Development of accident prediction models for pedestrian crossings

Piotr Olszewski[1,*], *Beata* Osińska[1], *Piotr* Szagała[1], and *Paweł* Włodarek[1]

[1]Warsaw University of Technology, Faculty of Civil Engineering, ul. L. Kaczyńskiego 16, 00-637 Warsaw, Poland

Abstract. In large Polish cities like Warsaw, pedestrians constitute almost 60% of road fatalities. Although traffic safety situation in general is improving, the numbers of pedestrians hit when crossing a road have not significantly decreased over the last six years. A negative binomial model was estimated for predicting accidents at unsignalised pedestrian crossings based on accident data from 52 crossings in Warsaw. A total of 58 pedestrian accidents were recorded at these crossings during the last seven years. The model shows that the number of accidents is less-than-proportional to both pedestrian and motorised traffic daily volumes. Other risk factors affecting pedestrian safety are: higher proportion of heavy vehicles and location in a mixed land use area. The model can be used with the Empirical Bayes method for an unbiased identification of high risk locations.

1 Introduction

Poland has one of the highest pedestrian fatality rates in the EU, with 23 persons killed per year per million population. There were 8436 pedestrian accidents in 2016 in which 868 pedestrians were killed. Pedestrian deaths constitute 28,6% of all traffic accident fatalities. Many of these accidents occur at marked pedestrian crossings – both with and without traffic signals. Pedestrian safety is the main road safety concern in large Polish cities such as Warsaw, where pedestrians constitute almost 60% of road fatalities [1,2]. A slow improvement of the safety situation has been observed since 2006. However, the number of pedestrians killed at road crossings has remained more or less at the same level since 2009.

Many pedestrian fatalities at unsignalised crossings occur on divided roads, where traffic speeds are higher and pedestrians have to cross two or more lanes. Divided road type and higher speed limit were identified as significant risk factors for unsignalised crossings in Poland in a previous study [2]. For crossings located at signalised intersections, many pedestrian accidents occur at the intersection exit roadway where pedestrians are in conflict with vehicles turning right or left during the same signal phase. Accidents occurring at the intersection approach roadway mostly involve a conflict between vehicle going straight with pedestrian crossing the street on red and often result in death due to high speed of the vehicle [3]. For the above reasons, field studies conducted in Warsaw for the European project InDeV

* Corresponding author: p.olszewski@il.pw.edu.pl

focused on unsignalised pedestrian crossings and on crossings at exit roadways at signalised intersections.

In order to improve the safety situation, identification of high-risk pedestrian crossing locations is of great importance. However, it is well known that identifying black spots based on accident counts alone is subject to the regression-to-the-mean bias [4]. This is due to random fluctuation of accident numbers at any one location which makes it likely that sites with a high number of accidents in any one year will show a decrease in accidents in the following year. Therefore, the recommended approach to identifying black spots, known as the Empirical Bayes method, involves calculation of the expected number of accidents by combining their observed and predicted numbers [5,6]. For this method to work, an accident prediction model is necessary. Accident prediction models relate the number of accidents to measures of exposure (volumes of pedestrian and motor traffic) and possibly other variables describing the road site (e.g. geometry, traffic control).

The objective of the paper is to present accident prediction models developed for pedestrian crossings based on data from the Warsaw accident database. The models can help to identify safety risk factors and allow to apply the Empirical Bayes method for obtaining an unbiased ranking of high risk locations.

2 Accident statistics and explanatory data

2.1 Accident data

A sample of 52 unsignalised and 50 signalised pedestrian crossings in Warsaw was selected for accident analyses within project InDeV. The sample consisted of fairly uniform set of crossings – all the crossings were located on four lane roads with a median or a pedestrian refuge island. Thus, at every site pedestrians had to cross two lanes of traffic moving in the same direction. The speed limit at all sites was 50 km/h. It should be noted that the sample of sites was not representative of all crossings in the city. At the selected signalized crossings pedestrians were not fully protected by traffic control – i.e. during the green signal they were in conflict with turning vehicles.

Police records show that during the period of seven years (2009-2016) at the selected sites 58 injury accidents occurred at unsignalised and 22 at signalized locations. Further analysis is focused on unsignalised crossings – among the 58 accidents there was one fatal injury and 15 serious injuries. The distribution of unsignalised sites by number of accidents is shown in Fig. 1. At 25 sites no accidents were recorded during seven years. The mean number of accidents was 1.12 per site and the variance was 2.22. Thus, variability of accident numbers is greater than predicted by the Poisson distribution and the problem of overdispersion has to be taken into consideration.

