Investigating the underreporting of pedestrian and bicycle crashes in and around university campuses − a crowdsourcing approach

Aditya Medury[a,*], Offer Grembek[b], Anastasia Loukaitou-Sideris[c], Kevan Shafizadeh[d]

[a] Safe Transportation Research and Education Center, University of California, Berkeley 2614, Dwight Way #7374, Berkeley, CA 94720, USA
[b] Safe Transportation Research and Education Center, University of California, Berkeley, CA, USA
[c] Department of Urban Planning, University of California, Los Angeles, CA, USA
[d] College of Engineering and Computer Science, California State University, Sacramento, CA, USA

ARTICLE INFO

ABSTRACT

In this paper, the non-motorized traffic safety concerns in and around three university campuses are evaluated by comparing police-reported crash data with traffic safety information sourced from the campus communities themselves. The crowdsourced traffic safety data comprise of both self-reported crashes as well as perceived hazardous locations. The results of the crash data analysis reveal that police-reported crashes underrepresent non-motorized safety concerns in and around the campus regions. The spatial distribution of police-reported crashes shows that police-reported crashes are predominantly unavailable inside the main campus areas, and the off-campus crashes over-represent automobile involvement. In comparison, the self-reported crash results report a wide variety of off-campus collisions not involving automobiles, while also highlighting the issue of high crash concentrations along campus boundaries. An assessment of the perceived hazardous locations (PHLs) reveals that high concentrations of such observations at/near a given location have statistically significant association with both survey-reported crashes as well as future police-reported crashes. Moreover, the results indicate the presence of a saturation point in the relationship between crashes and PHLs wherein beyond a certain limit, an increasing number of traffic safety concerns may not necessarily correlate with a proportional increase in the number of crashes. These findings suggests that augmenting our existing knowledge of traffic safety through crowdsourcing techniques can potentially help in better estimating both existing as well as emerging traffic safety concerns.

1. Introduction

Sustained multi-modal mobility patterns are often observed near university and college campuses where walking, bicycling and transit can provide a quick, inexpensive and flexible way of travel (Balsas, 2003). Studies also indicate that mode shares of non-motorized modes among campus communities are likely to be higher than in the general population (Nelson and Allen, 1997; Khattak et al., 2011; Buehler and Pucher, 2012; Schneider and Stefanich, 2015). However, the increased presence of pedestrians and bicyclists around campus environments also present several challenges for transportation practitioners. The vulnerable road users, i.e., pedestrians and bicyclists, are required to navigate through diverse types of built environments—ranging from a less motorized on-campus setting to the surrounding city's more motorized traffic. Additionally, campus-related movement is regulated and controlled by multiple planning and enforcement agencies, which makes it difficult to estimate the actual burden of transportation safety

as well as plan for improving it. These issues further compound the well-documented concerns of underreporting of pedestrian and bicycle crashes worldwide (Elvik and Mysen, 1999), wherein empirical evidence indicates that reporting a crash is less likely if there are no injuries, there is little property damage, or only one party is involved (Stutts and Hunter, 1998; Sciortino et al., 2005; Loo and Tsui, 2007).

This paper seeks to supplement police-reported pedestrian and bicycle crash data in a campus setting with traffic safety information obtained through crowdsourcing. Crowdsourcing is a mechanism to leverage information from a distributed network of individuals (Brabham, 2009). Also known as "citizen science," a subset of generalized crowdsourcing encourages the general public to help professional researchers expand the field of study through the broader collection of data (Goodchild, 2008). While the broad collection of scientific data by the general public has its drawbacks, namely in that it relies on anonymous, unidentified, untrained individuals (i.e., the "crowd") to voluntarily supply scientific data, many of the general
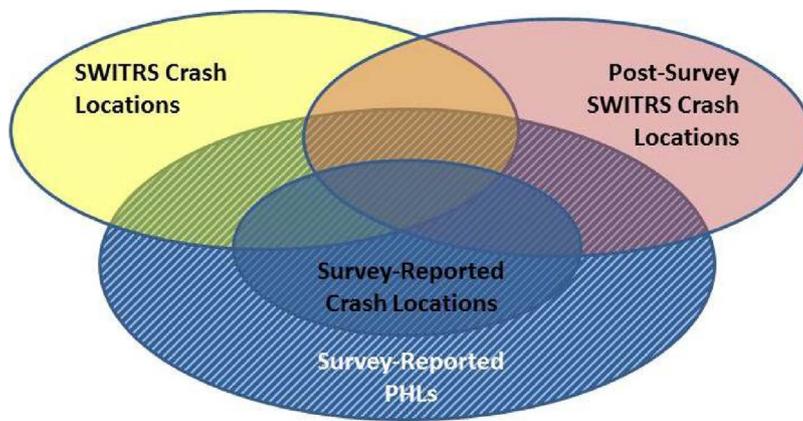
Fig. 1. A Venn diagram illustrating potential interactions between the different datasets.

merits and drawbacks associated with crowdsourcing have been identified and well documented by others (Howe, 2006; Doan et al., 2011; Saxton et al., 2013).

Despite its drawbacks, crowdsourcing creates new opportunities for data collection that may not have been feasible through more traditional data collection methods. Web-based mapping applications or GPS-enabled mobile devices provide a platform whereby large amounts of spatial and often temporal data can be collected. As a result, crowdsourcing is becoming increasingly used by engineers and planners to collect data in a wide range of transportation domains, such as bicycle mapping and route choice planning (Erickson, 2010; Krykewycz et al., 2012; Hood et al., 2011; Jestico et al., 2016), transit planning and operations analysis (Steinfeld et al., 2011), travel demand management for special events (Pereira et al., 2014), public participation in the transportation planning process (Misra et al., 2014), and pothole identification (City of Calgary City of Calgary-Pothole Repair, 2015). In the context of transportation safety, researchers can leverage information from the community at large by either directly soliciting information from the targeted demographics (young adults, pedestrians, etc.) (Begg et al., 1999; Arthur et al., 2001; Schneider et al., 2001; Transportation Alternatives, 2013), or monitoring reported incidents on social networks (Mai and Hranac, 2013; Schulz et al., 2013). In the context of university campuses and their neighboring regions, such an approach can be potentially effective as campuses have a focused, bounded territory, a well identified population that interacts with and knows the surrounding territory well, and can be easily notified through official campus contact information.

One of the objectives of this paper is to evaluate the different ways in which information about transportation safety can be requested and utilized. A commonly used option is to facilitate self-reporting of crashes, wherein a survey respondent can document the details of a crash they experienced in the past (Begg et al., 1999). Alternatively, the respondents can also be asked if they perceive certain locations to be hazardous to pedestrians and/or cyclists, and ascertain specific factors which lead them to feel so (New York City Vision Zero, 2017; Sanders, 2015; Blanc and Figliozzi, 2017). The former is more directly related to the overall goal of identifying under-reported crash data. In comparison, the latter has the advantage of not necessarily requiring respondents to have experienced a crash, but rather relying on their experiences as road users to document their grievances regarding traffic safety. Thus, it is likely that asking people about their safety perceptions may yield a larger volume of responses, which is increasingly being leveraged by transportation agencies working towards Vision Zero initiatives (New York City Vision Zero, 2017; Hinkle, 2017). However, whether these concerns correlate well with, not just historical crashes trends, but also emerging/underreported crash occurrences, requires further investigation.

