Transportation Research Part A 85 (2016) 53–64

Contents lists available at ScienceDirect

Transportation Research Part A

journal homepage: www.elsevier.com/locate/tra

Improving strategic policies for pedestrian safety enhancement using classification tree modeling

Soyoung Jung [a,*], Xiao Qin [b,1], Cheol Oh [a,2]

[a] Department of Transportation and Logistics Engineering, Hanyang University Erica Campus, 55 Hanyangdaehak-ro, Sangnok-gu, Ansan 426-791, Republic of Korea
[b] Department of Civil and Environmental Engineering, University of Wisconsin-Milwaukee, NWQ4414, P.O. Box 784, Milwaukee, WI 53201, USA

ARTICLE INFO

Article history:
Received 21 January 2014
Received in revised form 21 October 2015
Accepted 4 January 2016
Available online 22 January 2016

Keywords:
Pedestrian safety enhancement
Injury severities
Classification tree
Contributing factors
Strategic policies

ABSTRACT

Pedestrian safety enhancement is a key component in reducing traffic fatalities in the Republic of Korea. The purpose of this study was to review, validate, specify, and prioritize Korea's strategic policies for pedestrian safety enhancement using the classification tree method to model pedestrian injury severities. The findings show that pedestrian age and movement type are the two primary variables contributing to pedestrian fatalities and severe injuries. Traffic operation, road class, crash location, driver violation, and at-fault vehicle type are all secondary variables associated with pedestrian fatalities and severe injuries. Factors that contributed to crashes were compared with strategic polices for senior zones and school zones, road safety facilities, safe walking environments, and legal obligations of the driver in order to understand why certain polices are ineffective versus effective. Consequently, this study provides prescriptive analysis and specific insights pertaining to strategic policies for pedestrian safety enhancements, which can be employed in other countries for the similar purpose. For further research, this study suggests combining several other data-mining techniques with nationwide data collection.

1. Introduction

In 2008, the road fatality rate per billion vehicle-kilometers in the Republic of Korea was 20.1, which is the highest road fatality rate among all countries in the Organization for Economic Cooperation and Development (OECD) (OECD International Transport Forum, 2010). In a recent annual report released by the OECD, Korea was ranked first among all OECD countries in road fatalities per billion vehicle-kilometers (OECD International Transport Forum, 2013). Korea has experienced substantial economic losses and significant casualties due to lingering road safety problems.

Generally, pedestrians have a high potential of sustaining severe injuries or being killed during a collision with a motor vehicle due to their physical fragility. In the previous decade, the majority of the counties that belong to the International Traffic Safety Data and Analysis Group (IRTAD) have experienced a 50% reduction in traffic fatalities in collisions that involved motor vehicle occupants. However, pedestrians have experienced 30% reduction in traffic fatalities within the same time frame because they are not protected by the safety features that were designed to safeguard vehicle occupants (OECD International Transport Forum, 2013).

* Corresponding author. Tel.: +82 31 400 4508; fax: +82 31 436 8147.
E-mail addresses: jung2@hanyang.ac.kr (S. Jung), qinx@uwm.edu (X. Qin), cheolo@hanyang.ac.kr (C. Oh).
[1] Tel.: +1 414 229 7399; fax: +1 414 229 6958.
[2] Tel.: +82 31 400 5158; fax: +82 31 436 8147.

http://dx.doi.org/10.1016/j.tra.2016.01.002
0965-8564/

Korea is experiencing a similar trend in pedestrian fatalities. As shown in Fig. 1, a steady number of fatalities due to vehicle–pedestrian crashes have occurred from 2007 to 2012. Pedestrian fatalities caused by traffic accidents account for 38% of all fatalities in Korea during the same time period (OECD International Transport Forum, 2013). Fifty percent of these pedestrian fatalities involve the elderly, who are particularly vulnerable (OECD International Transport Forum, 2014).

To deal with emerging pedestrian safety issues, the Korean government enacted pedestrian safety legislation that addresses the following items (Korea Ministry of Government Legislation, 2013):

- Pedestrians should walk on sidewalks where the walkway is separated from the driving area.
- Pedestrians should keep to the right on sidewalk.
- On roads in which the walkway is not separated from the driving area, pedestrians should walk along the roadside or the roadside zone.
- Pedestrians should cross roads at crossing facilities if they are installed.
- On roads in which crosswalks are not installed, pedestrians should cross using the shortest path across the road.
- Pedestrians should not cross a road where crossing is prohibited by traffic signs.
- Children less than six years of age should cross roads using safety protection.

In addition to this pedestrian safety legislation, the Korean government has increased efforts to enforce the new regulations. For example, a vehicle that violates a crosswalk stop line is subject to a fine and penalty points. Another example is the school zone speed limit in Korea, which is now less than 30 km/h. In addition, the Korean government began promoting a project that is known as the "cutting traffic fatalities in half" plan that was intended to reduce the annual number of fatal vehicle crash injuries by 10% every year from 2008 to 2012. The Korean government made pedestrian safety improvements a top policy-making priority to achieve its goals. Strategic policies for improving pedestrian safety have been implemented by the Ministry of Land and Transport and Maritime Affairs (MLTM). The most relevant policies are described as follows:

- Construct or repair road safety facilities:
  - Install medians, crash barriers and crosswalks on roads with high pedestrian crash occurrences due to jaywalking
  - Construct or repair crosswalk lighting systems, especially for nighttime use
  - Separate pedestrian walkways from driving areas
- Intensify driver legal obligations, such as maintaining a safe driving speed, prohibiting parking/stopping, or obeying traffic signals and any other traffic rules within pedestrian protection zones
- Promote a safe walking environment:
  - Construct and widen sidewalks
  - Flatten road surfaces
  - Create access roads for pedestrians, agricultural machines, and livestock in rural areas
- Designate senior/school zones and construct safety facilities within these zones (senior/school zones are defined by the Korea Welfare of the Aged Act and Road Traffic Act as the area within a radius of 300 m from certain facilities, such as residential, medical, and welfare centers, and elementary schools that are frequented by elderly people or children to protect them from potential traffic accidents.)
- Investigate causal factors for pedestrian-related crashes using data analysis

The United States' pedestrian safety records have been comparable to the pedestrian safety records in Korea. Americans have continually demanded safer methods of transportation, including walking, bicycling, and public transit. As a result, the United States Department of Transportation and the Federal Highway Administration (FHWA) have adopted an "integrated" approach to improve pedestrian safety, which involves engineering, enforcement, education and emergency services (FHWA Safety Program, 2013). For example, the FHWA's Office of Safety developed a strategic plan for pedestrian safety in 2010. The plan includes the following aspects: an evaluation of a community-focused approach to pedestrian safety; general pedestrian crash information; tools to diagnose and solve problems; education, outreach and a campaign for pedestrian safety; a pedestrian safety guide for transit agencies; legislation and guidelines; and pedestrian-safety-related research (FHWA safety program, 2013). Based on field experiments and data analysis, the FHWA recommended research-proven countermeasures, such as medians and pedestrian crossing islands, hybrid pedestrian beacons, and roadway reconfiguration (Zegeer et al., 2005; Fitzpatrick and Park, 2010; Persaud et al., 2010). As a result, pedestrian fatalities in the United States have declined. The average number of pedestrian fatalities between 2007 and 2011 accounts for 12.41% of all road fatalities in the United States, which is approximately a third of the road fatalities in Korea (Korean Statistics Information Service, 2013; National Highway Traffic Safety Administration, 2013).

