Article:
Connors, RD, Maher, MJ, Wood, A et al. (2013) Methodology for fitting and updating predictive accident models with trend. Accident Analysis and Prevention, 56. pp. 82-94. ISSN: 0001-4575

https://doi.org/10.1016/j.aap.2013.03.009

Methodology for fitting and updating predictive accident models with trend

Richard Connors[a], Mike Maher[a], Alan Wood[b], Linda Mountain[b] and Karl Ropkins[a]

[a] Institute for Transport Studies, University of Leeds, UK

[b] School of Engineering, University of Liverpool, UK

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.

Keywords: predictive accident models; negative binomial; overdispersion; MCMC methods; prediction uncertainty.

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 (which is the effect on the mean accident rate of variables other than those in the predictive model), and to propose the use of a negative binomial (NB) error structure in the regression modelling. This approach has since 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 1974-79 (Maycock and Hall 1984), and those for rural priority junctions on data from 1979-83 (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; DfT 2010b).

While it seems unlikely that the PAMs still in use but derived using data from 20-30 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.

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 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 13590. Virtually all links had a carriageway width of less than 9m; 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* (2012).

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:

$$\mu_i = a\, T\, Q_i^{\alpha} L_i \exp\left(\frac{2b}{L_i}\right) \qquad (1)$$

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, α = 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 20m 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 20m the correction term has the effect of multiplying the

predicted number of accidents by 317; whilst for a length of 50m, 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 50m in length, so these are excluded from the data in our analyses.

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 calculate the scaling factor as the ratio of the observed number of accidents and the predicted number. A more extensive approach is to re-fit the existing models: that is, use the same explanatory variables and functional form as in the original model but obtain new estimates for each of the parameters (a, α and b in (1)), possibly with the inclusion of an additional term to allow for trend within the period of the new data. These two cases, and the issues arising in them, will be dealt with separately in the following sections of the paper.

Re-calibration of the existing model

Estimation of long-term trend and goodness-of-fit criteria

Suppose we have a set of predictions μ_i from the TRL model for a set of sites (i = 1, .. N) , and observed accident frequencies y_i. To update the model, it is thought appropriate to modify the constant a in the model, and hence estimate a factor k by which the model predictions should be scaled to make them produce reliable predictions of the current frequencies. A number of summary statistics are available, to measure the goodness-of-fit and to compare the performance of different models, some of which have been suggested by Lord and Park (2008). Examples include the root mean squared error (RMSE), the root mean squared relative error (RMSRE), scaled deviance (SD) and the mean absolute deviation (MAD). These statistics will generally provide differing estimates of the optimal scaling factor k. For example, if we choose to scale so as to give unbiased predictions, the estimate will be: $k_1 = \sum y_j / \sum \mu_j$, as recommended firstly by Harwood *et al* (2000), by Persaud *et al* (2002), and then in the Highway Safety Manual (AASHTO, 2010). This minimises the absolute value of the mean error (AME): $\left|\sum(y_j - k_1\mu_j)\right|/N$ by making it zero. If instead we minimise the RMSE, it can be shown that this gives: $k_2 = \sum y_j\mu_j / \sum \mu_j^2$. Again, if we minimise the RMSRE, it can be shown that this gives: $k_3 = (1/N)\sum y_j/\mu_j$. Next if we maximise the log likelihood in a NB fit, this minimises the (NB) scaled deviance. (This is equivalent to the re-calibration procedure proposed by Sawalha and Sayed (2006)). Finally we could choose to minimise the MAD, so that k_5 minimises $\sum \left|y_j - k_5\mu_j\right|/N$.

To illustrate, consider the 341 rural single carriageway links, and the TRL model (1), with no scaling factor. As explained earlier, because of the vastly disproportionate effect of the correction term $\exp(2b/L)$ for short links referred to earlier, it was decided to omit those seven links of length of 50m or less. There are 996 observed accidents, whilst the sum of the (modified) predictions from the TRL model in (1) is 840.8. The scatter plot of y versus μ is shown in Figure 1, where y is the observed number of accidents at the 334 sites over $T = 5$ years, and μ is the predicted number of accidents from equation (1). As expected there is a considerable amount of scatter about the solid line ($y = \mu$) and it is not easy to tell visually whether the TRL model provides accurate predictions.