In addition, while most campus-related traffic safety studies have focused on-campus environments (Schneider et al., 2001; Schneider et al., 2004), evaluating the traffic safety situation at the campus boundary as well as its adjoining areas is even more critical, considering that on-campus traffic conditions are typically designed to better cater to the needs of non-motorized traffic.

Thus, to investigate the strengths and shortcomings of crowdsourced data collection in a campus setting, the study focuses on three university campuses in California: University of California, Berkeley (UCB), University of California, Los Angeles (UCLA), and California State University, Sacramento (CSUS). At UCB, roughly 83% of faculty, staff, and students commute either by walking, bicycling, ride sharing, or public transit (Berkeley Parking and Transportation, 2013). At UCLA, 46% of all commutes involve non-automobile modes (UCLA Transportation, 2011). Finally, CSUS, a regional commuter campus, has around 79% of the campus population driving to campus (Shafizadeh, 2013). However, bicycle activity at the CSUS campus is also associated with the proximity of a recreational trail which can easily be accessed by a footbridge restricted to pedestrians and bicyclists.

2. Comparing crowdsourced and police-reported data

The emphasis of this work is in conducting a comparative analysis of traffic safety data obtained from crowdsourced versus police-reported sources. Fig. 1 illustrates the potential interactions between the different self-reported and police-reported datasets that are used in this study. SWITRS corresponds to California's Statewide Integrated Traffic Records System, which is a repository of all police-reported injury crashes in California, while crowdsourced safety information is documented via surveys through self-reported crashes and perceived hazardous locations (PHL) that capture traffic safety perceptions.

In Fig. 1, an overlapping region can be interpreted as two or more datasets documenting similar types of traffic safety concerns. When evaluating police-reported crash data across space and time, issues of systematic underreporting (such as in tribal areas (Bailey and Huft, 2008) as well as regression-to-mean (Abbess et al., 1981) have been well documented. However, from the perspective of traffic safety management, establishing overlaps, or lack thereof, between crowdsourced and police-reported crashes is an important pre-cursor to developing mechanisms for integrating these differently sourced crash databases. Similarly, it is imperative to better understand the extent of the association between traffic safety perceptions (i.e. PHLs) and self-reported/police-reported crash data, prior to incorporating them into traffic safety management practices. Thus, the objective of this analysis is to evaluate the ways in which survey-reported crashes and perceived hazardous locations can aid in better identifying existing as well as emerging traffic safety concerns.

2.1. Crash data collection: police records and crowdsourcing

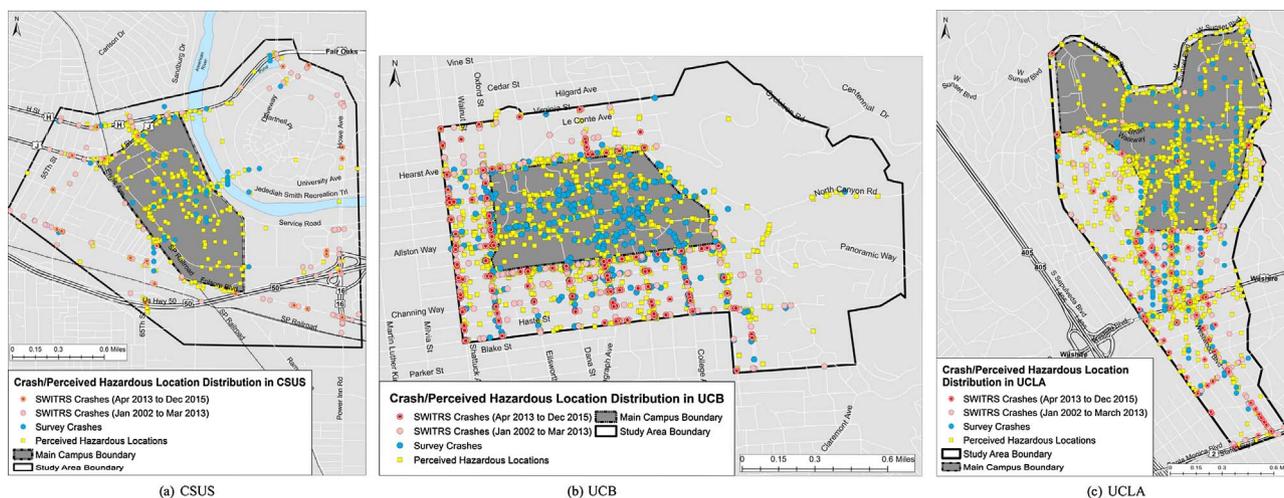The study used online surveys as a crash data collection technique,

Fig. 2. Maps of the three campuses indicating the locations of SWITRS and survey data.

which allowed respondents (students, faculty and staff) to self-report location information, crash narrative, and other attributes associated with non-motorized crashes in which they were involved. Since most stakeholders within a college community can be contacted via official campus email addresses, it can be argued that the barrier of using an online self-reporting mechanism such as in this study may be less formidable.

To assess pedestrian and bicycle safety at each campus, the study focused on the main campus area and a buffer region surrounding the campus, as shown in Fig. 2. An off-campus buffer was created to include neighboring areas, frequented by the campus community.

For the selected study areas, pedestrian and bicycle crashes were obtained through SWITRS for the period of January 2002 to March 2013. SWITRS yielded 143, 665, and 316 crashes for CSUS, UCB, and UCLA, respectively. The police-reported crash data included crashes of varying levels of injury severity—from fatal collisions to property damage only (PDO) crashes. However, it is important to note that local agencies in California are not required to document all PDO crashes (California Highway Patrol, 2012). More importantly, as can be seen in Fig. 2, SWITRS crashes are largely absent from the main campus area.

To supplement these records with more crash information, an online survey was developed and administered within each campus between February and March 2013 (Loukaitou-Sideris et al., 2014). Survey respondents were asked to share the following types of traffic safety concerns:

i. Crashes that they had experienced as, or with, bicyclists or pedestrians,

ii. Locations which they perceived to be hazardous for cyclists and/or pedestrians.

The web tool developed by the research team included a built-in Google Maps interface, which provided the respondents with the ability to mark the crash location on a map themselves. In addition, the survey respondents provided supplemental crash information such as time of day, injury severity, other modes involved, etc.

The self-reported crashes yielded 62, 346, and 217 pedestrian/bicycle crashes at CSUS, UCB and UCLA, respectively. In comparison, the questions regarding perceived hazardous locations yielded 477, 1769 and 2416 responses at CSUS, UCB and UCLA, respectively. The difference in the order of magnitude between the two types of survey-reported traffic safety concerns corroborates the a-priori hypothesis that a survey respondent is more likely to have experienced a larger set of hazardous events as a road user than having experienced a crash.