Unlike pedestrian safety enhancement policies in the United States, Korea's strategic policies for improving pedestrian safety are not supported by data-driven analysis at either the national level or the local level. Few quantitative studies have been conducted to investigate vehicle–pedestrian crashes in Korea. The strategic policies in Korea failed to identify and target the key contributors of pedestrian–vehicle crashes prior to implementation and lack specifications for both pedestrian safety legislation and the MLTM's strategic policies.

The strategic policies for pedestrian safety enhancement can be modified for improved effectiveness but any changes should be based on a rigorous study. A thorough review of available safety data and data-driven approaches should be con-

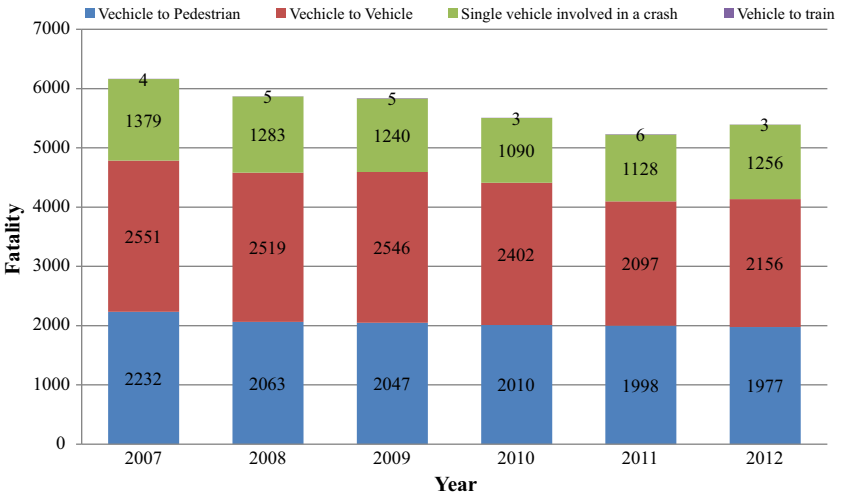

Fig. 1. Road fatalities in Korea by crash type (Korean Statistics Information Service, 2013). *Note.* Fatalities by vehicle-to-train crashes are low in number, as indicated by the data label on top of each bar.

ducted to strengthen the current policies. As a response, this study seeks to quantitatively explore the relationships among the variables that affect severe pedestrian injuries; verify the suitability of current pedestrian safety legislation and the MLTM's strategic policies; and identify and prioritize specific policy adjustments for pedestrian safety via classification tree modeling. Driven by safety data, the classification tree modeling approach was employed to examine three years of vehicle–pedestrian crashes that occurred in the Gyeonggi province of Korea.

2. Literature review

Previous studies have investigated the extensive array of factors that contribute to pedestrian safety. Griswold et al. applied a graphical cross-tabulation approach to explore the relationships among collision variables using fatal single-vehicle and pedestrian collision data from 1998 to 2007 (Griswold et al., 2011). The results revealed that the majority of fatal pedestrian collisions occur during twilight and the first hour of darkness and those weekly patterns vary by time of year due to the seasonal changes in sunset time. Research by Sullivan and Flannagan investigated these seasonal changes by focusing on how pedestrian fatalities are affected by the specific effect of the magnitude of darkness (Sullivan and Flannagan, 2001). The study discovered that pedestrian fatalities that occurred in dark conditions were related to posted speed limits and were significant on high-speed/limited-access roadways. The combination of high speed and limited sight distance multiplied the number pedestrian fatalities. Rifaat et al. investigated roadway characteristics and pedestrian safety using three years of pedestrian and bicycle crash data from the City of Calgary (Rifaat et al., 2011). The authors applied a multinomial logit model to estimate the effect of various factors on the severity of pedestrian and bicycle crashes and identified that the design of loops and lollipops in street patterns increased the probability of injury during a crash.

Researchers began considering the unobserved heterogeneities among explanatory variables, which enabled more insightful discoveries in pedestrian safety studies. Aziz et al. developed random parameter logit models to estimate pedestrian injury severity in New York City (Aziz et al., 2013). The authors revealed that road, traffic, and land use characteristics were statistically significant in affecting the severity of pedestrian injuries. A study by Wang et al. applied region-specific heterogeneity and correlation across response types via a Poisson-based multivariate conditional auto-regressive framework to examine the relationships in a three-year pedestrian crash count dataset from Austin, Texas (Wang and Kockelman, 2013). The results suggested that a greater mix of residential and commercial land use is associated with a higher potential for pedestrian crashes and that sidewalk provision is associated with lower rates of severe crashes. A study by Kim et al. analyzed the effect of several factors on pedestrian injury severity in pedestrian–vehicle crashes in North Carolina using a mixed logit model (Kim et al., 2010). In the study, drunk driving, speeding, and elderly pedestrians were found to increase the probability of fatal injuries in pedestrians. Eluru et al. proposed a generalized ordered response logit model to examine the effect of factors that contribute to non-motorist injury severity in traffic accidents in the United States. The study enabled flexibility in capturing the effect of the contributing factors (Eluru et al., 2008). The findings suggested that the most influential variables in non-motorist injury severity were the age of the individual, the speed limit, the location of the crash, and the time of day.