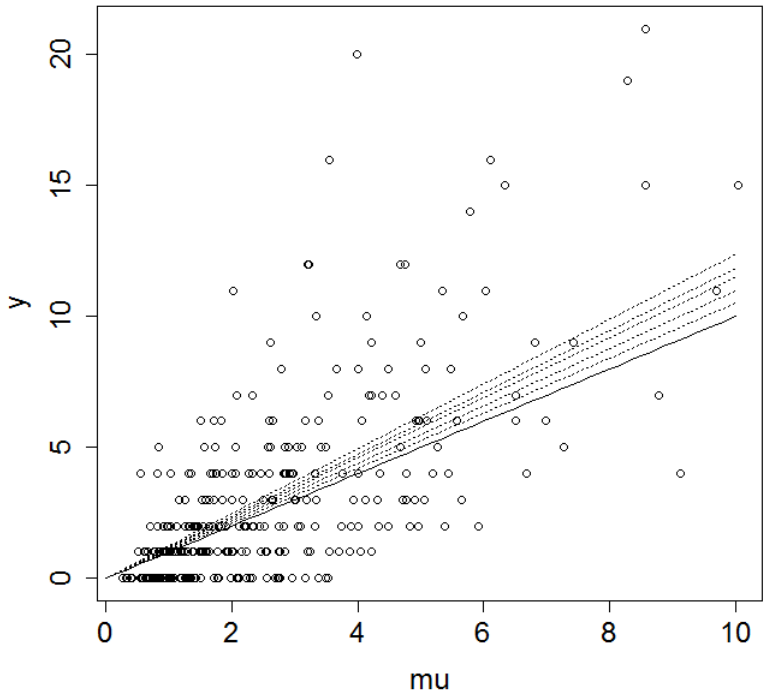

Figure 1: Scatter plot of observed accidents versus those predicted by the TRL model

Table 1 shows the values of all five scaling factors, and the goodness of fit measures obtained when using each scaling factor. The optimum values are italicised, and by definition occur on the main diagonal. The modified predictions from applying these five different scaling factors are represented in Figure 1 by the five dotted lines. It is not easy to tell visually which of these gives the best fit.

Table 1: Comparison of goodness-of-fit measures for five estimators of scaling factor

Scaling Factor	AME	RMSE	RMSRE	SD	MAD
$k_1 = 1.185$	*0.000*	2.662	1.107	0.561	1.851
$k_2 = 1.238$	0.134	*2.657*	1.113	0.564	1.873
$k_3 = 1.093$	0.217	2.692	*1.104*	0.562	1.830
$k_4 = 1.151$	0.085	2.671	1.105	*0.560*	1.841
$k_5 = 1.052$	0.334	2.719	1.105	0.566	*1.826*

There is no unique, best way of determining the scaling factor, because the different methods each optimise a different criterion for goodness of fit, although in many cases there is only a small difference between the values for the five estimators. All of the criteria are sensible and desirable (*eg* unbiassedness, minimum RMSE, max likelihood *etc*). The same considerations hold when we wish to compare the performance of alternative models: which is the best fitting model will depend on the criterion used to measure the goodness-of-fit of the models.

As well as summary statistics, graphical methods have been proposed for evaluating or comparing model performance. One such method is the CURE plot (Hauer and Bamfo, 1997) as used, for example, in Persaud *et* al (2002), and Lord and Park (2008). The CUmulative REsiduals $(y_i - \mu_i)$ are plotted against the ordered explanatory variables (or the fitted values) to examine how closely the plot follows the zero-residual line, or horizontal axis. The main benefit of this plot is that it aggregates the data, damping random fluctuations in individual residuals. Substantial deviations from the horizontal axis indicate a systematic weakness in the model.

