How to formulate an accident prediction model for urban intersections

A. Granà
Department of Road Infrastructure Engineering, Palermo University, Italy

Abstract

It is generally accepted that accident rates tend to be higher at intersections than on through sections of a road. This is particularly frequent in urban area where roads are characterized by intersections in close succession; moreover, the safe and effective operations of the urban road system can be significantly affected by safety conditions at intersections.

In this paper models and methods designed to understand and to predict the accident process at urban intersections are reviewed. In particular, this study intends to show why the statistical modelling approach is useful for accident analysis and how it can be applied to provide some general advice for conducting safety evaluations with accident data.

An exploratory example describes how to formulate an accident model for urban intersections by the analysis of significant explanatory variables affecting accident phenomenon and by the modelling of accident and traffic data for urban intersections.

Finally, in view of the recent great interest in the safety problems of urban roads with particular regard to intersections, the research intends to summarize the main features of the accident intersection models and their part in developing quantitative safety effectiveness measures for installation design improvements.

Keywords: accident, intersection, predictive models.

1 Introduction

The development of road safety principles and models as a branch of learning has happened in recent years and it now demands priority with both an innovative approach to scientific research on road safety and a systematic approach to implementing road safety countermeasures. Road safety researchers

WIT Transactions on The Built Environment, Vol 107, © 2009 WIT Press
www.witpress.com, ISSN 1743-3509 (on-line)
doi:10.2495/UT090521

are so called to describe suitably safety problems, but this requires one to revise the theories and the models used up to now; moreover, they have to be able to select among various models (also used in other field) the best way in which the specific unsafe situation can be treated through statistical accident modelling.

Researchers, as well as potential users and practitioners, have to deal with different tasks in most safety works. First of all they have to describe the present situation, collecting accident data (i.e. number of accidents, injuries and fatalities, geometric design features and other factors that are the consequences of the risk characterizing road traffic) from different data sources (i.e. police reports, hospital and insurance company statistics). Several efforts should be made to use the best information from sources of reference to evaluate accident data.

Another task regards the definition of exposure measurements specifically for road safety issues: traffic counts, travel habit surveys, local exposure measurements and fuel consumption. This is important because the risk is the relationships between accidents and exposure in terms of magnitude of activities generating road safety problems. Moreover accidents and exposure measures can be expressed in different ways. So, the term risk may be correlated to the units used in a specific study and it has to be used with attention, particularly when comparisons are discussed.

A key task is to show how the presence of uncertainties in traffic data used in describing the road safety conditions can influence the results and their interpretation. Uncertainties and inaccuracies can be caused by many reasons: the type of data sources, underreporting and misclassification in the collection of data and the time lag between the processing and the reporting of information, which may vary for different data. Table 1 shows categories of factors which may influence aggregate accident data.

Multivariate statistic models (i.e. econometric modelling), as well suited for accident analysis as they are for economics, can be used both to explain the effect of the systematic variable and to eliminate the effects of the first four factors. According to Hauer accident occurrence is best modelled using a multivariate statistic model [1]. By way of these methods it is often possible to evaluate the effect of a countermeasure.

The above considerations highlight that the interpretation of the traffic accident phenomenon requests the correct formulation of a predictive model in order to models are pictures of the complex reality characterizing road accidents.

In this paper, in view of the recent great interest in safety problems of urban roads with particular regard to intersections, an excursus of models and methods designed to understand and to predict the accident process are reviewed and discussed. In particular this study intends to show why the statistical modelling approach is useful for accident analysis and how it can be applied to provide some general advices for conducting safety evaluations with accident data.

By means of the formulation of an accident model for urban intersections, the analysis of significant explanatory variables affecting accident phenomenon, as well as modelling of traffic accident data for these infrastructures, is illustrated.

Table 1: Categories of factors influencing accident counts.

Categories of factors	description
autonomous factors	Determined outside the (national) social system: the weather, the natural endowment, the population size and structure, the state of technology and other factors that can hardly be influenced (not in the very long term) by any (single) government.
socio-economic conditions	Subject to political intervention, but rarely with the explicit purpose of promoting road safety.
the size and structure of the transportation sector	Not usually intended as an element of road safety policy: road infrastructure, public transportation, level-of-service, overall travel demand, modal choice, fuel and vehicle tax rates, size and structure of vehicle park, penetration rates.
the system of data collection	Accident underreporting and changes in the reporting routines can produce fictitious changes in the accident counts.
randomness	Inexplicable source of variation particularly prominent in small accident counts; for larger accident counts, the law of large numbers is prevailing and produces (in analogy with the dice game) an astonishing degree of long-run stability.
accident countermeasures	Measures intended to reduce the risk of being involved or injured in a road accident.

Finally, the research intends to summarize the main features of the accident intersection models and their contribution in developing quantitative safety effectiveness measures.

