Pergamon

Accid. Anal. and Prev., Vol. 28, No. 3, pp. 281–296, 1996

0001-4575/96 $15.00 + 0.00

0001-4575(95)00059-3

A COMPREHENSIVE METHODOLOGY FOR THE FITTING OF PREDICTIVE ACCIDENT MODELS

MICHAEL J. MAHER* and IAN SUMMERSGILL

Transport Research Laboratory, Crowthorne, Berkshire RG11 6AU, U.K.

(Accepted 16 August 1995)

Abstract—Recent years have seen considerable progress in techniques for establishing relationships between accidents, flows and road or junction geometry. It is becoming increasingly recognized that the technique of generalized linear models (GLMs) offers the most appropriate and soundly-based approach for the analysis of these data. These models have been successfully used in the series of major junction accident studies carried out over the last decade by the U.K. Transport Research Laboratory (TRL). This paper describes the form of the TRL studies and the model-fitting procedures used, and gives examples of the models which have been developed. The paper also describes various technical problems which needed to be addressed in order to ensure that the application of GLMs would produce robust and reliable results. These issues included: the low mean value problem, overdispersion, the disaggregation of data over time, allowing for the presence of a trend over time in accident risk, random errors in the flow estimates, the estimation of prediction uncertainty, correlations between predictions for different accident types, and the combination of model predictions with site observations. Each of these problems has been tackled by extending or modifying the basic GLM methodology. The material described in the paper, then, constitutes a comprehensive methodology for the development of predictive accident models.

Keywords—Generalized linear models, Poisson, Negative binomial, Prediction

1. INTRODUCTION

The relationship between traffic accidents and traffic volumes has been the subject of investigation for many years. Satterthwaite (1981) carried out an extensive review of work in this area, listing, even at that time, over 80 previous papers on the subject. In the last decade or so, there have been several studies whose aim has been to establish relationships between traffic accidents and road geometry, as well as traffic volumes, with the purpose of determining the effect of road and junction design on the frequency of accidents. For example, Zegeer et al. (1987, 1990), Okamoto and Koshi (1989), Joshua and Garber (1990), Miaou et al. (1992), Miaou and Lum (1993) and Miaou (1994) have used a variety of regression-based methods to establish relationships between vehicle accidents and the geometric design of road sections, such as road width, horizontal curvature, gradient etc., whilst in a continuing series of studies the U.K. Transport Research Laboratory (TRL) has also established relationships between accidents and design at various forms of road junctions (Maycock and Hall 1984; Pickering et al. 1986; Hall 1986).

The early work used multiple linear regression modelling, with its assumption of normally distributed errors and homoscedacity, but there has been a steady realization that the nature of the occurrence of traffic accidents is such that it is far better to model the process using the assumption of a Poisson distribution for the frequency of accidents in a given period of time at any one site. For example, Jovanis and Chang (1986) used a Poisson model to relate accident frequency to mileage and environmental variables. Comparative studies by, for example, Joshua and Garber (1990) and Miaou and Lum (1993) have confirmed the advantages of the Poisson model over the standard regression model. The desired forms of relationships, using the Poisson model, can then be developed using the technique of generalized linear models (GLMs) (see, for example, McCullagh and Nelder 1983), as has been done in the TRL studies. In a slightly different context, Dionne et al. (1993) have used the Poisson model to investigate the effect of medical conditions and exposure variables on truck drivers' accidents.

However, the Poisson model, although representing a significant advance in accurate and reliable

*Author for correspondence, presently at: Department of Civil and Transportation Engineering, Napier University, Edinburgh EH10 5DT, U.K.

modelling capability, is not without its weaknesses and technical difficulties which must be overcome if it is to be used effectively. One of these difficulties concerns the phenomenon of "overdispersion", whereby the assumption of a pure Poisson error structure can be seen to be inadequate. Maycock and Hall (1984) recognized this and showed how the negative binomial (NB) model could be used as an extension to the pure Poisson. Miaou and his co-authors have also proposed and used the NB model.

The purpose of this paper is threefold:

(a) To report on the progress of the continuing TRL junction accident studies, which have now covered 4-arm roundabouts, rural T junctions, urban 4-arm traffic signals, urban links and T junctions, urban crossroads, 3-arm signals, mini roundabouts, rural crossroads, rural single carriageways and rural dual carriageways; to describe the methodology adopted in these studies; and to give examples of the types of models which have been developed, with their implications for the effect of design on junction safety.
(b) To confirm the findings of some other authors concerning the advantages of the use of GLMs with Poisson error structure.
(c) To extend these findings, by describing the ways in which various technical problems in the use of the Poisson and NB models have been successfully overcome, so as to provide what is now believed to be a comprehensive and reliable methodology for the development of predictive accident models.

The structure of the paper is as follows. We firstly describe the form of the GLM with the basic assumption of a pure Poisson error, the methods available for the fitting of such a model, and the approach taken to the building of appropriate predictive accident models. The following section goes on to report on the series of TRL junction accident studies, and some of the findings from the application of the models. Subsequent sections then deal with a number of technical problems which have arisen in the TRL studies, and describe the ways in which the basic methodology has been extended or adapted to overcome these problems, which include: the low mean value problem, overdispersion, the estimation of prediction uncertainty, the disaggregation of data over time, uncertainty of flow estimates, aggregation of predictions (over different time periods or accident types), and their combination with observed values at particular sites through the use of the empirical Bayes method.

2. THE PURE POISSON MODEL

The basic form of model which we are considering is one in which the observed number of accidents at a site i is Y_i, which is assumed to be Poisson distributed about a mean of μ_i, which in turn is assumed to be proportional to the length of the observation period T_i. The expected number of accidents per year, λ_i (with $\mu_i = \lambda_i T_i$), is then related to the explanatory variables $\mathbf{x}_i$ (the traffic flows and physical characteristics of the site) through a log link function (the site subscript is omitted where no ambiguity should be caused):

$$\mu = \lambda T = \exp(\eta) = \exp(\boldsymbol{\beta}^T \mathbf{x}) \quad (1)$$

where η is known as the linear predictor and the vector $\boldsymbol{\beta}$ contains the parameters which are to be estimated by the fitting process. The vector $\mathbf{x}$, containing the values of the explanatory variables, has as its first term a 1, so that the first term in the vector $\boldsymbol{\beta}$ is the "intercept" or constant. In applications to sites which are lengths of road rather than junctions, it will usually be assumed that μ_i is also proportional to the section length L_i as well as to the time period, so that λ_i is then in terms of accidents per km per year.

The advantages arising from the use of the Poisson model over the more conventional multiple linear regression model with normally distributed errors have been listed by Joshua and Garber (1990) and by Miaou et al. (1992), in terms both of the theoretical justification and appropriateness of the model assumptions and of the improvement in the fit which can be obtained using real data. We shall not, therefore, dwell on that point here, but take this basic Poisson model as our starting point in what is to follow.

The fitting of a GLM can be achieved by the application of an iterated weighted least squares (WLS) method applied to the model in eqn 1, with the weight attached to a point at any iteration being inversely proportional to the fitted value $\hat{\mu}_i$ at the previous iteration (since, in the pure Poisson model, the variance is equal to the mean). Hence, fitting can be carried out using any statistical package which contains a routine for the application of WLS regression analysis. However, there are some advantages in the use of programs which have been specifically designed for the fitting of GLMs: for example, multilevel factors can be specified without the need for the explicit construction of the necessary dummy variables, and the iterative adjustment of weights in the WLS process is carried out internally. The programs

GLIM (Payne 1985) and GENSTAT (Lane et al. 1988) have been used extensively in the TRL studies.

The quality of the fit between the observed values y_i and the fitted values $\hat{\mu}_i$ can be measured by a number of statistics, the two best known of which are the scaled deviance (SD) and the Pearson X^2 statistic:

$$\mathrm{SD} = \sum_i 2\left[y_i \log\left(\frac{y_i}{\hat{\mu}_i}\right) - (y_i - \hat{\mu}_i)\right]$$

$$X^2 = \sum_i \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i} \qquad (2)$$

Both of these take the value zero when there is perfect agreement and are positive otherwise. As SD is based on the log likelihood function, and the estimation of the parameter estimates is obtained through the maximization of the likelihood, SD is the one which is more commonly used.

The development of models is achieved, typically, by the inclusion of extra terms (single variables or multi-level factors) one at a time, and by testing for their significance, using either the drop in SD or by the t ratio (the ratio of the estimated coefficient to its standard error). The drop in SD should be compared with a χ^2 distribution with as many degrees of freedom as there are extra parameters in the model. Comparison of different models can also be achieved by means of Akaike's Information Criterion (AIC) (Akaike 1973). For a well-fitting, or adequate, model the value of SD and X^2 should, according to standard theory, come from a χ^2 distribution with $(N-p)$ degrees of freedom, where N is the number of observations and p is the number of parameters which have been estimated. The theory underlying the fitting of such GLMs is well covered in, for example, McCullagh and Nelder (1983). We shall see, later in this paper, that some aspects of the standard theory cannot be justified in certain circumstances, with the consequence that the standard model building and testing methodology may give misleading conclusions. For the moment, however, we shall take the Poisson model as a useful and reliable tool for the development of predictive accident relations.