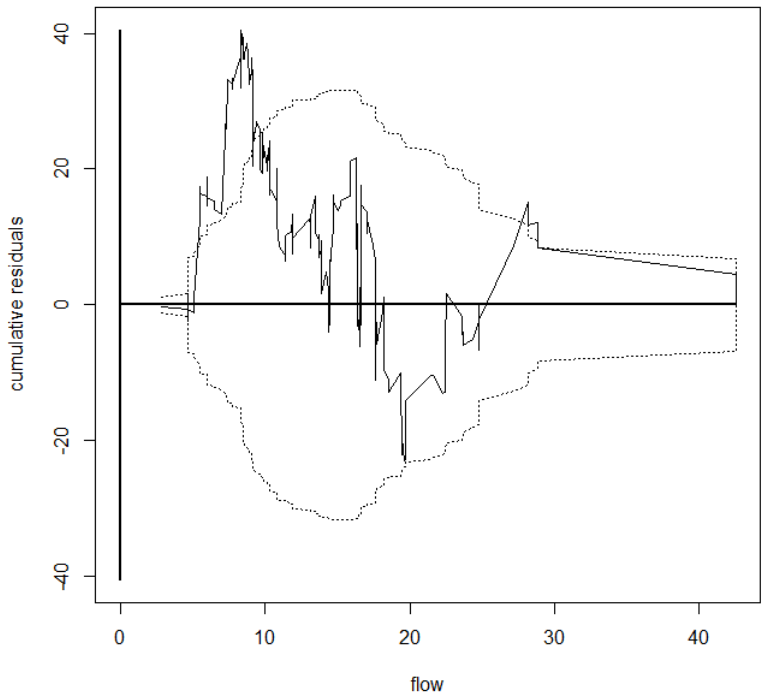

Figure 2: CURE plot of cumulative residuals against flow (AADT in 000s), with ±2 standard error lines (dotted)

A plot of raw residuals versus the flow values for the rural links data set is not very informative, because of the large amount of variability in each individual point. Figure 2 shows a CURE plot for the same data (with the predictions scaled by a factor of $k_1 = 1.185$, in order to give unbiassedness) and contains more useful information due to the degree of

aggregation (the code for this is given in section A4 of the Appendix). For example, it can be seen that for flow values between 5 and 8.5 (AADTs, expressed in thousands), there is a steady accumulation of positive residuals; followed by a steadily decline for flows between 8.5 and 11; then a shallow rise between 11 and 16; a steep fall between 16 and 20; and then a steady rise from there to 29 or so. Interpretation is therefore through the *rates of change* in the graph rather than the deviations from the horizontal axis: the steady rises and falls indicate a run of observations where the residuals are mainly positive or mainly negative.

