of built-environment characteristics on pedestrian crash frequency. Transportation Research Record: Journal of the Transportation Research Board, 2011; (2237):98–106.
18. Abdel-Aty M, Lee J, Siddiqui C, Choi K. Geographical unit based analysis in the context of transportation safety planning. Transportation Research Part A: Policy and Practice, 2013; 49(0):62–75.
19. Wang Y, Kockelman KM. A Poisson-lognormal conditional-autoregressive model for multivariate spatial analysis of pedestrian crash counts across neighborhoods. Accident Analysis & Prevention, 2013; 60:71–84.
20. Lee J, Abdel-Aty M, Choi K, Huang H. Multi-level hot zone identification for pedestrian safety. Accident Analysis & Prevention, 2015; 76(0):64–73.
21. Xie K, Wang X, Ozbay K, Yang H. Crash frequency modeling for signalized intersections in a high-density urban road network. Analytic Methods in Accident Research, 2014; 2: 39–51.
22. Peters A, MacDonald H. Unlocking the Census with GIS. Redlands, CA: Esri Press, 2009.
23. Kim K, Brunner I, Yamashita E. Influence of land use, population, employment, and economic activity on accidents. Transportation Research Record: Journal of the Transportation Research Board, 2006; (1953):56–64.
24. Gladhill K, Monsere C. Exploring traffic safety and urban form in Portland, Oregon. Transportation Research Record: Journal of the Transportation Research Board, 2012; (2318):63–74.
25. Deacon JA, Zegeer CV, Deen RC. Identification of hazardous rural highway locations. Transportation Research Record, 1975; 543:16–33.
26. Barker J, Baguley C. A road safety good practice guide. In Proceedings of the Good Practice Conference. Bristol, UK, 2001.
27. Huang HL, Chin HC, Haque MM. Empirical evaluation of alternative approaches in identifying crash hot spots naive ranking, empirical Bayes, and full Bayes methods. Transportation Research Record, 2009; (2103):32–41.
28. Persaud B, Lan B, Lyon C, Bhim R. Comparison of empirical Bayes and full Bayes approaches for before-after road safety evaluations. Accident Analysis and Prevention, 2010; 42(1):38–43.
29. Hauer E. Observational Before/After Studies in Road Safety. Estimating the Effect of Highway and Traffic Engineering Measures on Road Safety. Bingley, UK: Emerald Publishing Ltd., 1997.

30. Hauer E. Identification of sites with promise. Transportation Research Record: Journal of the Transportation Research Board, 1996; (1542):54–60.
31. Elvik R. State-of-the-Art Approaches to Road Accident Black Spot Management and Safety Analysis of Road Networks. Oslo, Norway: Transportøkonomisk institutt, 2007.
32. Huang H, Chin H, Haque M. Empirical evaluation of alternative approaches in identifying crash hot spots: Naive ranking, empirical Bayes, and full Bayes methods. Transportation Research Record: Journal of the Transportation Research Board, 2009; (2103):32–41.
33. Miaou SP, Song JJ. Bayesian ranking of sites for engineering safety improvements: Decision parameter, treatability concept, statistical criterion, and spatial dependence. Accident Analysis and Prevention, 2005; 37(4):699–720.
34. Hauer E, Kononov J, Allery B, Griffith MS. Screening the road network for sites with promise. Transportation Research Record: Journal of the Transportation Research Board, 2002; (1784):27–32.
35. Persaud B, Lyon C, Nguyen T. Empirical Bayes procedure for ranking sites for safety investigation by potential for safety improvement. Transportation Research Record: Journal of the Transportation Research Board, 1999; (1665):7–12.
36. Miranda-Moreno L. Statistical models and methods for the identification of hazardous locations for safety improvements. Ph. D. Thesis, Department of Civil Engineering, University of Waterloo, 2006.
37. Moons E, Brijs T, Wets G. Identifying hazardous road locations: Hot spots versus hot zones. Pp. 288–300 in Gavrilova MI, Tan CJK (eds). Transactions on Computational Science VI. Berlin, Heidelberg: Springer, 2009.
38. Flahaut B, Mouchart M, San Martin E, Thomas I. The local spatial autocorrelation and the kernel method for identifying black zones: A comparative approach. Accident Analysis & Prevention, 2003; 35(6):991–1004.
39. Yu H, Liu P, Chen J, Wang H. Comparative analysis of the spatial analysis methods for hotspot identification. Accident Analysis & Prevention, 2014; 66:80–88.
40. Anderson TK. Kernel density estimation and k-means clustering to profile road accident hotspots. Accident Analysis & Prevention, 2009; 41(3):359–364.
41. Anselin L. Local indicators of spatial association—Lisa. Geographical Analysis, 1995; 27(2):93–115.
42. Wikipedia. City block [cited 2016 Oct. 24].
43. Johnston K, Ver Hoef JM, Krivoruchko K, Lucas N. Using ARCGIS Geostatistical Analyst Redlands, CA: Esri Press, 2001.
44. Wang X, Abdel-Aty M, Nevarez A, Santos J. Investigation of safety influence area for four-legged signalized intersections: Nationwide survey and empirical inquiry. Transportation Research Record: Journal of the Transportation Research Board, 2008; (2083):86–95.
45. Xie K, Ozbay K, Yang H, Holguín-Veras J, Morgul EF. Modeling the safety impacts of off-hour delivery programs in urban areas. Transportation Research Record: Journal of the Transportation Research Board, 2015; 4784.
46. National Safety Council. Estimating the costs of unintentional injuries, 2012. Available at: http://www.nsc.org/NSCDocuments_Corporate/Estimating-the-Costs-of-Unintentional-Injuries-2014.pdf.
47. Leskovec J, Rajaraman A, Ullman JD. Mining of Massive Datasets. Cambridge: Cambridge University Press, 2012.
48. Guttman A. R-trees: A dynamic index structure for spatial searching. In Proceedings of the 1984 ACM SIGMOD international conference on Management of data, Boston, MA, 1984.
49. Cloud AEC. Amazon web services. Retrieved November, 2011; 9:2011.
50. Kurkcu A, Morgul EF, Ozbay K. Extended implementation methodology for virtual sensors: Web-based real time transportation data collection and analysis for incident management. Transportation Research Record, 2015; 3374.
51. Tobin J. Estimation of relationships for limited dependent variables. Econometrica: Journal of the Econometric Society, 1958:24–36.
52. Anastasopoulos PC, Tarko AP, Mannering FL. Tobit analysis of vehicle accident rates on interstate highways. Accident Analysis & Prevention, 2008; 40(2):768–775.
53. Chen F, Ma X, Chen S. Refined-scale panel data crash rate analysis using random-effects tobit model. Accident Analysis & Prevention, 2014; 73:323–332.
54. Greene WH. Econometric Analysis. Pearson Education India, Hoboken, NJ: John Wiley & Sons, 2003.
55. Draper NR, Smith H. Applied Regression Analysis, 2nd ed. 1981.
56. Akaike H. A new look at the statistical model identification. IEEE Transactions on Automatic Control, 1974; 19(6):716–723.
57. Schwarz G. Estimating the dimension of a model. Annals of Statistics, 1978; 6(2):461–464.
58. Burnham KP, Anderson DR. Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach. New York: Springer-Verlag, 2002.
59. Kass RE, Raftery AE. Bayes factors. Journal of the American Statistical Association, 1995; 90(430):773–795.
60. O'Brien RM. A caution regarding rules of thumb for variance inflation factors. Quality & Quantity, 2007; 41(5):673–690.