3. TRL JUNCTION AND LINK ACCIDENT STUDIES

The TRL has, over the last decade, conducted a continuing series of studies of accidents for particular components of the road network. The studies have been commissioned by customers within the Road Safety Division of the U.K. Department of Transport and within the Roads Engineering and Environmental Division of the Highways Agency of the U.K. Department of Transport. The Transportation Research Group of the University of Southampton have made substantial contributions to most of the studies. These studies have led to the need to develop model building methodology beyond that required for the pure Poisson model described in the previous section. Later sections explain how the model building methodology was developed. This section briefly describes the TRL studies and presents examples of some of the models.

3.1. *Scope of the studies*

The main TRL studies have covered the following types of junctions and links:

(a) 4-Arm roundabouts (Maycock and Hall 1984); 1427 injury accidents over a 6 year period at 84 junctions.
(b) 3-Arm major–minor priority junctions on rural single carriageway roads (Pickering et al. 1986); 674 injury accidents over 5 years at 302 junctions.
(c) 4-Arm signalized junctions on urban single carriageway roads (Hall 1986); 1772 injury accidents over 4 years at 177 junctions.
(d) 3-Arm major–minor priority junctions on urban single carriageway roads; 2699 injury accidents over 5 years at 980 junctions.
(e) Urban single carriageway "pure" links on both two-way and one-way roads; 1590 injury accidents over 5 years on 970 links.
(f) 4-Arm major–minor priority junctions on urban single carriageway roads; 2917 injury accidents over 5 years at 300 junctions.
(g) 3-Arm signalized junctions on urban single carriageway roads; 2262 injury accidents over 7 years at 221 junctions.
(h) Mini-roundabouts; 2100 injury accidents over 7 years at 305 junctions.
(i) Links between major junctions taken as a unit, including all the internal minor junctions and sections of "pure" links.
(j) 3-Arm and 4-arm major–minor priority junctions and 3-arm and 4-arm signalized junctions on urban single carriageway roads with one-way arms.
(k) 4-Arm major–minor priority junctions on rural single carriageway roads.
(l) Rural single carriageway "pure" links on modern English trunk roads.
(m) Rural dual carriageway "pure" links on modern English trunk roads.

Whilst findings have been reported in the literature for the first three studies, those from the next four studies are being finalized and have not yet been reported, except in the form of unpublished internal

reports, and the final six studies are in progress but have not yet reached the stage of producing results.

3.2. Site selection and data collection

The same empirical approach was adopted in all of the studies. An extensive national reconnaissance survey was conducted as part of each study to identify suitable sites, from which the sample of sites for study was drawn. The survey usually included about twice to three times the number of sites required for study. The sample of sites was selected so as to be stratified according to important variables, which were the vehicle and pedestrian flows and the main features of the layout. The sample was selected at random within each stratum so as to avoid bias. Care was taken to ensure that only those sites that had not been modified during the period over which the accident data were to be collected, were included in the sample.

The accident data consisted of records of all reported injury accidents occurring at the sites. In the case of junctions "at the site" meant at or within 20 metres of the junctions. This is not to say that this is the limit of the extent of influence of the junction but rather that this is the definition used in the UK STATS19 accident database. Traffic flows were measured on a weekday, avoiding public and school holidays. Turning flows by class of vehicle were obtained at each junction, usually measured over a 12 hour period from 0700 hr to 1900 hr. On the link sites, a 12 hour directional count by vehicle class was made. The vehicle counts were factored for each class of vehicle and manoeuvre to provide estimates of the annual average daily traffic (AADT) over the study period. The flows of pedestrians according to the direction in which they crossed the road were counted over the same periods of time as the vehicle flows. For a few of the studies, a subset of the sites had vehicle and pedestrian flows counted during only four 15 minute periods of the day, though covering the a.m. and p.m. peaks and the off peak. An extensive set of layout dimensions, signal timings and other variables was measured at each site, with the aim of including all those that might have an effect on accidents.

3.3. Analysis

Models were developed at three different levels. Level 1 models are coarse models which relate total accidents, and a limited disaggregation of these into vehicle-only accidents and pedestrian accidents, to some simple flow function. Variables representing major features (especially those used to stratify the sample) are also tested at this level and included as simple multiplicative factors.

At level 2, the accidents are disaggregated into a number of categories, according to the nature of the conflict. In the 4-arm roundabout study, for example, 5 types were used and in the 4-arm signal study 7 vehicle types and 4 pedestrian types were used. In general, each accident type refers to a specific conflict between one or more streams of vehicles or pedestrians. Exceptions are single-vehicle accidents, and types which include several forms of conflict, each of which has too few accidents to model separately. Where appropriate, the accidents were disaggregated according to the arm of the junction on which they occurred. The models relate accident frequency by type and arm to the relevant turning flows. Variables representing major features are also tested at this level and included as simple multiplicative factors.

The level 2 models are essentially a stage in the development of level 3 models. The latter retain the same flow function as the corresponding level 2 model, but in addition include all the relevant geometric, signal, and other variables. These higher level models were more demanding of data, but it was believed that through disaggregation, it would be easier to establish genuine and better-fitting relationships between accidents and flows and geometry.

Regression analysis is a powerful tool for identifying the variables that affect accidents, but it should not be used blindly. Engineering judgement was always an essential part of the model building process. The following criteria were taken into account in developing the level 3 models:

(a) The level of statistical significance. This was by far the dominant criterion. No variables were accepted at less than the 5% level, whilst none were rejected at the 1% level or better without very careful consideration.
(b) The stability of the model. If variables are associated with each other, then introducing one will tend to strongly affect the model parameters for the other. Since causal models are sought, such instability was carefully investigated. Care was taken at the site selection stage to minimize where possible the correlation between variables that were likely to appear in the models.
(c) The comprehensibility of the effect. It is desirable that the effect of a variable is in some sense understandable and that the models have a logical structure. For example, models for total accidents should not have the vehicle and pedestrian flows simply as a product, since this implies that total accidents tend to zero as pedestrian flow tends to zero. Models with estimated coefficients of the "wrong" sign were exam-

ined carefully to see whether the finding was robust.

(d) The size of the effect and ease of measurement. Variables that had a large effect on accidents in relation to their range and which were straightforward for the engineer to measure were preferred.

It would not be possible or sensible to give the fitted model forms from all the many TRL studies here, but in order to illustrate their general nature we shall give just a small selection of the level 1 and level 3 models.

3.4. Level 1 models

The first example is the model for total accidents at 3-arm major–minor junctions on rural single carriageway roads:

$$\lambda_{\text{total}} = 0.19\text{QN}^{0.53} \tag{3}$$

where λ_{total} is the expected number of accidents per year for the whole junction and QN is the encounter products flow function. (Note that the vector of explanatory variables **x** in eqn 1 consists, in this case, of the log of the constant, 0.19, and log(QN)). A product of each pair of crossing flows, each pair of merging flows, and each pair of diverging flows is formed and all are summed to give QN. The flows are expressed as annual average daily totals in units of one thousand. None of the factors representing basic features, such as traffic islands and the provision of lanes for the acceleration and deceleration of vehicles joining and leaving the main road ghost island on the major arms, appeared in the model. There were very few pedestrian accidents at these junctions and therefore separate models for vehicle-only and for pedestrian accidents were not developed.

For urban junctions and links, accidents involving pedestrians form about 40% of the total. Separate models were therefore developed for vehicle-only and for pedestrian accidents, as well as for total accidents. Several forms of vehicle and pedestrian flow functions were used, and these are illustrated below. Variables representing basic features of the sites often appear in the models, sometimes involving interaction terms with each other and more especially with the flows. In order to simplify the presentation, only the flow functions will be further considered here.

The models for vehicle-only, pedestrian and total accidents at 3-arm major–minor junctions on urban single carriageway roads were:

$$\begin{aligned}
\lambda_{\text{veh-only}} &= 0.049\ \text{QMA}^{0.80}\ \text{QMI}^{0.36} \\
\lambda_{\text{ped}} &= 0.052\ \text{QMA}^{0.51}\ \text{QMI}^{0.16}\ \text{PT}^{0.46} \\
\lambda_{\text{total}} &= 0.049\ \text{QMA}^{0.71}\ \text{QMI}^{0.30} \exp(0.68\ \text{PT}^{0.20})
\end{aligned} \tag{4}$$

where $\lambda_{\text{veh-only}}$, λ_{ped} and λ_{total} are the expected numbers of accidents per year for vehicle-only, pedestrian and total junction accidents respectively. QMA is the sum of the vehicle inflows on the major arms, and QMI is the inflow on the minor arm. PT is the total flow of pedestrians across the arms and the centre of the junction summed over both directions of crossing, and measured in thousands of pedestrians per 12 hour period. The pedestrian flow appears in the exponential term for total junction accidents so that a non-zero number of accidents is predicted as PT tends to zero.