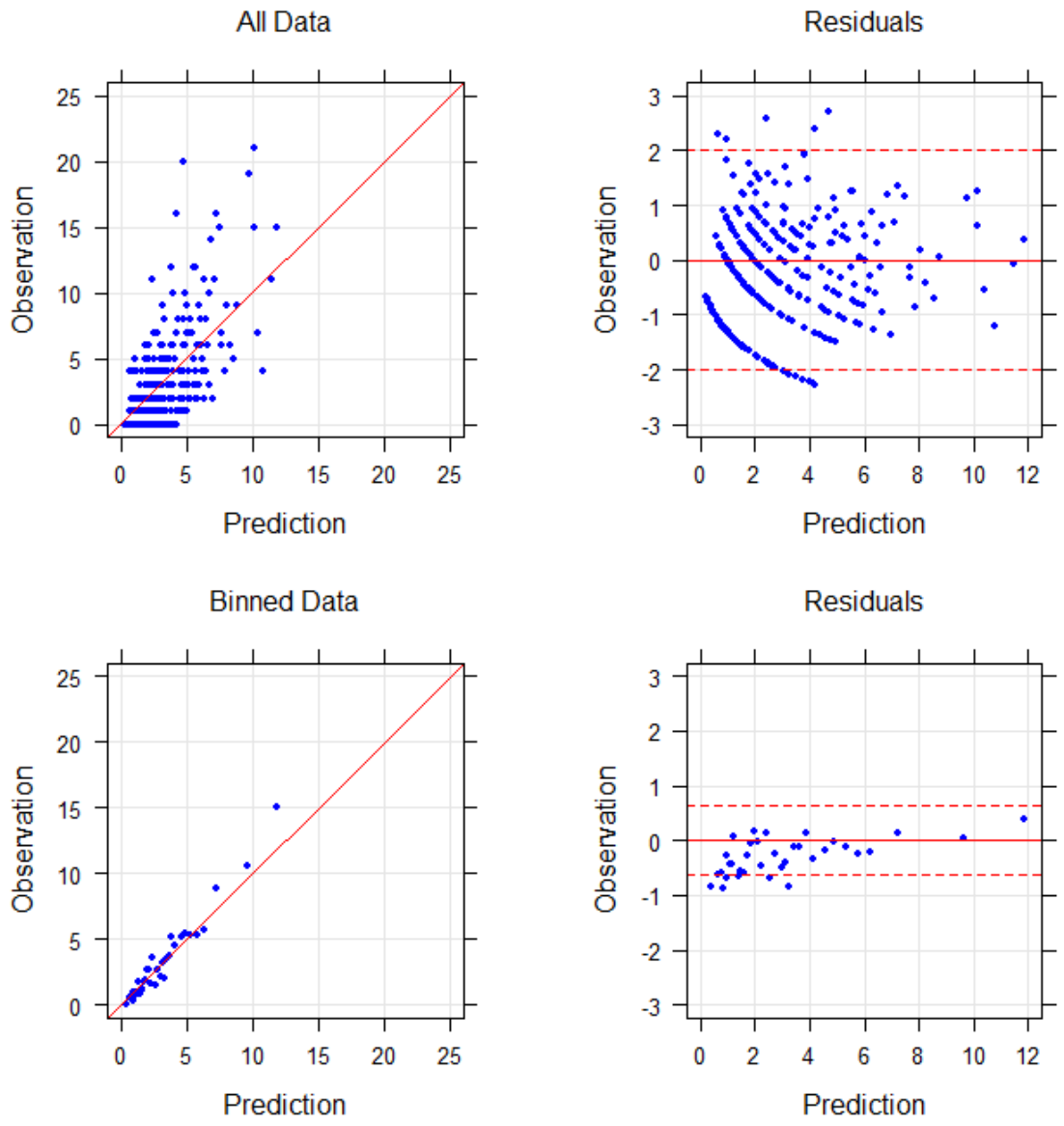

Figure 3: Plot of observed versus predicted and standardised residuals versus predicted (for individual sites and in bins of size 10)

An alternative, and simpler, approach is to use "binning", in which the sites are sorted by ascending order of the predicted value and grouped into bins containing equal numbers of successive sites. The top left-hand plot in Figure 3 shows the observed value versus the predicted, for all 334 individual sites; and the top right plot is of the standardised residuals

versus the predicted values. The lower plots are similar but for the grouped data, using 34 "bins" each containing nine or ten sites. In the left-hand plot, the mean of the observed values in each group is plotted against the mean of the predictions, and in the right-hand plot the mean of these standardised residuals is plotted against the mean predicted value for each group. The horizontal dotted lines in the residual plots indicate approximate 95% confidence bands. It can be seen that, by aggregating the data, a lot of the variability is removed and this permits a more informative view of the data and any systematic deviations.

Re-fitting the existing model

Distributional assumptions for overdispersion

Equation (1) gives a prediction for the expected number of accidents μ_i at site i in terms of the flow and design variables. Typically it is assumed that the observed number of accidents y_i is Poisson distributed about the mean m_i for site i, where the difference between m_i and μ_i is due to other variables, not in the model, that affect the actual value of the mean at that site and thereby contribute to what is known as overdispersion.

Conventionally it is assumed that m_i follows a gamma distribution about μ_i, so that the combined distribution of y_i is negative binomial (NB). This has been motivated by computational convenience, as there is no particular reason to suppose that the overdispersion truly follows a gamma distribution. Availability of suitable and easy-to-use software inevitably plays a role in the formulation of models and most statistical software packages include a routine for fitting regression models with NB errors.

For some time now it has been known that alternative distributions for the overdispersion are equally plausible and whilst less straightforward they can be fitted using either Markov Chain Monte Carlo (MCMC) methods or methods involving numerical integration and maximum likelihood. There is potentially a long list of possible distributions that are appropriate to describe continuous and non-negative variables but it is not practicable to consider them all. Distributions that have been proposed and used include the lognormal, Weibull, variable shaped gamma and others, in addition to the standard gamma (see Maher and Summersgill (1996), Maher and Mountain (2009), Lord and Park (2008), Lord and Geedipally (2011), Cheng *et al* (2012) for example).

The general form of model can be formulated as: y_i is Poisson distributed with a mean m_i and $m_i = f_i\mu_i$ $(i = 1,..n)$ where the site factors f_i are randomly and independently drawn for each site i from an appropriate overdispersion distribution scaled so that its mean is 1. Each overdispersion distribution has a parameter that measures its variability or spread: for example the shape parameter r in the gamma, the standard deviation σ in the lognormal, and v in the Weibull. Table 2 shows the density functions for these distributions, the parameterisation used, and the expressions for the mean and the coefficient of variation, C_v (the ratio of standard deviation to mean).