Recently, non-parametric models, such as the classification and regression tree (CART) model or clustering analysis, have become popular. Using the CART method, Chang and Wang examined the factors that affect injury severity levels in traffic accidents using one year of traffic accident data from the Taipei area (Chang and Wang, 2006). In their study, the probability that a pedestrian will incur injuries during traffic accidents was higher than the probability that a vehicle driver will incur injuries. Montella et al. focused on pedestrian crashes that occurred in Italy and identified contributing factors to fatal

crashes using a classification tree and association rules (Montella et al., 2011). In this study, fatal crashes were associated with the following factors: pedestrians older than 75 years of age, pedestrians older than 65 years of age, and crashes that occurred at night. Mohamed et al. also identified factors of fatal crashes and obtained results that were similar to the results of the Montella et al. study. The authors combined a cluster analysis with regression modeling to classify homogeneous accidents (Mohamed et al., 2013) and discovered that the following factors influence the likelihood of a fatal crash in New York City: older pedestrians, pedestrians under five years of age, and pedestrians that cross in the absence of a signal or crosswalk.

The clustering approach can also be combined with a traditional parametric method to model pedestrian–vehicle conflicts or crashes. A study by Thompson et al. employed clustering analysis and random effect regression models to evaluate how human factors impact the risk of intersection crossings (Thompson et al., 2013). Both road geometric types and human factors were considered. The study identified that distracting activities, such as text messaging, increase crossing times and are associated with the highest risk. A large amount of research has been conducted to investigate the factors that contribute to pedestrian safety risk. Risk is frequently defined as the probability that someone will be involved in a hazardous event and the severity of the event (Hakkert and Braimaister, 2002). Dai combined a spatio-temporal clustering technique with a traditional logistic regression to identify clusters of injured pedestrians and investigate the influence of personal and environmental factors on cluster-based injuries (Dai, 2012). The study revealed that the following factors significantly contribute to pedestrian injuries: suburban high-activity corridors, summertime, weekends, nighttime, age, pedestrian maneuvers, and inadequate lighting.

These findings should be applied to make more informed decisions and develop policies that better guide safety improvements. Clifton et al. employed a generalized ordered probit model to examine the impact of environmental factors on the severity of injuries that were sustained in pedestrian–vehicle crashes in Baltimore City (Clifton et al., 2009). The study showed that transit access and greater pedestrian connectivity are negatively associated with injury severity and should be properly considered when evaluating and planning for pedestrian safety. A study by Richmond et al. utilized a zero-inflated Poisson regression analysis to explore how a dedicated streetcar right-of-way (ROW) impacted pedestrian–vehicle collisions in Toronto (Richmond et al., 2014). The results suggested that streetcar ROW should be considered when developing policies for appropriate street environments because streetcars are a safer alternative for pedestrians compared with a mixed-traffic streetcar route. Chen et al. utilized a negative binomial model and a generalized estimating equation to evaluate the relative effectiveness of four signal-related pedestrian countermeasures on reducing pedestrian crashes in New York City (Chen et al., 2014). The study suggested that a longer cycle length be implemented at intersections with wide streets that are used by a higher percentage of elderly pedestrians.

Thus, previous studies have attempted to quantitatively identify the effect of factors on pedestrian safety with several analytical methods. However, few studies have employed the resultant findings in their analysis to review, validate, specify, and prioritize existing policy strategies for pedestrian safety enhancements, which is our research goal. We aim to accomplish this goal by providing specific and data-driven answers.

3. Data

This study employed Traffic Accident Analysis System (TAAS) of the Korea Road (KoRoad) Traffic Authority to collect data on vehicle–pedestrian crashes that occurred in Gyeonggi province from 2008 to 2010. According to the TAAS, the highest rate (more than 25%) of pedestrian injuries in Korea occurred in the Gyeonggi province (KoRoad Traffic Authority, 2013). Pedestrian injury level was the only outcome considered in this study. Property-damage-only (PDO) crashes were excluded to obtain more reliable results, because the number of PDO crashes in the dataset was less than 30.

All TAAS data fields were considered in this study. TAAS data contain a response variable with four injury levels and 21 explanatory variables. The number of crash observations is 21,904, which includes 1031 fatal injuries, 12,203 severe injuries, 8230 minor injuries, and 440 reported injuries. Table 1 lists the variable and descriptive statistics.

A large number of variables were evaluated. The majority of the variables retained their original format. For example, the posted speed limits in the crash dataset were treated as a numerical variable. The grade at each crash location involved three categorical levels: upgrade, downgrade, and flat.

To facilitate the analysis and comparison, a few numeric variables were transformed to categorical variables. The blood alcohol concentration (BAC) of the at-fault vehicle driver was classified by the Korean law enforcement levels. The pedestrian age was grouped into nine age groups, including children younger than six years of age and people who were 60 years of age or older. The hour-of-the-day variable was categorized into six time periods and a part of the time periods involved nighttime. These converted variables are consistent with the pedestrian policies, and the analysis results can be directly employed to measure the effectiveness of the targeted strategies.

4. Methods

4.1. Model selection

The variable association with pedestrian injuries and the variable importance is needed to validate and improve the current MLTM strategic policies. Although traditional statistical modeling approaches, such as ordered probability or a nominal

Table 1
Descriptive statistics for variables.

Variable	Category	Coding and description	Sample proportion (%)[a]
Pedestrian injury (response)	Fatality	1: Fatality within 30 days after crash occurrence	4.7
	Severe injury	2: Injury required more than 3 week hospital care	55.7
	Light injury	3: Injury required hospital care between 5 days and 3 weeks	37.6
	Injury report	4: Injury required less than 5 days of hospital care	2.0
1. Pedestrian age (years)		1: <6	2.4
		2: 6–13	13.9
		3: 14–19	9.0
		4: 20–29	12.2
		5: 30–39	12.1
		6: 40–49	16.7
		7: 50–59	13.3
		8: 60–69	10.3
		9: ⩾70	10.1
2. Pedestrian gender		1: Female	46.8
		2: Male	53.2
3. Pedestrian movement type		1: Walking along road edge	11.0
		2: Passing sidewalk	7.3
		3: Walking along a road	16.7
		4: Crossing crosswalk	65.0
4. At-fault indication		1: Pedestrian's fault	0.2
		2: Driver's fault	99.8
5. Driver age (years)		1: 20–29	18.3
		2: 30–39	23.7
		3: 40–49	30.6
		4: 50–59	19.7
		5: 60–69	6.4
		6: ⩾70	1.3
6. Driver gender		1: Female	19.4
		2: Male	80.6
7. Driver's driving experience (years)		1: No license	3.5
		2: <1	5.2
		3: 2–4	12.7
		4: 5–9	21.0
		5: 10–14	17.2
		6: ⩾15	40.4
8. At-fault driver's BAC (%)		1: Not under effect	94.3
		2: <0.05	0.1
		3: 0.05–0.09	1.3
		4: 0.10–0.14	1.9
		5: 0.15–0.19	1.5
		6: 0.20–0.24	0.6
		7: ⩾0.25	0.3
9.Weather at the crash time		1: Rain	8.4
		2: Snow	1.0
		3: Fog	0.3
		4: Cloudy	5.2
		5: Clear	85.1
10. Season at the crash time		1: Spring,	26.1
		2: Summer	24.7
		3: Autumn	26.7
		4: Winter	22.5
11. Day of the week		1: Weekdays	73.6
		2: Weekend	26.4
12. Hour of the day at the crash time (h)		1: Nighttime (0–5)	12.4
		2: AM peak (6–8)	10.2
		3: AM non-peak (9–12)	13.5
		4: PM non-peak (13–17)	26.3
		5: PM peak (18–19)	15.0
		6: Nighttime (20–23)	22.6

