A bilevel model for multivariate risk analysis of pedestrians' crossing behavior at signalized intersections

PLEASE CITE THE PUBLISHED VERSION

http://dx.doi.org/10.1016/j.trb.2014.03.003

PUBLISHER

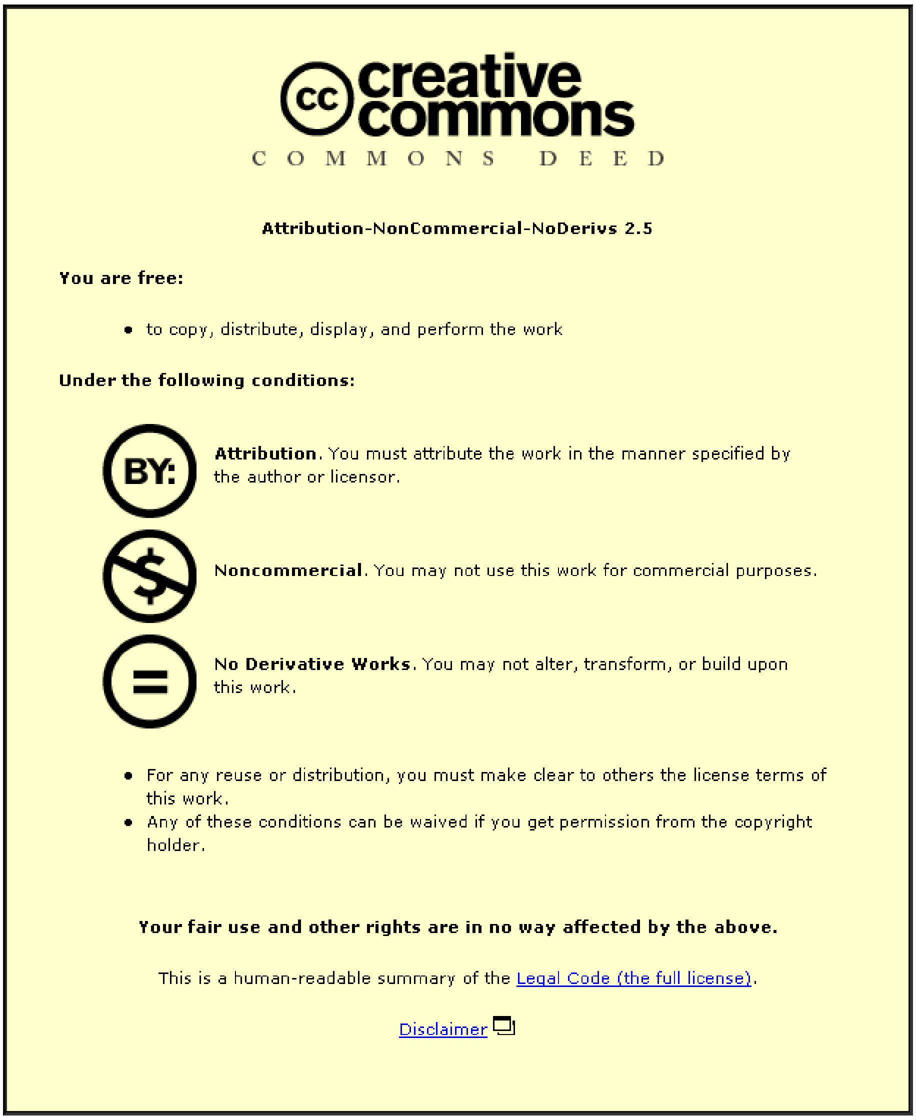
© Elsevier

VERSION

AM (Accepted Manuscript)

REPOSITORY RECORD

Li, Baibing. 2014. “A Bilevel Model for Multivariate Risk Analysis of Pedestrians' Crossing Behavior at Signalized Intersections”. Loughborough University. https://hdl.handle.net/2134/14646.



A Bilevel Model for Multivariate Risk Analysis of Pedestrians' Crossing Behavior at Signalized Intersections

Baibing Li

School of Business & Economics
Loughborough University
Loughborough LE11 3TU, United Kingdom
E-mail: b.li2@lboro.ac.uk

ABSTRACT

Pedestrians who cross streets during the red-man phase of traffic light signals expose themselves to safety and health hazards and hence are considered to be at risk. Pedestrians' street-crossing behavior is in general the outcome of interaction between pedestrians and vehicles: the gaps between vehicles provide an opportunity for pedestrians to cross the street, and pedestrians may or may not accept the street-crossing risk during the red-man phase. In this paper, we propose a multivariate method to investigate pedestrians' risk exposure associated with unsafe crossings. The proposed method consists of two hierarchically interconnected generalized linear models that characterize two different facets of the unsafe crossing behavior. It gauges pedestrians' attitudes toward risk-taking and also measures the impact of potential risk factors on pedestrians' intended waiting times during the red-man phase of the traffic lights. A Bayesian approach with the data augmentation method is used to draw statistical inference for the parameters associated with risk exposure. The proposed method is illustrated using field traffic data.

Keywords: Bayesian inference; Pedestrian's street-crossing behavior; Risk analysis; Signalized intersection.

1. Introduction

Pedestrians are vulnerable road users. Every year there are an enormous number of pedestrian-involved fatalities all over the world, in particular in the emerging-economic countries due to the rapid motorization and urbanization. Even in the developed countries, the percentage of pedestrian-involved fatalities is high, especially in the urban areas of large cities such as London.

Pedestrians crossing streets during the red-man phase of traffic light signals are considered to be at risk. In the traffic literature, considerable attention has been paid to risk analysis for pedestrians' street crossings, as this is usually the time of intensive interaction between pedestrians and vehicles. Keegan and O'Mahony (2003), Yang et al. (2006), Lipovac et al. (2012), among many others, investigated pedestrians' risk exposure associated with street crossings at signalized intersections in different countries, namely Ireland, China, and Bosnia and Herzegovina. They used questionnaires/interviews to identify the factors that may influence pedestrians' street-crossing behavior during the red-man phase of traffic light signals. Their surveys involved several different research issues including: (a) which particular group of pedestrians tends to be risk takers; and (b) for how long on average they intend to wait. Hamed (2001), Tiwari et al. (2007), and Wang et al., (2011), on the other hand, focused on one particular facet of the problem, i.e. the waiting time, with a more sophisticated analytical approach. In their studies, pedestrians' waiting times were treated as time-to-event data and the Cox proportional hazard model was applied to identify the risk factors affecting pedestrians' waiting times during their unsafe crossings. This multivariate approach enabled the researchers to gain further insights into risk analysis of pedestrians' crossing behavior by simultaneously revealing multiple traffic variables and social-economic factors that affected pedestrians' waiting times in Jordan, India, and China respectively.

From a methodological perspective, we note that the scenario considered in many of these empirical studies were a special case of the general situation. An example can be seen from the study in Tiwari et al. (2007) where the duration of the entire cycle time of traffic light signals was set to be long, with three to five phases, at the seven intersections under investigation. The length of the cycle time averaged over the seven intersections in the study was over three minutes with the longest cycle time of 285 s; see Table 1 in Tiwari et al. (2007). As a consequence, pedestrians in the samples tended not to wait for the green-man phase of the traffic light signals. Correspondingly, the waiting time distributions in Tiwari et al. (2007) had a shape similar to that of the exponential distribution, showing that as the waiting time increased, the pedestrians became less likely to wait. The reader is referred to Figure 6 in Tiwari et al. (2007) for the survival functions, both of which resemble an exponential function. A similar example can be found in Figure 2 in Wang et al. (2011). For these problems, given the fact that the waiting time distributions were close to the exponential distribution, it seemed reasonable to apply the Cox proportional hazard model in the risk analysis to investigate pedestrians' waiting times.

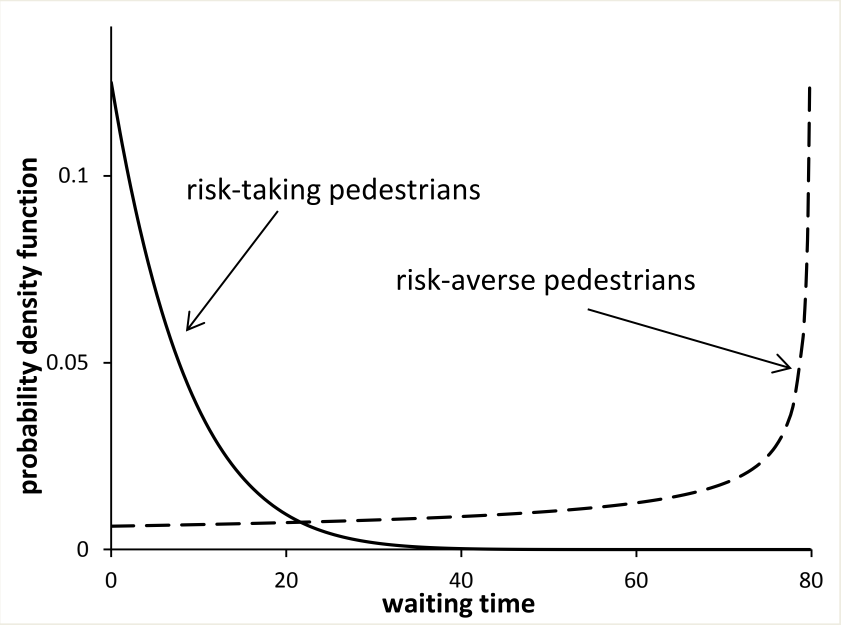


Figure 1. Illustration of pedestrians' street-crossing behavior. Risk-taking pedestrians tend to value their time highly and hence become less likely to wait as the waiting time increases; whereas risk-averse pedestrians tend to be law-abiding. The longer they wait, the less likely that the risk-averse pedestrians will cross the street during the red-man phase.