To assess the accuracy of the self-reported locations, observations picked at random were manually checked by comparing the marked locations with location-related descriptions entered by the survey respondents. These assessments revealed that the marked locations may have an error of 10–20 m. However, to alleviate the concerns vis-à-vis location accuracy, this study focuses on comparisons based on aggregated data.

Finally, to contrast the crash/perceived hazardous locations documented at the time of the survey with additional non-motorized crashes that have taken place since then, SWITRS data were collected from April 2013 to December 2015. The post-survey SWITRS crash dataset comprised of 27, 146 and 88 crashes for CSUS, UCB, and UCLA respectively.

2.2. Survey-reported crashes versus SWITRS crashes

Survey respondents were asked whether they had reported the crashes they were documenting in the survey to the campus or municipal police. The results revealed that more than 90% of the survey-reported crashes had not been reported to the police. When the respondents were asked to indicate the reasons for failing to report crashes to the authorities, the most common response across all crash types, for all modes, and on all campuses was that the crash was minor (72.6%), or the respondents did not believe that the police would do anything about it (25.8%).

Thus, given the low crash reporting rates, it can be hypothesized that the survey-reported crashes and SWITRS crashes available at the time of the survey may not have a strong overlap with each other.

2.2.1. Annual distribution

Fig. 3 shows the annual distribution of crashes reported in SWITRS and the surveys for the time period of January 2002 to March 2013, which reveals that the majority of the survey crashes took place within two-three years from the time of the survey. In addition, a significant percentage of the survey respondents did not remember the year when the crash occurred (7.9% at CSUS, 10.2% at UCB, and 14.9% of respondents at UCLA), suggesting that the recollection of specific crash details was challenging for some respondents. In comparison, the SWITRS records were more uniformly distributed across years.

2.2.2. On-campus versus off-campus presence

As Fig. 2 shows, police-reported crashes have a negligible on-campus presence. While the absence of on-campus crashes can be attributed to jurisdictional limitations of city police, the SWITRS database also contains separate jurisdictions for all California State Universities and University of California campuses. The absence of sufficient records within either city or campus jurisdictions indicates that campus-related crashes are not being entered into the state crash database. In
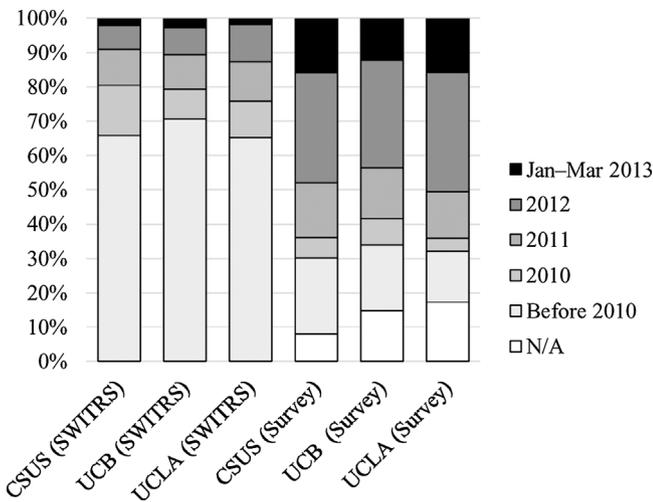
Fig. 3. Annual Distributions of SWITRS and Survey-Reported Crashes.

comparison, the survey-reported crashes included a substantial number of both on-campus and off-campus collisions.

2.2.3. Comparison of off-campus crash data

Since the crash dynamics inside the campuses are possibly different from off-campus locations and SWITRS contains negligible on-campus collisions, the two databases are further compared using only the off-campus crashes. In addition, considering the significant variation in the temporal distribution of the survey crashes, the comparative analysis of the two crash datasets is further limited to all off-campus crashes which took place between 2010 and 2013. Such a time period is also consistent with a typical four-year undergraduate cycle.

Table 1 compares the data sources across four categories: individual modes involved, mode combinations, injury severity levels, and the crash density around the campus boundary. Since the datasets have different sample sizes, a two-sample test of proportions was utilized to assess whether sample estimates of SWITRS and survey data for identical time periods can be interpreted to be similar or not (Washington et al., 2010). Herein, the null hypothesis states that the two sample proportions are equal. The two-sample test of proportions helps assess whether the distribution of a variable of interest within the survey and SWITRS crashes differs significantly from a pooled set of crash observations, which is reasonable considering that most of the self-reported crashes were not reported to the police. A significance level of 0.05 was chosen, and the attributes indicating a p-value lower than 0.05 are highlighted in bold.

Table 1 indicates that SWITRS crashes had a relatively higher representation of automobile-related collisions when compared to survey crash data at all three campuses. In comparison, the surveys captured a greater fraction of collisions that did not involve automobiles. For instance, the surveys captured a wide variety of bicycle-related collisions, which in some cases were higher not just in percentages but also in absolute numbers. A substantial number of survey-reported collisions also reported no injuries, which are not always required to be collect by police officials. Considering these differences in the crash attributes, it can be further argued that the data obtained from police- and survey-reported sources represent different parts of the overall crash spectrum of these locations.

The results shown in Table 1 also highlight some differences across the campuses. While the sample size of CSUS is limited, it can be observed that it has the highest shares of bicycle-related collisions across both SWITRS and survey databases, which can be explained by the bicycle trails in the adjoining areas of the campus. In comparison, UCLA has the highest automobile representation in the survey-reported crashes across all campuses, which is reflective of the auto-friendly

Table 1
Comparison of the collision attributes of SWITRS and survey records (number of collisions shown in parenthesis).