(*continued on next page*)

Table 1 (*continued*)

Variable	Category	Coding and description	Sample proportion (%)[a]
13. At-fault driver violation		1: Pedestrian's fault	0.2
		2: Speeding	0.3
		3: Improper intersection pass	0.4
		4: No pedestrian protection	12.5
		5: Illegal turning	0.6
		6: Signal violation	10.1
		7: Safe distance violation	0.2
		8: Failure in driving duty[b]	71.8
		9: Overtaking violation,	0.1
		10: Centerline violation,	1.1
		11: Other violation of driver regulation	2.7
14. At-fault vehicle type		1: Passenger car	66.3
		2: Bus	12.3
		3: Motorcycle	5.6
		4: Bicycle	3.2
		5: Truck	11.8
		6: Machine	0.8
15.Usage of at-fault vehicle		1: Commercial	22.8
		2: Non-commercial	77.2
16. Road functional class		1: Expressway	0.3
		2: National highway	24.8
		3: Rural principal road	13.7
		4: County road	1.7
		5: City road	59.5
17. Signal operation		1: Signalized road	34.4
		2: Non-signalized road	65.6
18.Type of crash location		1: At intersection,	12.2
		2: Near intersection	16.9
		3: On crosswalk	16.9
		4: Near crosswalk	3.9
		5: On single-route road	49.5
		6: Bridge or tunnel	0.6
19. Curve		1: Curve to the left	2.1
		2: Curve to the right	2.3
		3: Straight	95.6
20. Grade		1: Upgrade	4.8
		2: Downgrade	6.9
		3: Flat	88.3
21. Posted speed limit (km/h)		60–100 (Mean = 69, SD. = 11)[c]	–

[a] % indicates the proportion of sample size involved in each variable category to total sample size of 21,904.
[b] Failure in driving duty includes distracted and reckless driving as well as poor maneuvering.
[c] Standard deviation.

response model, can partly identify variable association with variable interaction terms, this parametric modeling approach has limited ability to express high-order interactions between explanatory variables and the rank of variable importance (Aziz et al., 2013; Chen et al., 2014; Clifton et al., 2009; Dai, 2012; Eluru et al., 2008; Kim et al., 2010; Richmond et al., 2014; Rifaat et al., 2011; Thompson et al., 2013).

A tree-based approach via the CART model, however, can capture non-additive behaviors (Chambers and Hastie, 1993), which offers the ability to highlight sophisticated relationships that are otherwise difficult to discover (Washington, 2000). The tree-based method is also well-suited for a case with a large number of discrete variables (Washington, 2000). Explanatory variable correlation problems are not problematic when the CART method is employed, and outliers are isolated into a node and eventually pruned with no effect on splitting. The method also has practical applications for large-scale data (Chang and Wang, 2006; Kashani and Mohaymany, 2011; Chang and Chien, 2013).

Although CART has several advantages, the model can be unstable and the resultant trees are various. The method may not adequately examine the marginal effect of the explanatory variables on the target response variable (Chang and Chien, 2013; Kashani and Mohaymany, 2011).

However, the variations in the resultant tree growth in this study were small, considering the prediction accuracy and the variable importance measures. The corresponding CART results in this study also revealed consistent rankings of the values of the variable importance measures. The variations in the resultant tree growth are quantitatively supported in Section 5. One of the objectives of this study is to identify the interactions among the factors that contribute to pedestrian injury

severities for developing strategic policy specifications. An examination of the marginal effect of the contributing factors was not a priority; therefore, the CART approach was selected.

4.2. Classification tree model

Relationships between variables can be identified during the classification process by splitting a large data set into more homogenous subsets to reveal patterns and relationships in the data. The classification tree begins from the topmost node, where statistical tests are run against every attribute in the data set. Each branch of the tree represents a test outcome, and each leaf node (or terminal node) contains a class label. The topmost node in the tree, which is the root node, shows the maximum reduction in variability for the response variable when the data are divided at the right value of the right explanatory variable. The greater is the distance from the root node, the higher is the order of interaction (Washington, 2000). After several iterations, the result is the terminal node, which is considered to be homogeneous or "pure."

Information gain and the Gini index were compared to identify the proper purity indicator. Although both variables returned consistent results in terms of the list of important explanatory variables and their relationships, the Gini index was selected because its validity has been proven in previous studies (Chang and Chien, 2013; Kashani and Mohaymany, 2011).

$$\mathrm{Gini}(t) = 1 - \sum_{k=1}^{k} P_{tk}^2 \tag{1}$$

where P_{tk} = the number of observations in class k divided by all observations in node t.

If all observations in node t belong to one class, Gini(t) is zero, which indicates the greatest possible purity. The tree grows from the root node, which contains all observations, until only similar observations exist at each terminal node. A cost-complexity algorithm is applied to "prune" the classification tree to prevent overfitting of the data (Han et al., 2012). The algorithm considers the cost complexity of a tree as a function of the number of leaves on the tree. The error rate of the tree is the percentage of misclassification in the dataset. The pruning process begins at the bottom of the tree, where cost complexity is computed at each internal node with and without the sub-tree. The sub-tree is pruned only if a smaller cost complexity is attained. The pruned tree is assessed by how well the classification tree structure generalizes to a given population.

A simple split-sample validation method was employed for its computational ease, its moderate sample size, and its proven validity (Chang and Chien, 2013; Kashani and Mohaymany, 2011; Han et al., 2012; Stewart, 1996). The data were randomly split into 70% and 30% for learning and testing, respectively. The learning samples were randomly selected by a computer. This practice was repeated many times, and a chi-squared test was additionally performed each time for both the learning samples and the testing samples to determine whether a significant difference in the pedestrian injury severity distributions existed.