Li (2013) has recently proposed a U-shaped distributional model to characterize pedestrians' intended waiting time during the red-man phase of traffic light signals at signalized intersections. This model was developed based on the interaction between pedestrians and vehicles, and hence it helps understand the mechanism of pedestrians' street-crossing behavior. The U-shaped distributional model shows that the exponential distribution or the like reflects the crossing behavior of risk-taking pedestrians only; the waiting time of a population that consists of both risk-taking and risk-averse pedestrians in general exhibits a U-shaped distribution, as illustrated in Figure 1. The reason behind this is that law-abiding pedestrians behave differently: they are aware that the red light will soon change, making a safe crossing possible, and hence the longer they wait, the *less* likely that they will cross the street during the red-man phase. This argument was supported by many existing empirical studies. For instance, Keegan and O'Mahony (2003) found that a substantial proportion of the pedestrians in their survey *always* wait for the green-man signal before crossings. In other studies (e.g. Hamed (2001), Ahuja et al. (2005)), it is found that elderly and/or female pedestrians tend to be more law-abiding. In addition, pedestrians tend not to accept higher risk if they involved in a traffic accident in the past, or they are accompanied by children, or they have heavy luggage, or their mobility is impaired.

The motivation of this paper is two-fold. First, we note that there is a real need to extend the univariate approach in Li (2013) to a multivariate method. The U-shaped distribution in Li (2013) is a univariate model that involves one variable only, i.e. pedestrians' intended waiting time. In many traffic safety studies, pedestrian safety is of great importance, and it is imperative to develop a multivariate method to assess risk factors associated with pedestrians' unsafe crossing behavior at signalized intersections.

Secondly, in order to better understand pedestrians' street crossing behavior, it is of interest to unpack pedestrians' street-crossing risk and investigate different facets of risk separately, i.e. (a) to gauge the pedestrians' attitudes toward risk-taking and investigate what factors influence their decision; and (b) to measure the impact of risk factors that affect the pedestrians' waiting times.

For this end, we will develop a bilevel multivariate approach which consists of two interconnected generalized linear models: one focuses on pedestrians' intended waiting time conditional on pedestrian type, and the other is based on the marginal distribution of pedestrian type that gauges pedestrians' attitudes toward risk-taking. Statistically, the two models are closely linked to each other. As a whole, the bilevel model measures different facets of risk exposure associated with pedestrians' street-crossings during the red-man phase of traffic light signals.

This paper is structured as follows. In the next section we propose a multivariate approach to risk analysis for pedestrians' unsafe street-crossing behavior at signalized intersections. In Section 3 we investigate Bayesian inference and develop Markov chain Monte Carlo (MCMC) algorithms. To illustrate this model, a practical example is given in Section 4. Finally, concluding remarks are offered in Section 5.

2. A bilevel multivariate model for pedestrians' street-crossing behavior

In this section, we develop a bilevel multivariate approach to risk analysis for pedestrians' street-crossing behavior at signalized intersections. We first introduce the notation used in this paper. Then we provide a brief summary for the univariate model proposed in Li (2013), and finally we extend the univariate model to multivariate risk analysis.

2.1 Notation

Pedestrians' street-crossing behavior is the outcome of interaction between pedestrians and vehicles. We first introduce the notation that describes traffic flow, pedestrians and traffic signal setting respectively.

Li (2013) argues that, when vehicular speed is capped by a relatively low speed limit in urban areas (especially in the city/town centers), the most important traffic variable that affects pedestrians' street-crossing behavior is vehicle time headway because it characterizes the gap between two consecutive vehicles and hence provides a measure of opportunity for a pedestrian to cross the street during the red-man phase. In this paper, we use H to denote the random variable of vehicle time headway and the lower case h to denote its realization.

A commonly used headway model is the two-component model developed in Cowan (1975) that consists of a 'tracking' component and a 'free' component:

$$f_H(h;\rho,\lambda) = \rho\delta(h-\tau) + (1-\rho)\lambda^{-1}\exp\{-(h-\tau)/\lambda\} \qquad \text{for } h \geq \tau, \tag{1}$$

where τ is the minimum headway of traffic. ρ is the proportion of vehicles associated with the 'tracking' component. Vehicles associated with the "tracking" component have a minimum headway τ. $\lambda+\tau$ is the average headway of the 'free' component. Hence, parameter λ represents the difference in the average headway between the “tracking” component and the ‘free’ component. $\delta(x)$ denotes the Dirac delta function of x. It is a generalized function on the real line that is zero everywhere except at zero, with an integral of one over the entire real line.

As for pedestrians, Li (2013) shows that the effective critical headway (ECH), defined to be the minimum vehicle time headway required by a pedestrian to cross a street safely, also plays an important role in the modeling of pedestrians’ behavior. A pedestrian will cross a street only if the current vehicle time headway is greater than his/her ECH. In the literature, the ECH is shown to be related to the crosswalk length and a pedestrian’s average walking speed (Ishaque and Noland, 2008). In reality, the time that a pedestrian requires for street crossings is usually larger than this level. This is because he/she also takes into consideration of safety issue before crossings, and there is a perception-response time of human beings to a reasonably clear stimulus (Dewar and Olson, 2007).

Besides these physical factors, pedestrians’ street-crossing behavior is also related to their risk attitude. Risk attitude is a concept based on the behavior of humans while exposed to uncertainty. It is used in a number of disciplines such as business and economics to describe the choice made by a given individual or group in the face of a particular risky situation. Although risk attitude exists on a continuous spectrum in the sense that there are an infinite variety of possible responses to risk which can be displayed by a particular individual or group, individuals are usually clarified into a few risk attitude groups in the literature.

Following Li (2013), we consider two broad categories of pedestrian in this paper, i.e. risk averse and risk taking, according to whether or not their ECH is greater than the minimum headway τ or not. Risk-averse pedestrians tend not to trade safety with time and/or have less mobility. In general, they have a higher average level of ECH, so they tend to wait until they are sure it is safe to cross. Risk-taking pedestrians, on the other hand, value their time highly so they have a lower average level of ECH and tend to cross the street whenever possible. Let μ_1 and μ_2 denote the average ECH of risk-taking and risk-averse pedestrians respectively with $\boldsymbol{\mu} = [\mu_1, \mu_2]^T$. By definition we have $\mu_1 \leq \tau$ and $\mu_2 > \tau$.

In the literature, there are many factors that are related to characterizing the risk-averse and risk-taking groups but none of them can uniquely determine their risk attitudes. In terms of demographic characteristics, age and gender are widely considered to be factors affecting gap acceptance and pedestrian compliance (Ishaque and Noland, 2008). Many other factors have also been identified, including trip purposes (commuting, shopping, etc.), education level of a pedestrian, physical restriction on mobility, past experience involving road accidents (see, e.g. Hamed (2001), Ahuja et al. (2005)).

Finally, throughout this paper, we use W to denote a pedestrian's intended waiting time (random variable) and the lower case w to denote its realization. In addition, we use C to denote the duration of the red-man phase of the traffic light signals.

2.2 A brief summary of the univariate model

Now we introduce the univariate model developed in Li (2013). For a pedestrian of type P (P =1 for risk-taking and 2 for risk-averse pedestrians) with a given headway $H = h$, Li (2013) considered the following conditional probability model for intended waiting time W during the red-man phase:

$$W|(H = h, P) \sim \begin{cases} f_{BP}(w; \theta_P, C) & if\ h > \mu_P \\ \delta(w - C) & if\ h \le \mu_P \end{cases} . \tag{2}$$

Model (2) simply says that for a given headway h, the pedestrian's intended waiting time W during the red-man phase follows a bounded Pareto distribution $f_{BP}(w; \theta_P, C)$ if the vehicle time headway is greater than his/her ECH; otherwise he/she has to be prepared to wait for up to the entire red-man phase. The probability density function of the bounded Pareto distribution is given by

$$f_{BP}(w; \theta, C) = (\theta/C)(1 - w/C)^{\theta - 1} \qquad \text{for } 0 \le w \le C,$$

with θ the parameter. As shown in Li (2013), $f_{BP}(w; \theta, C)$ is strictly decreasing (or increasing) if $\theta > 1$ (or $0 < \theta < 1$). $f_{BP}(w; \theta, C)$ with $\theta > 1$ is used to model the scenario where a pedestrian highly values his/her time and becomes more impatient as he/she waits longer, whereas $f_{BP}(w; \theta, C)$ with $0 < \theta < 1$ describes the scenario where a pedestrian tends to be law-abiding and not to risk his/her safety. Hence, the longer he/she waits, the less likely that he/she will cross the street during the red-man phase. See Figure 1 for an illustration of $f_{BP}(w;\theta,C)$ with $\theta = 10$ for risk-taking pedestrians and $\theta = 0.5$ for risk-averse pedestrians.