Table 2: Density functions of distributions, and expressions for $E(X)$, C_v

Distribution	Density function: $f(x)$	$E(X)$, C_v
Gamma $X \sim G(r, \beta)$	$\frac{1}{\Gamma(r)} \beta^r x^{r-1} \exp(-\beta x)$	$E(X) = \frac{r}{\beta}$, , $C_v = \frac{1}{\sqrt{r}}$ $E(X) = 1$ needs $r = \beta$
Lognormal $log(X) \sim N(d, \sigma^2)$	$\frac{1}{x}\sqrt{\frac{1}{2\pi\sigma^2}} \exp\left(-\frac{(\ln(x) - d)^2}{2\,\sigma^2}\right)$	$E(X) = \exp\left(d + \frac{\sigma^2}{2}\right)$ $C_v = \sqrt{\left(e^{\sigma^2} - 1\right)}$ $E(X) = 1$ needs $d = -\frac{\sigma^2}{2}$
Weibull $X \sim W(v, \lambda)$	$v\lambda x^{v-1} \exp(-\lambda x^v)$	$E(X) = \frac{1}{\lambda^{1/v}} \Gamma\left(1 + \frac{1}{v}\right)$ $C_v = \sqrt{\frac{\Gamma\left(1 + \frac{2}{v}\right)}{\Gamma^2\left(1 + \frac{1}{v}\right)} - 1}$ $E(X) = 1$ needs $\lambda = \left(\Gamma\left(1 + \frac{1}{v}\right)\right)^v$

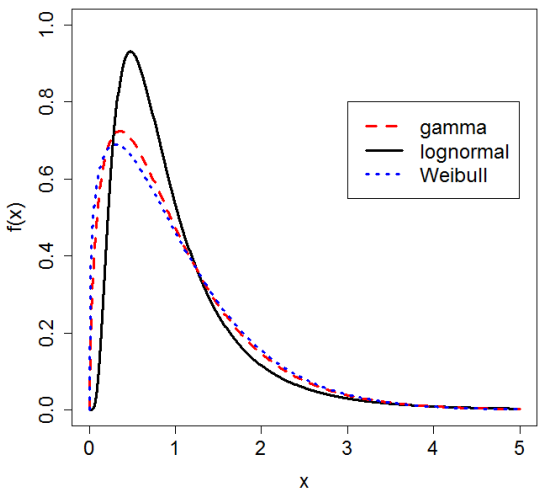

Figure 4: gamma, lognormal and Weibull density functions, with $E(X) = 1$ and $C_v = 0.8$.

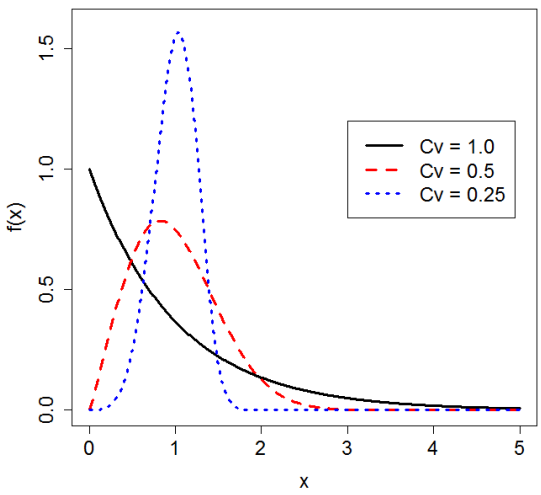

Figure 5: Weibull density functions, with $E(X) = 1$ and C_v = 1.0, 0.5 and 0.25.

Figure shows a comparison of these density functions when all have a mean of 1, and C_v values of 0.8. To achieve this, the parameter values required are: $r = \beta = 1.563$ (gamma), $d = -0.247$ and $\sigma = 0.703$ (lognormal), and $\lambda = 0.913$ and $v = 1.258$ (Weibull). Figure 5 shows Weibull densities for three different C_v values, all with unit mean, showing very different shapes. Clearly the choice of distribution may play a role in determining the goodness of fit achievable from any model for overdispersion, but the degree of dispersion (C_v value) is the dominant factor.

