Accident Analysis and Prevention 66 (2014) 136–146

Contents lists available at ScienceDirect

Accident Analysis and Prevention

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

Applying quantile regression for modeling equivalent property damage only crashes to identify accident blackspots

Simon Washington[a,1], Md. Mazharul Haque[b,*], Jutaek Oh[c,2], Dongmin Lee[c,3]

[a] Civil Engineering and Built Environment, Science and Engineering Faculty and Centre for Accident Research and Road Safety (CARRS-Q), Faculty of Health, Queensland University of Technology, 2 George Street, GPO Box 2434, Brisbane, QLD 4001, Australia
[b] Centre for Accident Research and Road Safety (CARRS-Q), Faculty of Health and Civil Engineering and Built Environment, Science and Engineering Faculty, Queensland University of Technology, 130 Victoria Park Road, Kelvin Grove, QLD 4059, Australia
[c] The Korea Transport Institute, 2311 Daehwa-dong, Ilsanseo-gu, Goyang-si, Gyeonggi-do 411-701, Republic of Korea

ARTICLE INFO

Article history:
Received 28 August 2013
Received in revised form 10 January 2014
Accepted 10 January 2014

Keywords:
Hot spots
Network screening
Non-parametric models
Quantile regression
Empirical Bayes'
Negative binomial regression

ABSTRACT

Hot spot identification (HSID) aims to identify potential sites—roadway segments, intersections, crosswalks, interchanges, ramps, etc.—with disproportionately high crash risk relative to similar sites. An inefficient HSID methodology might result in either identifying a safe site as high risk (false positive) or a high risk site as safe (false negative), and consequently lead to the misuse the available public funds, to poor investment decisions, and to inefficient risk management practice. Current HSID methods suffer from issues like underreporting of minor injury and property damage only (PDO) crashes, challenges of accounting for crash severity into the methodology, and selection of a proper safety performance function to model crash data that is often heavily skewed by a preponderance of zeros. Addressing these challenges, this paper proposes a combination of a PDO equivalency calculation and quantile regression technique to identify hot spots in a transportation network. In particular, issues related to underreporting and crash severity are tackled by incorporating equivalent PDO crashes, whilst the concerns related to the non-count nature of equivalent PDO crashes and the skewness of crash data are addressed by the non-parametric quantile regression technique. The proposed method identifies covariate effects on various quantiles of a population, rather than the population mean like most methods in practice, which more closely corresponds with how black spots are identified in practice. The proposed methodology is illustrated using rural road segment data from Korea and compared against the traditional EB method with negative binomial regression. Application of a quantile regression model on equivalent PDO crashes enables identification of a set of high-risk sites that reflect the true safety costs to the society, simultaneously reduces the influence of under-reported PDO and minor injury crashes, and overcomes the limitation of traditional NB model in dealing with preponderance of zeros problem or right skewed dataset.

1. Introduction

Once operational, the transportation system (consisting of road segments, intersections, ramps, interchanges, etc.) does not perform homogenously with respect to safety due to both random and systematic influences. Not surprisingly, heterogeneity in the driving population, roadside features, weather, traffic conditions, driver behavior, and design features leads to heterogeneity in crash frequencies. Because of a desire and mandates to provide a safe driving environment, professionals are charged with identifying and improving "high risk" locations. Once potential high risk sites are identified—say the top 10% of all urban intersections in a city—safety engineers conduct safety audits of the sites to identify and rectify operational or geometric deficiencies.

There is a fairly extensive literature focused on methods for the identification of "black spots", "high risk sites", "sites with promise", or "hotspots" (HSID). The term "network screening" is also synonymous with HSID. A variety of methods have been proposed, presented, and applied. Previous research (e.g., Hauer, 1997; Persaud, 1986, 1988; Persaud and Hauer, 1984) reported that methods relying on a simple ranking of crash counts or crash rates, due to the random fluctuation of crashes from year to year, can produce large number of false positives (safe sites falsely identified as unsafe) and false negatives (truly hazardous sites escape

* Corresponding author. Tel.: +61 7 3138 4511.
E-mail addresses: simon.washington@qut.edu.au (S. Washington), m1.haque@qut.edu.au, mmh@alumni.nus.edu.sg (M. M. Haque), jutaek@koti.re.kr (J. Oh), dmlee@koti.re.kr (D. Lee).
[1] Tel.: +61 7 3138 9990.
[2] Tel.: +82 31 910 3174.
[3] Tel.: +82 31 910 3011.

0001-4575/$ – see front matter © 2014 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.aap.2014.01.007

identification). These errors result in inefficient use of federal and/or state aid and local government resources applied for safety improvements.

A discussion in the literature regarding the continuous nature of crash risk across sites and the tradeoffs between false negatives and positives is also conspicuously short. The most "risky" *X*% of sites is typically determined by available resources and mandated safety management practices. These *X*% of sites are marginally 'less safe' than the sites just below the mostly artificial crash rate or frequency threshold. The difference between "safe" and "risky" sites is not deterministic, i.e., the presence or absence of a particular feature. Thus, the entire process of HSID is a somewhat artificial separation of "good" and "bad" sites on a continuous crash risk measurement scale. As a result, tradeoffs between "false positives" and "false negatives" will be impacted by the choice of *X*, or the percentage of sites determined to be "risky". For example, all sites performing worse than average could be considered to represent unacceptable risk, resulting in *X* near 50%. Conversely, only the top 1% of sites could be considered to have unacceptable risk. These decisions significantly impact all aspects of HSID and influence the performance results of the methods.