The parameter θ_P in model (2) is linked to the headway and the pedestrians' ECH as follows:

$$\theta_P = \beta_P \max(h - \mu_P, 0) \qquad \text{with} \qquad P = 1 \text{ and } 2, \tag{3}$$

where $\beta_P > 0$ is termed sensitivity coefficient. It reflects how sensitive a certain type of pedestrian is to a given excess gap $h - \mu_P$. Risk-taking pedestrians are usually more sensitive, i.e. $\beta_1 > \beta_2$. Let $\boldsymbol{\beta} = [\beta_1, \beta_2]^T$ denote the vector of the sensitivity coefficients.

Equation (2) is a model conditional on both headway and pedestrian type. To derive an unconditional distribution model, Li (2013) assumed that the binary random variable P, i.e. pedestrian type, follows a Bernoulli distribution:

$$\Pr\{P = 1\} = \pi \quad \text{and} \quad \Pr\{P = 2\} = 1 - \pi, \tag{4}$$

where π is the probability that a pedestrian is risk-taking. Li (2013) also assumed that vehicle time headway follows model (1). On the basis of the two marginal models for pedestrian type and vehicle time headway, i.e. (4) and (1), Li (2013) derived the following unconditional distribution function of pedestrians' intended waiting time:

$$F_W(w) = r_1 G(w; A_{RT}, B_{RT}, C) + r_2 G(w; A_{RT}, 0, C) + r_3 G(w; 0, B_{RA}, C) + r_4 G(w; 0,0, C), \tag{5}$$

where $G(w; A, B, C)$ (for $A \geq 0$, $B \geq 0$) is a distribution function given by

$$G(w; A, B, C) \ = 1 - \left(1 - \frac{w}{C}\right)^A \{1 - B \ln\left(1 - \frac{w}{C}\right)\}^{-1} \quad \text{if} \quad w \in [0, C].$$

In model (5), $r_1, \dots, r_4$ are mixing probabilities of the four components with $\sum_{j=1}^{4} r_j = 1$. A_{RT}, B_{RT}, and B_{RA} are the parameters of the individual component distributions. Conceptually, all these parameters depend on $\boldsymbol{\mu}$, $\boldsymbol{\beta}$ and the parameters in model (1). In empirical studies, however, they are estimated directly from observed waiting time data.

2.3 A conditional multivariate model for pedestrians' intended waiting time

Next, we extend the univariate distributional model to a bilevel multivariate model so that the risk factors of interest can be accounted for. This is carried out in two steps: first, conditional on pedestrian type, we develop a generalized linear model for pedestrians' intended waiting time; then we use a logit model to gauge pedestrians' attitudes toward risk-taking.

We first note that the unconditional distribution (5) is complicated. It has six independent parameters. Potentially all of them can depend on the risk factors of interest. If each of these parameters is explicitly linked to multiple risk regressors, the entire model would likely suffer from a serious over-fitting problem. Hence, it is difficult, if not impossible, to investigate the impact of risk factors with model (5).

The conditional model, equation (2), however, is much more mathematically convenient because the bounded Pareto distribution belongs to the exponential distribution family (see, e.g., McCullagh and Nelder, 1989). More specifically, the bounded Pareto distribution can be expressed in the following form:

$$f_{BP}(w;\theta,C) = (1/C)\exp\{\theta log\left(1-\frac{w}{C}\right) + \log(\theta) - log\left(1-\frac{w}{C}\right)\} \quad \text{for } 0 \le w \le C,$$

with the mean function $\partial\{-\log(\theta)\}/\partial\,\theta = -1/\theta$. Li (2013) thus considered a simple generalized linear model, equation (3), to approximate the complicated relationship and related the parameter θ_P to the only attribute, the excess headway $h - \mu_P$.

Clearly, when risk factors are to be taken into consideration explicitly, model (3) needs to be further extended. Specifically, let $\mathbf{x}$ denote a vector of M risk covariates that may potentially affect pedestrians' intended waiting times for street-crossings at signalized intersections. These may include any factors in the following categories (see, e.g. Hamed (2001); Ishaque and Noland (2008)):

- current traffic condition: traffic volume, vehicle speed, etc.
- roadway layout: location of the crossing point, number of lanes, width of the street, if there is a raised median refuges (Yes/No), etc.
- settings of the traffic light signal: total cycle length, number of phases, etc.
- time (peak or off-peak time), day (weekday or weekend) and weather condition.
- social-economic characteristics associated with a pedestrian: gender, age, education, physical disability, trip purpose, pedestrian's speed, pedestrians' group size, if the pedestrian was involved or witnessed an accident in the past (Yes/No), etc.

In this paper, we propose the following generalized linear model to account for these risk factors that may affect pedestrians' intended waiting time:

$$\theta_P = \beta_P \exp(\boldsymbol{\gamma}^T \mathbf{x}) \max(h - \mu_P, 0) \qquad \text{with } P = 1 \text{ and } 2, \tag{6}$$

where $\boldsymbol{\gamma} = [\gamma_1, \dots, \gamma_M]^T$ is an M-vector of coefficients associated with the risk vector $\mathbf{x}$.

Under the generalized linear model (6), the sensitivity to the excess headway $h - \mu_P$ is captured by $\beta_P \exp(\boldsymbol{\gamma}^T \mathbf{x})$. Hence, it describes how risk covariates $\mathbf{x}$ affect pedestrians' sensitivity to the excess headway and impact on their intended waiting times. Equation (6) reduces to the conditional model (3) with sensitivity β_P if $\mathbf{x} = \mathbf{0}$.

2.4 A marginal model for gauging risk attitude

Model (2) with (6) is a probabilistic model conditional on pedestrian type P. To complete the model specification, we need to further consider a marginal model for random variable P. This also provides an approach to gauging pedestrians' attitudes toward risk-taking.

Let vector $\mathbf{z}$ include all K risk covariates that may affect pedestrians' risk attitudes, as well as 1 that corresponds to the intercept. On the basis of equation (4) that assumes that binary random variable P follows a Bernoulli distribution, we specify the following logit model for pedestrian type:

$$\pi = 1/\{1 + \exp(-\boldsymbol{\alpha}^T \mathbf{z})\}, \tag{7}$$

where $\pi = \Pr\{P = 1\}$. $\boldsymbol{\alpha} = [\alpha_0, \alpha_1, \dots, \alpha_K]^T$ is a (K+1)-dimensional vector of coefficients. Note that it usually includes intercept α_0.

2.5 A bilevel multivariate model

The conditional model (2) and marginal distribution (7) can be pooled together using the total probability theorem:

$$W|(H = h) \sim f_{MBP}(w; \boldsymbol{\theta}, \pi, C) \ , \tag{8}$$

with $\boldsymbol{\theta} = [\theta_1, \theta_2]^T$. $f_{MBP}(w; \boldsymbol{\theta}, \pi, C) = \pi f_{BP}(w; \theta_1, C) + (1-\pi) f_{BP}(w; \theta_2, C)$ is the probability density function of the mixture distribution of the bounded Pareto distributions for risk-averse and risk-taking pedestrians.

Equation (8) is a bilevel model with $\boldsymbol{\theta}$ and π specified by equations (6) and (7) respectively, each reflecting a different facet of risk exposure associated with pedestrians' unsafe crossing behavior. Pedestrians may have different perceptions to risk-taking and also there may be a number of factors affecting their risk attitudes. This is characterized by equation (7). Equation (6), on the other hand, describes the waiting time that a pedestrian is willing to spend, given the risk attitude. The two generalized linear models are hierarchically linked to each other, one conditional on the other.

In terms of statistical inference, we note that although logit model (7) has the standard form of multinomial discrete choice models reflecting pedestrians' individual decision-making (see, e.g., Train, 2009; Li, 2011), vector $\boldsymbol{\alpha}$ cannot be estimated solely based on model (6) because the pedestrian type is not directly observable in practice.

In general, risk factors that are included in equation (6) may differ from the risk factors in (7). In empirical analysis, the following modeling strategy can be used to identify the relevant risk factors for each facet. Initially, the risk covariates in vector $\mathbf{z}$, *if applicable*, may be chosen the same as the risk covariates in vector $\mathbf{x}$. Then in the subsequent statistical analysis, we can identify which particular risk covariates are associated with each facet of risk exposure via a variable selection process. This will be illustrated in the empirical study later in the paper.

Finally, we consider the interpretation of the risk coefficients, $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$. First, we note that vector $\boldsymbol{\alpha}$ in logit model (7) can be interpreted in terms of odds ratio as does in logistic regression analysis.

In general, consider two outcomes with probabilities of π and $1-\pi$ respectively. The odds are defined to be $\pi/(1-\pi)$. The ratio of two odds is termed odds ratio.