The models for urban single carriageway "pure" links are somewhat different to those for junctions. (By a "pure" link, we mean a stretch of road between junctions. Such a link contains no junctions —only driveways and minor accesses). The models for vehicle-only, pedestrian and total accidents illustrate this:

$$\begin{aligned}
\lambda_{\text{veh-only}} &= 0.103\ \text{SL}\ \text{QT}^{0.78} \exp(0.75\ \text{PTSL}^{0.20}) \\
\lambda_{\text{ped}} &= 0.180\ \text{SL}\ \text{QT}^{0.72}\ \text{PTSL}^{0.44} \\
\lambda_{\text{total}} &= 0.083\ \text{SL}\ \text{QT}^{0.74} \exp(1.63\ \text{PTSL}^{0.15})
\end{aligned} \tag{5}$$

where again $\lambda_{\text{veh-only}}$, λ_{ped} and λ_{total} are the expected numbers of accidents per year for vehicle-only, pedestrian and total link section accidents. SL is the length of the link section measured in kilometres, QT is the vehicle flow, and PTSL is the pedestrian density (thousands of pedestrians crossing the road per kilometre per 12 hr day). For all the models, the number of accidents is directly proportional to the length of the link section, and depends on the vehicle flow and the pedestrian density. Pedestrian density has a direct effect on the numbers of pedestrian accidents, but only an indirect effect on the numbers of vehicle-only accidents. Placing the pedestrian density in the exponential term ensures that a non-zero number of vehicle-only accidents are predicted as PTSL tends to zero. Pedestrian density also has a direct effect on total link section accidents, but the pedestrian density is placed in the exponential term so that accidents are still predicted even when PTSL becomes zero.

3.5. Level 3 models

An example of a level 3 model is that for injury accidents involving a collision between a vehicle turning right (remember that this is under U.K. rules of the road and therefore equivalent to turning left in North America and other European countries) and another vehicle moving straight ahead from the opposite arm at 4-arm traffic signals. These are referred to as principal right turn accidents:

$$\lambda_{\text{prin-rt}} = 0.179\ Q_3^{0.59}\ Q_8^{0.48}\ K_1K_2K_3K_4 \tag{6}$$

where $\lambda_{\text{prin-rt}}$ is the expected number of principal right-turn accidents per year for one arm, Q_3 is the

right turning vehicle flow and Q_8 is the ahead vehicle flow for the opposite arm. The vehicle flows are expressed as annual average daily totals in units of one thousand. K_1, K_2, K_3 and K_4 are multipliers.

$K_1 = \exp(-0.017\theta - 0.1\text{DISP} + 2.76\text{PT}_8)$ in which θ is the angle between the opposite arm and the right hand arm (degrees), DISP is the absolute value of the centre line displacement of the arm in relation to the opposite arm (m), and PT_8 is the proportion of two-wheeled vehicles in the ahead flow Q_8. Fig. 1 illustrates the definition of these geometric variables. $\lambda_{\text{prin-rt}}$ is reduced by increasing θ and DISP and by decreasing PT_8.

$K_2 = 1.32$ for arms where there is a central island, but otherwise $K_2 = 1$. $K_3 = 0.1$ if the right turning flow has a separate signal stage; $K_3 = 0.6$ if there is early cut-off or late release; $K_3 = 1$ if there is no special stage for right turning vehicles. Early cut-off and late-release are terms used in the UK for two different control arrangements in which traffic has a part-opposed and a part-fully protected right turn.

$K_4 = \exp(0.85C_{18} + 0.12C_{12})$ in which C_{18} is the flow of vehicles from the arm into the junction per second of green time, and C_{12} is the intergreen time; that is, the time interval between the end of the green for the vehicle stage on the arm and the beginning of the green on the next vehicle crossing stage. $\lambda_{\text{prin-rt}}$ is reduced by decreasing the value of either of these control variables.

Flows of vehicles or pedestrians which do not take a direct part in the accident occasionally appear in the level 3 models. For example, at 4-arm major–minor junctions on urban single carriageway roads, the model for accidents between vehicles turning right from the major arm and those travelling ahead from the opposite major arm, also includes the flow of vehicles making a right turn from the opposite major arm. Such flows which can have only an indirect effect on the numbers of accidents are included in an exponential term in the later studies.

4. THE LOW MEAN VALUE PROBLEM

The standard theory of GLMs prescribes that, for a perfect model (that is, one in which the predicted value is the true mean value μ), the discrepancy between the observed value y and the predicted $\hat{\mu}$ at each site, as measured by its contribution to the scale deviance (SD) is asymptotically χ^2_1 distributed (McCullagh and Nelder 1983). As a consequence, the value of the total SD from an adequate model should be χ^2 distributed with $(N-p)$ degrees of freedom, where N is the number of data points and p is the

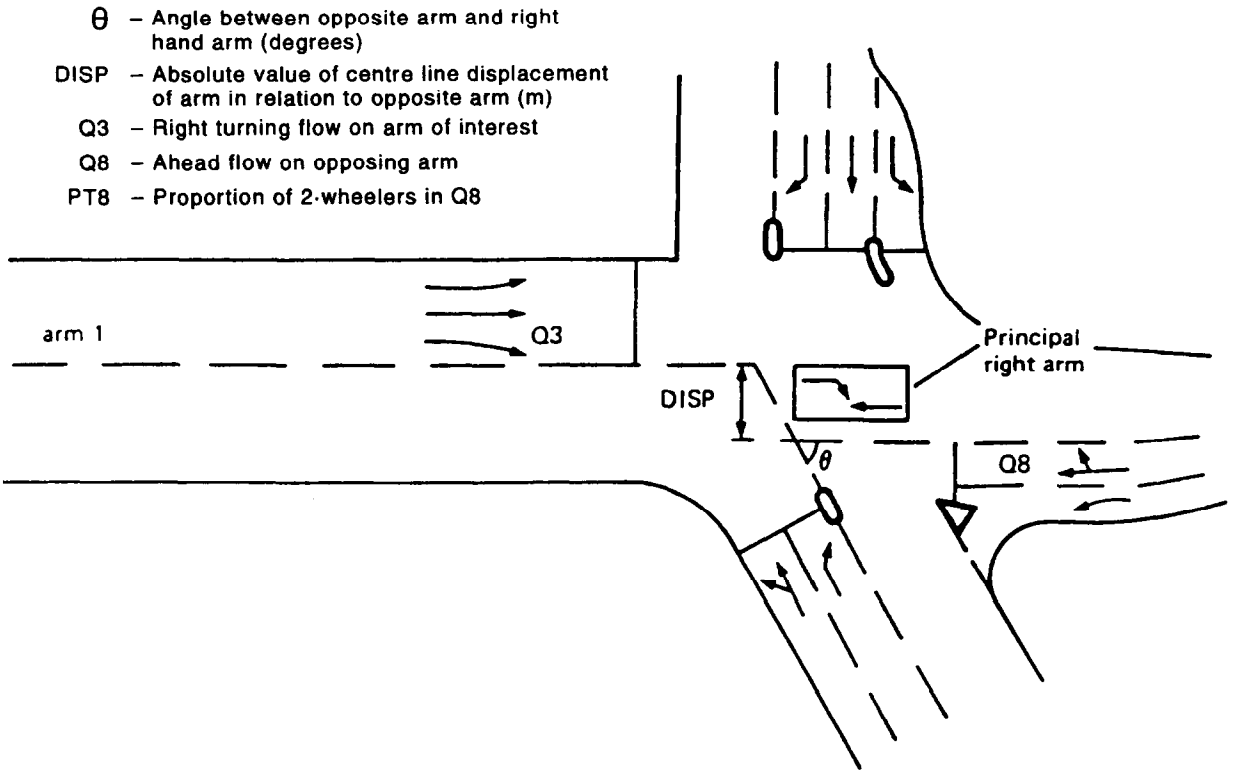

Fig. 1. Principal right turn accidents at 4-arm signals (arm 1 only) showing relevant geometric parameters.

number of parameters which have been fitted. Therefore, if the value of SD is close to $(N-p)$, it is taken as an indication that the model is adequate, and that the inclusion of further terms in the model is not called for.

However, this result is based on asymptotic theory and cannot necessarily be relied upon for finite samples. In fact, it is commonly observed when fitting pure Poisson GLMs that, when fitting even quite coarse models with very few explanatory variables, the value of SD can often be well below $(N-p)$. The reason for this was first pointed out (at least, in the context of predictive accident models) by Maycock and Hall (1984). They showed that, for a perfect model, the expected value of SD for a single point was not constant at a value of 1 for all values of the mean μ, as the asymptotic results might have one believe. Instead, $E(\mathrm{SD})$ varies appreciably with μ and, in particular, when the mean value is low (<0.5, for example), the value of $E(\mathrm{SD})$ can be very much less than unity. The consequence of this is that when the data set contains a high proportion of sites for which the fitted value $\hat{\mu}_i$ is low, the use of $(N-p)$ as a "target" figure for the value of SD can be very misleading, since it will often imply that the fitted model is adequate even when it is plainly not.

There are two possible remedies for this problem. One is to use, instead of SD, the Pearson X^2 statistic, since the expected value of the contribution to X^2 from a single point is not so dependent upon the value of the true mean. Consequently, the use of $X^2/(N-p)$ should provide a better indication of model adequacy in these circumstances.

The other approach is, instead of accepting the asymptotic value of 1 for the expected value from each point, to compute it directly, using the fitted values $\hat{\mu}_i$ obtained following the fit of the model:

$$E(SD)=\sum_i \sum_{y_i=0}^{\infty} \frac{\exp(-\hat{\mu}_i)\hat{\mu}_i^{y_i}}{y_i!}\ 2\left[y_i \log\left(\frac{y_i}{\hat{\mu}_i}\right)-(y_i-\hat{\mu}_i)\right] \tag{7}$$

This provides a more realistic target value for the actual SD. Simulation tests, using data generated artificially according to a variety of model forms (all with pure Poisson error structure), have confirmed that the resulting value is indeed a useful and approximately unbiased estimate of $E(\mathrm{SD})$. Results from one set of simulation tests are displayed in Fig. 2. Here, we compare three statistics: (i) $\mathrm{SD}/(N-p)$, (ii) $X^2/(N-p)$ and (iii) $\mathrm{SD}/E(\mathrm{SD})$ with $E(\mathrm{SD})$ computed using eqn 7. It can be seen that, as P, the proportion of data points with low mean value (here, less than 0.5) increases, the estimate from $\mathrm{SD}/(N-p)$ falls appreciably. The other two estimators are far more reliable, even when P becomes very high. There is little to choose between the estimators $X^2/(N-p)$ and $\mathrm{SD}/E(\mathrm{SD})$ in these tests.