Crash Characteristics		CSUS Survey (Off-Campus)	CSUS SWITRS	CSUS P-Value	UCB Survey (Off-Campus)	UCB SWITRS	UCB P-Value	UCLA Survey (Off-Campus)	UCLA SWITRS	UCLA P-Value
Modes Involved	Ped	27% (4/15)	35% (17/49)	0.56	50% (62/125)	44% (86/196)	0.32	30% (24/81)	61% (67/110)	0
	Bicycle	93% (14/15)	65% (32/49)	**0.03**	76% (95/125)	59% (116/196)	0	73% (59/81)	39% (43/110)	0
	Auto	20% (3/15)	100% (49/49)	0	56% (70/125)	84% (164/196)	0	83% (67/81)	96% (106/110)	0
	Others	40% (6/15)	0% (0/49)	0	17% (21/125)	2% (3/196)	0	12% (10/81)	0% (0/110)	0.83
Mode combinations	Ped-Bicycle	27% (4/15)	0% (0/49)	0	22% (27/125)	3% (5/196)	**0.02**	1% (1/81)	1% (1/110)	0.83
	Ped-Auto	0% (0/15)	31% (15/49)	**0.01**	24% (30/125)	37% (72/196)	0.74	22% (18/81)	59% (65/110)	0
	Bicycle-Auto	20% (3/15)	63% (31/49)	0	40% (50/125)	42% (82/196)	0.46	54% (44/81)	36% (40/110)	**0.01**
	Bicycle-Other	27% (4/15)	0% (0/49)	0	17% (21/125)	14% (27/196)	**0.01**	12% (10/81)	1% (1/110)	0
	Bicycle-Bicycle	20% (3/15)	0% (0/49)	0	3% (4/125)	0% (0/196)	0.65	1% (1/81)	0% (0/110)	0.24
	Multiple modes	7% (1/15)	6% (3/49)	0.94	6% (7/125)	5% (10/196)	0.99	6% (5/81)	3% (3/110)	0.24
Injury Severity (Fatals excluded)	Severe	0% (0/13)	0% (0/45)	0.23	3% (4/125)	6% (11/195)	0	4% (3/81)	3% (3/110)	0.7
	Minor	47% (7/15)	89% (40/45)	0	53% (66/125)	91% (178/195)	0	52% (42/81)	91% (100/110)	0
	No Injury	53% (8/15)	2% (1/45)	0	42% (52/125)	3% (6/195)	0	44% (36/81)	6% (7/110)	0
Time of Day	Night	0% (0/13)	12% (6/49)	0.18	2% (3/123)	14% (27/196)	0	4% (3/79)	15% (17/110)	**0.01**
	Day	100% (13/13)	88% (43/49)	0.18	98% (120/123)	86% (169/196)	0	96% (76/79)	85% (93/110)	**0.01**
Crashes within 100 m of campus boundary		33% (5/15)	10% (5/49)	**0.03**	58% (72/125)	34% (67/196)	0	19% (15/81)	25% (28/110)	0.26
Crashes within 200 m of campus boundary		73% (11/15)	20% (10/49)	0	66% (82/125)	42% (83/196)	0	35% (28/81)	28% (31/110)	0.35

infrastructure of its off-campus regions. The multimodal interactions around the UCB campus appear to be the most complex with a wide variety of automobile- and non-automobile-related collisions featuring in both SWITRS and survey crashes, albeit in varying proportions.

A comparison of the time-of-day distributions indicates that survey-reported crashes over-represent daytime events relative to SWITRS crashes, wherein the time period corresponding to nighttime is between 11 PM and 7 AM for the survey crashes and midnight and 6 AM for the SWITRS crashes. This relative absence of nighttime crashes in the self-reported data could be potentially linked to campus-related activity being less prominent during nighttime. The time-of-day comparison also excludes two survey crashes in each of the campuses since the time-of-day information were not reported in those cases.

Finally, the survey crash data revealed a relatively high percentage of crashes taking place along the campus boundary at UCB and CSUS. The boundary represents the interface between the on- and off-campus transportation infrastructures. If the two areas have distinct traffic patterns, then the non-motorized traffic must be appropriately protected as it travels from one environment to another. In this case, both CSUS and UCB have restricted automobile access to the campus. UCLA, on the other hand, has thoroughfares providing access to different on-campus parking lots, which could explain the less significant impact at the campus boundary.

2.3. Perceived hazardous versus crash locations

The ratio of perceived hazardous locations to all self-reported crashes in the survey data is around 7.5, which jumps to over 20 when comparing PHLs to crashes reported between 2010 and 2013. Considering the vast difference in the sample sizes, it appears that (i) information regarding PHLs is easier to crowdsource than self-reporting crashes; and (ii) the community at large is evidently concerned about non-motorized safety in and around campus locations.

However, a pertinent concern about utilizing safety perceptions is whether high concentrations of perceived hazardous locations in a given region coincide with high concentrations of crashes. In particular, these concerns may be spatially correlated with locations where crashes have already taken place (as observed through historical survey and/or SWITRS crashes), and/or locations where crashes may occur in the future (as observed through the post-survey SWITRS crashes). Crash prediction in particular can be a desirable feature for policymakers, especially if a high concentration of PHLs is more likely to reveal an emerging traffic safety concern. Thus, the challenge, when interpreting perceived hazardous locations, is to recognize under what concentration thresholds these concerns can be reliably linked to crash occurrences.

To quantify the presence/absence of these associations better, a spatial overlap analysis was undertaken to ascertain if locations with PHLs also have crashes from police- and survey-reported datasets nearby. More specifically, in each campus, the off-campus PHL data was compared to 2009–2013 survey and SWITRS crash data, as well as the post-survey SWITRS crash data available up until end of 2015.

2.3.1. Spatial overlap analysis

To evaluate the spatial overlap of the different datasets at a comparable resolution, the analysis was undertaken as follows:

- Identifying a unit of observation: In order to consolidate points that are close to each other, a grid with cells of 0.001 by 0.001 ° (based on latitude/longitude co-ordinates) overlapping the study area was developed (Fig. 4). For purposes of comparison, a cell of 0.001 by 0.001 ° is equivalent to a region of 110 m × 110 m, which is comparable to other safety influence areas in traffic safety studies (Wang et al., 2008). Moreover, this level of aggregation makes the analysis more robust to any inaccuracies in the marked locations within the self-reported data, which may be around 10–20 meters.
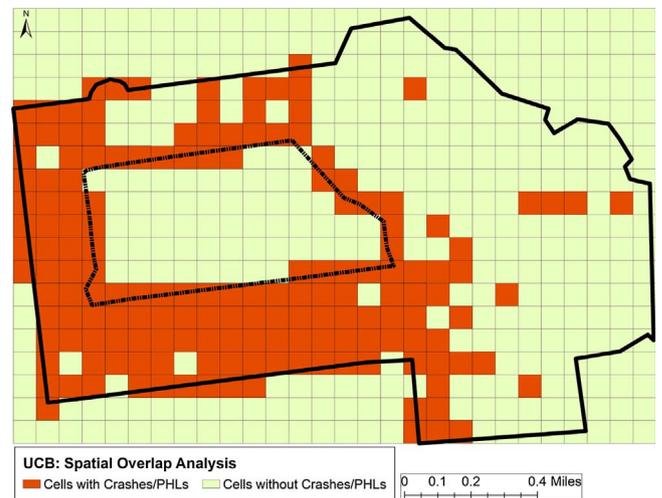

Fig. 4. Aggregating crash and PHL data to analyze spatial overlaps.

- Aggregating crash and PHL data: All points (PHLs, police- and self-reported crashes) located within each cell were aggregated. As Fig. 4 illustrates, some cells may not include even a single observation of any dataset, and were thus excluded from the remainder of the analysis.

