ACCIDENT PREDICTION MODELS FOR URBAN AREAS A STATE-OF-THE-ART

Sandra Vieira Gomes*, João Cardoso*, Carmen Carvalheira**, Luís Picado Santos**

* LNEC - Laboratório Nacional de Engenharia Civil
Lisbon, Portugal
sandravieira@lnec.pt, joao.cardoso@lnec.pt

** Department of Civil Engineering, University of Coimbra
Coimbra, Portugal
carmen@dec.uc.pt, picsan@dec.uc.pt

Abstract

This paper summarizes the results of the bibliographic study on accident prediction models applied to urban areas. Several types of models where analyzed, namely for accidents involving pedestrians or cyclists, collisions between motorized vehicles, total accidents, non injury accidents, accidents with fatalities, injury accidents, night time accidents and accidents involving only vehicles. Models with different levels of disagregation were also studied: aggregated models describe general safety trends (national or regional); disaggregated models represent specific changes in the transportation system.

This study is part of the "IRUMS – *Safer Roads in Urban Areas*" project, carried out at the National Laboratory of Civil Engineering and at the Department of Engineering of the University of Coimbra, financed by the Foundation for Science and Technology. This project intends to develop methods for safety management of urban road networks. Procedures for the estimation of expected accident frequencies, identification of sites with a promise and selection of efficient corrective measures are being developed as well. The case study is being applied in Lisbon.

INTRODUCTION

According to official statistics, an important percentage of accidents and injuries are reported in urban areas: in Portugal, during the period from 2004 to 2006, 70% of all injury accidents, and 44% of all fatalities occurred inside urban areas.

To tackle this problem, municipal road administrations need tools for the quantification of safety levels and the explicit consideration of safety issues in the road management process. A better knowledge of the relations between accident frequencies and variables describing the urban road environment will allow a more efficient selection of priorities for intervention and safety funding. This can be achieved by means of accident prediction models adapted to the urban context where they are applied.

Estimating the number of accidents that may result for a given highway design is a matter of great interest to the highway engineering community. Several studies were developed in this area aiming to determine the effects of different design elements on road safety. Since safety is a primary consideration in highway design, the safety consequences of highway design features are very important.

Accident prediction models are developed to provide a realistic estimate of the expected number of accidents or victims, as a function of explanatory variables such as the traffic volumes and road geometry characteristics. They consist of relations between independent variables and accidents, through mathematical functions. This procedure allows to quantify the variation in the safety level due to changes on each considered variable. The development of these estimates is a fundamental component for safety considerations in road planning.

ACCIDENT PREDICTION MODELS

Bases for development

The development of accident prediction models must be carefully made, so that the results and interpretations that they provide are suitable, in what concerns: the choice of the exploratory variables and type of model; the specification of functional relations; the evaluation of the adjustment (validation); the causal interpretation of relations; the evaluation of the performance of the model in the forecast; and the evaluation of potential causes of errors.

Several recommendations for the development of accident prediction models were recently proposed in the RIPCORD-ISEREST project (Reurings et al, 2005):

1. The probabilistic distribution of accidents in the original data set must be identical to the one of the residual terms of the model.
2. Models must be disaggregated by level of severity (fatal accidents, accidents with victims and property-damage-only accidents), by type of road element (road section, intersections, bridges, tunnels, curves and railroad crossings) and by class of vehicles (trucks, cars, two wheelers, pedestrians and cyclists).
3. The correlations between the explanatory variables must be analyzed in detail, with justification for the functional forms chosen, as well as all the causal relations. All the variables with high correlation between them, as well as the ones that are considered confounding, must be eliminated. The possibility of omitted variable bias must be taken into account, since it is not feasible to create an accident prediction model with all the variables that influence accident occurrence.
4. The overall goodness-of-fit of the model must allow the decomposition of the variation of the number of accidents in: b) random variation, b) systematic variation explained by model, and c) systematic variation not explained by the model. This last one must be analyzed, to decide if the over-dispersion can be described by a simple parameter or if it must be modelled by a variable parameter.
5. The predictive performance of the model must be tested through its application on a data set that has not been used for its development.