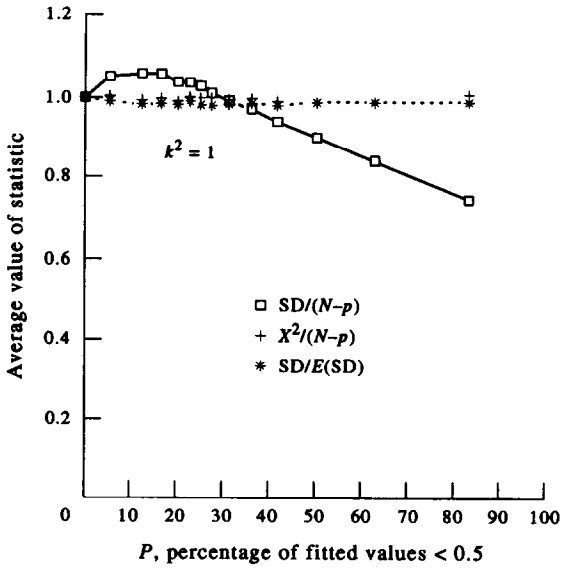

Fig. 2. Comparison of three statistics (pure Poisson model).

Having noted that, with a high proportion of sites with low mean values, the value of SD can often be much lower than that expected on the basis of the standard theory, a further possible problem arises. As was discussed earlier, it is common to use the drop in SD as the basis of testing the significance of extra terms in the model. If the overall level of the SD values may be lower than expected, could it then follow that the drops in SD are also on a different scale? If so, this could have an appreciable influence on the outcome of the model building process. To investigate this, simulation experiments were again carried out, generating many data sets from a known model and calculating, for each data set, the drop in SD when an irrelevant term (essentially a "treatment" which in fact had no effect) was brought into the model. The experiments showed that, even when there is a high proportion of points with low mean values, the average drop in SD accorded reasonably well with the value of 1 which would be anticipated on the basis of the drop in SD being χ_1^2 distributed. Therefore, our conclusion was that, for pure Poisson models, even in those circumstances when the value of SD itself could be appreciably below the standard target value of $(N-p)$, the drop in SD could be used (albeit with some caution) as the basis of a significance test to decide on the inclusion of extra terms in the model.

5. MODELLING OVERDISPERSION

In the TRL junction accident studies, every reasonable attempt was made to find a full explanation

of the variation in accidents between sites. A variety of explanatory variables were tried consistent with engineering judgement. For each accident type, the most important term in the model was the relevant flow or flows (for example, particular turning flows at the junction), followed by explanatory variables which measured relevant physical characteristics of the site (such as, for example, entry width or entry path curvature) and control variables (such as which movements receive green together at traffic signals). Nevertheless, despite such painstaking efforts, it was virtually inevitable that the final models should be, in the technical sense, "inadequate" (taking account, of course, of the considerations discussed in the previous section of this paper relating to the possible low mean value problem). That is to say, the explanatory variables do not provide a complete explanation of the between-site variation, so that the residual variation is more than would be expected on the basis of the pure Poisson model. There are several possible reasons for this:

(a) There are other, unobserved, explanatory variables at work which effectively add to the random error, or "noise".
(b) There are errors in some of the explanatory variables, most particularly the flows. (The flow estimates, taken to be representative of the flow across the whole of the observation period, are often merely "snapshots", taken on one occasion).
(c) The model may be mis-specified.

Miaou (1994) has similarly commented on the occurrence of, and reasons for, overdispersion.

5.1. *The quasi-Poisson model*

There are a number of ways in which the basic pure Poisson model may be modified in order to take account of, or correct for, overdispersion. Wedderburn (1974) suggested the form of model often referred to as the "quasi-Poisson" (QP) model, in which the variance of Y_i is given by $k^2\mu_i$. In principle, the parameter k^2 can be estimated by any of the three statistics $SD/(N-p)$, $X^2/(N-p)$ or $SD/E(SD)$.

The parameter estimates resulting from the QP model are identical to those from the pure Poisson model. The only difference is in the magnitude of their standard errors, which are inflated by a factor of k. The effect in model building terms is that some variables would not be deemed significant under the QP model, because their t ratios would not achieve the necessary level. To carry out a significance test on an extra term using the drop in SD, the drop should be divided by a reliable estimate of k^2; the resulting statistic should then be approximately χ^2 distributed.

A further set of simulation tests were carried out, using data generated from a QP model (with a true value of k^2 of 1.5), in order to investigate the reliability of the three estimators of k^2. (Generation from a QP can be achieved by generating from a NB with the value of the shape parameter proportional to the mean—see later in section 5.3). Some results are shown in Fig. 3, from which it can be seen that, once again, $SD/(N-p)$ behaves very poorly, as it falls off quite rapidly from its nominal value of 1.5 as P, the proportion of data points with low mean value, increases. $SD/E(SD)$ performs only a little better. The Pearson estimator $X^2/(N-p)$ is clearly the best of the three, but even that consistently underestimates the scale factor, especially when P is above 60% or so.

As with the pure Poisson case, experiments were carried out to investigate the average drop in SD when an extra (irrelevant) term is added into the model. Once again the average drop is reasonably close to the nominal mean value of 1, (in fact, slightly above 1), and the drop does not appear to be influenced by the value of P.

5.2. *The standard negative binomial model*

A second way to model overdispersion is to use the NB model. This is equivalent to assuming that (i) Y is Poisson distributed about a true mean of μ ($= \lambda T$ where T is the length of the observation period and λ is the true rate per year), and (ii) λ is gamma distributed with a mean of $\hat{\lambda}$ (the estimate based on the known values of the explanatory variables, as in eqn 1: $\hat{\lambda} = \exp(\hat{\boldsymbol{\beta}}^T\mathbf{x})$ and a shape of α. In this, the

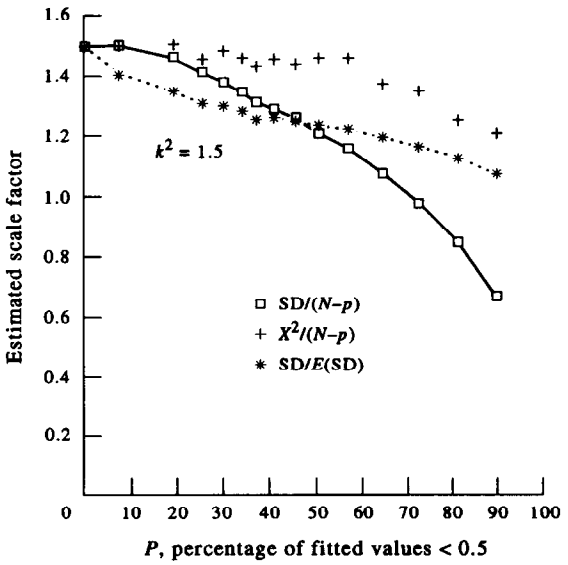

Fig. 3. Comparison of three statistics (quasi-Poisson model).

variability of λ about $\hat{\lambda}$ represents the variability due to other, unobserved explanatory variables (that is, those not included amongst the vector **x**). The shape parameter describes the spread of the gamma distribution, with low values implying a large amount of overdispersion. The error variance for Y, given the vector **x**, is:

$$\text{Var}(Y|\mathbf{x}) = \hat{\lambda}T + \frac{(\hat{\lambda}T)^2}{\alpha} \tag{8}$$

As α becomes large, the amount of overdispersion decreases and the NB model tends towards the pure Poisson. The NB model may be fitted, as may the pure Poisson, using iterated WLS, for any given value of α. Using an outer iterative loop, the maximum likelihood estimate (MLE) of α can be found, together with the associated parameter estimates $\hat{\boldsymbol{\beta}}$.

We have, then, two alternative ways to allow for overdispersion: the QP and the NB models. It has been found from extensive use of the two models in the TRL studies that the parameter estimates which result are only very slightly different. Therefore, the fitted values are almost identical and the predictions which would result from their use will again be almost identical. It may be tempting to conclude, therefore, that it hardly matters which model form is chosen. However, appreciable differences can occur when one comes to consider the uncertainty which should be attached to predictions, as measured, for example, by the prediction error variance. This is a subject to which we shall return later, but for the moment we merely state that there are good reasons for making a choice between the QP and the NB models.