One of the most useful outputs from CART is the independent variable importance measure (VIM) that was proposed and specified by Breiman et al. (1984) in Eq. (2).

$$\mathrm{VIM}(x_j) = \sum_{t=1}^{T} \frac{N_t}{N} \Delta\mathrm{Gini}(\mathrm{Sx}_j, t) \tag{2}$$

where T is the total number of nodes, N_t represents the observations in the dataset that belong to node t, N is the total number of observations, ΔGini (Sx_j, t) is the reduction in the Gini index at node t achieved by splitting variable x_j formulated by $\mathrm{Gini}(t) - \frac{NtR}{Nt}\mathrm{Gini}(t_R) - \frac{NtL}{Nt}\mathrm{Gini}(t_L)$, and N_{tR} and N_{tL} are the number of observations at the child node t_R and the number of observations at the child node t_L, respectively, from the parent node t. The VIM was produced for the variable that adds the reduction of the Gini index, or ΔGini, for each variable with a weight of the sample size at node t over the tree, and CART ranks each predictor variable according to its importance to the model.

5. Results and discussion

The likelihood ratio statistic of independence employed in this study reveals a large number of highly correlated explanatory variables, which is why this study used a classification tree method as opposed to a statistical modeling technique.

For pedestrian–vehicle collisions, the pedestrian injury was originally categorized as: fatal (1031 records), severe injury (12,203 records), light injury (8230 records), or injury report (440 records). The number of fatalities accounts for five percent or less of the total number of crashes, so are the injury reports. The disproportion of the number of fatalities or injury reports to the other categories suggests that the current pedestrian injury severity levels may not be ideal for developing a robust classification tree. For example, Strobl et al. stated that the classification trees may be biased toward the categories with more observations in a dependent variable if there is a wide variation in the number of observations among categories (Strobl et al., 2007). To mitigate the bias by selection, it is suggested to convert the response variable to a binary target variable class (Allwein et al., 2001; Delen et al., 2006; Dissanayake and Lu, 2002; Kashani and Mohaymany, 2011; Tax and Duin, 2002).

In this study, the injury severity was re-categorized as fatality/severe injuries versus minor injuries/injury report. This dichotomy classification emphasizes a distinctive feature of injuries with the potential of death. Furthermore, this new classification offered the highest overall prediction accuracies for the learning and testing samples in all response category classifications according (Aziz et al., 2013; Kashani and Mohaymany, 2011).

Fig. 2 shows the classification tree that was developed after applying the fatality/severe injuries versus minor injuries/injury report. In Fig. 2, the classification tree generated eight splitters and 14 terminal nodes. Note that the splitter is defined as a variable that produces the largest reduction in diversity for the dependent variable classification. The splitter variables are in bold in Fig. 2.

Terminal nodes, 3, 8, 13, 14, 15, 20, and 21 have sizable observations and higher potential of causing pedestrian fatalities or severe injuries than those for light injuries or injury reports, and were therefore more thoroughly examined as follows.

The tree was first split by pedestrian age, then by pedestrian movement. To the left, at node 3, the classification tree shows an 81.2% chance of pedestrian fatality or severe injury for pedestrians aged 50 or greater who are hit when crossing a crosswalk or passing a sidewalk. Further down the tree, at node 8, 76.2% of cases were likely to involve pedestrian fatalities or severe injuries if a pedestrian aged 70 or greater is hit when walking along a road or along the edge of a road. Road class appears to be related to pedestrian severe injuries. At node 14, fatal and severe injuries were more likely to be observed when a pedestrian aged 50–69 is hit when walking along the road/road edge of high-speed road (e.g. expressway, national highway, rural principal). The findings from the left side of tree imply that pedestrian age and movement type are factors that increase the potential of severe injuries in pedestrians. Accordingly, road safety measures such as traffic refuge islands or sidewalk barriers, walkway separation from driveway, and speed limit are particularly effective for improving elderly pedestrian safety.

On the right side of the tree, pedestrian age and movement type were the first two splitters, which is consistent with the first two splitters in the left. Pedestrian age and movement type interacted with other variables such as at-fault vehicle types, posted speed limit, signalization, crash location, and driver violation. At node 15, the classification tree shows a 66.7% potential for fatality or severe injury pedestrians aged 40–49 who are hit by a large vehicle such as a bus or a truck. Node 21 shows that when younger pedestrians aged 40 or less are hit by larger vehicles, the possibility of a fatal or severe injury is also attributed to at-fault vehicle driver violations such as speeding, signal violation, failure in duty, and overtaking in particular. Policy makers can therefore target those under the age 40 for education/campaign programs regarding safe crossing of roads and enhanced police enforcement at intersections or midblock crosswalks for reckless driving behavior.

When node 20 is compared with node 15, the potential for fatality and severe injuries in pedestrians aged 40–49 was associated with movement type, at-fault vehicle size, and speed limit, but in very different ways. Middle-aged pedestrian fatalities and severe injuries were associated with large-vehicles when crossing crosswalks, whereas the 14 or less age group was associated with speed limit during walking along the road or road edge and passing a sidewalk. At the terminal node 20, pedestrians aged 14 or less were associated with both walking along the road/road edge, passing a sidewalk, and a speed limit of 70 km/h or higher, which caused approximately 58.4% fatal and severe injury occurrences. Node 15 supports the need for educating middle-aged pedestrians on proper use of sidewalks and crosswalks. The finding at node 20 supports the current MLTM strategic policy for school zone designation and road safety facilities for walkway separation. Findings suggest enforcing a 30 km/h speed limit and installing a road safety facility (e.g. sidewalk crash barrier) within school zones.

In Table 2, the classification tree model for at least severe injury vs. minor injury showed approximately 62% and 61% in overall prediction accuracy for learning and testing samples, respectively. This predictive power of CART model depicted in Fig. 2 is comparable to past studies (Aziz et al., 2013; Kashani and Mohaymany, 2011). Specifically, the true positive rates were found to be higher than overall prediction accuracies in both learning and testing samples and the true negative rates were exceeded. The results imply that the CART model produced reasonable prediction power to correctly identify pedestrian fatalities and severe injuries.

The variable importance measure (VIM) further confirms the aforementioned circumstances that affect pedestrian fatalities and severe injuries identified in the classification tree growth. According to the rank of standardized VIM in Table 2, a comparatively higher VIM group includes pedestrian age, pedestrian movement, signal operation, posted speed limit, functional road class, crash location, at-fault driver violation, and at-fault vehicle type. The rest of the variables have a less than 5% of pedestrian age in VIM.