- Comparing crash concentrations across different PHL thresholds: Based on the aggregation, a wide distribution of PHLs (ranging from 0 to 171), were observed within 164 and 199 cells in UCB and UCLA respectively. To evaluate how crash concentrations vary with increasing values of PHLs, a variety of different PHL thresholds were considered. For each PHL threshold, two types of groups were identified: cells with at least as many PHLs as the threshold, and cells with PHLs less than the threshold. In order to compare the corresponding crash concentrations between these groups, a Kolmogorov-Smirnov test at 95% confidence interval was conducted, to test whether groups with higher number of PHLs also had a higher number of crashes. The non-parametric Kolmogorov-Smirnov test was utilized to compare the aggregated crash distribution across the two groups, so as to accommodate skewed crash distributions.

2.3.2. Results

Figs. 5–7 plot the relationship between PHL thresholds and 2010–2013 survey, 2010–2013 SWITRS, and 2013–2015 SWITRS crash concentrations aggregated to the cell level. In each figure, the x-axis corresponds to the PHL threshold, the y-axis corresponds to the average number of crashes within cells that meet the PHL threshold, and the marker size is weighted by the number of cells that satisfy the PHL threshold. Finally, the marker colour indicates the results of the Kolmogorov-Smirnov test for the given PHL threshold.

The results of the spatial overlap analysis reveal that, for a wide range of PHL thresholds, cells that contain the requisite number of PHL observations also exhibit a higher concentration of crashes, when compared to the cells that do not meet the threshold. Moreover, as the PHL threshold increases, the average number of crashes within the cells that meet the threshold also increase with statistical significance. This result demonstrates that safety perception concentrations are likely to be positively correlated with actual crashes. The PHL thresholds that do not yield statistically significant differences in crash distribution are typically low PHL values (e.g., 1, 2), which satisfy cells with very few PHL observations, or high PHL thresholds (> 80) wherein the sample size of cells meeting the threshold is too small to yield any statistical inference.

Another promising finding emerging from the plots is that the association between PHLs and overlapping crashes appears to fade away

Survey-Reported Crashes vs PHLs
(Markers weighted by number of cells that satisfy a given PHL threshold)

UCLA: Survey-Reported Crashes vs PHLs
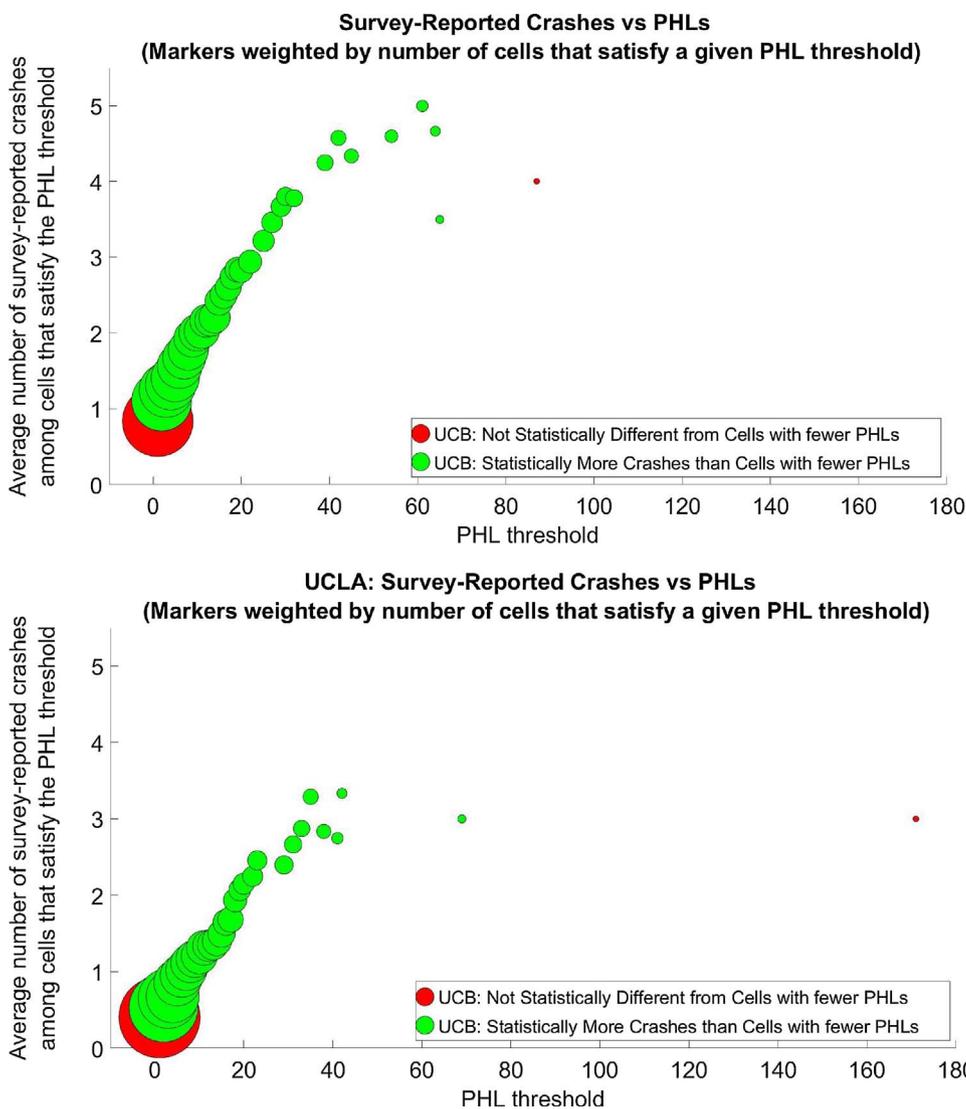**(Markers weighted by number of cells that satisfy a given PHL threshold)**

Fig. 5. Comparing survey-reported crash concentrations vs PHLs.

beyond a certain threshold, wherein a higher PHL concentration does not necessarily yield a commensurate increase in crash concentration. This plateauing trend could be a manifestation of smaller sample sizes associated with high PHL thresholds. However, it is likely that there may be a potential saturation effect in crowdsourced information, wherein beyond a certain point, more crowdsourced safety perception observations (in this case, PHLs) do not translate to the presence of more crashes. This point of saturation, that hints to an underlying piecewise-linear relationship, can vary for different campus-crash type combinations, ranging from 20 to 40 PHL thresholds. From a policy perspective, the saturation point can be a potential threshold choice for hot spot identification using PHLs, as it suggests a diminishing value of information from crowdsourced perception data.

Furthermore, when comparing the spatial overlap with the different crash datasets, the association between PHLs and survey crashes appears strongest, which reflects some internal consistency between the two related data sources. However, the positive association between PHL thresholds and crash concentration persists in the case of SWITRS datasets as well. This finding reveals that high concentrations of traffic safety perceptions are spatially well-correlated not just with historical crashes, but also under-reported as well as persistent/emerging crashes.

A low sample size of survey responses at CSUS (177 off-campus PHLs) made it impossible to successfully identify any statistically significant PHL thresholds. However, the comparisons between UCB and UCLA indicate a higher number of crashes per PHL threshold for UCB.