5.3. *A more general negative binomial model*

In fact, the QP and NB models can be seen as two special cases of a family of models (see Cameron and Trivedi 1986). In the standard NB model, the underlying gamma distribution for λ has a constant shape parameter α, regardless of the mean $\hat{\lambda}$. In the more general family, the shape (now denoted by ν) is dependent upon the estimate of the mean in the following manner:

$$\nu = \alpha\hat{\lambda}^n \tag{9}$$

Then it follows that the error variance for Y given **x** is:

$$\text{Var}(Y|\mathbf{x}) = \hat{\lambda}T + \frac{\hat{\lambda}^{2-n}T^2}{\alpha} \tag{10}$$

Therefore, when $n = 0$, we have the standard NB model, with $\nu = \alpha$ (constant shape) and eqn 10 reduces to eqn 8. When $n = 1$, the shape is proportional to the mean (implying that there is relatively less spread when the mean is large), and eqn 10 reduces to the QP form in which the variance is proportional to $\hat{\lambda}$ with the scale parameter $k^2 = 1 + T\alpha^{-1}$. If T is constant (or nearly so) across the sites, then, the QP model is a special case of this family of NB models.

For any value of n, therefore, the MLEs of α and the $\boldsymbol{\beta}$ can be found and the resulting maximized value of the log likelihood, $\log L_{\max}(n)$, found. A plot of $\log L_{\max}(n)$ versus n will indicate that value of n which is optimal; that is, which member of the family of models fits the data best. Such a plot is shown in Fig. 4, for the case of approaching accidents in the TRL 4-arm roundabout study (Maycock and Hall 1984). This shows that the best value of n is around 0.25, closer to the standard NB model ($n = 0$) than to the QP model ($n = 1$). In fact, a 95% confidence interval on n can be determined from the plot as that range of values of n for which $2\log L_{\max}(n)$ is within 3.84 of the maximum value. It can be seen that the standard NB model falls well within this, but the QP model does not (the difference between $2\log L(0.25)$ and $2\log L(1)$ is almost 6).

Because of the fact that the fitted values from any of the models within this family are very similar, an approximate but efficient method has been developed to determine the optimal value of n. Model building is carried out using the QP model (using GLIM, typically, in the TRL studies) and, once the final model form has been established, the observed values y_i and fitted values $\hat{\mu}_i$ are written out into a file, which is then used as input to a small FORTRAN77 program which can calculate the log

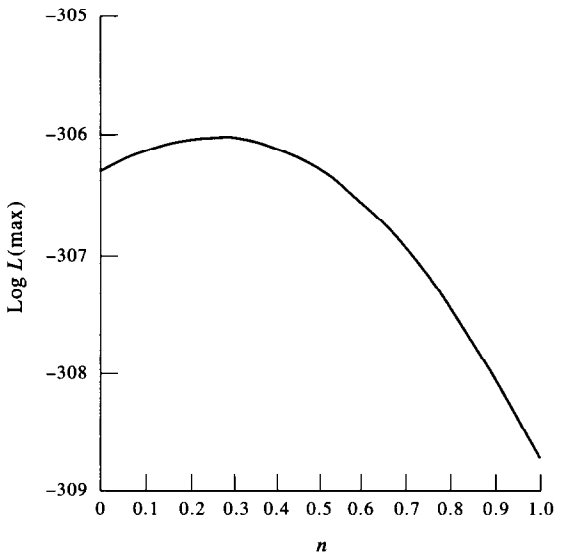

Fig. 4. Log L(max) versus n (approaching accidents).

likelihood for any given values of n and α:

$$\log L(n,\alpha) = \sum_i \log\left[\frac{\Gamma(y_i+\nu)}{y_i!\Gamma(\nu)}\left(\frac{\nu}{\nu+\hat{\mu}_i}\right)^{\nu}\left(\frac{\hat{\mu}_i}{\nu+\hat{\mu}_i}\right)^{y_i}\right] \quad (11)$$

and, for any value of n, determine the MLE of α in $\nu = \alpha\lambda^n$.

This technique of embedding the QP and NB models within this more general family has been applied to all the models developed in the various TRL junction accident studies. It is not always possible to discriminate between the models (that is, the 95% confidence interval covers the whole range of values for n between 0 and 1), but where it has, it has always been the case that it is the NB model which is the better choice; often clearly so. The ability to discriminate depends very much on the sample size in the data set for that particular accident type.

The conclusion from a large number of empirical studies, then, is that the NB model is the most appropriate way by which to model overdispersion.

6. ESTIMATING THE UNCERTAINTY OF PREDICTIONS

Once the most appropriate model has been fitted to the data, and the parameter estimates $\hat{\boldsymbol{\beta}}$ obtained, it is possible to consider the question of how much uncertainty there should be attached to predictions which are made using the model. As we have already seen, the prediction (for the mean number of accidents per year) will be:

$$\hat{\lambda} = \exp(\hat{\eta}) = \exp(\hat{\boldsymbol{\beta}}^T\mathbf{x}) \quad (12)$$

As the $\hat{\beta}_i$ are merely estimates of the true parameter values, they have standard errors, which are calculated routinely in the model fitting process. The $\hat{\beta}_i$ are generally correlated (sometimes quite highly) and the correlations between them can also be obtained. If the correlation matrix is denoted by $\{r_{ij}\}$, the variance of the linear predictor is given by:

$$\mathrm{Var}(\hat{\eta}) = \sum_i \sum_j r_{ij}s_is_jx_ix_j \quad (13)$$

where s_i denotes the standard error of the parameter estimate $\hat{\beta}_i$.

Uncertainty in the $\hat{\beta}_i$ leads to uncertainty in the linear predictor and hence to uncertainty in the prediction $\hat{\lambda}$ since they are linked by eqn 12. It can be shown that the uncertainty of the prediction, measured by its error variance, is approximately given by:

$$\mathrm{Var}(\hat{\lambda}) \doteq \mathrm{Var}(\hat{\eta})\hat{\lambda}^2 \quad (14)$$

There is however a second component to be allowed for. The uncertainty of the estimate of the true mean λ consists of the regression effect (uncertainty in λ) and the overdispersion effect (uncertainty in λ about λ):

$$\mathrm{Var}(\lambda) = \mathrm{Var}(\lambda|\hat{\lambda}) + \mathrm{Var}(\hat{\lambda}) \quad (15)$$

in which the first term is given by the fact that λ is gamma distributed about a mean of $\hat{\lambda}$ and the second by eqn 14, giving:

$$\mathrm{Var}(\lambda) = \frac{E(\hat{\lambda}^{2-n})}{\alpha} + \mathrm{Var}(\hat{\eta})\hat{\lambda}^2 \quad (16)$$

For the two special cases, this reduces to the following:

Quasi-Poisson model

$$\mathrm{Var}(\lambda) = (k^2-1)\hat{\lambda} + \mathrm{Var}(\hat{\eta})\hat{\lambda}^2 \quad (17)$$

Negative binomial model

$$\mathrm{Var}(\lambda) = \hat{\lambda}^2\left[\frac{1}{\alpha} + \mathrm{Var}(\hat{\eta})\left(1+\frac{1}{\alpha}\right)\right] \quad (18)$$

Figure 5 shows plots (on a log–log scale) of the prediction error variance calculated for each of the data points used in the fitting of the model for approaching accidents in the TRL 4-arm roundabouts study. Approaching accidents are injury accidents which involve a collision between vehicles on the approach to the roundabout; mostly rear-end shunts when one vehicle runs into the back of another, but also including accidents where a vehicle is changing lanes. Two sets of points are displayed, for the QP and the NB models. Two particular aspects of the plots should be noted:

(a) The estimated prediction error variances are very different under the two models, especially for the more extreme points. Therefore, although the choice of model has little or no effect on the form of the fitted model and on the predictions made from it, it does have an effect on the estimate of the uncertainty of those predictions.

(b) For the NB model, the points lie almost exactly on a straight line (of slope two in the log–log plot), indicating that $\mathrm{Var}(\lambda)$ is virtually proportional to $\hat{\lambda}^2$. Although, in eqn 18, $\mathrm{Var}(\hat{\eta})$ may vary from point to point, this variation is small compared with the magnitude of the other terms within the square brackets in eqn 18 which remain constant.

This latter point suggests that the uncertainty of the prediction can be summarized in a very simple fashion, in the form of a coefficient of variation C_v (ratio of prediction standard error to the prediction