Assuming one can identify all the systematic variation, it is possible to consider that road accident occurrence follows a Poisson distribution, with useful properties, namely the fact of its

expected value is equal to the variance. In case this is not possible, a dispersion parameter can be calculated, that allows to consider the omitted variables and to know whether the systematic variation of the accidents is conveniently explained or not.

Types of models

Accident prediction models can be classified in different ways (Cardoso, 2007):

- According to the technique used to estimate the effect on safety indicators: through before-after studies or through the adjustment of mathematical equations;
- According to the consideration of "time" in the model: parametric (cross-sectional models) or variable (time series models);
- According to level of disaggregation of the variables used: aggregated and disaggregated models.

Before-after studies and adjustment of mathematical equations

Although before-after studies cannot be considered real accident prediction models, they include on its methodology mathematical expressions that allow estimating the effect on safety of changes in the transport system. They are considered quite efficient as long as the disturbing factors are controlled and the sample dimension and the analysis techniques used are adequate. However, the applicability of the defined relations is restricted, since they are specific for the context where they were adjusted, a fact that imperils its generalization.

The adjustment of mathematical equations through statistical methods, allows relating data on accidents or victims with a series of explanatory variables, creating the so called accident prediction models. The main advantage of this type of models relies on the possibility for a direct use in the evaluation of the effect on the safety indicator of changes on the exploratory variables. Its usefulness increases when the number of explanatory variables is high, when the number of confounding variables is high (and they cannot be treated through the consideration of control groups) or when the sample (of accidents or victims) is small.

Cross-sectional models and time series models

Cross-sectional models allow representing the variation between variables that characterize different road entities and its level of safety, for the same instant. These models explore the variation between different entities in the same time period, relating accidents to the variation of characteristics of different entities (any geographic unit or physical element - people, vehicles or groups with similar characteristics). In the development of this type of models it is important to ensure that the road entities are similar and that all variables with influence on accident occurrence are considered. This type of models is normally applied when data sets of considerable dimension are available and when the possible explanatory variables are independent and with low co-variation.

Time series models comprise several observations of the same element in time. In this type of modelling, variations are very small, especially when time series disclose a considerable colinearity between its potential independent variables. These models generally show less correlation between successive confounding terms, associated to the fact of not being more

difficult to include all the relevant explanatory variables. The development of this type of models is considered easy, since several tools are available to deal with problems of self-correlation and auto-regression, among others (Cardoso, 2007; OECD, 1997).

Aggregated models and disaggregated models

Aggregate models allow describing general safety trends on the regional or national level, making possible the development of short-term safety evolution estimations, as a function of traffic and macro-economic variables. Estimates can be improved by including descriptive factors of the impact of safety measures. The use of this type of models, does not allow, however, to evaluate the effect of changes in parts of the transportation system, neither the global impact of safety interventions on specific groups of users, since variables are generalist.

Disaggregated models may be used to represent the effect of changes in specific parts of the transportation system or in the safety of specific user groups. They are used in the evaluation of the application of safety policies and in road safety estimations, being therefore considered a support instrument to road safety management at a macroscopic level (regional or national). The usual types of disaggregation are by transport mode, age group, sex and type of road (Cardoso, 2007; OECD, 1997).

Functional form of mathematical equations

The functional form of mathematical equations (models) can be defined through several statistical methods, as described in the following sub-chapters.

Linear regression models

A linear model expresses its systematic component as a linear function of the following parameters

$$y_i = \sum_{j=1}^{J} \beta_j x_{ji} + u_i \quad (1)$$

Where:
yi - variable dependent;
xij - variable independent;
ui - random error.

The linear models can be developed through the squared minimums estimation techniques (simple or generalized).