The magnitude of overdispersion is expressed by parameter φ which shows how much a given distribution of accidents differs from the Poisson distribution. The value of φ is given by the equation:

$$\varphi = \frac{Var(A) - Mean(A)}{Mean^2(A)} \quad (1)$$

Mean(A) = mean number of accidents per site during Y years,
Var (A) = variance of the number of accidents for all the sites during Y years.

For the Poisson distribution the value of overdispersion parameter is zero. For the 52 unsigalised sites in Warsaw φ was equal to 0.877.

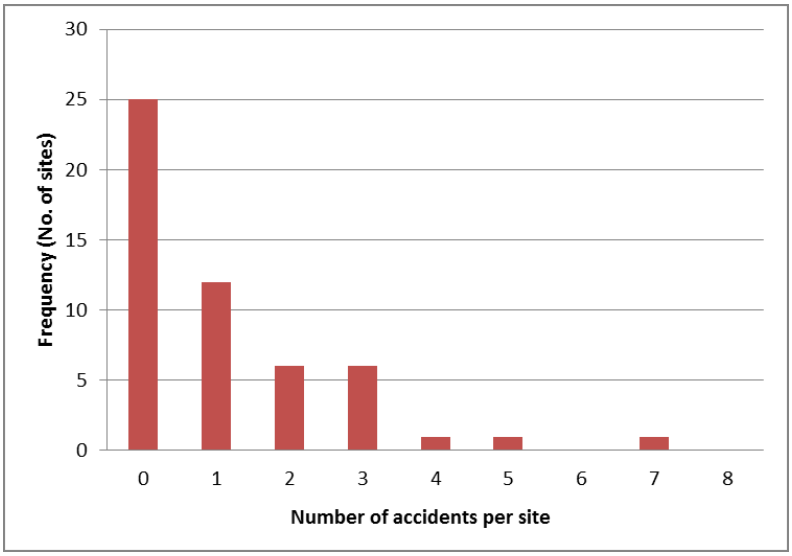

Fig. 1. Distribution of unsignalised crossings by the number of accidents in 7 years.

2.2 Daily volume models

In accident prediction modelling the basic explanatory variables are daily (24 h) traffic volumes of vehicles and pedestrians [6]. However, as these are difficult to obtain for a large number of sites, an estimation method was used. Hourly volumes of pedestrians and motor vehicles were counted at each site for three hours: 7-8, 12-13 and 16-17. In addition, at selected sites, 24-hour volume counts were carried out or obtained from other studies. Daily volumes (DV_i) for each road user type (i) were estimated by linear regression using the equation:

$$DV_i = b_{1i} \times V7_i + b_{2i} \times V12_i + b_{3i} \times V16_i \quad (2)$$

$V7_i$ = observed hourly volume (7-8) for road user type i (pedestrian, motor vehicle),
$V12_i$ = observed hourly volume (12-13) for road user type i,
$V16_i$ = observed hourly volume (16-17) for road user type i,
b_{1i} , b_{2i} , b_{3i} = regression coefficients for road user type i.

Table 1 shows the estimation results. A very good fit was obtained (adjusted $R^2 > 0.91$, all variables significant at 0.01 level) which proves that daily volumes can be estimated based on 3 hourly counts.

Table 1. Daily traffic volume regression models.

Variable	**Pedestrian**		**Motor vehicle**	
	coefficient	p-value	coefficient	p-value
V7	2.249	0.002	3.419	0.000
V12	9.100	0.000	7.058	0.001
V16	3.140	0.002	4.841	0.000
n	18		15	
Adjusted R^2	0.928		0.915	
Std error	9.0%		4.77%	

2.3 Explanatory variables

Table 2 shows explanatory variables which were considered and used in the study as well as their basic statistics. Number of accidents was used as the dependent variable.