2 Review of accident models

Predictive accident models can be differentiated in cross-sectional models and in time-series models. Cross-sectional models consider the (spatial) variation between different entities observed at the same time. It notes that an “entity” could be a geographically defined unit, or an identifiable physical or institutional object (i.e. a person, a family, a company, a vehicle, a car make, or a group of such units with specific common characteristics). In time-series modelling the unit of observation is a period in time (hour, day, month, year.); this approach involves repeated observations of the same physical or institutional object.

In cross-sectional analysis data, sets are often characterized by lots of variation without strong covariation between the independent variables. Cross-sectional accident models assume that only the variables entering the model explain differences in the units of observations.

In time-series modelling, the units of observation may differ because of little variation, because time series data sets tend to show considerable collinearity between potential regressors. Moreover, time series models show correlation between successive disturbance terms because all the relevant variables have not been included in the set of regressors.

Lots of highly specialized time series analytical techniques have been developed to tackle autocorrelation (i.e. the dependent variable is a function of previous representations of itself). Compared to cross-sectional studies, in time-series modelling these techniques allow one to discuss the autocorrelation as additional information to be exploited. Different estimates on the same parameter can be obtained through models based on cross-sectional or on time-series data sets. Different time horizons can explain the difference in estimates: time-series models provide estimates of short term effects; the cross-sectional models provide parameters with a long term effect interpretation. In many cases it is not obvious which approach provides the correct outcome and combinations of cross sections and time series data sets can represent a good source of information.

In multivariate statistics the linear regression model is the functional form where the systematic part is a linear function of the parameters, but it is not necessarily linear in the variables:

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

where:

$y_i =$ dependent variable;

$x_{ji} =$ independent variables, or a transformation of a set of independent variables (logarithmic, quadratic, cubic, or trigonometric functions);

$u_i =$ random error term.

Regressors should primarily be chosen starting from the theory used and the question to be answered; the multiple correlation and curve fitting ambition may lead to a good fit of data but to results with little value in terms of understanding and almost impossible to generalise outside the specific sample used. Two main methods can be used for estimating the parameters in the model: the least squares method and the maximum likelihood method.

In presence of autocorrelation the covariance between the error terms related to different time points is non-zero; moreover, heteroskedastic (i.e. in statistics a measure that refers to the variance of the errors over the sample) is present if the error variance is not constant across the sample. Autoregression means that the dependent variable depends on previous representations of itself:

$$y_t = \alpha_1 y_{\text{t-1}} + \alpha_2 y_{t\text{-}2} + \ldots + \alpha_p y_{t\text{-}p} + x_t + u_t \tag{2}$$

where y_t is p^{th} order autoregressive if $\alpha_p \neq 0$ and a $\alpha_j = 0 \ \forall \ j > p$.

The typical framework of the generalised linear models is usually represented as follows:

$$h(\lambda_i) = \sum_{j} \beta_j x_{ji} \tag{3}$$

in which h is the link function: i) monotonic functions link the expected value of the dependent variable to a linear regression term; ii) the disturbance term

(y_i - λ_i) can be very close to any one of the so-called exponential family of probability distributions.

Literature counts now a large number of examples on applications of this methodology to accident analysis: Poisson and negative binomial models can fit into category of count data models. To explain the development of aggregate exposure, accidents and their severity over time another category of models was developed: DRAG model [2] taking into account a very large number of explanatory variables. To control the linearity assumption usually included in a regression model a DRAG model uses Box-Cox-transformations. A special case of the Box-Cox regression model is Log-linear model. The current formulation for Log-linear model is:

$$\ln\left(y_t\right) = \sum_{j=1}^{J} \beta_j x_{tj} + u_t \tag{4}$$

in which the logarithm of the dependent variable is a linear function of the coefficients and the error term. This model is different from the Poisson specification, where the disturbance term is multiplicative rather than additive.

Poisson models are often referred to analyze accidents (i.e. as count data models) because the dependent variable - following the Poisson distribution or its generalization - is a non-negative integer (or a count variable).

In general, the (generalized) Poisson model is suitable to analyze small accident counts. The distribution within the Poisson modelling framework is normal. So for large accident counts Gaussian models (i.e., normally distributed disturbance terms) can be applied.

So to formulate an accident model the analyst has to have a good notion of the nature of the probability distribution governing the random “disturbance” term, because the efficiency of alternative estimation techniques can depend on the distributional characteristics of this term. This problem together with the choice of the general functional form of the model, the determination of the set of explanatory (independent) factors, and the estimate the parameters entering the function will be discuss in the following paragraph.

3 Determination of an accident prediction model

An accident prediction model, or a Safety Performance Function, is a mathematical model that predicts estimates of expected accident frequency for a given entity (i.e. a road section, an intersection). The model - an equation or a set of equations - links the expected accident frequency to measurable road traits: e.g. traffic volume and roadway geometries (lane width, number of lanes, etc). The determination of these models, directed to summarize the previous knowledge on safety of entities similar to those considered, represents a critical component in the consideration of safety in road design and in safety evaluations.