While UCLA received more PHL responses from the survey, the number of off-campus perceived hazardous locations reported at UCLA (1355) and UCB (1333) were similar. Thus, the consistent difference between the two campuses may be reflective of a higher number of self-reported/SWITRS crashes in UCB (as observed in Table 1).

3. Discussion

3.1. Strengths of crowdsourced data

The results shown in Figs. 5–7 indicate that a significant clustering of perceived hazardous locations can be utilized as a potential surrogate for underreported/emerging risk assessment of non-motorized modes. Considering that PHL data can be collected with much greater ease than self-reported crashes, the potential of systematically encouraging civic participation to proactively assess traffic safety concerns warrants further research.

The comparisons between survey- and police-reported crashes reveal that crowdsourcing can also help capture underreported crash types, such as pedestrian-bicycle, bicycle–bicycle and bicycle-only collisions. Even though a significant percentage of these collisions involved no injuries or property damage, the repeated occurrences of such events raise legitimate traffic safety concerns, and these underrepresented parts of the crash spectrum must also be accounted for in the planning and operations of multimodal traffic.
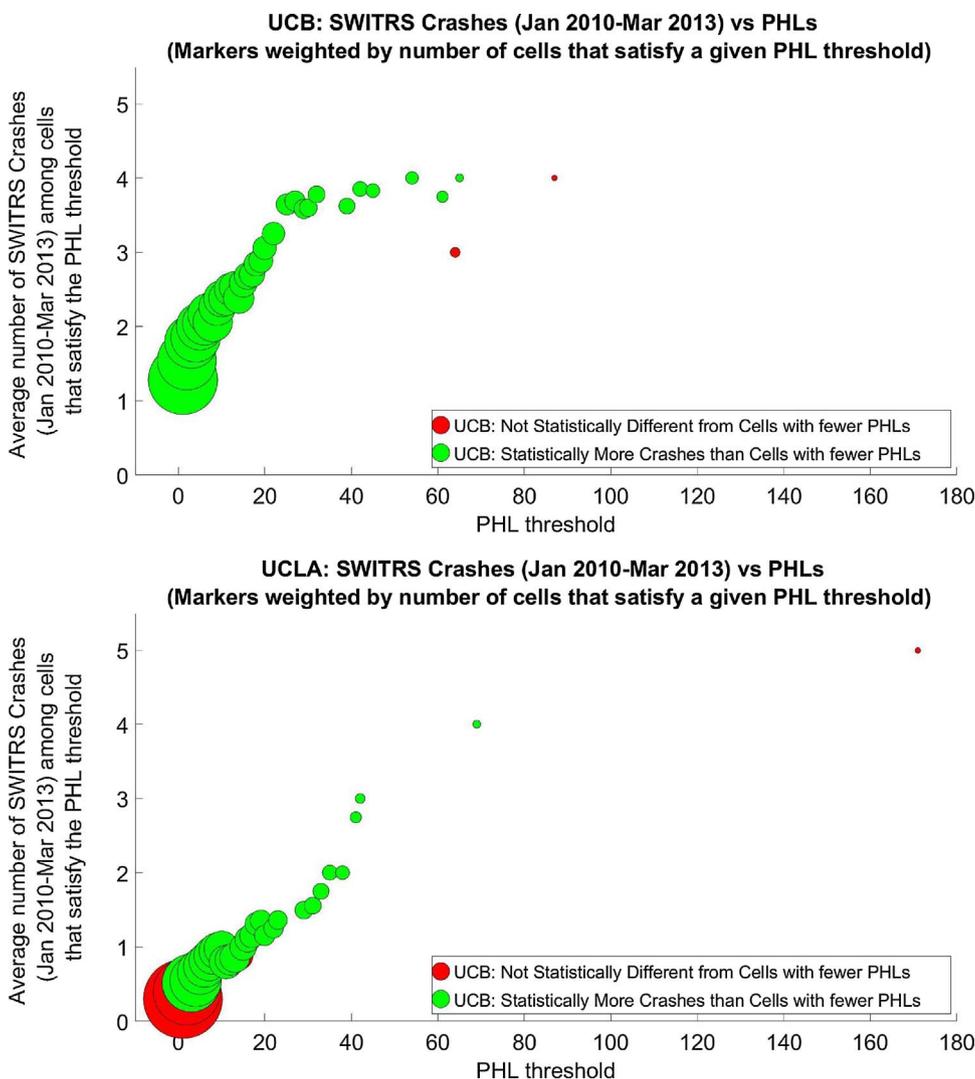
UCB: SWITRS Crashes (Jan 2010-Mar 2013) vs PHLs
(Markers weighted by number of cells that satisfy a given PHL threshold)

UCLA: SWITRS Crashes (Jan 2010-Mar 2013) vs PHLs
(Markers weighted by number of cells that satisfy a given PHL threshold)

Fig. 6. Comparing Jan 2010–Mar 2013 SWITRS crash concentrations vs PHLs.

More generally, crowdsourcing provides a flexible, open-ended, data collection framework, which can be customized to the needs of the study, while reaching out to the community directly, and not being hampered by jurisdictional and administrative considerations (campus versus city police). At the same time, crowdsourcing traffic safety data can be effective in bounded territories with a well-identified population of interest groups that knows the territory well, such as campuses, or tribal areas, especially in the absence of sufficient police-reported crash data.

3.2. Shortcomings of crowdsourced data

Crowdsourcing crash data suffers from limitations, which include recall bias (inability to recollect older details), participation bias (certain stakeholders having greater interest in the survey or access to the internet), and reporting bias (inability to verify the claims made in the surveys). However, it is possible to mitigate some of these concerns by collecting responses periodically, and in the long run, creating a dedicated crash reporting system that can facilitate continuous, instantaneous reporting of collisions from both within and beyond the campus community. These platforms can include mechanisms to improve the reliability of reported data by allowing users to import photos/videos (e.g., Blanc and Figliozzi, 2017), as well as cross-validate self-reported incidents (e.g., users up/down-voting incidents on the community-based traffic and navigation app, Waze).

Methodologically, self-reported crash data differ from police-reported data in one distinctive way—they are reported solely from the perspective of one of the parties involved in the collision. Consequently, the information given is not assessed by the other parties or by trained professionals. This introduces a reporting bias, which limits the insights that can be obtained from an individual incident. However, when large quantities of crowdsourced data are collected, trends can emerge and reveal valuable insights.

Finally, while the documentation of perceived hazardous locations is an inherently simpler process, its usefulness also relies on successfully obtaining large volumes of observations, which was a challenge at CSUS.

4. Summary

In this paper, the non-motorized traffic safety concerns in and around three university campuses are evaluated by comparing police-reported crash data with traffic safety information sourced from the campus communities themselves. The crowdsourced traffic safety data comprised of both self-reported crashes as well as perceived hazardous locations.

