Accident Analysis and Prevention 56 (2013) 82–94

Contents lists available at SciVerse ScienceDirect

Accident Analysis and Prevention

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

Methodology for fitting and updating predictive accident models with trend

Richard D. Connors[a,*], Mike Maher[a], Alan Wood[b], Linda Mountain[b], Karl Ropkins[a]

[a] *Institute for Transport Studies, University of Leeds, UK*
[b] *School of Engineering, University of Liverpool, UK*

ARTICLE INFO

Article history:
Received 4 July 2012
Received in revised form 8 January 2013
Accepted 7 March 2013

Keywords:
Predictive accident models
Negative binomial
Overdispersion
MCMC methods
Prediction uncertainty

ABSTRACT

Reliable predictive accident models (PAMs) (also referred to as Safety Performance Functions (SPFs)) have a variety of important uses in traffic safety research and practice. They are used to help identify sites in need of remedial treatment, in the design of transport schemes to assess safety implications, and to estimate the effectiveness of remedial treatments. The PAMs currently in use in the UK are now quite old; the data used in their development was gathered up to 30 years ago. Many changes have occurred over that period in road and vehicle design, in road safety campaigns and legislation, and the national accident rate has fallen substantially. It seems unlikely that these ageing models can be relied upon to provide accurate and reliable predictions of accident frequencies on the roads today. This paper addresses a number of methodological issues that arise in seeking practical and efficient ways to update PAMs, whether by re-calibration or by re-fitting. Models for accidents on rural single carriageway roads have been chosen to illustrate these issues, including the choice of distributional assumption for overdispersion, the choice of goodness of fit measures, questions of independence between observations in different years, and between links on the same scheme, the estimation of trends in the models, the uncertainty of predictions, as well as considerations about the most efficient and convenient ways to fit the required models.

1. Introduction

Reliable predictive accident models (or safety performance functions) have a wide variety of uses in traffic safety analysis and modelling. For scheme appraisal, when it is necessary to consider the likely effects of alternative transport proposals, this includes the effect on accidents. For example, PAMs can be used in the design of junctions to estimate the effects of any proposed design on safety as well as on operational measures such as capacity or average queues and delays. In trying to identify sites in need of remedial treatment, rather than focus on sites with the highest number of accidents in recent years, it is more efficient to compare the observed number of accidents with the number expected from a site of that type, carrying that amount of traffic. In order to estimate the effectiveness of any treatment, it is natural to carry out before and after comparisons of the accident frequencies. However, simple comparisons are known to suffer from the regression to mean effect that, if not corrected for, can lead to exaggerated estimates of the treatment effectiveness. One way to overcome this problem is through the use of the empirical Bayes (EB) method, which requires a reliable PAM (see, for example, Mountain et al., 2005; Persaud and Lyon, 2007). The widespread importance of PAMs is therefore clear; meanwhile the availability of high quality models is rather less certain.

A PAM is derived, for any given type of site, by the fitting of a regression model using data from a large number of such sites. These models relate the expected number of accidents at a site to the flows passing through the site and, possibly, to variables that describe the design, or geometry of the site. In the case of the UK, following a review by Satterthwaite (1981), the Transport Research Laboratory (TRL) carried out a series of large-scale studies for various junction and link types in the 1980s and 1990s, starting with 4-arm urban traffic signals (Hall, 1986) and 4-arm roundabouts (Maycock and Hall, 1984). The models were at various levels of detail, from models relating total accidents to an overall measure of total flow, through to models for specific accident types in terms of relevant flows and various design variables. These models are widely-used in the UK for scheme appraisal, being incorporated in software such as ARCADY, PICADY and OSCADY for the design of roundabouts, priority junctions and signalised junctions respectively.

These TRL studies were amongst the first to recognise the need to model *overdispersion* and to propose the use of a negative binomial (NB) error structure in the regression modelling instead of the Poisson. It has since been widely recognised that a pure

* Corresponding author. Tel.: +44 (0)113 3435325.
E-mail address: R.D.Connors@its.leeds.ac.uk (R.D. Connors).

0001-4575/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.aap.2013.03.009

Poisson regression model is inadequate and generally underestimates the number of sites with zero accidents. Hauer (2001) gives a good explanation of why and how overdispersion occurs, concluding that the root cause of overdispersion is that “entities with the same represented traits have different means”: that is, there are omitted variables: factors affecting the site mean that are not included in the fitted model. Lord et al. (2005) propose a fundamental mechanism of the occurrence of crashes as a Bernoulli process where the (small) probability of a crash by any vehicle passing through the site varies due to heterogeneity of the site risk and driver behaviour. Through a set of simulation experiments they demonstrate that only when conditions are homogeneous does the Poisson model provide a good fit; in heterogeneous conditions there are generally excess zeroes and the fit is poor. They conclude that the root causes of excess zeroes, and overdispersion, are “(1) spatial or time scales that are too small; (2) under-or mis-reporting of crashes; (3) sites characterised by low exposure and high risk; and (4) important omitted variables describing the crash process” and that a negative binomial regression model can provide a superior fit to the Poisson.