Variations shown in the resultant tree growth were small in this study considering prediction accuracy and variable importance measures. During many trials of CART development, total prediction accuracies ranged approximately from 59% to 62%, which is a very small interval. The corresponding CART results also showed consistent rankings in the values of variable importance measurement. As the most important variable group, the importance measurement value for pedestrian age was found to be the highest and the importance measurement value for pedestrian movement type was the second highest by approximately half of the pedestrian age importance in every trial of CART development. As the second most important variable group, signal operation, posted speed limit, road class were also commonly observed with importance measurement values ranging 10–20% compared to the importance measurement values for pedestrian age. Types of location, driver violation, at-fault vehicle type were also commonly observed as the third most important variable group with variable importance measurement values ranging 4–9% compared to the importance measurement values for pedestrian age.

Three key findings are presented in Table 2. First, VIM values for pedestrian age groups and movement types were remarkably large compared to VIM values for other variables. Since these two variables were observed as the first two splitters and were appeared repeatedly during tree growth, pedestrian age and pedestrians' movement type are the most influ-

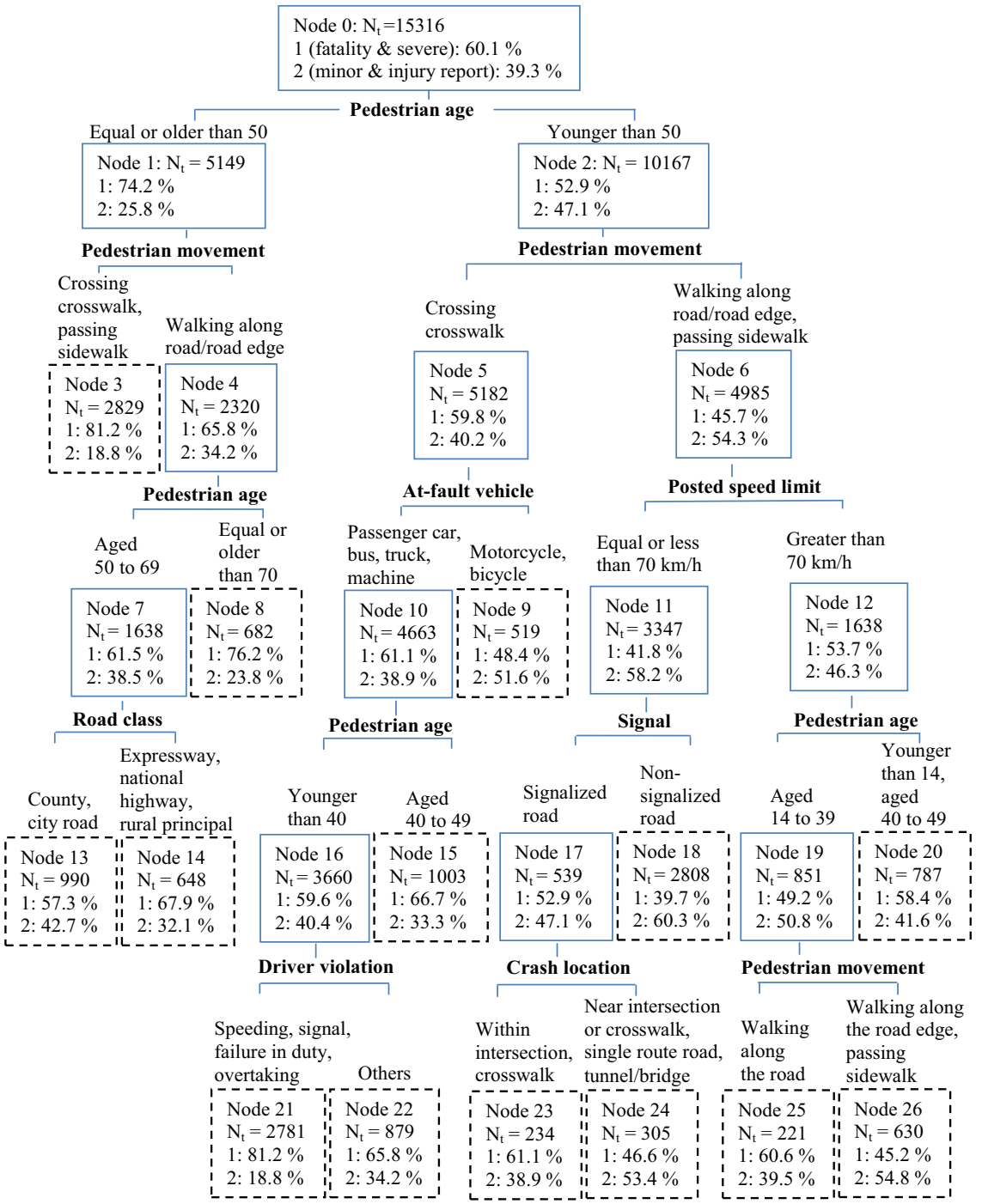

Fig. 2. Classification tree for two-level pedestrian injury severity. *Note.* Terminal nodes are presented with dashed box.

ential factors identifying pedestrian injury severity levels. Based on one of the study objectives, a certain pedestrian age group and pedestrian movement type should be considered top priorities for enhancing pedestrian safety when specifying additional strategies such as education program or road safety facility construction.

The second finding involves the critical role of traffic operation in enhancing pedestrian safety. Signal operation, speed limit, and road geometric characteristics such as road functional class or crash location, are variables with VIM values that are greater than 10%. Therefore, traffic operation and roadway geometry improvement strategies should be considered in conjunction with pedestrian age and pedestrian's movement type. Clear road design and appropriate road safety facilities

Table 2
Prediction performance and variable importance.

			Confusion matrix Predicted	
			Fatalities/severe injuries (1)	Light injuries/injury report (0)
Actual	Learning	1[a]	5903	3298
		0[b]	2495	3620
		Precision	70.3%	
		Recall	64.2%	
		Specificity	59.2%	
		Accuracy	62.2%	
	Testing	1	2497	1536
		0	1013	1542
		Precision	71.1%	
		Recall	62.0%	
		Specificity	60.4%	
		Accuracy	61.3%	

Variable	Standardized VIM [c] (%)	Variable	Standardized VIM (%)
Variable importance			
Pedestrian age	**100**	Weather	1.7
Pedestrian movement type	**47.7**	Pedestrian gender	1.6
Signal operation	**21.8**	Hour of the day	1.5
Posted speed limit	**16.1**	Driver age	1.2
Road class	**15.5**	Day of the week	0.6
Type of crash location	**10.1**	Curve	0.5
At-fault driver violation	**9.9**	Grade	0.3
At-fault vehicle type	**6.3**	At-fault indication	0.2
BAC	3.0	Season	0.2
Usage of at-fault vehicle	2.0	Driver gender	0.0
Driving experience	2.0		

[a] Fatalities/severe injuries.
[b] Light injuries/injury report.
[c] To compare the importance of each predictor between all predictors, the VIM was standardized compared with the highest value of variable importance. Variables with the standardized VIM in bold indicates splitters shown in classification tree.

should influence pedestrian behavior and protect them against motor vehicle crashes. Pedestrian crossing refuges and flash beacon along with marked crosswalk on high-speed highway facilities would be the relevant example.