The results of the crash data analysis reveal that police-reported crashes underrepresent non-motorized safety concerns in and around the three campuses studied in this paper. The spatial distribution of police-reported crashes shows that police-reported crashes are predominantly unavailable inside the campus regions, and the off-campus crashes over-represent automobile involvement. In comparison, the
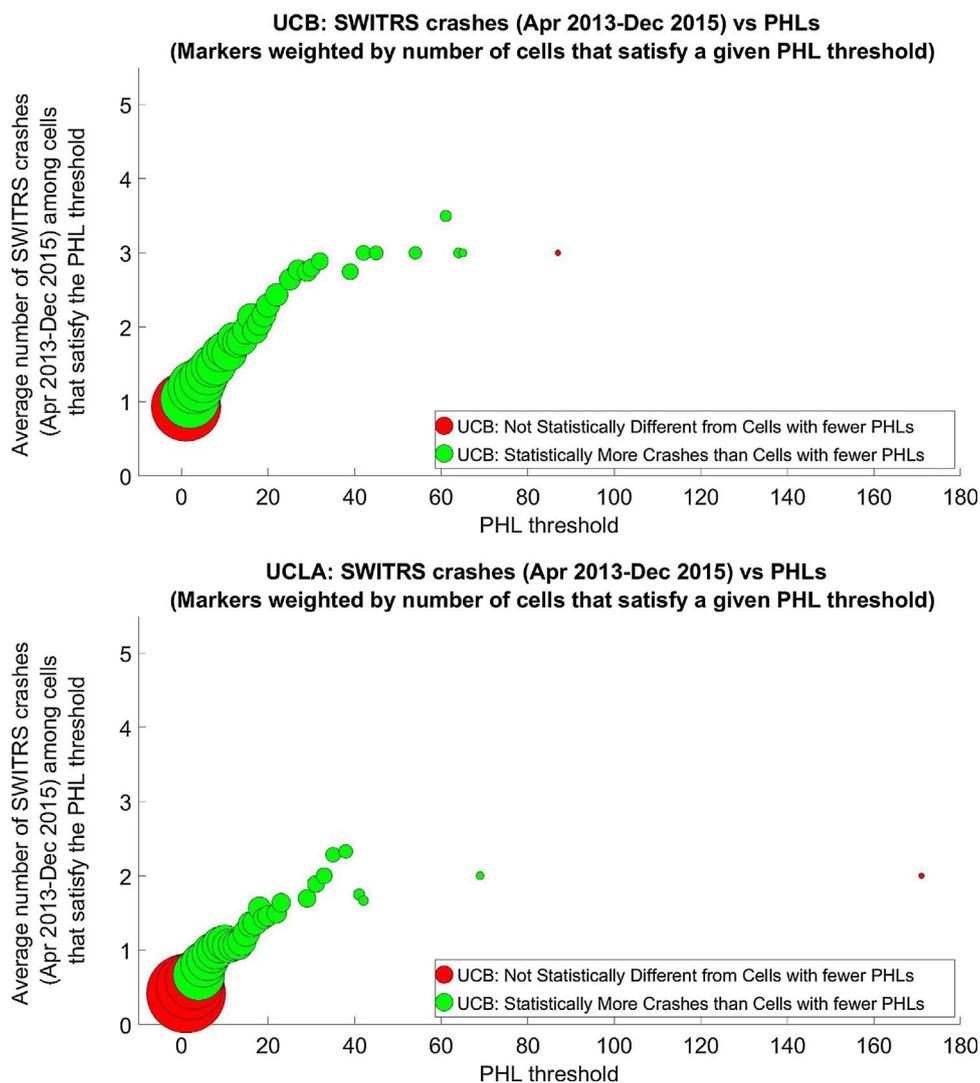
self-reported crash results report a wide variety of off-campus collisions not involving automobiles, while also highlighting the issue of high crash concentrations along campus boundaries. A comparison of the temporal factors indicates underrepresentation of less recent and nighttime crashes in the self-reported data, while an assessment of the injury severity distributions reveals that self-reported crashes report substantially more crashes with low/no injuries. These findings reveal that the self-reported crashes documented in this study may not have a significant overlap with the police-reported crashes. However, it is important to assess whether these trends persists in a long-term crowdsourcing effort of crash data.

An assessment of the perceived hazardous locations reveals that high concentrations of PHLs in a region have statistically significant association with higher self-reported as well as police-reported crashes. Moreover, the results indicate the presence of a saturation point in the relationship between crashes and PHLs wherein beyond a certain limit, an increasing number of traffic safety concerns may not necessarily correlate with a proportional increase in the number of crashes. The identification of such a breakpoint in the crowdsourced data may be helpful in gauging how long should long-term crowdsourcing efforts be, since beyond a certain point collecting more data may not be as informative. While this finding is still preliminary, it emphasizes the need for more research to understand the nuances of the relationship between traffic safety perceptions and crashes.

It is important to recognize that crowdsourced data cannot replace official crash records, that verify, document and analyze traffic incidents with much greater scrutiny. However, given the various jurisdictional, political, and resource-related constraints associated with the ecosystem of crash documentation, it is difficult for police-reported crashes to capture the wide spectrum of collisions occurring within an urban environment. At the same time, it is important for decision-makers to be cognizant of these limitations and take measures to fill in the data gaps. In this context, systematically collecting crowdsourced data can be both helpful and relatively inexpensive for planners and policymakers to gain a more complete understanding of non-motorized traffic safety concerns. For future research, a methodical procedure of fusing police-reported crashes, hospital data and crowdsourced information must be undertaken to enrich traffic safety studies and overcome the biases of both data sources.

Acknowledgment

References

Abbess, C., Jarrett, D., Wright, C.C., 1981. Accidents at blackspots: estimating the

effectiveness of remedial treatment, with special reference to the regression-to-mean effect. Traffic Eng. Cont. 22 (HS-032 693).

Arthur, W., Tubre, T., Day, E.A., Sheehan, M.K., Sanchez-Ku, M.L., Paul, D., Paulus, L., Archuleta, K., 2001. Motor vehicle crash involvement and moving violations: convergence of self-report and archival data. Human Factors: J. Human Factors Ergonom. Soc. 43 (1) 1–1.

Bailey, L., Huft, D., 2008. Improving crash reporting: study of crash reporting practice on nine Indian reservations. Transport. Res. Rec.: J. Transport. Res. Board 2078, 72–79.

Balsas, C.J., 2003. Sustainable transportation planning on college campuses. Transp. Policy 10 (1), 35–49.

Begg, D.J., Langley, J.D., Williams, S.M., 1999. Validity of self reported crashes and injuries in a longitudinal study of young adults. Inj. Prev. 5 (2), 142–144.

UC Berkeley Parking and Transportation, 2013. Sustainable Transportation URL 2017. date accessed: 2015-10-06. http://sustainability.berkeley.edu/transportation.

Blanc, B., Figliozzi, M., 2017. Safety Perceptions, Roadway Characteristics, and Cyclists' Demographics: A Study of Crowdsourced Smartphone Bicycle Safety Data.