A simple linear regression model describes the relation between a quantitative independent variable X and a quantitative dependent variable Y, in the following terms (Guimarães et al, 1997):

$$Y_n = \alpha + \beta \times \left(X_n - \overline{X}\right) + E_n \quad (2)$$

Where:
n = index of the observation;

α, β = fixed parameters to estimate from the linear relation between X and Y

En = random error associated to the observed value Yn

Regression's theory is based on the verification of a set of hypotheses:

1. All variables must be measured, with precision and error-free;
2. Errors must have expected value null and constant variance;
3. Errors must be mutually independent, meaning that error of two sets of values of the explanatory variables shouldn't be correlated;
4. Errors are normally distributed.

A model of multiple linear regression describes a relation between a set of independent quantitative variables X_j (j=1,2...,J) and a dependent quantitative variable Y, through the expression (Guimarães et al, 1997):

$$Y_n = \alpha + \beta.\left(X_{1n} - \overline{X}\right) + ... + \beta.\left(X_{jn} - \overline{X}\right) + E_n \quad (3)$$

Where:

n = index of the variable $X_1...X_J$ and Y (n=1... N);

$(X_{1n}...X_{Jn}, Y_n)$ = n^{th} of the variable $X_1...X_J$ and Y;

$\overline{X}_j$ = arithmetic mean of the observations of the variable X_j ($\overline{X}_j = \frac{1}{N}.\sum_n X_{jn}$);

$\alpha, \beta_1, ..., \beta_j$ = fixed parameters to estimate from the linear relation between $X_1...X_J$ and Y;

E_n = associated random error to the observed value Y_n.

The underlying hypotheses to this model are identical to the ones considered in the simple linear regression models.

Generalized linear models

Generalized linear models are an extension of multiple regression linear models: the dependent variable follows a distribution from the exponential family (Normal, Poisson, Binomial, or Gamma, for example). The relation between the average value of the response variable and the explanatory variables can be established by any monotonous and differentiable function.

The general structure of a generalized linear model has three components: the systematic, the random and the linking function between the systematic component and the random component. The generalized linear models take normally the form of (Wichert et al, 2006):

$$h(\lambda_i) = \sum_j \beta_j x_{ji} \quad (4)$$

Where:

h = linking function (the expected value of the dependent variable is linked to a linear regression through a monotonous function)

β = parameters to estimate

x = explanatory variables

The main characteristic of these types of models are (Wichert et al, 2006):

- The response y is observed independently of the fixed values of the explanatory variables x_1..., x_p.
- The explanatory variables can only influence y distribution through a linear function named linear predictor:

$$\eta = \beta_1 x_1 + ... + \beta_p x_p \tag{5}$$

Where:
β = parameters to estimate
x = explanatory variables

- The distribution of y has density of the form:

$$f(y_i;\theta_i;\varphi) = \exp\left[A_i\left\{y_i\theta_i - \gamma(\theta_i)\right\}/\varphi + \tau(y_i,\varphi(A_i))\right] \tag{6}$$

Where:
φ = Scale parameter (possibly known)
A_i = Known prior weight
θ_i = Depends on the linear predictor
Therefore, the distribution of y must come from the exponential family.

- The mean μ is a smooth invertible function of the linear predictor:

$$\mu = m(\eta), \eta = m^{-1}(\mu) = l(\mu) \tag{7}$$

- Inverse function *l* is called link function, and it describes how the mean of the response variable depends on the linear predictors (explanatory variables).

The estimates are calculated through the iterative weighed least squares technique, since explicit expressions for the maximum likelihood are not usually available.

Poisson generalized Model

In the stochastic accident analysis it is usual to admit that the occurrence of accidents is controlled by a stationary process of Poisson. The justification for the consideration of this hypothesis relies not only in the good adjustment of the observed values, but also in the high number of opportunities for accident occurrence associated with a low probability of happening each one of these opportunities (Wichert et al, 2006).