Table 2. Variable description and statistics

Description	Variable	Mean	Std. dev.	Minimum	Maximum
No of accidents (7 years)	ACC	1.115	1.503	0.0	7.0
Daily Pedestrian Volume	DPV	1063.3	909.0	149.0	3696.0
Daily Motorised Traffic Vol.	DTV	9306.1	4696.8	2735.0	21223.0
Location at intersection/ not	INTERS	0.7692	0.4254	0.0	1.0
Public transport stop nearby	PTSTOP	0.5961	0.4954	0.0	1.0
Land use: residential/ mixed	LUSE	0.2692	0.4478	0.0	1.0
Proportion of HV	HGV	0.0918	0.0350	0.0410	0.1690
Ped. peak to off-peak ratio	PPEAK	1.5237	0.6026	0.6348	3.8000
Traffic peak to off-peak ratio	TPEAK	1.2416	0.2222	0.8154	1.8742

Three dummy variables were used to characterise the sites:

- location of the crossing at a intersection (INTERS=1) or at midblock,
- public transport stop nearby (PTSTOP=1) or not,
- land use type: residential or mixed (LUSE=1).

Three additional variables reflected the character of traffic:

- proportion of heavy vehicles (trucks and buses) in the traffic stream (HGV),
- average peak hour volume to off-peak hour volume ratio (*V7+V16*)/(2**V12*) which was calculated both for vehicles (TPEAK) and pedestrians (PPEAK).

Variables TPEAK and PPEAK were introduced to reflect differences in traffic peaking characteristics among the sites: in some cases the peaks were very pronounced, while in others peak hour volume was actually lower than the off-peak volume (see Table 2).

3 Method

3.1 Empirical Bayes Method

Accurate identification of high-risk pedestrian crossings (black spots) is possible using the Empirical Bayes method based on expected rather than observed number of accidents [6]. Expected number of accidents is a linear combination of the observed and the predicted numbers of accidents:

$$E(A) = wA_{pre} + (1 - w)A_{obs} \tag{3}$$

$E(A)$ = estimated expected number of accidents per year,
A_{pre} = number of accidents per year predicted by the accident model for similar sites,
A_{obs} = number of accidents per year at the site,
w = statistical weight:

$$w = \frac{1}{1+Y\,A_{pre}\,\varphi} \tag{4}$$

Y = number of years for which accident observations are made,

φ = overdispersion parameter associated with the accident prediction model.

The value of parameter φ is estimated together with parameters of the accident prediction function using the numbers of accidents observed at similar sites during the period of several years. Statistical weight w controls the relative importance of model predictions versus the observed number of accidents in eq. (3). It is argued [7] that if data used to calibrate the model show little dispersion (low φ value), weight w will be larger and thus more emphasis will be given to the model predictions.

3.2 Accident Prediction Models

In general, a traffic accident prediction model relates the number of accidents to a measure of exposure (traffic volume) and several variables describing characteristics of the road site (geometry, traffic control). This function is also called the Safety Performance Function (SPF) in American Highway Safety Manual [8]. To ensure that the predicted accident numbers are non-negative, a multiplicative model form is used. The general model form for intersections can be written as follows:

$$A_{pre} = \alpha Q_1^{\beta_1} Q_2^{\beta_2} e^{\Sigma \gamma_i x_i} \tag{5}$$

where:

A_{pre} = predicted number of accidents per year,
Q_1 = first daily traffic volume (e.g. major road) entering the intersection,
Q_2 = second daily traffic volume (e.g. minor road or pedestrian) at the intersection,
x_i = set of risk factors associated with the site,
α, β_1, β_2, γ_i = model parameters.

Daily volumes of pedestrians and vehicles are the main explanatory variables in the analysis. Other risk factors which may be considered include: type of control (signalization, priority junction), intersection geometry (no. of legs, no. of lanes, intersection angle), surrounding land use type (residential, commercial, industrial, mixed), presence of public transport stops, percentage of heavy vehicles, peaking characteristics, etc.

4 Results

Statistical model given by equation (5) can be calibrated by regression assuming that the underlying probability distribution is either Poisson or negative binomial (NB). NB model takes into considerations the problem of overdispersion (greater variability than predicted by the Poisson model) [9]. The estimated model coefficients and their p-values are presented in Table 3. Logarithms of both daily pedestrian volume (LnDPV) and daily motorised traffic volume (LnDTV) are statistically significant at the 1% level of significance for the Poisson model and at 5% level for the negative binomial model. The predicted number of accidents per year (A_{pre}) is a function of daily pedestrian volume (DPV) and daily motorised traffic volume (DTV):

$$\text{Poisson model 1:} \quad A_{pre} = 8.513 \cdot 10^{-8} DPV^{0.864} DTV^{0.944} \tag{6}$$

$$\text{Negative binomial model:} \quad A_{pre} = 2.748 \cdot 10^{-7} DPV^{0.791} DTV^{0.871} \tag{7}$$

The Akaike's Information Criterion (AIC) points to NB model as better fitted. On the other hand pseudo R-squared value and significance of explanatory variables are slightly better for the NB model. The Cameron-Trivedi test for overdispersion is not conclusive ($p > 0.1$), which means that Poisson distribution cannot be rejected.