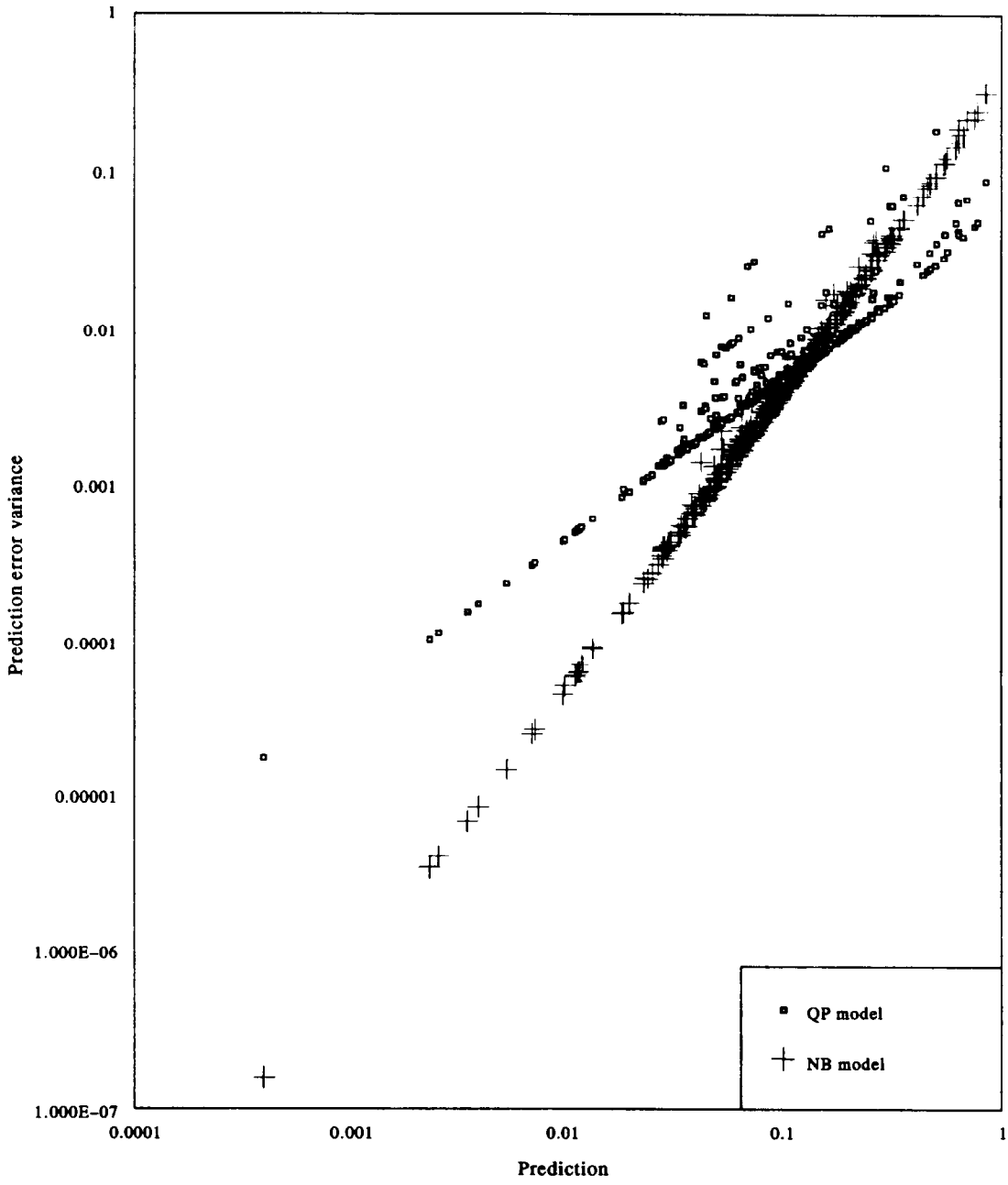

Fig. 5. Prediction error variance.

itself). For the case of approaching accidents, C_v takes the value 0.64. Therefore, if a particular site is predicted to have 2 approaching accidents per year, the standard error of this prediction would be 1.28. Although this might be felt to be surprisingly large, it should be remembered that this is for just one accident type. For most practical purposes, the important prediction will be for the total accidents at the junction. When the predictions are aggregated over all accident types and perhaps over all arms of the junction, the coefficient of variation for the total prediction will be appreciably smaller. This will be discussed further in section 9.

7. DISAGGREGATION OF DATA OVER TIME

The accident data may be disaggregated in a number of ways: by accident type, by arm of the junction (in some instances) and by time period. The TRL studies typically obtained accident data over a 4 or 5 year period (the same period for all sites) and

generally models were fitted using as the unit of data the total accidents over the whole period. However, it might be asked whether the data could, or should, be disaggregated, so that each year/site combination provides a unit of data.

In fact, if the parameter estimates are obtained through the technique of maximum likelihood, it does not make any difference to those estimates whether one fits using the aggregate or disaggregate form of the data. Where it can make a difference, however, is in the modelling of overdispersion. For example, if the QP model is used, the value of the scale factor k^2 is typically found to be lower when the disaggregate data is used than when the aggregate data is used. A similar finding is obtained when the NB model is used. This can clearly have an effect on the model building process, since the standard errors of the β_i are affected, leading in some cases to a term appearing to be significant when the disaggregate model is used but non-significant when the aggregate model is used.

The reason for this phenomenon is that one of the causes of overdispersion is the influence of explanatory variables which are not included in the model. These can be thought of largely as a "site effect", being effectively random between sites, but remaining more or less constant at any one site from year to year. Using the disaggregate data, the errors cannot be regarded as all independent; the errors for the same site in different years are likely to be very highly correlated. The error structure would be a mixture of within-site and between-site errors. Therefore, we have concluded that it is better to use the aggregate form of the data, in that the form of the model then allows what is believed to be a plausible interpretation of the actual error structure.

Where we have deviated somewhat from this approach in the various TRL studies is in some instances in which the accident and flow data are drawn from a rather longer time period (for example from 1979 to 1990), and where data is available for different parts of this time period for different sites. Furthermore, it has been felt to be important, given the length of the whole period, to allow for the possibility of a change in the level of accident risk over this time. That is, the "intercept" term β_0 in eqn 1 may vary over time.

To illustrate, let us consider a simple form of model like that in eqn 3, in which the mean number of accidents per year depends only on the flow Q:

$$\lambda = cQ^{\beta} \tag{19}$$

The constant c may be regarded as a measure of "risk" (accidents per unit of exposure Q^{β}). Suppose now that (i) the flow varies over the period of observation and is denoted in year t by Q_t, and (ii) the risk parameter c in this relationship varies steadily over time (in fact, in a geometric manner). Then μ, the total mean number of accidents at that site over the whole of its observation period (between years r and s), is given by:

$$\mu = c_0 \sum_{t=r}^{s} \gamma^t Q_t^{\beta} \tag{20}$$

in which c_0 is the risk in year 0 and γ is the factor by which the risk changes from one year to the next. Using the usual log link function, the linear predictor is then:

$$\eta = \log(c_0) + \log\left[\sum_{t=r}^{s} \gamma^t Q_t^{\beta}\right] \tag{21}$$

The parameters to be estimated in eqn 21 are c_0, β and γ. Unfortunately, eqn 21 is not in standard form if one is to implement MLE using the iterated WLS approach, since the terms cannot be set out in additive linear form. However, an iterative approach can be developed, using what are known as "constructed variables" (see McCullagh and Nelder 1983). Suppose that, initially, rough estimates $\hat{\beta}_0$ and $\hat{\gamma}_0$ are known (these could both be 1, for example). Then the technique is to expand eqn 21 using a Taylor series expansion about the rough estimates, thereby obtaining a form of the linear predictor in which the parameters $\log(c_0)$, $(\beta - \hat{\beta}_0)$ and $(\gamma - \hat{\gamma}_0)$ appear in an additive, linear form:

$$\eta = \log(c_0) + X_1 + (\beta - \hat{\beta}_0)X_2 + (\gamma - \hat{\gamma}_0)X_3 \tag{22}$$

in which:

$$X_1 = \log\left[\sum \hat{\gamma}_0^t Q_t^{\hat{\beta}_0}\right] \quad X_2 = \frac{\sum \hat{\gamma}_0^t \log(Q_t) Q_t^{\hat{\beta}_0}}{\sum \hat{\gamma}_0^t Q_t^{\hat{\beta}_0}}$$

$$X_3 = \frac{\sum t \hat{\gamma}_0^{t-1} Q_t^{\hat{\beta}_0}}{\sum \hat{\gamma}_0^t Q_t^{\hat{\beta}_0}} \tag{23}$$

in each of which the summations are over the observation period (r, s) for that site.

In eqn 22, then, X_1 acts as an "offset", whilst X_2 and X_3 act as explanatory variables associated with the parameters $(\beta - \hat{\beta}_0)$ and $(\gamma - \hat{\gamma}_0)$ which are the changes which should be made to the current estimates. X_1, X_2 and X_3, known as "constructed variables", are calculated from a knowledge of the flow values Q_t and the current parameter estimates $\hat{\beta}_0$ and $\hat{\gamma}_0$. Generally, we have found that only a couple of iterations are required in order to reach the point at which the changes $(\beta - \hat{\beta}_0)$ and $(\gamma - \hat{\gamma}_0)$ are sufficiently small, indicating that convergence has occurred.

To illustrate the use of the method, we generated some artificial data, representing 100 sites, following the model in eqn 20. The value of c_0 was set to 0.2, and that of β to 1. The decline in accident risk each year was represented by a value for γ of 0.9. The

lengths of the periods for which there were accident and flow data varied from site to site, but the flows followed a general, but not systematic, increasing trend over time. Initial values of $\hat{\beta}_0 = \hat{\gamma}_0 = 1$ were set, and the values of X_1, X_2 and X_3 calculated and used, together with the accident frequencies Y_i as input to GLIM, in order to fit a Poisson error structure GLM with the linear predictor shown in eqn 22. The estimates obtained were:

$$\hat{c}_0 = 0.169 \quad \hat{\beta} - \hat{\beta}_0 = 0.08674 \quad \hat{\gamma} - \hat{\gamma}_0 = -0.1327$$

With the revised estimates of $\hat{\beta}_1 = 1.08674$ and $\hat{\gamma}_1 = 0.8673$, a second iteration was carried out, re-calculating the values of X_1, X_2 and X_3 and re-entering GLIM to fit eqn 22. The output consists of:

$$\hat{c}_0 = 0.174 \quad \hat{\beta} - \hat{\beta}_1 = 0.00031 \quad \hat{\gamma} - \hat{\gamma}_1 = 0.0249$$

Further iterations produce virtually no change in the parameter estimates, indicating that convergence has occurred. The final (rounded) estimates would therefore be: $\hat{c}_0 = 0.174$, $\hat{\beta} = 1.09$ and $\hat{\gamma} = 0.89$.