Lastly, at-fault driver behavior and vehicle type were shown as splitters in Fig. 2 and as high VIM-value (i.e., greater than 5%) factors in Table 2. This finding suggests a need for more enforcement of safety policies, specifically with regard to commercial vehicle drivers in particular and reckless driving behavior. Nevertheless, a combination of two or more strategies should be considered for locations that are appropriate for a comprehensive pedestrian safety treatment.

6. Conclusions and future studies

The Korean government has made pedestrian safety a top policy-making priority since its 2008 campaign to reduce road fatalities. Despite the enactment of several strategic policies that are aimed enhancing pedestrian safety, the flat trend of pedestrian fatalities suggests the ineffectiveness of these policies. This study intends to investigate the causes of the high pedestrian severities to assist policy-making by comparing the current strategic policies with pedestrian safety data and to enhance the policy specifications based on data-driven results. This study is the first attempt to quantitatively review current pedestrian policies and provide directive insights into the strategic policies for pedestrian safety improvements in Korea. The main findings are summarized as follows:

- Pedestrian's age and pedestrian's movement type were identified as the most important factors, followed by the posted speed limit and the road functional class. Driver violation and at-fault vehicle type were ranked third.
- Fatalities and severe injuries for pedestrians who are at least 50 years of age were frequently observed in events such as crossing a crosswalk and walking along a sidewalk. Fatalities and severe injuries for elderly pedestrians who are at least 70 years old were closely associated with walking along roads/road edges.
- For pedestrians whose ages range from 50 to 69, fatalities and severe injuries were associated with walking along rural roads/road edges. Fatalities and severe injuries for pedestrians who are 13 years of age or younger were associated with a posted speed limit of 70 km/h or higher when walking along roads/road edges.
- The potential for severe injuries in pedestrians who are 40 years of age or younger was related to large vehicles and driver violations, such as speeding, signal violation, failure in driving duty, and overtaking when traversing a crosswalk.

- The potential for severe injury for pedestrians whose ages range from 40 to 49 was associated with at-fault vehicle types when the pedestrians cross a crosswalk. When the pedestrians walk along roads/road edges or walk on a sidewalk, the potential for severe injury was related to a posted speed limit of 70 km/h or higher.

Considering pedestrian age groups and movement types, the MLTM's current strategic policies, including the designation of senior/school zones, the construction/repair of road safety facilities, and the promotion of safe walking environments are reasonable and appropriate. The findings related to driver violations also support the current MLTM strategic policy of driver legal obligation intensification. However, the existing policies should be prioritized based on their urgency and importance. Implementation of the senior/school zone policy should be regarded as the top priority among all pedestrian safety polices. Current policies that promote safe walking environments and help pedestrians use crosswalks, pass sidewalks, walk along roads or road edges should also be considered top-priority. Strict law enforcement related to driver legal obligation should be considered third priority. These recommendations are based on the pedestrian safety data and statistical analysis.

Reflecting on the variable interactions that were identified in the CART model, the policies regarding the MLTM strategies are specified and ranked in order of variable importance:

1. Senior/school zones should be warranted in rural areas that have a high population of pedestrians or a significant number of pedestrian crashes.
2. Senior/school zones warrant additional or upgraded road facilities or safe walking environments, such as flash beacons, speed zones, sidewalk crash barriers, widened sidewalks, marked road surfaces, and pedestrian refuge areas for crossing.
3. Middle-aged pedestrians can benefit from education or campaigns regarding inappropriate movement types, specifically with regard to larger vehicles, walking along roads/road edges, and walking on sidewalks.
4. Pedestrians under 40 years of age and drivers of large vehicles are potential target groups for education with respect to safe crossing methods. Police can enforce and guide safer road crossing.
5. Traffic calming should combine a reduced speed limit and road safety facilities, such as road diets, narrower lane widths, tight curb radii, and/or extended curbs in school zones.
6. Drivers of large vehicles (e.g., buses or trucks) in particular should be educated to remind of their legal obligations with regard to pedestrian safety.
7. Certain driver violations, such as speeding, violation of signals, and improper performance of driving duties (e.g., overtaking near crosswalks) should be targeted with increased legal obligations and increased fines.

Future research that combines non-parametric data-mining with parametric statistical modeling approaches can identify the specific impact of each factor on injury severity and the relationships among the variables within a specific cohort of pedestrians. For example, the contributing factors of injury severity by pedestrian age or intersection type can be identified using cluster-based statistical modeling techniques to evaluate safety policy strategies. Pedestrian characteristics and behaviors were top ranked factors in the VIM, which suggests that human factors, such as age, height, and gender, can contribute to more specific pedestrian safety policies.

The data utilized in this study were collected from pedestrian–vehicle crashes in the Gyeonggi province of the Republic of Korea, which is an urban area that includes some rural sections. Future studies should expand this study using nationwide data analyses to consider additional circumstances, especially in rural areas, that may cause severe pedestrian injuries.

Acknowledgements

This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT and Future Planning (2013R1A1A3006898). This work was supported by a National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (NRF-2010-0029449). The database center at the Korea Road Traffic Authority provided the pedestrian–vehicle crash dataset.