In a generalized Poisson model it is assumed that the response variable follows a Poisson distribution. This type of model is preferably used in situations with small accident counts. For accident data sets of large dimension, Gauss models may be used, considering a normal distribution (Wichert et al, 2006). The model of regression of Poisson (or log-linear model) has usually the following form:

$$f(y_i \mid \mu_i) = e^{-\mu_i}.\frac{\mu_i^{y_i}}{y_i!} = \exp\left\{-e^{z_i^T\beta} + y_i z_i^T\beta - \log y_i!\right\},\ y_i = 0,1,... \tag{8}$$

Where:
y_i = independent responses modelled by a Poisson distribution;
μ_i = mean, and $\log(\mu) = z_i^T\beta$;
β = parameters to estimate.

Accident prediction models in intersections and road links

This section summarizes the result of the bibliographic study on accident prediction models applied to urban areas. Several types of models where analyzed, namely for accidents involving pedestrians or cyclists, collisions between motorized vehicles, total accidents, non injury accidents, accidents with fatalities, injury accidents, night time accidents and accidents involving only vehicles. Models with different levels of disaggregation were also studied.

Aggregated models - Regional level

Washington et al (2006) developed accident prediction models for: total accidents, fatal accidents, fatal and injured accidents, injured accidents, pedestrian accidents, cyclist's accidents, night time accidents and accidents without victims. The standard form off all models is a log linear regression model, which included general variables regarding: population (total, by age groups, by area, by means of transportation, etc), road length (total, main roads, motorways, urban/rural roads, etc), vehicles kilometres travelled, number of intersections per km, average income and number of housing units (total and per area):

$$Log(Accidents+1)=a_0+a_1\times Variable_1+a_2\times Variable_2+...$$

Disaggregated models

Turner et al (2007) developed several accident prediction models regarding specifically pedestrians in intersections (signalized intersections, roundabouts and T-junctions). They disaggregated their multiplicative models also by type of movement: crossing, left-turn, and right-turn, using traditional variables (motorized traffic volumes and pedestrian volumes):

$$Accidents_{pedestrians} = a_0 \times Traffic_1^{a_1} \times Traffic_2^{a_2} \times Pedestrians^{a_3}$$

but also specific variables associated with conflicting movements, namely: the proportion of pedestrians that cross with the "green-man", the average crossing distance and number of lanes that vehicles that turn left have to cross:

$$Accident_{pedestrians} = a_0 \times Traffic^{a_1} \times Pedestrians^{a_2} \times Crossing_{dist}^{b_3} \text{ (for example)}$$

Pedestrian accidents at intersections were also modelled by other authors, namely Brüde and Larson (1993), Maher and Summersgill (1996) and Gårder (2004) (for roundabouts). All of them used multiplicative models which included motorized vehicles and pedestrian volumes as explanatory variables:

$$Accident_{pedestrians} = a_0 \times Traffic^{a_1} \times Pedestrians^{a_2}$$

(Brüde and Larson; Maher and Summersgill)

$$Accident_{pedestrians} = a_0 \times (Traffic \times Pedestrians)^{a_1}$$

(Gårder)

Maher and Summersgill developed accident prediction models for pedestrians in urban T-junctions without median, with a desegregation of motorized traffic by major and minor roads of the intersection:

$$Accident_{pedestrians} = a_0 \times Traffic_1^{a_1} \times Traffic_2^{a_2} \times Pedestrians^{a_2}$$

Additionally they included accident prediction models for other types of accidents in the same type of intersection, namely: total accidents, property damage only accidents; and also for road links, namely: total accidents, pedestrian accidents and property damage only accidents. The general form of the mentioned models is:

$$\text{Accident}_{\text{pedestrians-road links}} = a_0 \times \text{Road Lenght} \times \text{Traffic}^{a_1} \times \text{Pedestrians}^{a_2}$$

Total accident frequency is the most common response variable in accident modelling. Several authors developed mathematical functions to explain total accident occurrence at intersections. The common explanatory variables used were also motorized vehicles and pedestrian volumes, sometimes desegregated by major and minor legs:

- Lars Leden, 2002 (intersections)

$$Total\ accidents = a_0 \times Traffic^{a_1} \times Pedestrians^{a_2}$$

- Sayed and Rodriguez, 1999 (non-signalized urban intersections controlled by STOP signs), Greibe, 2003 (urban intersections with three or four legs with and without signals) and Mountain and Fawaz, 1996 (intersections with different types of traffic control)

$$Total\ accidents = a_0 \times Traffic_1^{a_1} \times Traffic_2^{a_2}$$

(Only injury accidents for Sayed and Rodriguez and Mountain and Fawaz)

- Lord and Persaud, 2000 (signalized urban intersections with four legs) and Persaud et al (2002) (three or four legs intersections with and without signals)

$$Total\ accidents = a_0 \times Traffic_1^{a_1} \times Traffic_2^{a_2} \times e^{a_3 \times Traffic_2}$$

- Bauer and Harwood, 2000 (collisions in urban and rural intersections - four legs with STOP; three legs with STOP and four legs with signal lights)

$$Total\ collisions = e^{a_0} \times Traffic_1^{a_1} \times Traffic_2^{a_2} \times e^{a_3 \times Traffic_2 + \ldots + a_n \times Traffic_n}$$

Although a high percentage of accidents in urban areas occur at intersections, accident prediction models were also developed for road links. This was the research subject of several authors, in what concerns total accidents. Motorized vehicles traffic volumes were the basic explanatory variables used, but other variables like road length, driveway density, number of minor intersections existent, pedestrian volumes, road width, number of lanes, speed, were also used:

- Turner et al, 2003

$$Total\ accidents = a_0 \times Traffic^{a_1}$$

- Mountain, Fawaz and Jarret, 1996

$$Total\ accidents = a_0 \times Traffic^{a_1} \times Road\ length^{a_2}$$

- Bonneson e McCoy, 1997

$$Total\ accidents = Traffic^{a_0 + a_1 \times Variable_1 \times Variable_2} \times Road\ length^{a_2} \times e^{a_3 + a_4 \times Variable_3 + \ldots + a_n \times Variable_n}$$

- Abo-Qudais, 2001

$$Total\ accidents = a_0 \times Variable_i^3 \times Variable_i^2 \times Variable_i$$

$$Total\ accidents = a_0 \times Speed^2 \times Speed$$

$$Total\ accidents = a_0 \times Traffic^{a_1}$$

- Greibe, 2003

$$Total\ accidents = a_0 \times Traffic^{a_1} \times a_2 \times ... \times a_n$$

FINAL NOTES

This paper summarizes the results of the bibliographic study on accident prediction models applied to intersections and road links in urban areas. APM are mathematical functions that describe the relation between the road safety and explanatory variables, as traffic, road length and width, number of intersections, etc. Its common form is expressed as the following multiplicative expression: $A = \alpha \times T_1^{\beta} \times T_2^{\beta} \times e^{\sum \gamma_i . x_i}$, where A is the expected value of the number of accidents, which varies with the traffic volume (T) and with other factors (Xi). The effect of traffic in the accident's occurrence is modelled through the power ß. The effect of the several risk factors that usually influence accident probability is modelled through an exponential function of base *e* and raised to the sum of the product of the γ_i coefficients by the risk factors, x_i.

In what concerns disaggregated models for intersections, the use of variables such as traffic volume, and for accidents with pedestrians, the average distance crossed or lane with are frequently used, which reveals its high significance in the explanation of these phenomenon. Regarding road links, traffic volumes, road width, speed and road length were the most frequent significant explanatory variables used to model road accidents.

Considerable progress has been made in the techniques for establishing the relationship between accidents, traffic volumes and road geometry. Specific problems such as low mean value, overdispersion, disaggregation of data over time and random errors were already identified by several authors; the solutions they outlined will help the development of accident prediction models for Portuguese urban areas, calibrated with data from the city of Lisbon.