8. RANDOM ERROR IN THE FLOW ESTIMATES

In GLM, as in more conventional multiple linear regression, one of the important assumptions is that random error occurs only in the Y_i, and that the explanatory variables **x** are known without error. For the geometric and control variables, this presents no problem but it has to be accepted that for the flows this is certainly not strictly true. Ideally, the flows should be the average AADTs over the whole of the time period for which the accident data are taken. However, typically, the flow estimates will be obtained as a "snapshot", through counts taken on perhaps just a single day. In the TRL study of 4-arm roundabouts, classified 16 hour counts were made at each site and in the study of 4-arm urban signals the counts were carried out over 12 hours. In such studies, one of the major costs is that of carrying out these flow counts and therefore it is important to have an appreciation of the effect of the length of the flow counts on the accuracy of the models which will be developed.

8.1. Randomization experiments

Let us denote the true (but unknown) AADT for a particular flow at a site by z_i, the log of which would then ideally appear as one element in the vector **x** of explanatory variables in eqn 1 which determines μ_i. However, z_i is not known and instead an estimate is obtained from a count Q_i carried out over a portion of a day, using standard scaling factors to allow for the period of the day over which the count was made, the day of the week and the month. Uncertainty in this estimate then arises not only from Poisson sampling, but also from non-Poisson variation (due to, for example, between-day variation). However, for simplicity and compactness, we express this uncertainty through the notion of an effective period t_i over which the count was made.

To gain some initial understanding of the magnitude of the bias which might be involved, some simulation tests were carried out based on a specific model which had been obtained from the study of 4-arm signals (Hall 1986). This related the total accidents at the junction to the total inflow over a 24 hour period. The relationship between μ and z was of the form: $\mu = c\, z^{1.38}$.

Here the observed values of total inflow ranged from 7000 to 35000 with a mean around 21000 and a standard deviation of approximately 7000. A large number of artificial data sets for Y and Q were generated and in each case a standard form of GLM fitted. When the sampling period t_i was taken to be 1, representing a whole day flow count, the bias in the estimate of the power in the relationship was only 0.04%. When the sampling period was only 15 minutes, the bias increased to 4%.

As a next step, a second series of simulation tests was carried out, this time using some of the more detailed models developed in the 4-arm signals study. For example, the model which was developed for principal right turn accidents contained the two relevant junction turning flows and seven other explanatory (geometric and control) variables. There were 668 observations. The technique adopted in these simulations was to perturb one or both flow variables, by generating a randomized value of it (or each) from a lognormal distribution with mean equal to the actual observed value and a variance to mean ratio of k. All other variables were unmodified. The randomized data set was then input to GLIM and the resulting parameter values noted: in particular that representing the coefficient of the perturbed flow. When the first flow was randomized using $k = 0.1$, the effect was a drop of 21% in the coefficient associated with that flow and a drop of 3% in the coefficient associated with the other flow. When the second flow was randomized, the drops were 0.4% and 9% respectively. When both flows were randomized simultaneously, the drops were 22% and 12%. The conclusions from this and other similar investigations were:

(a) Randomizing one flow variable leads to a bias (underestimate) in its coefficient.
(b) Randomizing one variable has a small but non-zero "cross-over" effect leading to a

bias in the coefficients associated with other flow variables.

(c) The effect of randomizing two variables simultaneously is the sum of the two separate effects.

These simulation experiments indicated that, with the 12 or 16 hour counts employed in previous TRL studies, the magnitude of the bias was sufficiently small to be ignored. However, it had been planned to use considerably shorter counting periods in some future studies and, as a consequence, it was felt important to be able to take proper account of the errors in the flow estimates in the modelling process. To this end, the standard form of model in eqn 1 was extended to include not only random variation in the Y_i but also in the Q_i. The extended version is known as a "functional model".

8.2. A formal functional model

To illustrate how this was achieved, consider a case in which just one of the explanatory variables is a flow, and let us modify the model form eqn 1 so as to separate the flow (with true AADT z_i) from the other variables:

$$\mu_i = \lambda_i T_i = T_i \exp[\boldsymbol{\beta}^T \mathbf{x} + \gamma \log z_i] \tag{24}$$

in which Y_i, the number of accidents, is Poisson distributed with mean μ_i and Q_i, the traffic flow count, is Poisson distributed with mean $z_i t_i$. Then the log likelihood can be split into two parts, one relating to the accidents and the other to the flows:

$$\log L(\boldsymbol{\beta},\gamma,\mathbf{z}|\mathbf{y},\mathbf{q}) = \sum_i (-\mu_i + y_i \log \mu_i) + \sum_i (-z_i t_i + q_i \log(z_i t_i)) \tag{25}$$

Maximizing the log likelihood with respect to the parameters $\boldsymbol{\beta}$, γ and $\mathbf{z}$ leads to two sets of conditions. The first of these relates to the accident model and corresponds to the usual model, conditional on flow estimates $\mathbf{z}$. The second relates to the flow model and can be written as:

$$q_i - z_i t_i = \gamma(\mu_i - y_i) \quad (i = 1,2,....) \tag{26}$$

This immediately suggests an iterative form of determination of the ML solution:

(1) Initial estimates of the true AADTs are calculated: $z_i = q_i/t_i$.

(2) The latest flow estimates are used, together with the other explanatory variables, to fit the accident model, and hence obtain estimates of the $\boldsymbol{\beta}$ and γ.

(3) The latest estimates of $\boldsymbol{\beta}$ and γ, and hence of the μ_i, are used, together with q_i, t_i and y_i to produce a new estimate of z_i for each site, using eqn 26. Return to step 2.

The process continues, carrying out alternate applications of the accident model and flow model, until convergence is reached, normally after just a couple of iterations.

Following preliminary work by Maher (1989), the above approach has been developed as a formal algorithm (Wright and Barnett 1991), which has been implemented, not just for one, but for two and three flow variables, in the form of GLIM and GENSTAT macros. Simulation tests have been carried out which have demonstrated that the biases which would arise through the use of the basic, or "primitive", model are very much reduced when the modified, functional model is applied. The tests also show that the standard error estimates from the primitive model can appreciably underestimate the true standard errors, but that those from the functional model are much closer.

9. AGGREGATION OF PREDICTIONS

As was explained earlier, modelling was carried out in the TRL studies at a number of levels of disaggregation. The most detailed models were in the form of predictive relationships between accidents of a particular type and certain flows and geometric variables. Having developed these separate models for the different accident types, prediction for the total number of accidents per year, λ_{total}, at the junction is obviously achieved by summation over the various accident types:

$$\lambda_{\text{total}} = \sum_{\text{type } i} \lambda_i \tag{27}$$

In some cases, the summation would also be over arms of the junctions as well as accident types. If these predictions were independent, the overall prediction error variance could be obtained by summing the separate variances. However, it has been found that this assumption of independence cannot be justified. Analysis of the data and predictions from each of the TRL studies has consistently revealed that there are considerable correlations between the separate prediction errors. The reasons for this are likely to be as follows: as was explained earlier when describing overdispersion and the use of the NB model to take account of it, there are, for each accident type, other explanatory variables, absent from the model, which contribute to the discrepancy between the observed and predicted value. It appears, then, that there is some commonality between the "missing variables" (for example, the proportion of the time that the road surface is wet or dry) for the different accident types, so that there is correlation

between the extra random error terms induced by these missing terms. It is also likely that the flow estimates used in the models for the different accident types will contain errors (as has been discussed at some length in section 8 and that, due to the sampling of these flows at each site on one particular day, these errors will tend to be positively correlated.

Taking account of these correlations, the prediction error variance for λ_{total} will then be of the form:

$$\text{Var}(\lambda_{\text{total}}) = \sum_{\text{type } i} \sum_{\text{type } j} \rho_{ij}^{(\text{type})} [\text{Var}(\lambda_i)\text{Var}(\lambda_j)]^{1/2} \tag{28}$$

In those cases where aggregation is to be carried out over accident types and arms, a two stage process was used. Firstly, aggregation was carried out over types within each arm, to give predictions for the number of accidents per year on arm k, $\lambda_k^{(\text{arm})}$. Then, aggregation is done over arms, taking account of the correlations between the $\lambda_k^{(\text{arm})}$:

$$\text{Var}(\lambda_{\text{total}}) = \sum_{\text{arm } i} \sum_{\text{arm } j} \rho_{ij}^{(\text{arm})} [\text{Var}(\lambda_i^{(\text{arm})})\text{Var}(\lambda_j^{(\text{arm})})]^{1/2} \tag{29}$$

The magnitudes of these correlations has been found to differ somewhat from one study to another. A summary of the findings is as follows:

(a) 4-arm roundabouts (Maycock and Hall 1984): $\rho_{ij}^{(\text{type})} = 0.2$ between the five different accident types on the same arm, and $\rho_{ij}^{(\text{arm})} = 0.4$ between arms.
(b) Rural T junctions (Pickering et al. 1986): $\rho_{ij}^{(\text{type})} = 0.5$ between each of the 13 accident types.
(c) Urban 4-arm signals (Hall 1986): $\rho_{ij}^{(\text{type})} = 0.2$ between the 11 accident types, and $\rho_{ij}^{(\text{arm})} = 0.75$.
(d) Urban T junctions: $\rho_{ij}^{(\text{type})} = 0.5$ between each pair of the 16 accident types.
(e) Urban links: $\rho_{ij}^{(\text{type})} = 0.4$ between each pair of the nine accident types, and $\rho_{ij}^{(\text{side})} = 0.85$ between the two sides of the link.