Brabham, D., 2009. Crowdsourcing the public participation process for planning projects. Plann. Theory 8 (August (3)), 242–262.

Buehler, R., Pucher, J., 2012. Cycling to work in 90 large American cities: new evidence on the role of bike paths and lanes. Transportation 39 (2), 409–432.

California Highway Patrol. 2012 SWITRS Part 2: Property Damage Only Collisions — Preface. URL: http://www.chp.ca.gov/switrs/pdf/2012_ PDO_Preface.pdf, last accessed: 2015-02- 03.

City of Calgary City of Calgary-Pothole Repair, URL: http://www.calgary.ca/ Transportation/Roads/Pages/Road-Maintenance/Pothole-repair.aspx; 2015; date accessed: 10/09/2015.

Doan, A., Ramakrishnan, R., Halevy, A., 2011. Crowdsourcing systems on the world-Wide web. Commun. ACM 54 (4), 86–96.

Elvik, R., Mysen, A., 1999. Incomplete accident reporting: meta-analysis of studies made in 13 countries. Transp. Res. Rec.: J. Transport. Res. Board 1665, 133–140.

Erickson, T., 2010. Geocentric Crowdsourcing and Smarter Cities: Enabling Urban Intelligence in Cities and Regions. In 1st International Workshop on Ubiquitous Crowdsourcing. UbiComp, Copenhagen, Denmark.

Goodchild, M., 2008. Assertion and authority: the science of user-Generated geographic ContenEarth. Earth 2008t, 1–18.

Hinkle, N. NearKilled.Me. URL: https://nearlykilled.me/about.html last accessed: 2017-03-29.

Hood, J., Sall, E., Charlton, B., 2011. A GPS-based bicycle route choice model for San Francisco, California. Transport. Lett.: Int. J. Transport. Res. 3, 63–75.

Howe, J., 2006. The Rise of Crowdsourcing. Wired, Conde Nast Digital, No. 14 .06 June.

Jestico, B., Nelson, T., Winters, M., 2016. Mapping ridership using crowdsourced cycling data. J. Transp. Geography. 52, 90–97.

Khattak, A., Wang, X., Son, S., et al., 2011. Travel by university students in Virginia: is this travel different from travel by the general population. Transport. Res. Rec.: J. Transport. Res. Board 2255, 137–145.

Krykewycz, G., Pollard, C., Canzoneri, N., He, E., 2012. Web-based crowdsourcing approach to improve areawide bikeability scoring. Transport. Res. Rec. 2011 1–7. http://dx.doi.org/10.3141/2245-01.

Loo, B.P., Tsui, K., 2007. Factors affecting the likelihood of reporting road crashes resulting in medical treatment to the police. Inj. Prev. 13 (3), 186–189.

Loukaitou-Sideris, A., Medury, A., Fink, C., Grembek, O., Shafizadeh, K., Wong, N., Orrick, P., 2014. Crashes on and near college campuses: a comparative analysis of pedestrian and bicyclist safety. J. Am. Plann. Assoc. 80 (3), 198–217.

Mai, E., Hranac, R., 2013. In: Twitter Interactions as a Data Source for Transportation Incidents. InProc. Transportation Research Board 92nd Annual Meeting. (No. 13–1636).

Misra, A., Gooze, A., Watkins, K., Asad, M., Le Dantec, C., 2014. Crowdsourcing and its application to transportation data collection and management. Transp. Res. Rec. 2414, 1–8. http://dx.doi.org/10.3141/2414-01.

Nelson, A., Allen, D., 1997. If you build them, commuters will use them: association between bicycle facilities and bicycle commuting. Transport. Res. Rec.: J. Transport. Res. Board 1578, 79–83.

New York City Vision Zero, Share your Safety Issue. URL: http://www1.nyc.gov/assets/ visionzero/maps-data/vz-input-map.shtml. last accessed: 2017-03-29.

Pereira, F.C., Rodrigues, F., Ben-Akiva, M., 2014. Using data from the web to predict public transport arrivals under special events scenarios. J. Intelli. Transport. Syst. 19 (3), 1–16.

Sanders, R.L., 2015. Perceived traffic risk for cyclists: the impact of near miss and collision experiences. Acc. Anal. Preven. 75, 26–34.

Saxton, G.D., Oh, O., Kishore, R., 2013. Rules of crowdsourcing: models, issues, and systems of control. Infor. Sys. Manage. 30:1, 2–20. http://dx.doi.org/10.1080/ 10580530.2013.739883.

Schneider, R.J., Stefanich, J., 2015. In: Neighborhood Characteristics That Support Bicycle Commuting: Analysis of the Top 100 US Census Tracts. Transportation Research Board 94th Annual Meeting.

Schneider, R., Khattak, A., Zegeer, C., 2001. Method of improving pedestrian safety proactively with geographic information systems: example from a college campus. Transport. Res. Rec. 1773, 97–107.

Schneider, R.J., Ryznar, R.M., Khattak, A.J., 2004. An accident waiting to happen: a spatial approach to proactive pedestrian planning. Acc. Anal. Preven. 36 (2), 193–211.

Schulz, A., Ristoski, P., Paulheim, H., 2013. I See a Car Crash: Real-time Detection of Small Scale Incidents in Microblogs. The Semantic Web: ESWC Satellite Events. Springerpp. 22–33.

Sciortino, S., Vassar, M., Radetsky, M., et al., 2005. San Francisco pedestrian injury surveillance: mapping, under-reporting, and injury severity in police and hospital records. Acc. Anal. Preven. 37 (6), 1102–1113.

Shafizadeh, K., 2013. Sacramento State Campus Intercept Survey: 2007-present. 2013; date accessed: 2015-10-06. URL:http://www.csus.edu/indiv/s/shafizadehk/CSUS_ Intercept_Survey_mode_share.html.

Steinfeld, A., Zimmerman, J., Tomasic, A., Yoo, D., Aziz, R., 2011. Mobile transit information from universal design and crowdsourcing. Transp. Res. Rec. 95–102. http://dx.doi.org/10.3141/2217-12.

Stutts, J.C., Hunter, W.W., 1998. Police reporting of pedestrians and bicyclists treated in hospital emergency rooms. Transp. Res. Rec.: J. Transp. Res. Board 1998 (1635), 88–92.

Transportation Alternatives., Crashstories NYC. URL:http://www. crashstories. org/ date accessed 2014–06 - 2013; 22.

UCLA Transportation, 2011. State of the Commute Report. 2011; date accessed: 2015-10-06. URL: http://www.transportation.ucla.edu/portal/pdf/ 2011StateoftheCommuteReport.pdf.

Wang, X., Abdel-Aty, M., Nevarez, A., Santos, J., 2008. Investigation of safety influence area for four-legged signalized intersections: nationwide survey and empirical inquiry. Transp. Res. Rec. 2083, 86–95.

Washington, S., Karlaftis, M., Mannering, F., 2010. Statistical and Econometric Methods for Transportation Data Analysis. CRC press.