The expected number of accidents is not a constant but it varies with site and time: this variation attributable to causal factors is systematic. Two components split the total variation in accident numbers: systematic and random variation:

$$\mathrm{var}\ (y) = E\ [\mathrm{var}\ (y|\mathbf{x})]+\mathrm{var}\ [E(y|\mathbf{x})] \tag{5}$$

where the first term is the random variation and the second term is the systematic variation.

In multivariate analysis a linear model specifies the relationship between a response variable Y and a set of explanatory variables, as follows:

$$Y = b_0 + b_1X_1 + b_2X_2 + ... + b_kX_k \tag{6}$$

where b_0 is the regression coefficient for the intercept and the b_i are the regression coefficients computed from the data.

A simple linear equation is not able to summarize lots of relationships because the dependent variable may have a non-continuous distribution, and the predicted values should also follow the respective distribution (i.e. any other predicted values are not logically possible). Moreover, the effect of the predictors on the dependent variable may not be linear in nature.

A generalized linear model differs from the general linear model in two cases: i) the distribution of the dependent or response variable cannot be continuous (it can be binomial, multinomial, or ordinal multinomial; ii) the dependent variable values are predicted from a linear combination of predictor variables, which are connected to the dependent variable by a link function. The general linear model for a single dependent variable can be considered a special case of the generalized linear model [3].

In the general linear model a response variable is linearly associated with values on the X variables:

$$Y = (b_0 + b_1X_1 + b_2X_2 + ... + b_kX_k) + e \tag{7}$$

where *e* is the error variability and the expected value of *e* is assumed to be 0.

The relationship in the generalized linear model is:

$$Y = g\ (b_0 + b_1X_1 + b_2X_2 + ... + b_kX_k) + e \tag{8}$$

where *e* is the error, and *g(...)* is a function. Formally, the inverse function of *g(...)* is called the link function. Table 2 shows that various link functions can be chosen, depending on the assumed distribution of the y variable values.

Table 2: Link functions.

Normal, Gamma, Inverse normal, and Poisson distributions	Identity link: *f(z) = z* Log link: f(z) = *log(z)* Power link: *f(z) = z^a*, for a given a
Binomial, and Ordinal Multinomial distributions	Logit link: *f(z)= log(z/(1-z))* Probit link: *f(z)=invnorm(z)* where invnorm is the inverse of the standard normal cumulative distribution function. Complementary log-log link: *f(z)=log(-log(1-z))* Log-log link: *f(z)=-log(-log(z))*
Multinomial distribution:	Generalized logit link: $f(z_1|z_2,...,z_c) = log(x_1/(1-z_1-...-z_c))$ where the model has c+1 categories.

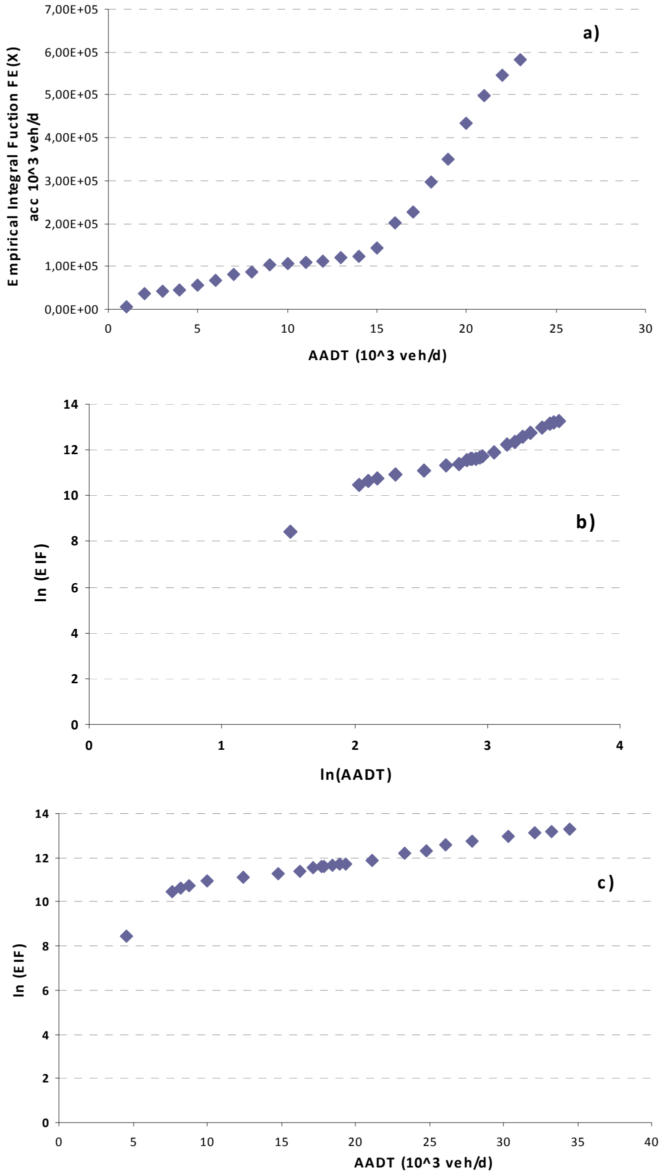

Figure 1: Functional form: a) EIF for data; b) power function F_1; c) exponential form F_2.