After aggregation, then, predictions can be obtained for the total site accident rate, λ_{total}, and an estimate of the prediction error variance. It has been found that there is a strong and simple relationship between the two quantities, so that the uncertainty which should be attached to the predictions can be conveniently and accurately summarized in the form of a coefficient of variation C_v. The most disaggregate form of modelling gave the overall prediction methods with the lowest values of C_v, demonstrating the benefits to be derived from this more detailed level of modelling, with its greater demands in terms of data. However, the magnitude of this benefit was somewhat less than had been anticipated. The values of C_v arising from the level 3 modelling in the various TRL studies are: 0.30 (4-arm roundabouts), 0.36 (rural T junctions), 0.29 (4-arm signals), 0.53 (urban T junctions) and 0.51 (urban links).

10. COMBINING PREDICTIONS WITH SITE OBSERVED VALUES

Having established the models by means of the fitting data set, they can be used to provide predictions for any other sites, once the values of the site geometric and flow variables are known. The work described in section 9 has shown that the uncertainty to be attached to this prediction can be summarized by a C_v value. An alternative way of interpreting this is to say that the true mean accident rate per year λ is gamma distributed about the prediction $\hat{\lambda}$, with a shape parameter of C_v^{-2}. Therefore, since we assume that Y, the actual number of accidents in T years is Poisson distributed with a true mean of λT, it follows that the unconditional distribution of Y is NB.

The gamma distribution may be regarded, in Bayesian terms (see, for example, Hauer 1986, Hauer 1992; Morris et al. 1989; Brüde and Larsson 1993), as the prior distribution for λ. If we now observe y, the number of accidents in T years at the site, we can update the prior into a posterior distribution, also gamma, with a mean of:

$$\hat{\lambda}^{\text{post}} = \frac{\hat{\lambda}\dfrac{\alpha}{\hat{\lambda}^2} + \dfrac{y}{T}\dfrac{T}{\hat{\lambda}}}{\dfrac{\alpha}{\hat{\lambda}^2} + \dfrac{T}{\hat{\lambda}}} = \frac{\hat{\lambda}(\alpha + y)}{\alpha + \hat{\lambda}T} \tag{30}$$

and a shape of $(\alpha + y)$. From this, it can be seen that as $T \to 0$ (and therefore $y \to 0$ also), the posterior mean becomes simply $\hat{\lambda}$ (that is, the prior mean value), whereas as $T \to \infty$ it becomes simply y/T (the observed rate). The weights attached to the two estimates are equal, so that the posterior mean is the arithmetic mean of $\hat{\lambda}$ and y/T, if $\hat{\lambda}\, T = \alpha$. For example, if the prediction method has a value of C_v of 0.3, then, for a site at which it is predicted there will be 5 accidents per year, the prediction has the same weight as $0.3^{-2}/5$ or approximately 2 years of accident data. However, if it is predicted that there would only be 0.5 accidents per year, it is equivalent to approximately 20 years of data. The greatest benefit from the prediction method, then, is for junctions with low accident rates.

11. SUMMARY AND CONCLUSIONS

Considerable progress has been made in recent years in techniques for establishing the relationship between accidents, flows and geometry. It has now been generally recognized that the use of GLMs with

Poisson error structure is far more appropriate than conventional multiple linear regression models, based on least squares. GLMs have been applied successfully in the series of TRL junction accident studies, which have now covered most of the types of road junctions and links. In each study, the analysis has been carried out using disaggregated data, fitting separate models to each accident type. The models have been fitted using maximum likelihood estimation, generally through the use of the GLIM and GENSTAT statistical programs which contain routines for the fitting of GLMs.

However, there are certain technical problems which need to be addressed in order to ensure that the application of GLMs will produce robust and reliable results. In this paper we have dealt with a number of such problems: the low mean value problem, overdispersion, the disaggregation of data over time, allowing for the presence of a trend over time in accident risk, random errors in the flow estimates, aggregation of predictions for different accident types by allowing for the correlation between the prediction errors, and the combination of model predictions with site observations.

In each case, we have proposed some extension or modification to the basic methodology. Taken altogether, we believe that what we have described in the paper constitutes a comprehensive methodology for the development of predictive accident models.

Acknowledgements—The authors would like to express their thanks to the many colleagues at the Transport Research Laboratory who, over recent years, through their work in applying in practice the models described in this paper and in many discussions, have contributed significantly to the development of the methodology. Particular mention should be made of Geoff Maycock, Rod Kimber, Ray Vincent, Janet Kennedy, Roger Layfield, Marie Taylor, Ian Burrow and David Walmsley. Thanks also go to the referees who made helpful comments and suggestions for improvement of an earlier version of the paper.

REFERENCES

Akaike, H. Information theory and an extension of the maximum likelihood principle. In: Second International Symposium on Information Theory. Petrov, B.N.; Czaki, F. (Eds). Budapest: Akademiai Kiado; 1973: 267–281

Brüde, U.; Larsson, J. Models for predicting accidents at junctions where pedestrians and cyclists are involved. How well do they fit? Accid. Anal. Prev. 25: 499–509; 1993.

Cameron, A. C.; Trivedi, P. K. Econometric models based on count data: comparisons and applications of some estimators and tests. J. Appl. Econometrics 1: 29–53; 1986.

Dionne, G.; Desjardins, D.; Laberge-Nadeau, C.; Maag, U. Medical conditions, risk exposure and truck drivers' accidents: an analysis with count data regression models. Presented at the 37th Annual Meeting of the Association for the Advancement of Automotive Medicine, San Antonio, Texas; 1993.

Hall, R. D. Accidents at four-arm single carriageway urban traffic signals. Contractor Report CR65. Crowthorne, Berks, U.K.: Transport Research Laboratory; 1986.

Hauer, E. On the estimation of the expected number of accidents. Accid. Anal. Prev. 18: 1–12; 1986.

Hauer, E. Empirical Bayes approach to the estimation of "unsafety": the multivariate regression method. Accid. Anal. Prev. 24: 457–477; 1992.

Joshua, S. C.; Garber, N. J. Estimating truck accident rate and involvements using linear and Poisson regression models. Transport. Planning Technol. 15: 41–58; 1990.

Jovanis, P. P.; Chang, H. L. Modelling the relationship of accidents to miles travelled. Transport. Res. Record 1068: 42–51; 1986.

Lane, P. W.; Galwey, N. W.; Alvey, N. G. GENSTAT 5: An Introduction. Oxford University Press: 1988.

Maher, M. J. Fitting predictive equations to accident data with uncertainty in the flow estimates. PTRC Summer Annual Meeting. University of Sussex; 1989.

Maycock, G.; Hall, R. D. Accidents at 4-arm roundabouts. Laboratory Report LR1120. Crowthorne, Berks, U.K.: Transport Research Laboratory; 1984.

McCullagh, P.; Nelder, J. A. Generalized Linear Models. London: Chapman and Hall; 1983.

Miaou, S-P. The relationship between truck accidents and geometric design of road sections: Poisson versus nngative binomial regressions. 73rd Annual Meeting Transportation Research Board, Washington, DC; 1994.

Miaou, S-P; Lum, H. Modeling vehicle accidents and highway geometric design relationships. Accid. Anal. Prev. 25: 689–709; 1993.

Miaou, S-P.; Hu, P. S.; Wright, T.; Rathi, A. K.; Davis, S. C. Relationships between truck accidents and highway geometric design: a Poisson regression approach. Transport. Res. Record 1376: 10–18; 1992.

Morris, C. N.; Bishop, M. K.; Scaff, C. L.; Pendleton, O. J. Empirical Bayes methodology in traffic accident analyses. American Statistical Association Winter Conference, San Diego, California; 1989.

Okamoto, H.; Koshi, M. A method to cope with the random errors of observed accident rates in regression analysis. Accid. Anal. Prev. 21: 317–332; 1989.

Payne, C. D. (Ed.) The GLIM System Release 3.77 Manual. Oxford: Numerical Algorithms Group; 1985.

Pickering, D.; Hall, R. D.; Grimmer, M. Accidents at rural T-junctions. Research Report RR65. Crowthorne, Berks, U.K.: Transport Research Laboratory; 1986.

Satterthwaite, S. P. A survey of research into relationships between traffic accidents and traffic volumes. Supplementary Report 692. Crowthorne, Berks, U.K.: Transport Research Laboratory; 1981.

Wedderburn, R. W. M. Quasi-likelihood functions, generalized linear models and the Gauss–Newton method. Biometrika 61: 439–447; 1974.

Wright, D. E.; Barnett, V. Fitting predictive accident models in GLIM with uncertainty in the flow estimates. Contractor Report CR286. Crowthorne, Berks, U.K.: Transport Research Laboratory; 1991.

Zegeer, C. V.; Hummer, J.; Reinfurt, D.; Herf, L.; Hunter, W. Safety effects of cross-section design for two-lane roads, Vols I and II; prepared for the Federal Highway Administration and Transportation Research Board by the University of North Carolina; 1987.

Zegeer, C. V.; Stewart, R.; Reinfurt, D.; Council, F.; Neumann, T.; Hamilton, E.; Miller, T.; Hunter, W. Cost effective geometric improvements for safety upgrading of horizontal curves; prepared for the Federal Highway Administration by the University of North Carolina; 1990.