Hauer and Persaud (1984) drew an analogy between the first stage of identification of black-spots and a sieve, and discussed how to measure the performances of various methods of identifying hot-spot sites. Based on this study, Higle and Hecht (1989) conducted a simulation experiment to evaluate and compare techniques for the identification of hazardous locations in terms of crash rates. Subsequent work by Hauer (1997) and others (e.g., Bauer and Harwood, 2000; Hadayeghi et al., 2003; Miaou and Lord, 2003) showed that safety performance functions might be curvilinear with respect to VMT, and therefore should not in general be used to rate the risk of various sites.

The Empirical Bayes' (EB) method, formally introduced by Hauer (1997), has been adopted as the state of the practice HSID. The application of EB for HSID has received a great deal of attention as it accounts for both crash history and expected crashes on similar sites—two essential clues to safety at a site (Persaud, 1999). It follows that the safety of a site is affected by not only some common measurable factors shared by a corresponding reference population (generally captured in the safety performance function) but also some unique characteristics associated with the site (reflected in its crash history). In EB method, the safety of a site is estimated by a weighted average of observed crash count of the subject site and expected crashes of similar sites, where the weight is determined by the variance in estimating expected crashes of the reference sites. Hauer et al. (1988) applied Empirical Bayes' (EB) method to estimate the safety at signalized intersections, Persaud (1991) evaluated crash potential of Ontario road sections and Higle and Witkowski (1988) presented a Bayesian technique making use of crash rates. In a carefully controlled Monte Carlo simulation study comparing crash rate ranking, frequency ranking, accident reduction potential, and EB methods, Cheng and Washington (2005) showed that under controlled experimental conditions the EB method is in general superior to all other methods available for identifying high risk sites—revealing the lowest percentage of false positive and false negative errors. In subsequent work Cheng and Washington (2008) developed new criteria under which HSID methods can be evaluated and again the EB method yielded superior performance.

In a few studies on HSID methods researchers have attempted to tackle the complex issue of crash severity. For example, the Missouri Department of Transportation identified seven methods for identifying high crash locations, two of which acknowledged the importance of crash severity (MDT, 1999). They detailed a crash severity method that weighed injury and fatal crashes, dictated by 'local policy' that appears to be somewhat arbitrary, by a factor of (for example) 6 compared to property damage only (PDO) crashes to obtain an EPDO estimate. The severity-rate method they identified takes the EPDO estimate and divides by exposure across locations to obtain an EPDO based rate.

Tarko and Kanodia (2003) recommended the index of crash frequency and index of crash cost as the 'best' methods for conducting HSID after conducting a thorough review. The index of crash frequency method in simple terms estimates safety performance functions by location (rural multi-lane roads, rural interstates, etc.) and compares the observed to expected total crash frequencies (divided by the standard deviation of the difference estimate) to rank sites for potential improvement. This method does not account for severity nor does it account for possible regression to the mean effects. Their second recommended method is similar to the first except that it uses crash costs to incorporate severity. Count models are estimated separately for PDOs and injuries and fatalities (I/Fs) (Tarko et al., 2000). Then, the average costs for PDOs and I/Fs (and other ancillary statistics) are used to calculate a severity-weighted index. This method accounts for severity, but requires as many regression models as there are severity classes which becomes cumbersome and requires estimation on increasingly smaller samples sizes. Ma et al. (2008) proposed even a more complex multivariate Poisson-lognormal model to consider severity and frequency in a safety performance function simultaneously. While this approach is extremely capable of accounting both frequency and severity for HSID, it is cumbersome for practitioners and safety managers to apply due to its significant complexity and time commitment to estimate.

A well specified safety performance function is the key to efficiently identify sites with high risks. The preponderances of zeros in the crash data led researchers to apply zero-inflated count model by considering the possibility of the existence of dual-state data-generating process: one state is the "zero state" where the probability of an event is so low that it cannot be distinguished from zero and the other state is the "normal-count state" that includes the zeros and positive integers (e.g., Chin and Quddus, 2003; Shankar et al., 1997). The application of zero-inflated models for modeling motor vehicle crashes has been proved inappropriate mainly because of theoretical inconsistencies (e.g., Lord et al., 2007, 2005). To account for over-dispersion and various other types of heterogeneities in the crash data, researchers have tried different modeling options like generalized estimating equation models (Lord and Persaud, 2000), finite mixture regression model (Park and Lord, 2009), three-processes count model (Washington and Haque, 2013), random effects model (Shankar et al., 1998), random parameters model (Anastasopoulos and Mannering, 2009), Bayesian hierarchical models (e.g., Haque et al., 2010; Huang et al., 2009), artificial neural network model (Chang, 2005), quantile regression models (e.g., Liu et al., 2013; Qin, 2012; Qin and Reyes, 2011) and many others (see Lord and Mannering, 2010 for details). Recently, Qin et al. (2010) have applied a non-parametric quantile regression model to account for heterogeneities and skewed distribution of crash data, but they did not take into account crash severity into the methodology of identifying high risk sites.

In summary, challenges remain in HSID methodologies and include: (1) incorporating crash severity and costs into the hot spot identification technique, (2) underreporting of minor injury and property damage only (PDO) crashes, and (3) lack of a reliable of safety performance function that can deal with the crash data heavily skewed by a preponderance of zeros partially caused by crash underreporting issues. Addressing these challenges, this paper proposes a combination of PDO equivalency calculation and quantile regression technique to identify sites with high risks in a transportation network. The method applies a non-arbitrary weighting scheme to account for crash costs, is not analytically cumbersome, and can be applied relatively quickly and efficiently.