Equation (7) can be rewritten as

$$log \frac{\pi}{1-\pi} = \boldsymbol{\alpha}^T \mathbf{z},$$

where π (or $1-\pi$) represents the probability of being a risk-taking (or risk-averse) pedestrian. Consider one particular covariate z_j in vector $\mathbf{z}$ and its corresponding coefficient α_j in vector $\boldsymbol{\alpha}$. When

z_j is increased by one unit, the odds are multiplied by $\exp(\alpha_j)$. Hence, each exponentiated coefficient $\exp(\alpha_j)$ is the ratio of two odds. For further discussion on the interpretation of the coefficients of logistic regression, see, e.g. Gelman and Hill (2007, Chapter 5) and Christensen (1997) in the general situation, and Washington et al. (2010, Chapter 12) in the context of transportation and traffic studies.

Next, we turn to consider the interpretation of vector $\boldsymbol{\gamma}$ in equation (6). We note that conditional on pedestrian type, equation (2) with (6) is a proportional hazard model. To see this, we write out the hazard function of the bounded Pareto distribution below:

$$h(w;\theta,C)=\frac{f_{BP}(w;\theta,C)}{1-F_{BP}(w;\theta,C)}=(\theta/C)(1-w/C)^{-1},$$

where $F_{BP}(w;\theta,C)=1-(1-w/C)^{\theta}$ is the cumulative distribution function of $f_{BP}(w;\theta,C)$. Substituting equation (6) into the above equation, we obtain

$$h(w;\theta_P,C)=\tilde{h}_P(w;C)\exp(\boldsymbol{\gamma}^T\mathbf{x}),$$

where $\tilde{h}_P(w;C)=\beta_P\max(h-\mu_P,0)(C-w)^{-1}$ is defined to be the baseline hazard function of pedestrian type P (P =1 and 2). Therefore, conditional on pedestrian type P, each risk coefficient in vector $\boldsymbol{\gamma}$ can be, in principle, interpreted in a similar manner as does in Cox's proportional hazard models in time-to-event analysis; see, e.g., Washington et al. (2010, Chapter 10) and Collett (2003), for an introduction to Cox's proportional hazard models.

Specifically, consider pedestrians of type P and a 0-1 binary covariate x_j (say $x_j=1$ for males and 0 for females). For the two pedestrian sub-groups with $x_j=1$ and $x_j=0$ respectively, with the same excess headway $h-\mu_P>0$ and all other risk factors being equal, $\exp(\gamma_j)$ may be interpreted as the hazard ratio for the sub-group of $x_j=1$ (males) versus the sub-group of $x_j=0$ (females). Note that the benchmark for the comparison of hazards differs for different pedestrian type because $\tilde{h}_P(w;C)$ depends on P.

The above interpretation applies to all risk-taking pedestrians (P=1). Care must be taken, however, for risk-averse pedestrians (P=2): the interpretation of $\exp(\gamma_j)$ as the hazard ratio applies only to the case where excess headway $h-\mu_2>0$; for any two risk-averse pedestrians with the same excess headway $h - \mu_2 < 0$, both pedestrians will wait until the green-man signal shows. In this case, they are not at risk at all because they do not expose themselves to safety and health hazards.

3. Statistical inference

In this section, we investigate statistical inference for the developed bilevel multivariate model via a Bayesian approach.

In practice, vehicle time headway may or may not be observed in a risk analysis, which has important implications to statistical inference. We hence differentiate two different scenarios: (a) measurements on vehicle time headway are available in analysis; and (b) vehicle time headway is *not* observed. We first focus on scenario (a). Then with the data augmentation method, scenario (b) will be investigated by a straightforward extension of the method used for scenario (a).

3.1 Statistical inference with observed vehicle time headways

In this subsection we investigate statistical inference when the measurements on vehicle time headway are available. We first discuss the likelihood function and specify the prior distribution. Then we derive the posterior distribution of the parameters and develop an MCMC algorithm for the Bayesian analysis.

3.1.1 Data and likelihood function

Pedestrians' waiting time data in empirical analysis usually contains a substantial number of censored values because the observation process of pedestrians' intended waiting times is often interrupted by the green-man signal: if the green-man signal appears before a pedestrian crosses, the observation of the intended waiting time is not available and hence considered to be censored (Tiwari et al., 2007; Li, 2013).

In general, we consider a random sample of *n* pedestrians, $\{(w_j, c_j, \mathbf{x}_j, \mathbf{z}_j, h_j)\}$ for *j*=1,…,*n*, where $\mathbf{x}_j$ and $\mathbf{z}_j$ are the vectors of risk factors associated with pedestrian *j*. h_j is the measured vehicle headway when pedestrian *j* crossed the street. w_j is the observed *actual* waiting time of pedestrian *j*, defined to be the time difference between arriving at the crossing point during the red-man phase and leaving the curb. c_j is the corresponding indicator, with $c_j = 0$ if the time that pedestrian j is willing to wait is observed, and $c_j = 1$ if the observation is interrupted by the green-man signal.

Now we turn to consider the likelihood function. When a piece of waiting time data, say the jth, is censored, the only information available is that the intended waiting time is longer than the observed duration. Hence, its contribution to the likelihood function is $1 - F_{MBP}\left(w_j; \boldsymbol{\theta}, \pi, C\right)$, where $F_{MBP}\left(w_j; \boldsymbol{\theta}, \pi, C\right)$ is the cumulative distribution function of $f_{MBP}(w; \boldsymbol{\theta}, \pi, C)$ defined in equation (8). On the other hand, if the waiting time of pedestrian j is observed, its contribution to the likelihood function is proportional to $f_{MBP}\left(w_j; \boldsymbol{\theta}, \pi, C\right)$. The likelihood associated with pedestrian j for the two different scenarios can be written in a unified form:

$$L_j(\boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\mu} | w_j, \mathbf{x}_j, \mathbf{z}_j, h_j) \propto \left[f_{MBP}\left(w_j; \boldsymbol{\theta}, \pi, C\right)\right]^{1-c_j}\left[1 - F_{MBP}\left(w_j; \boldsymbol{\theta}, \pi, C\right)\right]^{c_j}.$$

Note that headway h_j, risk covariates $\mathbf{x}_j$ and $\mathbf{z}_j$, as well as parameter vectors $\boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\mu}$, are linked to $\boldsymbol{\theta}$ and π via equations (6) and (7). In most risk analysis for pedestrians' street-crossings, we are interested in the coefficient vectors $\boldsymbol{\gamma}$ and $\boldsymbol{\alpha}$. Taking into consideration of all pedestrians, the likelihood is given by

$$L(\boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\mu} | \mathbf{w}, \mathbf{X}, \mathbf{Z}, \mathbf{h}) = \prod_{j=1}^{n} L_j(\boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\mu} | w_j, \mathbf{x}_j, \mathbf{z}_j, h_j)\,, \tag{9}$$

where $\mathbf{w} = [w_1, \dots, w]^T$, $\mathbf{X} = [\mathbf{x}_1^T, \dots, \mathbf{x}_n^T]^T$, $\mathbf{Z} = [\mathbf{z}_1^T, \dots, \mathbf{z}_n^T]^T$, and $\mathbf{h} = [h_1, \dots, h_n]^T$.

3.1.2 Choice for the prior

Now we specify the prior distribution. If there is some information available on the parameters of model (8) that was obtained in the previous studies, it may be used to form the prior distribution $p(\boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\mu})$. In this paper, we assume that there is no such information in the current analysis. Hence we use a non-informative prior for the parameters of primary interest:

$$p(\boldsymbol{\alpha}, \boldsymbol{\gamma}) \propto 1.$$

It is well known that Bayesian analysis for mixture models can experience identifiability difficulty in computation (see, e.g. Gelman et al. (2013, Chapter 18); Zucchini and MacDonald (2009, Chapter 7)). Prior knowledge that defines the sub-groups of the mixture models can substantially alleviate this problem. For the model considered here, the two pedestrian groups are defined by their ECHs and sensitivity coefficients. By definition, the ECH for risk-averse (or risk-taking) pedestrians is greater (or not greater) than the minimum headway τ. In addition, risk-taking pedestrians have a larger sensitivity coefficient. We thus specify the following prior:

$$p(\boldsymbol{\beta}, \boldsymbol{\mu}) \propto I(0 < \beta_2 < \beta_1) I(0 < \mu_1 \leq \tau) I(\tau < \mu_2 \leq C).$$

In practice, the minimum headway τ can be obtained using the current headway measurements $\mathbf{h} = [h_1, \dots, h_n]^T$ via model (1).