Development of the accident prediction models involves some determining steps: i) choice of the explanatory variables (regressors), and eventually their grouping; ii) search of the best functional form, that is, how variables should enter into the model; iii) estimate of regression parameters. This process also needs to know the nature of the probability distribution governing the random variation because the efficiency of the estimate technique depends on the distributional characteristic of this random term.

The Integrate-Differentiate method [4] has been proposed to recognize a suitable functional form for the model behind the empirical integral function; this is when the scatterplot of data is not discernible. In accordance with the ID method it assumes a link between the expected accident frequency and the total entering annual average daily traffic as exploratory variable [5]. The Empirical Integral Function (EIF) allows one to estimate the effective integral function of the functional relationship searched for (see Figure 1a). Figures 1b) and 1c) report the *ln(EIF)* as a function of ln(AADT) and of AADT. The ID method has showed that power function and exponential function can be assumed to represent the functional form of the model.

So the possible models are the following: $F_1 = \alpha\, AADT^{\beta}$ and $F_2 = \alpha\, e^{\beta AADT}$.

Table 3 reports the data fit only for the model F_2 as an example. The analogous data fit for model F_1 has not reported because it was less significant.

Least squares principles and (quasi) maximum likelihood method allow parameters to be estimated. In large samples maximum likelihood estimation method is efficient. On the contrary, the least squares methods can be applied even if the probability distribution is not specified (but the efficiency varies with the true probability distribution). In general, the aim of developing an accident model is not to obtain the best goodness of fit; moreover the goodness of fit is not a measure of model performance [6].

In presence of temporal correlation within responses an additional effort to consider the suitable data correlation structure has to be requested.

In these cases, the need to use non-traditional calibration procedures allows better estimates of unknown parameters [7].

4 Final considerations

Several safety prediction models and methods have been developed to estimate the relationship between the expected accident frequency and various urban intersection geometry and operational attributes (i.e. number of lanes, number of arms, functional classification of the major and/or the minor streets, etc.).

Table 3: Data fit for $F_2 = \alpha e^{\beta\, AADT}$.

Parameter	estimate	s.e.	t(*)	t pr.	antilog of estimate	R^2
Constant	1.35	0.145	9.31	<.001	3.843	*0,6*
AADT_1000	0.069	0.0056	12.42	<.001	1.072	
Distribution: Poisson; Link function: Log; Fitted terms: Constant, AADT_1000 *Note that s.e. are based on a dispersion parameter fixed at value 1.*						

The development of an accident model is a demanding task, particularly when different factors concur and several procedures have been designed to settle many safety problems.

The paper, absolutely not exhaustive, refers briefly starting from current literature on the specific topic some problems having to be faced to understand (and to predict) the accident process at intersections also through an explorative example. Moreover it intends to underline the important contribution of statistical modelling to the development of techniques for estimating the safety benefits of alternative designs that are not yet available at an advanced level.

References

[1] Hauer Ezra (2002). *Observational before - after studies in road safety*. Pergamon/Elsevier Science.

[2] Gaudry M., Fournier F. and R. Simard (1995). *Un modèle économétrique appliqué au kilometrage, aux accidents et à leur gravité au Québec: Synthèse des résultats*. Société de l'assurance automobile du Québec.

[3] McCullagh P. and J. Nelder (1983). *Generalized linear models*. Chapman and Hall, New York.

[4] Hauer, E. and J. Bamfo (1997). *Two Tools for Finding What Function Links the Dependent Variable to the Explanatory Variables*. International Cooperation on Theories and Concepts in Traffic Safety Conference, Lund, Sweden, 1997.

[5] Kaub. A. R. and Kaub J.A. (1999). *Predicting annual intersection accidents with conflict opportunities*. TRB Circular E-C019: Urban Street Symposium, pp. 1-12.

[6] Road Transport Research. *Road safety principles and models*. OECD/GD(97)153.

[7] Lord, D. and B. N. Persaud (2000). Accident Prediction Models With and Without Trend: Application of the Generalized Estimating Equations Procedure. In *Journal of the Transportation Research Board*, No. 1717, TRB, National Research Council, Washington, D.C., pp. 102–108.