Table 3. Results of model estimation.

Variable	Poisson 1		Negative Binomial	
	coeff.	p-value	coeff.	p-value
Const	-14.333***	0.0002	-13.161**	0.0237
LnDPV	0.864***	0.0000	0.791**	0.0128
LnDTV	0.944***	0.0028	0.871**	0.0452
Dispersion α			0.580	0.2118
Significance	0.00002		0.04179	
Pseudo R^2	0.130		0.029	
AIC	150.5		148.4	

significant at: ***0.01, **0.05, *0.1

It should be noted that in both models (eqs 6-7) exponents for DPV and DTV are smaller than one. This indicates that the predicted number of pedestrian accidents is less than directly proportional to both pedestrian and motorised traffic daily volumes. It follows that there is a weak "safety-in-numbers" effect. This result is different from that obtained by Elvik et al. for pedestrian crossings in Oslo [10], where a much stronger "safety-in-numbers" effect was found. However, that model included both signalised and unsignalised crossings.

In order to compare the goodness of fit of the models, patterns of cumulative residuals can be examined. According to the method proposed by Hauer and Bamfo [11], cumulative residuals can be plotted against the sum of pedestrian and vehicle traffic volumes. Such Cureplots were created for both Poisson 1 and negative binomial models. The results were very similar for both models. The cumulative residuals were within double standard error bands. Cureplot for the negative binomial model is presented in Fig. 2.

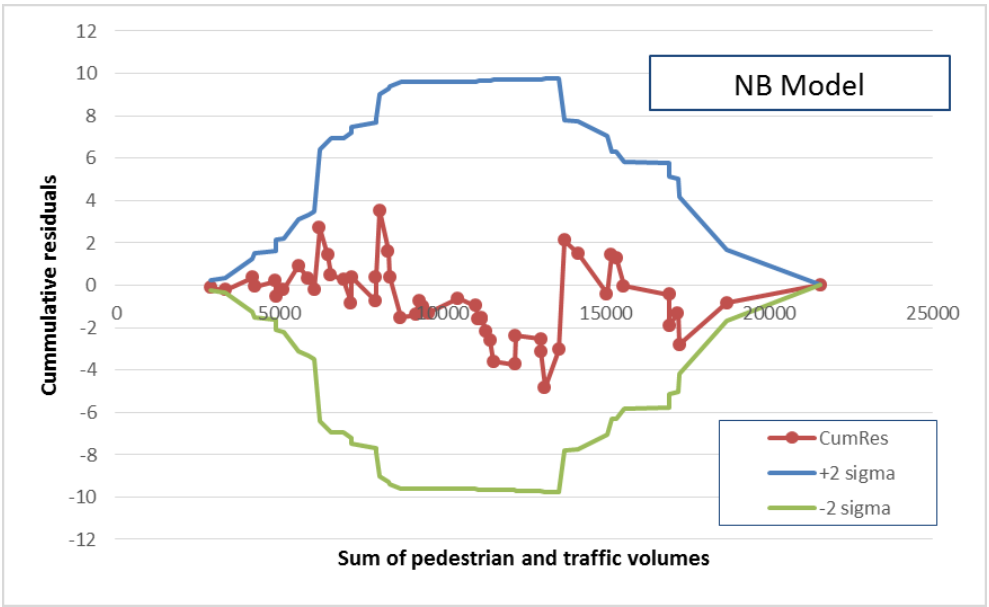

Fig. 2. Cureplot for relationship between total traffic volume (pedestrians and vehicles) and total number of accidents based on negative binomial model.

Both Poisson and NB models for unsignalised crossings perform well, based on goodness-of-fit measures and CURE plots. However, Negative Binomial model is generally preferred as it accommodates overdispersion of accidents. NB model is also more flexible and versatile.

In order to identify accident risk factors, an extended Poisson model was also estimated with all the explanatory variables presented in Table 2. The results of model estimation are shown in Table 4. Only two additional variables turned out to be significant: proportion of heavy vehicles (HGV) and land use type (LUSE). The final model is given by the formula:

$$A_{pre} = e^{-14.63} DPV^{0.90} DTV^{0.85} e^{(7.30HGV + 0.74LUSE)} \quad (8)$$

The calibrated function indicates that a "mixed" (as opposed to residential only) land use is an important risk factor: the predicted number of accidents is 2.1 times higher. Another strong risk factor is presence of heavy vehicles: with 10% HVs we get 2.1 times more accidents, with 20% HVs about 4.3 times more.