Combining the above prior distributions, the joint prior distribution is specified as

$$p(\boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\mu}) \propto I(0 < \beta_2 < \beta_1) I(0 < \mu_1 \leq \tau) I(\tau < \mu_2 \leq C). \tag{10}$$

3.1.3 The posterior distribution and MCMC algorithm

The posterior distribution for the parameters in model (8) can be derived straightforwardly by applying Bayes' rule that pools likelihood (9) with prior distribution (10):

$$p(\boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\mu} | \mathbf{w}, \mathbf{X}, \mathbf{Z}, \mathbf{h}) \propto L(\boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\mu} | \mathbf{w}, \mathbf{X}, \mathbf{Z}, \mathbf{h}) p(\boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\mu}) . \tag{11}$$

The above posterior distribution is analytically intractable. In empirical analysis, we have to obtain a numerical solution using Markov chain Monte Carlo (MCMC) simulation. The overall structure of the MCMC algorithm used in this paper is based on the Gibbs sampler where each block of the parameters is simulated, one at a time, during each iteration *k*, as outlined below:

```
Algorithm I:
Initialization: set an initial guess of $\boldsymbol{\beta}^{(0)}$, $\boldsymbol{\mu}^{(0)}$  $\boldsymbol{\alpha}^{(0)}$ and $\boldsymbol{\gamma}^{(0)}$;
For $k$=1: $N$
    - simulate parameters $\boldsymbol{\alpha}^{(k)}$ from (11) for given $\boldsymbol{\gamma}^{(k-1)}$, $\boldsymbol{\beta}^{(k-1)}$ and $\boldsymbol{\mu}^{(k-1)}$;
    - simulate parameters $\boldsymbol{\gamma}^{(k)}$ from (11) for given $\boldsymbol{\alpha}^{(k)}$ , $\boldsymbol{\beta}^{(k-1)}$ and $\boldsymbol{\mu}^{(k-1)}$;
    - simulate parameters $\boldsymbol{\beta}^{(k)}$ from (11) for given $\boldsymbol{\alpha}^{(k)}$, $\boldsymbol{\gamma}^{(k)}$ and $\boldsymbol{\mu}^{(k-1)}$;
    - simulate parameters $\boldsymbol{\mu}^{(k)}$ from (11) for given $\boldsymbol{\alpha}^{(k)}$, $\boldsymbol{\gamma}^{(k)}$ and $\boldsymbol{\beta}^{(k)}$.
End
```

The total number of iteration N in Algorithm I is specified sufficiently large to ensure the convergence. The first N_0 ($N_0 < N$) iterations are termed burn-in period and samples simulated in this period are discarded. Summary statistics (such as posterior means, posterior standard deviations, and credible intervals) are calculated using the samples simulated beyond the burn-in period. The reader who is not familiar with MCMC is referred to Gelman et al. (2013), Chapter 11, for a comprehensive introduction to MCMC.

Next we focus on the simulation of $\boldsymbol{\beta}$; the simulation of the other parameters can be undertaken in a similar manner.

We first note that the sensitivity coefficients are positive. Hence, instead of simulating $\boldsymbol{\beta} = [\beta_1, \beta_2]^T$, we apply a log-transformation $\gamma_{0P} = \log(\beta_P)$ (for P=1 and 2) and simulate γ_{0P} at each iteration k. From equation (10), we can obtain the prior distribution for γ_{01} and γ_{02}, i.e. $I(\gamma_{02} < \gamma_{01})$.

We focus on iteration k. Let $\gamma_{01}^{(k-1)}$ and $\gamma_{02}^{(k-1)}$ denote the draws of γ_{01} and γ_{02} obtained at iteration $k-1$ with $\boldsymbol{\beta}^{(k-1)} = [\exp(\gamma_{01}^{(k-1)}), \exp(\gamma_{02}^{(k-1)})]^T$. We now wish to simulate γ_{01} and γ_{02} at iteration k, denoted by $\gamma_{01}^{(k)}$ and $\gamma_{02}^{(k)}$. There is no simple way to simulate them directly from the posterior so we use the Metropolis-Hastings algorithm.

Specifically, let $\varphi(x; a, b^2, c)$ denote the probability density function of a normal distribution, right-truncated at c (where c can be finite or $+\infty$), with a location parameter a and scale parameter b respectively. We use the following random walk to simulate candidates of $\gamma_{0P}^{(k)}$:

$$\tilde{\gamma}_{0P}^{(k)} = \gamma_{0P}^{(k-1)} + \varepsilon_P \qquad \text{for } P=1, 2,$$

where ε_P follows $\varphi(x; 0, b^2, c_P)$. We choose $c_1 = +\infty$ (i.e. without truncation). To ensure $\gamma_{02}^{(k)} < \gamma_{01}^{(k)}$, we take $c_2 = \tilde{\gamma}_{01}^{(k)}$. b is a tuning parameter which can be tuned in the iteration process. Let $\tilde{\boldsymbol{\beta}}^{(k)} = [\exp(\tilde{\gamma}_{01}^{(k)}), \exp(\tilde{\gamma}_{02}^{(k)})]^T$ denote the candidate vector of the sensitivity coefficients.

Note that the joint proposal distribution for $(\tilde{\gamma}_{01}^{(k)}, \tilde{\gamma}_{02}^{(k)})$ is

$$p\left(\tilde{\gamma}_{01}^{(k)}, \tilde{\gamma}_{02}^{(k)}\right) = p\left(\tilde{\gamma}_{01}^{(k)}\right) p\left(\tilde{\gamma}_{02}^{(k)} | \tilde{\gamma}_{01}^{(k)}\right) = \varphi(\tilde{\gamma}_{01}^{(k)}; \gamma_{01}^{(k-1)}, b^2, +\infty)\ \varphi(\tilde{\gamma}_{02}^{(k)}; \gamma_{02}^{(k-1)}, b^2, \tilde{\gamma}_{01}^{(k)})$$

which is not symmetrical. Hence we calculate the ratio as follows:

$$r = \frac{p\left(\boldsymbol{\alpha}^{(k)},\boldsymbol{\gamma}^{(k)},\tilde{\boldsymbol{\beta}}^{(k)},\boldsymbol{\mu}^{(k-1)}\middle|\mathbf{w},\mathbf{X},\mathbf{Z},\mathbf{h}\right)/\varphi(\tilde{\gamma}_{02}^{(k)};\gamma_{02}^{(k-1)},b^2,\tilde{\gamma}_{01}^{(k)})}{p(\boldsymbol{\alpha}^{(k)},\boldsymbol{\gamma}^{(k)},\boldsymbol{\beta}^{(k-1)},\boldsymbol{\mu}^{(k-1)}|\mathbf{w},\mathbf{X},\mathbf{Z},\mathbf{h})/\varphi(\gamma_{02}^{(k-1)};\tilde{\gamma}_{02}^{(k)},b^2,\gamma_{01}^{(k-1)})} .$$

Following the Metropolis-Hasting algorithm, the candidate vector $[\tilde{\gamma}_{01}^{(k)}, \tilde{\gamma}_{02}^{(k)}]^T$ is accepted with probability of $\min(1, r)$, i.e. we draw a value v from the uniform distribution on interval [0, 1]. We then take $[\gamma_{01}^{(k)}, \gamma_{02}^{(k)}]^T = [\tilde{\gamma}_{01}^{(k)}, \tilde{\gamma}_{02}^{(k)}]^T$ if $v < r$; otherwise $[\gamma_{01}^{(k)}, \gamma_{02}^{(k)}]^T = [\gamma_{01}^{(k-1)}, \gamma_{02}^{(k-1)}]^T$.

3.2 Statistical inference when the measurements on vehicle time headway are unavailable

In practice, sometimes the measurements on vehicle headway are not available. In this subsection, we investigate statistical inference in this scenario.

Essentially, statistical inference in this case can be drawn in a similar manner as outlined in the previous subsection. However, since the headway measurements **h** are not available, they need to be integrated out from the posterior in equation (11). Usually, the dimension of **h** is extremely high, and hence directly evaluating the integral is difficult. Instead, the computation can be carried out using the data augmentation method: the headway vector **h** in each iteration of the MCMC simulation is simulated, and on the basis of the imputed vector **h**, all the parameters are drawn using the method outlined in Algorithm I. This process continues in an alternating manner until the convergence. We now discuss the details of this method.

3.2.1 Likelihood function

The data collected in this case includes a random sample of *n* pedestrians, $\{(w_j, c_j, \mathbf{x}_j, \mathbf{z}_j)\}$ for *j*=1,…,*n*. To make use of the algorithm in Section 3.1, we treat vehicle time headway *H* as a latent variable and carry out the analysis as if the observations were available in the analysis. To clarify the scenarios of with and without the headway measurements in the analysis, we distinguish two different types of likelihood, i.e. 'complete'-data and 'incomplete'-data likelihoods. The former is referred to the likelihood based on the data of form $\{(w_j, c_j, \mathbf{x}_j, \mathbf{z}_j, h_j)\}$ (*j*=1,…,*n*), whereas the latter is based on $\{(w_j, c_j, \mathbf{x}_j, \mathbf{z}_j)\}$ (*j*=1,…,*n*).