References

Allwein, E., Schapire, R., Singer, Y., 2001. Reducing multiclass to binary: a unifying approach for margin classifiers. J. Mach. Learn. Res. 1, 113–141.
Aziz, H., Ukkusuri, S., Hasan, S., 2013. Exploring the determinants of pedestrian–vehicle crash severity in New York city. Accid. Anal. Prev. 50, 1298–1309.
Breiman, L., Friedman, J., Olshen, R., Stone, C., 1984. Classification and Regression Trees. Wardworth Inc., Belmont (CA).
Chambers, J.M., Hastie, T.J., 1993. Tree-based models. Statistical Model in S. Chapman & Hall, New York (Chapter 9).
Chang, L., Chien, J., 2013. Analysis of driver injury severity in truck-involved accidents using a non-parametric classification tree model. Saf. Sci. 51, 17–22.
Chang, L., Wang, H., 2006. Analysis of traffic injury severity: an application of non-parametric classification tree techniques. Accid. Anal. Prev. 38, 1019–1027.
Chen, L., Chen, C., Ewing, R., 2014. The relative effectiveness of signal related pedestrian countermeasures at urban intersections—lessons from a New York City case study. Transp. Policy 32, 69–78.
Clifton, K., Burnier, C., Akar, G., 2009. Severity of injury resulting from pedestrian–vehicle crashes: what can we learn from examining the built environment? Transp. Res. Part D 14, 425–436.
Dai, D., 2012. Identifying clusters and risk factors of injuries in pedestrian–vehicle crashes in a GIS environment. J. Transp. Geogr. 24, 206–214.
Delen, D., Sharda, R., Bessonov, M., 2006. Identifying significant predictors of injury severity in traffic accidents using a series of artificial neural network. Accid. Anal. Prev. 38 (3), 434–444.

Dissanayake, S., Lu, J., 2002. Factors influential in making in making an injury severity difference to older drivers involved in fixed object-passenger car crashes. Accid. Anal. Prev. 34 (5), 609–618.
Eluru, N., Bhat, C., Hensher, D., 2008. A mixed generalized ordered response model for examining pedestrian and bicyclist injury severity level in traffic crashes. Accid. Anal. Prev. 40, 1033–1054.
FHWA Safety Program, 2013. Pedestrian & Bicycle Safety <http://safety.fhwa.dot.gov/ped_bike/>.
Fitzpatrick, K., Park, E., 2010. Safety Effectiveness of the HAWK Pedestrian Crossing Treatment, FHWA Report No. FHWA-HRT-10-042.
Griswold, J., Fishbain, B., Washington, S., David, R., Ragland, D.R., 2011. Visual assessment of pedestrian crashes. Accid. Anal. Prev. 43, 301–306.
Hakkert, A.S., Braimaister, L., 2002. The Use of Exposure and Risk in Road Safety Study. SWOV Institute for Road Safety Research, Leidschendam, The Netherlands, R-2002-12.
Han, Jiawei, Kamber, M., Pei, J., 2012. Data Mining: Concepts and Techniques, third ed. Morgan Kaufmann, Waltham, MA, USA.
Kashani, A., Mohaymany, A., 2011. Analysis of the traffic injury severity on two-lane, two-way rural roads based on classification tree models. Saf. Sci. 49, 1314–1320.
Kim, J., Ulfarsson, G., Shankarc, V., Mannering, F., 2010. A note on modeling pedestrian-injury severity in motor-vehicle crashes with the mixed logit model. Accid. Anal. Prev. 42, 1751–1758.
Korea Ministry of Government Legislation, 2013. Korea Road Traffic Act <http://www.law.go.kr/lsEfInfoP.do?lsiSeq=140211#0000>.
Korean Statistics Information Service, 2013. Statistics Korea <http://kosis.kr/themes/themes_04 List.jsp>.
KoRoad Traffic Authority, 2013. Traffic Accident Analysis System <http://taas.koroad.or.kr/Eng/indexMain.jsp>.
Mohamed, M., Saunier, N., Miranda-Moreno, L., Ukkusuri, S., 2013. A clustering regression approach: a comprehensive injury severity analysis of pedestrian–vehicle crashes in New York, US and Montreal, Canada. Saf. Sci. 54, 27–37.
Montella, A., Aria, M., D'Ambrosio, A., Mauriello, F., 2011. Data-mining techniques for exploratory analysis of pedestrian crashes. Transport. Res. Rec.: J. Transport. Res. Board (2237), 107–116, Transportation Research Board of the National Academies, Washington, D.C..
National Highway Traffic Safety Administration, 2013. Fatality Analysis Reporting System Web-Based Encyclopedia <http://www-fars.nhtsa.dot.gov/People/PeoplePedestrians.aspx>.
OECD International Transport Forum (ITF), 2010. A Record Decade for Road Safety Press Release.
OECD International Transport Forum (ITF), 2013. Road Safety Annual Report.
OECD International Transport Forum (ITF), 2014. Road Safety Annual Report.
Persaud, B., Lana, B., Lyon, C., Bhim, R., 2010. Comparison of empirical Bayes and full Bayes approaches for before–after road safety evaluations. Accid. Anal. Prev. 42 (1), 38–43.
Richmond, S., Rothman, L., Buliungd, R., Schwartz, N., Larsend, K., Howard, A., 2014. Exploring the impact of a dedicated streetcar right-of-way on pedestrian motor vehicle collisions: a quasi experimental design. Accid. Anal. Prev. 71, 222–227.
Rifaat, S., Tay, R., Barros, A., 2011. Effect of street pattern on the severity of crashes involving vulnerable road users. Accid. Anal. Prev. 43, 276–283.
Stewart, J., 1996. Applications of classification and regression tree methods in roadway safety studies. Transport. Res. Rec.: J. Transport. Res. Board (1542), 1–5, Transportation Research Board of the National Academies, Washington, D.C..
Strobl, C., Boulesteix, A., Zeileis, A., Hothorn, T., 2007. Bias in random forest variable importance measures: illustration, Source and a Solution. BMC Bioinform. 8 (1), 20.
Sullivan, J. M., Flannagan, M.J., 2001. Characteristics of Pedestrian Risk in Darkness, Transportation Research Institute, The University of Michigan, Report No. UMTRI-2001-33.
Tax, D., Duin, R., 2002. Using two-class classifiers for multiclass classification. In: International Conference on Pattern Recognition, pp. 124–127.
Thompson, L., Rivara, F., Ayyagari, R., Ebel, B., 2013. Impact of social and technological distraction on pedestrian crossing behaviour: an observational study. Injury Prevent. 19, 232–237.
Wang, Y., Kockelman, K., 2013. A Poisson-lognormal conditional-autoregressive model for multivariate spatial analysis of pedestrian crash counts across neighborhoods. Accid. Anal. Prev. 60, 71–84.
Washington, S., 2000. Iteratively specified tree-based regression: theory and trip generation example. J. Transport. Eng. 126 (6), 482–491.
Zegeer, C.V., Stewart, J. R., Huang, H.H., Lagerwey, P.A., Feaganes, J., Campbell, B.J., 2005. Safety Effects of Marked versus Unmarked Crosswalks at Uncontrolled Locations: Final Report and Recommended Guidelines, FHWA Report No. FHWA-HRT-04-100.