Table 4. Results of model estimation – extended model.

Variable	**Poisson 3 – full**		**Poisson 4 - extended**	
	coeff.	p-value	coeff.	p-value
Const	-14.289***	0.0006	-14.628***	0.0001
LnDPV	0.808***	0.0015	0.903***	0.0002
LnDTV	0.791**	0.0137	0.847***	0.0071
LUSE	0.999**	0.0443	0.740*	0.0600
HGV	7.328	0.150	7.300*	0.0782
INTERS	0.388	0.295		
PTSTOP	0.056	0.891		
PPEAK	-0.583	0.101		
TPEAK	0.979	0.274		
Significance	0.00009		0.00002	
Pseudo R^2	0.193		0.165	
AIC	152.0		148.7	

significant at: ***0.01, **0.05, *0.1

A joint model for both signalized and unsignalized crossings was also developed. However, the model did not perform very well: although lack of signalisation was a significant risk factor, pedestrian volume was not significant.

5 Conclusions

Accident prediction models for unsignalised pedestrian crossings, each across two lanes of traffic in the same direction, were developed based on a sample of 52 sites in Warsaw. A total of 58 pedestrian accidents were recorded there during seven years.

The recommended accident prediction model is based on negative binomial distribution. The predicted number of accidents per year is a function of daily pedestrian volume and daily vehicular traffic volume. The recommended model shows that the number of accidents is less-than-proportional to both pedestrian and motorised traffic volumes. This shows a weak "safety-in-numbers" effect. Statistically significant risk factors affecting pedestrian safety at marked crossings are: proportion of heavy vehicles, location not in a residential area and no traffic signal.

The model makes it possible to apply the Empirical Bayes method for identification of crossings with higher than expected number of accidents and thus to obtain an unbiased ranking of high risk locations. Accounting for seasonal variability (the transformation of the traffic data from DTV to AADT) and daily variability (peak ratio) can be possible future improvements. Modelling will be extended to other types of crossings (two lane roads, roads with different speed limits, etc.).

The study is part of research project InDeV sponsored by the European Commission under grant agreement No. 635895.

References

1. SEWIK – Polish database of road accidents and collisions (2017)
2. P. Olszewski, P. Szagała, M. Wolański, A. Zielińska, Pedestrian fatality risk in accidents at unsignalized zebra crosswalks in Poland, Accid. Anal. Prev. **84** pp. 83–91 (2015). doi:10.1016/j.aap.2015.08.008
3. P. Olszewski, B. Osińska, A. Zielińska, Pedestrian Safety at Traffic Signals in Warsaw, Transp. Res. Procedia. **14** pp. 1174–1182 (2016). doi:10.1016/j.trpro.2016.05.188
4. R. Elvik, State-of-the-art approaches to road accident black spot management and safety analysis of road networks (Transportøkonomisk Institutt, Oslo, 2007)
5. E. Hauer, Observational before-after studies in road safety - estimating the effect of highway and traffic engineering measures on road safety (Emerald, 1997)
6. R. Elvik, Traffic Safety - Chapter 27, in: Handb. Transp. Eng. Vol. 2, Appl. Technol. (Second Ed.), (McGraw-Hill, New York, 2011)
7. P. Olszewski, B. Osińska, P. Szagała, Road accident statistics and available analysis techniques, in: E. Polders, T. Brijs (Eds), How to Anal. Accid. Causation ? A Handb. with Focus Vulnerable Road Users, (Hasselt University, Diepenbeek, 2018)
8. Highway Safety Manual (HSM) (AASHTO, Washington DC, 2010)
9. J. Kamińska, Negative binomial distibution in accident prediction modelling (in Polish), Logistyka. **6** (2011)
10. R. Elvik, M.W.J. Sørensen, T.O. Nævestad, Factors influencing safety in a sample of marked pedestrian crossings selected for safety inspections in the city of Oslo, Accid. Anal. Prev. **59** pp. 64–70 (2013). doi:10.1016/j.aap.2013.05.011
11. E. Hauer, J. Bamfo, Two tools for finding what function links the dependent variable to the explanatory variables, in: Proc. ICTCT 1997 Conf., (Lund, Sweden, 1997)