Consider each pedestrian j. As shown in Section 3.1, the likelihood conditional on a given headway measurement h_j is

$$L_j(\boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\mu} | w_j, \mathbf{x}_j, \mathbf{z}_j, h_j) = [f_{MBP}(w_j; \boldsymbol{\theta}, \pi, C)]^{1-c_j}[1 - F_{MBP}(w_j; \boldsymbol{\theta}, \pi, C)]^{c_j}.$$

Hence, the joint distribution of (w_j, h_j) for given $(c_j, \mathbf{x}_j, \mathbf{z}_j)$ is

$$\tilde{L}_j(\boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\mu}, h_j | w_j, \mathbf{x}_j, \mathbf{z}_j) \mathrel{\widehat{=}} L_j(\boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\mu} | w_j, \mathbf{x}_j, \mathbf{z}_j, h_j)\, f_H(h_j; \rho, \lambda),$$

where $f_H(h; \rho, \lambda)$ is specified in equation (1). However, the Dirac delta function $\delta(h-\tau)$ in equation (1) can cause some difficulties in numerical computation. We circumvent this difficulty as follows: we use an exponential distribution $\lambda_0^{-1}\exp\{-(h-\tau)/\lambda_0\}$ with a very small λ_0 to replace the Dirac delta function $\delta(h-\tau)$ in (1). This is equivalent to using the headway model proposed in Griffiths and Hunt (1991):

$$f_H(h; \rho, \lambda_0, \lambda) = \rho\lambda_0^{-1}\exp\{-(h-\tau)/\lambda_0\} + (1-\rho)\lambda^{-1}\exp\{-(h-\tau)/\lambda\} \quad \text{for } h \geq \tau. \tag{1a}$$

Let $\boldsymbol{\phi} = [\boldsymbol{\beta}, \boldsymbol{\mu}, \rho, \lambda_0, \lambda]^T$. The ‘complete’-data likelihood can be rewritten as

$$L(\boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\phi}, \mathbf{h} | \mathbf{w}, \mathbf{X}, \mathbf{Z}) = \prod_{j=1}^{n} \tilde{L}_j(\boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\mu}, h_j | w_j, \mathbf{x}_j, \mathbf{z}_j) f_H(h_j; \rho, \lambda_0, \lambda)\,. \tag{12}$$

3.2.2 Model identifiability

We note that when both headway H and ECH μ_P are not observed, model (8) with (6) is no longer identifiable. This is because from equation (6), adding any constant value to both headway measurements h_j and ECH μ_P does not affect the statistical inference for the other parameters. Hence it is only the differences between headway h_j and ECH μ_P that can be estimated. As a consequence, the minimum headway τ can be set to any reasonable value without changing the estimates of the parameters of interest, i.e. $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$.

3.2.3 Choice for the prior

Now we specify the prior distribution. The prior for $\boldsymbol{\alpha}$, $\boldsymbol{\gamma}$, $\boldsymbol{\beta}$ and $\boldsymbol{\mu}$ is kept the same as equation (10). We specify the prior of parameters ρ, λ_0 and λ in equation (1a) as follows:

$$p(\rho, \lambda_0, \lambda) \propto p(\rho) I(\lambda_0 < \lambda) I(\lambda \leq \Lambda) I(\lambda_0 \leq \Lambda_0),$$

where the upper bound Λ for λ is a pre-specified hyper-parameter in statistical analysis. Note that $\lambda + \tau$ is the average time headway for the 'free'-component. Hence this upper bound Λ can be easily elicited in practice based on the traffic characteristics. In addition, as mentioned earlier, the upper bound Λ_0 for the 'tracking' component should be pre-specified as a small value. Finally, we choose prior $p(\rho)$ of ρ as a uniform distribution, i.e. $p(\rho) \propto 1$.

Combining the above prior distributions, we specify the joint prior distribution as

$$\begin{aligned} & p(\boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\phi}) \\ & \quad \propto I(0 < \beta_2 < \beta_1) I(0 < \mu_1 \leq \tau) I(\tau < \mu_2 \leq C) I(\lambda_0 < \lambda) I(\lambda \leq \Lambda) I(\lambda_0 \leq \Lambda_0). \end{aligned} \tag{13}$$

3.2.4 The posterior distribution and MCMC algorithm

The 'complete'-data posterior distribution can be derived straightforwardly by applying Bayes' rule that pools the 'complete'-data likelihood (12) with the prior distribution (13):

$$p(\boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\phi}, \mathbf{h} | \mathbf{w}, \mathbf{X}, \mathbf{Z}) \propto L(\boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\phi}, \mathbf{h} | \mathbf{w}, \mathbf{X}, \mathbf{Z}) p(\boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\phi}) \; . \tag{14}$$

The 'complete'-data posterior (14) depends on vector $\mathbf{h}$, and hence cannot be dealt with analytically. A commonly used method in statistical inference for problems with latent variables is data augmentation where the simulation is carried out in an alternating manner: first the values of the unobserved latent variable(s) are simulated for fixed values of the parameters; then with the imputed values of the latent variable(s), the parameters are drawn from the posterior distribution. This process continues until the convergence.

The MCMC algorithm with data augmentation used in this paper is outlined below:

```
Algorithm II:
Initialization: set an initial guess of $\boldsymbol{\phi}^{(0)}$, $\boldsymbol{\alpha}^{(0)}$ and $\boldsymbol{\gamma}^{(0)}$;
For k=1: N
    - Imputation step
      simulate vector $\mathbf{h}^{(k)}$ from (14) for fixed parameters $\boldsymbol{\alpha}^{(k-1)}$, $\boldsymbol{\gamma}^{(k-1)}$, and $\boldsymbol{\phi}^{(k-1)}$;
    - Posterior step
```

for the given headway vector $\mathbf{h}^{(k)}$, simulate all the parameters parameters $\boldsymbol{\alpha}^{(k)}, \boldsymbol{\gamma}^{(k)}$, and $\boldsymbol{\phi}^{(k)}$ from (14).

End

Clearly once vector $\mathbf{h}^{(k)}$ is imputed at each iteration *k*, the posterior step can be undertaken in a similar way as does in Algorithm I. So next we focus on the imputation step only.

We consider the simulation of headway vector $\mathbf{h}$ at iteration *k*. There is no convenient way to simulate the headway vector directly, so we use the Metropolis-Hastings algorithm. Let $\mathbf{h}^{(k-1)}$ denote vector $\mathbf{h}$ obtained at iteration $k-1$.

At iteration k, the simulation of $\mathbf{h}^{(k)}$ using the Metropolis-Hastings algorithm is carried out element-wise. Consider the simulation for the *j*-th element (for *j*=1,…,*n*). When *j*=1, we define the current headway vector as $\mathbf{h}_c = \mathbf{h}^{(k-1)}$. Note that $\mathbf{h}_c$ will be updated for each of *j*=2,…,*n*.

We suggest using distribution (1a) as the proposal distribution to generate a proposal, i.e., to draw each entry $\tilde{h}_j^{(k)}$ as

$$\tilde{h}_j^{(k)} \sim f_H\left(h; \rho^{(k-1)}, \lambda_0^{(k-1)}, \lambda^{(k-1)}\right) \qquad \text{for } j=1,\ldots,n.$$

Hence the jumping distribution is $J(\mathbf{h}) = \prod_{j=1}^{n} f_H\left(h_j; \rho^{(k-1)}, \lambda_0^{(k-1)}, \lambda^{(k-1)}\right)$. Let $\tilde{\mathbf{h}}_j^{(k)}$ be the current headway vector $\mathbf{h}_c$ except that the *j*-th element is replaced with $\tilde{h}_j^{(k)}$. The ratio for the *j*-th element in the Metropolis-Hastings algorithm is:

$$r_j = \frac{p\left(\boldsymbol{\alpha}^{(k-1)}, \boldsymbol{\gamma}^{(k-1)}, \boldsymbol{\phi}^{(k-1)}, \tilde{\mathbf{h}}_j^{(k)} \middle| \mathbf{w}, \mathbf{X}, \mathbf{Z}\right) / J(\tilde{\mathbf{h}}_j^{(k)})}{p\left(\boldsymbol{\alpha}^{(k-1)}, \boldsymbol{\gamma}^{(k-1)}, \boldsymbol{\phi}^{(k-1)}, \mathbf{h}_c \middle| \mathbf{w}, \mathbf{X}, \mathbf{Z}\right) / J(\mathbf{h}_c)} \quad .$$

The candidate $\tilde{h}_j^{(k)}$ is accepted with probability $\min(1, r_j)$, i.e.

$$h_j^{(k)} = \begin{cases} \tilde{h}_j^{(k)} & \text{if accepted} \\ h_j^{(k-1)} & \text{otherwise} \end{cases} \quad .$$

Then the current headway vector $\mathbf{h}_c$ is updated with the *j*-th element being replaced by $h_j^{(k)}$. We can then progress to draw the (*j*+1)-th element.

4. A practical example

To illustrate the proposed method, we consider a practical example in this section. The same data was examined in Li (2013) using the univariate model (5).

4.1 Data and models

The intersection considered in Li (2013) is located in a busy area of a Chinese metropolitan city, Kunming. It has four arms, with two major roads crossing. Around the intersection are a theater, a number of small shops, and several residential areas. The pedestrians' crossing behavior was observed at a crossing point of the north arm of the intersection. During the time period of data collection, the intersection was signalized with the standard three-phase cycle. The duration of the red-man phase was C=75 s.

The data used in Li (2013) includes measurements on intended waiting time, an indicator whether pedestrians crossed the street in the red-man phase or waited until the green-man signal showed, plus two factors, i.e. age (grouped into young, middle-aged, and elderly categories) and gender (males or females). In total, 283 observations were included in the analysis in Li (2013). The data shows that there were a considerable number of pedestrians who crossed the street immediately after their arrivals at the crossing point, and also a large proportion of the pedestrians who were willing to wait for the entire red-man phase. Overall, the data distribution was shown to be U-shaped; see Figure 1 in Li (2013).