The assumption of a negative binomial error structure has, then, become commonplace in accident modelling, though primarily for mathematical convenience. Indeed, it has been demonstrated that other error structures are equally plausible (see Maher and Mountain, 2009; Lord and Mannering, 2010) and possibly more appropriate. Modern statistical techniques and software have mostly overcome the need to restrict attention to the NB distribution for modelling overdispersion.

However, perhaps the most serious problem in the use of these models is the passage of time since they were developed and the data on which they were based was collected. Over these decades there have been changes in both road and vehicle design, in safety initiatives and legislation and in driver training, so that the relationships between expected accidents and the explanatory variables may well have changed. For example, the PAMs for 4-arm roundabouts are based on data from 1974 to 1979 (Maycock and Hall, 1984), and those for rural priority junctions on data from 1979 to 1983 (Summersgill et al., 1996). In the UK, the annual number of personal injury accidents fell by 30% from 1985 to 2009, whilst the annual total traffic (in veh-kms) increased by 61% (DfT, 2010a,b).

While it seems unlikely that the PAMs still in use but derived using data from 20 to 30 years ago should provide accurate predictions now, it is not necessarily practicable to repeat the large and expensive data collection and model development exercises carried out by TRL that would be required to derive entirely new models. A more frugal approach is to see how existing models may be updated rather than disposed of. This updating may be by re-calibration (that is by application of a simple multiplicative scaling factor), or by re-fitting the model parameters (that is, using the same explanatory variables and function form), and may also involve the inclusion of a trend term. This is the objective of the present research study, of which this paper is one part. To achieve this, a new database has been compiled containing recent data on accidents, flows and geometric design parameters. In this paper we use data for modern rural single carriageway roads.

2. Data

The database comprises 341 rural single-carriageway links distributed amongst 73 schemes, situated in various counties across England. A *scheme* refers to the largest structure studied, and is a section of road with similar flow characteristics, between two *major junctions* (where the traffic flow on the scheme has to give way). Each scheme is partitioned into *minor junctions* (defined as any other junction properly marked with a give way or stop line and a centre line on at least one junction arm), and *links* (the section of road between any two junctions). Most of the schemes were analysed across a five year period (2005–2009), with annual accident frequencies obtained from the STATS19 database or from local authorities, and annual flow measures from the DfT or local authorities. The STATS19 data provides OS grid references that allow crashes to be located to an accuracy of 10 m in either direction. Partial confirmation of location is provided by the text description of the road names/numbers, and Google Earth was then used to manually check the coordinates of each crash site.

The total length of the 341 links was 310 km, with lengths ranging from 0.01 km to 3.9 km There was a total of 996 accidents giving an average of 2.92 accidents per link, or 3.21 per km, over the five years. The flows (measured in two-way AADTs) ranged from 2887 to 42520, with a mean of 13,590. Virtually all links had a carriageway width of less than 9 m; 41% had a hardstrip; the mean bendiness (degs turned per km) was 45, with a standard deviation of 58; the hilliness (metres gained/lost per km) had a mean of 21, and a standard deviation of 14; and the mean access density (per km) was 43 and a standard deviation of 4.9. The total length of links and the total number of accidents in the data were 60% and 71% respectively of the totals in the corresponding TRL study from which the model in the next section was developed. Further details of the data gathered and comparisons with the data used in the original TRL studies can be found in Wood et al. (2013).

3. The TRL models

Similar methodological issues arise when fitting PAMs for each type of junction, link or scheme. For simplicity we restrict attention here to models for the total number of accidents on rural single carriageway links. One of the simpler TRL models for rural single-carriageway links (see Walmsley et al., 1998) has the expected number of accidents μ_i at site i over a period of T years given by:

$$\mu_i = aTQ_i^{\alpha} L_i \exp\left(\frac{2b}{L_i}\right) \tag{1}$$

where L_i is the link length (in km), and Q_i is the flow (two-way AADT in thousands). The parameter estimates obtained by TRL were: $a = 0.0552$, $\alpha = 0.831$, $b = 0.0576$. The exponential term is intended to account for any “spillover” effects from the junctions at the two ends of the link; the junction density is approximately $2/L_i$ (accidents occurring within 20 m of the junction, as determined by the police officer attending the accident, were excluded). However, the form of this correction term is not ideal as it tends to infinity as L_i tends to zero. For a link of length 20 m the correction term has the effect of multiplying the predicted number of accidents by 317; whilst for a length of 50 m, the factor is 10. The TRL data presumably did not include any short links, and hence TRL cannot have realised the effect of this term on short links. Our data set includes seven links that are less than 50 m in length, so these are excluded from the data in our analyses. Other researchers may use a different cut-off value to exclude excessively short links.

4. Aims of the study

Our objective was to decide how best to adjust the existing TRL model to allow it to be used as a predictive tool for modern data collected from a different set of sites. At its simplest the adjustment could be by re-calibration: that is, by application of a scaling factor so as to modify the value of a in (1), to take account of long-term trend since the original models were fitted, whilst keeping other parameter values and the functional form the same. The recommended method of re-calibration in the US Highway Safety Manual (AASHTO, 2010) is to apply the existing model to each site in the new data to obtain a predicted number of accidents, and then to