REFERENCES

Abo Qudais S. (2001). Urban road accident prediction models. Roads-Routes, AIPCR, World Road Association, n. 309-I.

Bauer, K. M.; Harwood, D. W. (2000). Statistical models for at-grade intersection accidents – Addendum. FHWA-RD-99-094 Report, U.S. Department of Transportation, Federal Highway Administration.

Bonneson, J. A.; McCoy, P. T. (1997). Effect of median treatment on urban arterial safety; An accident prediction model. In: Transportation Research Record nº 1581, p. 27-36.

Brüde, U.; Larson, J. (1993). Models for predicting accidents at junctions where pedestrians and cyclist are involved. How well do they fit? – Accident Analysis and Prevention, Vol. 25, No. 5, p. 499-509.

Cardoso, J. L. (2007). Métodos racionais de apoio à intervenção da engenharia em segurança rodoviária. LNEC, Lisboa.

Gårder, P.E (2004). The impact of speed and other variables on pedestrian safety in Maine – Accident Analysis and Prevention, Vol. 36, p. 533-542.

Greibe, P. (2003). Accident prediction models for urban roads. In: Accident analysis and prevention, 35, p. 273-285.

Guimarães, R. C.; Cabral, J. A. S. (1997). Estatística – McGraw-Hill.

Leden, L. (2002). Pedestrian Risk decreases with pedestrian flow. A case study based on data from signalized intersections in Hamilton, Ontario – Accident Analysis and Prevention, Vol. 34, p. 457-464.

Lord, D.; Persaud, B.N. (2000). Accident prediction models with and without trend. Application of the generalized estimating equations procedure – In: Transportation Research Record 1717, p. 102-108.

Maher, M. J.; Summersgill, I. (1996). A comprehensive methodology for the fitting of predictive accident models – Accident Analysis and Prevention, Vol. 28, no. 3, p. 281-296.

Mountain, L.; Fawaz, B.; Jarret, D. (1996). Accident prediction models for roads with minor junctions. In: Accident Analysis & Prevention, 28 (6), p. 695-707.

Mountain, L.; Fawaz, B. (1996). Estimating accidents at junctions using routinely-available input data – Traffic Engineering &Control, 37 (11), p. 624-628.

Persaud, B.; Lord, D.; Palmisano, J. (2002). Calibration and transferability of accident prediction models for urban intersections – In: Transportation Research Record 1784, p.57-64.

Reurings, M.; Janssen, T.; Eenink, R.; Elvik, R.; Cardoso, J.; Stefan, C. (2005). Accident Prediction Models and Road safety Impact Assessment: a state-of-the-art. European project RIPCORD-ISEREST.

OECD (1997). Road safety principles and models: Review of descriptive, predictive, risk and accident consequence models.

Sayed, T.; Rodriguez, F. (1999). Accident prediction models for urban unsignalized intersections in British Columbia – In: Transportation Research Record 1665, p. 93-99.

Turner, S.; Durdin, P.; Bone, I.; Jackett, M. (2003). New Zealand accident prediction models and their applications. In: Transport: our highway to a sustainable future: proceedings of the 21st ARRB and 11th REAAA Conference, Cairns, Queensland, Australia, 18-23 of May.

Turner, S. A.; Roozenburg, A. P.; Francis, T. (2006). Predicting Accident Rates for Cyclists and Pedestrians – Land Transport New Zealand Research Report 289, Christchurch, New Zealand.

Washington, S.; Meyer, M.; Schalkwyk, I.; Dumbaugh, E.; Mitra, S.; Zoll, M. (2006). Guidance: Incorporating safety into long-range transportation planning. National Cooperative Highway Research Program Report 546, Transportation Research Board of the National Academies. Washington, D. C.

Wichert, S.; Cardoso, J. L. (2006). Accident prediction models for Portuguese motorways. WP2 – Projecto Europeu RIPCORD-ISEREST.