In order to apply the bilevel multivariate approach, we created a couple of covariates, including: (a) a gender indicator u_1 for the investigation of gender effect: $u_1 = 1$ for male pedestrians and 0 otherwise; (b) two covariates u_2 and u_3 for the investigation of age effect, i.e. $u_2 = 1$ if a pedestrian was young and 0 otherwise; and $u_3 = 1$ if a pedestrian was middle-aged and 0 otherwise.

We will perform a multivariate analysis to examine risk factors. For this end, several models were explored and compared:

- Model I: $\mathbf{x} = u_1$ and $\mathbf{z} = [1, u_1]^T$; Equations (6) and (7) are specified as $\theta_P = \beta_P \exp(\gamma_1 u_1) \max(h - \mu_P, 0)$ and $\pi = 1/\{1 + \exp[-(\alpha_0 + \alpha_1 u_1)]\}$;
- Model II: $\mathbf{x} = [u_2, u_3]^T$ and $\mathbf{z} = [1, u_2, u_3]^T$; Equations (6) and (7) are specified as $\theta_P = \beta_P \exp(\gamma_2 u_2 + \gamma_3 u_3) \max(h - \mu_P, 0)$ and $\pi = 1/\{1 + \exp[-(\alpha_0 + \alpha_2 u_2 + \alpha_3 u_3)]\}$;
- Model III: $\mathbf{x} = [u_1, u_2, u_3]^T$ and $\mathbf{z} = [1, u_1, u_2, u_3]^T$; Equations (6) and (7) are specified as $\theta_P = \beta_P \exp(\gamma_1 u_1 + \gamma_2 u_2 + \gamma_3 u_3) \max(h - \mu_P, 0)$ and $\pi = 1/\{1 + \exp[-(\alpha_0 + \alpha_1 u_1 + \alpha_2 u_2 + \alpha_3 u_3)]\}$.

Model I focused on the gender effect, investigating if the male pedestrians behaved differently from the females when crossing the street during the red-man phase. Model II, on the other hand, was used to investigate the age effect, i.e. if the pedestrians in the different age groups had different street-crossing behavior. Finally, both gender and age effects were taken into account in Model III.

4.2 Settings in the MCMC algorithm

As vehicle time headway was not observed during the data collection, Algorithm II was used to simulate the posterior distribution.

There were a few hyper-parameters that needed to be set in the MCMC simulation. In the following numerical computation, we set the minimum headway $\tau = 2$s. As mentioned earlier, the choice for τ had no impact on the estimation of $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$. The two upper bounds for headway were set as $\Lambda = 10$ s and $\Lambda_0 = 10^{-6}$ s respectively. The choice of $\Lambda = 10$ s was reasonable for the problem under investigation because during the data collection the traffic was congested with relatively small vehicle time headway. The choice for Λ_0 was purely technical as mentioned in the previous section. The setting for the initial values of the parameters was explored using some randomly generated initial values. The following initial values were used in the final results reported below: $\boldsymbol{\alpha} = \mathbf{0}$, $\boldsymbol{\gamma} = \mathbf{0}$, $\boldsymbol{\beta} = [1, 1/2]^T$, $\boldsymbol{\mu} = [\tau/2, C/2]^T$, $\rho = 0.5$, $\lambda_0 = \Lambda_0$, and $\lambda = 1$.

The total number of iterations of the MCMC simulation was set to be 10,000. The first 5,000 iterations were considered the burn-in period so the samples obtained in this period were discarded. The results reported in the next sub-section were based on the remaining 5,000 iterations.

Finally, following Zucchini and MacDonald (2009, pp. 10) and Li (2013), we used a discretized probability mass function of the bounded Pareto distribution $f_{BP}(w;\theta,C)$ in the numerical computation with waiting times grouped into 75 time intervals of duration 1 s.

4.3 Empirical results and analysis

The results of statistical analysis using different models are displayed in Table 1. The multivariate analysis has revealed some interesting findings.

Table 1. The coefficients of the risk factors in the empirical study [*]

parameters	Model I	Model II	Model III	Model IV
α_0	-0.232 (-0.780, 0.357)	-0.837 (-1.522, -0.168)	-1.045 (-1.736, -0.178)	-0.866 (-1.485, -0.227)
α_1	0.314 (-0.374, 1.003)	-	0.400 (-0.221, 0.964)	-
α_2	-	0.963 (0.145, 1.715)	1.114 (0.269, 1.911)	1.171 (0.285, 2.050)
α_3	-	0.902 (0.114, 1.751)	0.827 (0.021, 1.556)	0.899 (0.103, 1.605)
γ_1	0.814 (0.288, 1.515)	-	0.817 (0.314,1.352)	0.877 (0.284, 1.442)
γ_2	-	0.002 (-0.722, 0.731)	0.265 (-0.579, 1.691)	-
γ_3	-	-0.271 (-1.092, 0.498)	0.172 (-0.647, 1.552)	-

[*] Posterior means and 95% credible intervals in parentheses

First, it can be seen from column two of Table 1 that the estimate of γ_1 in Model I is equal to 0.814 and is significant at 5% level. This indicates that the males tended to be more impatient and had shorter waiting time than females. However, as the estimate of α_1 is not significant, there was not enough evidence about whether the males were more likely to be risk-takers.

Now we turn to consider the age effect. It can be seen from column three of Table 1 that the estimates of α_2 and α_3 in Model II are equal to 0.963 and 0.902 respectively. Both of them are significant at 5% level, indicating that compared with the elderly pedestrians, the young and middle-aged pedestrians tended to be risk-takers. However, the estimates of both γ_2 and γ_3 are not significant at 5% level. Hence, there was not enough evidence about the age effect on pedestrians' intended waiting time.

Model III took into account both age and gender effects simultaneously. The results, as displayed in column four of Table 1, show that: by controling gender effect, the young and middle-aged pedestrians tended to be risk-takers; in addition, by controling age effect, the males tended to be less patient when waiting for the green-man signal.

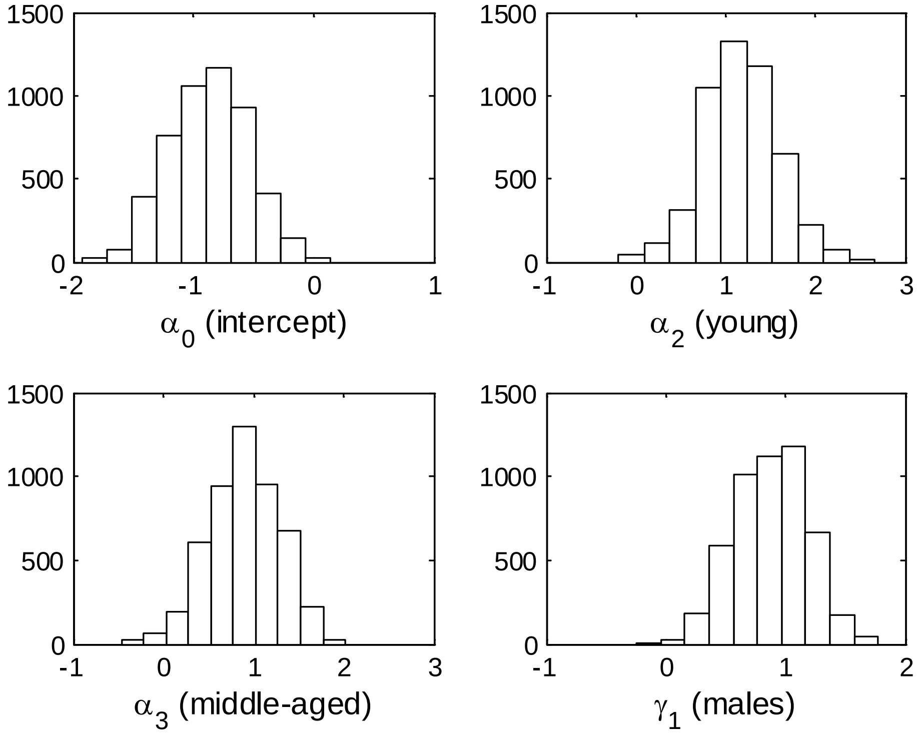


Figure 2. Posterior distributions of the risk coefficients in Model IV.

Finally, we removed the covariates that were not significant in Model III to work out the final model, Model IV. This model included the two age-related covariates to gauge the age effect on pedestrians' risk attitudes, i.e. $\mathbf{z} = [1, u_2, u_3]^T$. On the other hand, only gender indicator, x_1, was used to measure the impact on intended waiting time, i.e. $\mathbf{x} = u_1$. Hence, the two generalized linear models (6) and (7) in Model IV were specified as:

$$\theta_P = \beta_P \exp(\gamma_1 u_1) \max(h - \mu_P, 0) \quad \text{with } P = 1 \text{ and } 2,$$

$$\pi = 1/\{1 + \exp[-(\alpha_0 + \alpha_2 u_2 + \alpha_3 u_3)]\}.$$

The results are displayed in the last column of Table 1. Figure 2 displays the simulated posterior distributions of the parameters of interest in the final model.