Fitted models: fixed shape

The key question is whether, for any particular data set, the regression parameter estimates are sensitive to the distribution used, and whether one distribution fits significantly better than others. We fit the PAM in (1) to the links data (with the 5-year totals and the average flow) to obtain estimates of a, α and b, using gamma, lognormal and Weibull distributions, and denote these models by PG (Poisson-gamma), PLN (Poisson-lognormal) and PW (Poisson-Weibull) respectively. The PG model can be fitted exactly using negative binomial regression in R (R Development Core Team (2006)). The resulting parameter estimates (with standard errors in parentheses) are shown in the first row of Table 3. The PG, PLN and PW models were then all fitted using MCMC methods, in WinBUGS (Lunn *et al*, 2000), using a burn-in stage of 5000 iterations, followed by 25000 further iterations to collect the statistics on the parameters. Core parts of the WinBUGS code for the lognormal model are given in Appendix A1. The estimates from these are given in the next three rows of Table 3, denoted by PG-MCMC etc, with the goodness-of-fit given by the DIC (Deviance Information Criterion), a measure similar to the AIC (Akaike Information Criterion) – see Lunn *et al* (2000). Finally, the three models were fitted using numerical integration / maximum likelihood (NIML) techniques. In these the value of the likelihood for each observation, for any given values of the parameters, was achieved using numerical integration with 1000 steps (corresponding to evenly-spaced quantiles of the overdispersion distribution), and the log likelihood (logL) was then maximised using the `maxLik` optimisation function in R. Like MCMC methods, the NIML method can be quite time-consuming, so is not necessarily ideal for everyday use. It is however deterministic (unlike MCMC methods), and can be used to serve as a method that can be used for any form of overdispersion distribution. The estimates produced by this method are denoted in in Table 3 by PG-NIML, PLN-NIML etc. The fact that the estimates from the PG-exact and PG-NIML are virtually identical provides reassurance that the numerical integration method produces estimates of sufficient accuracy, so that we make take the estimates from the PLN-NIML and PW-NIML models, where we do not have any exact fitting method, as being reliable.

It can be seen that there is relatively little difference between the parameter estimates for the gamma, lognormal and Weibull models, and that the values of the dispersion parameters are such that the implied coefficients of variation C_v are again very similar, suggesting that whichever overdispersion distribution is assumed, the *amount* of overdispersion, as measured by the C_v value, is what is most important.

More noteworthy however is that if one was to choose between the three models on the basis of the MCMC fits, it is the Weibull that has distinctly the lowest value of the DIC and so would be the preferred model. However, when one looks at the results from the NIML fits, one sees that all three models have virtually the same logL value, so that there is nothing to choose between the three models. The reason for this would appear to be that whilst the DIC value is a valid way to choose between nested models with different numbers of parameters, it is less reliable when choosing between non-nested models as we have here. The DIC is based on the scaled deviance statistic which is twice the difference between the log likelihood of the fitted model and the log likelihood of the "full" model (in which there are as many

parameters as observations – see McCullagh and Nelder (1983)), with a penalty for the number of parameters used in the model. Here, the number of parameters is identical for all three models, but the value of the log likelihood of the full model is different for the three models, (being -465.1 for the PG, -457.9 for the PLN and -469.9 for the PW), and so this base value for the deviance is different. The conclusion is that choosing the best fitting model between PG, PLN and PW models cannot be done reliably using their DIC values from MCMC fits. The conclusion here, for this data set, is that the fits from the three models are virtually identical, so that it makes no difference which overdispersion distribution is assumed.

Table 3: Comparison of estimates from fixed-shape models

Model	a	α	b	dispersion	C_v	DIC/logL
PG-exact	0.08902	0.7210 (0.12)	0.0411 (0.013)	r = 2.95 (0.53)	0.5822	-652.6
PG-MCMC	0.09064	0.7352 (0.12)	0.0396 (0.013)	r = 2.976 (0.55)	0.5797	1233.57
PLN-MCMC	0.08726	0.7489 (0.11)	0.0385 (0.013)	σ = 0.5652 (0.05)	0.6135	1243.95
PW-MCMC	0.09248	0.7227 (0.11)	0.0405 (0.012)	v = 1.806 (0.18)	0.5731	1226.87
PG-NIML	0.08898	0.7212 (0.12)	0.0411 (0.013)	r = 2.95 (0.53)	0.5823	-653.0
PLN-NIML	0.08384	0.7473 (0.12)	0.0398 (0.013)	σ = 0.5603 (0.05)	0.6073	-653.0
PW-NIML	0.09324	0.7017 (0.11)	0.0418 (0.012)	v = 1.799 (0.17)	0.5751	-653.1

Fitted models: variable shape

Lord and Park (2008) illustrate the application of what they refer to as a generalised negative binomial distribution in which they allow the NB dispersion parameter (or reciprocal of the shape), as well as the mean μ, to be a regression function of the covariates instead of taking a fixed value for all sites. This form of model is rather similar to the variable-shaped NB, suggested by Cameron and Trivedi (1986), and applied by Maher and Summersgill (1996), in which r the shape parameter, instead of being constant for all observations, is a function of the predicted mean μ. The Lord and Park model was fitted to the links data, with the shape r having the same functional form as μ in (1) but with parameters a', α' and b' instead of a, α and b and a power s on the link length L. Code is given in section A2 of the Appendix to show how this was fitted in R. The technique adopted was to fit the fixed-shape negative binomial model first (using the function `glm.nb`), and use the estimates of the parameters a, α and b to provide initial values (together with $a' = r_0$, $\alpha' = 0$, $b' = 0$ and $s = 0$ where r_0 is the estimate of the fixed shape) in a maximum likelihood estimation routine using the function `maxLik`. The NB likelihood for a single observation is given in (2).

$$p(Y = y) = \frac{\Gamma(y+r)}{y!\Gamma(r)}\left(\frac{r}{r+\mu}\right)^r\left(\frac{\mu}{r+\mu}\right)^y \qquad (2)$$

The estimates obtained were: a = 0.0896, α = 0.7248, b = 0.0397, a' = 0.0697, α' = 0.7720, b' = 0.0281 and s = 0.5816, with logL = -649.71. As we know, the shape r and the coefficient of variation are directly related by $C_v = 1/\sqrt{r}$. Figure 6 shows a scatter plot of the fitted C_v value versus the predicted μ values, from which a clear and strong relationship can be seen. From this it would seem simpler to allow C_v to be a power function of μ, as this is a more

parsimonious and smoother way than the regression function as done by Lord and Park. Therefore we allow C_v to be a power function of μ: $C_v = c\mu^n$, The resulting fitted function for the shape has estimates c = 0.8701 and n = -0.3114, as indicated by the dotted line in Figure 3, with the value of the log likelihood (logL) = -650.13. So this representation is preferable, given that it uses fewer parameters, is smoother in form and gives almost as good a log likelihood value.

This variable-shape Poisson-gamma (VS-PG) model then includes, as special cases, the fixed shape PG model ($n = 0$) and the quasi-Poisson ($n = -½$) and gives a better fit than the fixed-shape model.

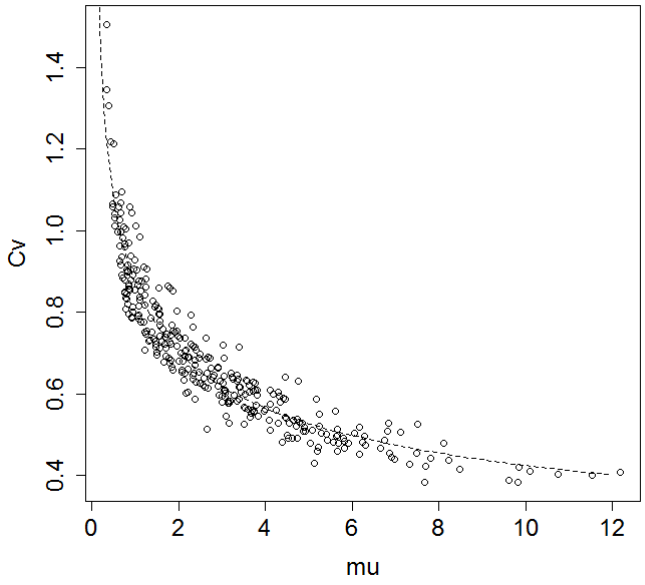

Figure 6: scatter plot of fitted C_v versus $\hat{\mu}$ from the Park and Lord generalized gamma model

It is possible to bring the same type of flexibility into the other two distributions. By inverting the expression for C_v in

Table 2 for the lognormal distribution, it is possible to see how the value of σ^2 for a site must be related to the predicted mean μ, for any particular values of c and n in $C_v = c\mu^n$.