The final column in Table 1 shows that the odds ratio for young pedestrians versus elderly pedestrians is $\exp(1.171) = 3.225$, and the odds ratio for middle-aged pedestrians versus elderly pedestrians is $\exp(0.899) = 2.457$. In addition, the hazard ratio for being males versus females is $\exp(0.877) = 2.404$.

The findings of this multivariate analysis seem reasonable and are consistent with our observation. Physically younger pedestrians are more able to accept smaller gaps during street crossings. Compared to elderly people, they usually have a much faster pace of life during the current rapid urbanization in China. Hence, in comparison with the elderly people, they tend to accept higher risk and to cross the street immediately after arriving at the crossing point. On the other hand, once the pedestrians stopped at the crossing point in the red-man phase, the results suggest that the males seemed to be more impatient and tended to find an opportunity to cross the street, leading to a shorter waiting time.

Overall the results obtained by using the bilevel multivariate model have confirmed what were revealed in Li (2013) using the univariate analysis: both the age and gender of pedestrians affected pedestrians' street-crossing behavior. The results are also in line with other empirical studies in the literature; see, e.g. Oxley et al. (1997), Keegan and O'Mahony (2003), Yang et al. (2006), among many others. However, with a univariate approach we are unable to consider potential factors simultaneously, i.e. investigate some risk factors by controlling the others, and unable to reveal different facets of risk exposure. The analysis in this paper shows that different risk factors can affect pedestrians in different ways: some of them primarily affect their attitudes toward risk-taking, whereas the others may impact on their intended waiting times.

Before we conclude this section, there are a couple of points we would like to make. First, we point out that this is an over-simplified example for risk analysis of pedestrians' unsafe crossings. It serves illustration purposes only. In practice, besides gender and age effect, pedestrians' crossing behavior is usually influenced by many other factors, some of which were listed in Section 2.3. As this paper focuses on the methodological development, a full investigation on the identification of these risk factors is beyond the scope of this paper. However, the method developed in this paper provides a useful tool for researchers to fully explore this important research issue.

We also note that pedestrians' perceptions to risk vary from time to time and vary from intersection to intersection. For instance, even at the same crossing point, the pattern in peak time of normal working days will in general differ from that in weekends. The duration of the cycle time and the layout of intersections may also have a great impact on pedestrians' crossing behavior. Care must be taken when generalizing the empirical findings in this paper to other scenarios.

Finally, an alternative way of identifying risk factors can be used in practice: we can initially set vector $\mathbf{z}$ in the way that includes the same covariates as that contained in $\mathbf{x}$, both having all potential risk factors. This leads to Model III for this particular example. We can then undertake backward variable selection by removing any insignificant risk covariates out of the model, one at a time, so that a refined model, Model IV in the example, can be obtained.

5. Concluding remarks

In this paper we have developed a bilevel multivariate modeling approach to risk analysis for pedestrians' unsafe street-crossing behavior at signalized intersections.

The developed bilevel multivariate method consists of two interconnected generalized linear models, each focusing on a particular facet of pedestrians' risk exposure associated with unsafe street-crossings: one gauges pedestrians' attitudes toward risk-taking, whereas the other measures the impact of the risk factors on pedestrians' waiting times. Statistically the two models are hierarchically linked to each other.

We have investigated two different approaches to statistical inference for the developed method: inference with and without measurements on vehicle time headway. We would like to emphasize that making use of headway data in the statistical analysis will greatly enhance the quality of the research. Given the rapid technological advances, traffic data collection has been becoming cheaper and cheaper. Efforts should be made in future empirical studies to make use of headway data.

As shown in the practical example, risk factors associated with the two facets of the developed method may differ, indicating that some risk factors may have larger effects on risk attitudes, whereas the others primarily on waiting times. Although the example given in this paper serves illustration purposes only and its findings may not be directly generalized to other intersections, the developed modeling methodology *per se* can be used in a much wider range of applications. The bilevel multivariate model developed in this paper can help us identify risk factors and thus better understand the potential risk exposure of pedestrians. In particular, following Hamed (2001), Tiwari et al. (2007), and Wang et al., (2011), further empirical studies could be done to investigate the potential risk factors for the scenarios where pedestrians' waiting times are U-shaped. The research focus could be to understand which risk factors affect which facets of the risk analysis: do they influence the intended waiting time or acceptance of risky crossing or both? In addition, the developed method can also be used to investigate the impact on pedestrians' unsafe street-crossings with different settings of traffic light signals (number of phrases, total cycle time, etc.) and different layouts of street, and hence improve on pedestrians' safety. For instance, to investigate the impact of a new street layout, before-and-after comparisons can be carried out to compare and contrast the changes in pedestrians' crossing behavior.

Before we conclude this paper, we would like to point out some limitations of the proposed method. The multivariate method in this paper is built on the basis of the univariate model in Li (2013). This univariate model assumes that pedestrians' intended waiting time, conditional on pedestrian type and headway, follows a bounded Pareto distribution. In practice, this assumption could be too restrictive. A more advanced non-parametric approach could be used in the multivariate method to overcome this limitation, where the bounded Pareto distribution is replaced with an unspecified underlying distribution function. In addition, the logit model for pedestrians’ choice bebavior (accept or not accept a gap) could be restrictive in some applications, and hence it can be replaced with a more general semi-parametric model developed in Li (2011) in future research. Finally, for both the univariate and multivariate methods, we consider two broad categories of pedestrians with respect to their attitudes to risk: risk-taking and risk-averse groups. This can be extended to the case of three or more pedestrian groups. For this end, equation (7) could be replaced with the ordinal logistic regression.

Acknowledgements

The author would like to thank the two reviewers for their helpful comments on the earlier version of this paper.

References

Ahuja, S., Hao, X., MacDonald, M., Bell, M., Phull, S., 2008. Pedestrian crossing behaviour at signalised crossings. In: Proceedings of European Transport Conference, Leiden, The Netherlands.

Christensen, R., 1997. Log-Linear Models and Logistic Regression. 2nd ed. New York: Springer.

Collett, D., 2003. Modelling Survival Data in Medical Research. 2nd ed. Boca Raton: Chapman & Hall/CRC.

Cowan, R. J., 1975. Useful headway models. Transportation Research 9 (6), 371-375.

Dewar, R. E., Olson, P. L., 2007. Human Factors in Traffic Safety. 2nd ed. Tucson: Lawyers & Judges Publishing.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., Rubin, D. B., 2013. Bayesian Data Analysis. 3rd ed. London: Chapman & Hall/CRC.

Gelman, A., Hill, J., 2007. Data Analysis Using Regression and Multilevel/Hierarchical Models. New York: Cambridge University Press.

Griffiths, J. D., Hunt, J. G., 1991. Vehicle headways in urban areas. Traffic Engineering Control 32 (10), 458 – 462.

Hamed, M. M., 2001. Analysis of pedestrians' behavior at pedestrian crossings. Safety science 38 (1), 63-82.

Ishaque, M. M., Noland, R., 2008. Behavioural issues in pedestrian speed choice and street crossing behavior: a review. Transport Reviews 28 (1), 61-85.

Keegan, O., O'Mahony, M., 2003. Modifying pedestrian behavior. Transportation Research Part A 37 (10), 889-901.

Li, B., 2011. The multinomial logit model revisited: A semi-parametric approach in discrete choice analysis. Transportation Research Part B 45 (3), 461–473.

Li, B., 2013. A model of pedestrians' intended waiting times for street crossings at signalized intersections. Transportation Research Part B 51, 17–28.

Lipovac, K., Vujanic, M., Maric, B., Nesic, M., 2012. Pedestrians' behavior at signalized pedestrian crossings. Journal of Transportation Engineering 139 (2), 165-172.

McCullagh, P., Nelder, J.A., 1989. Generalized Linear Models. 2nd ed. London: Chapman & Hall/CRC.

Oxley, J., Fildes, B., Ihsen, E., Charlton, J., Days, R., 1997. Differences in traffic judgements between young and old adult pedestrians. Accident Analysis and Prevention 29 (6), 839-847.

Tiwari, G., Bangdiwala, S., Saraswat, A., Gaurav, S., 2007. Survival analysis: pedestrian risk exposure at signalized intersections. Transportation Research Part F 10 (2), 77-89.

Train, K., 2009. Discrete Choice Methods with Simulation. 2nd ed. New York: Cambridge University Press.

Wang, W., Guo, H., Gao, Z., Bubb, H., 2011. Individual differences of pedestrian behavior in midblock crosswalk and intersection. International Journal of Crashworthiness 16 (1), 1-9.

Washington, S. P., Karlaftis, M. G., Mannering, F. L., 2010. Statistical and Econometric Methods for Transportation Data Analysis. 2nd ed. London: Chapman & Hall/CRC.

Yang, J., Deng, W., Wang, J., Li, Q., Wang, Z., 2006. Modeling pedestian' road crossing behavior in traffic systems micro-simulation in China. Transportation Research Part A 40 (3), 280-290.

Zucchini, W., MacDonald, I. L., 2009. Hidden Markov Models for Time Series. London: Chapman & Hall/CRC.