For the Weibull distribution it is not possible to invert the relationship between C_v and v. However, the plot of $\log(C_v)$ versus $\log(1/v)$ in Figure 7 reveals an almost linear relationship over the range of practical interest: $v > 1$, and $C_v < 1$. Therefore, if we allow $1/v$ to follow a power law with μ in the Weibull case, it is almost equivalent to assuming the same power law relationship between C_v and μ as used in the gamma and lognormal cases. Hence we can fit models VS-PG, VS-PLN and VS-PW that are each generalisations of the PG, PLN and PW cases. These three variable shape models may be again fitted using MCMC

methods (WinBUGS code illustrating the VS-PW model is shown in section A3 of the Appendix) or by the NIML method described earlier. The results are shown in Table 3.

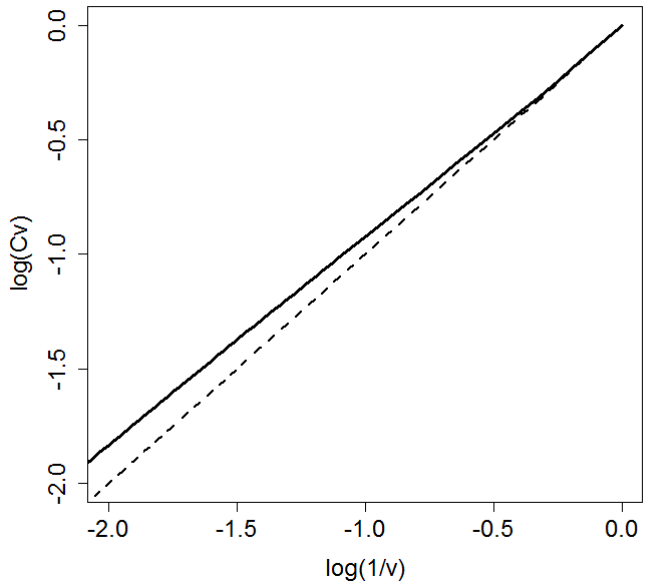

Figure 7: plot of $log(C_v)$ versus $log(1/v)$for the Weibull model, showing their close approximation

Table 4: Comparison of estimates from variable-shape models

Model	a	α	b	dispersion	DIC/logL
VS-PG-exact	0.08922	0.7278 (0.12)	0.0391 (0.014)	c = 0.8699 (0.15) n = -0.3114 (0.13)	-650.1
VS-PG-MCMC	0.08986	0.7417 (0.11)	0.0394 (0.014)	c = 0.8667 (0.16) n = -0.2966 (0.14)	1221.93
VS-PLN-MCMC	0.09084	0.7370 (0.11)	0.0416 (0.015)	c = 1.029 (0.28) n = -0.3669 (0.18)	1236.84
VS-PW-MCMC	0.09882	0.7024 (0.10)	0.0392 (0.014)	c = 0.8621 (0.17) n = -0.3299 (0.14)	1217.34
VS-PG-NIML	0.08924	0.7277 (0.12)	0.0391 (0.014)	c = 0.8698 (0.15) n = -0.3112 (0.13)	-650.1
VS-PLN-NIML	0.08704	0.7368 (0.11)	0.0409 (0.015)	c = 0.9762 (0.24) n = -0.3607 (0.17)	-650.8
VS-PW-NIML	0.09160	0.7180 (0.11)	0.0389 (0.014)	c = 0.8462 (0.16) n = -0.3260 (0.14)	-650.8

It can be seen firstly that the results for the VS-PG models from the exact method and from the NIML method are virtually identical, demonstrating the accuracy of the numerical integration method. Secondly we again can see that if the best-fitting model were to be selected on the basis of the DIC values from the MCMC fits, the VS-PW model would be the choice. However, when the logL values from the three models are compared from the NIML fits, it can be seen that they are almost identical, and that it is the VS-PG that is marginally fitting best. (The values of the log likelihood for the full models are -466.9 for the VS-PG, -457.8 for the VS-PLN and -471.0 for the VS-PW). Therefore again, as with the fixed-shape models, it can be seen that the choice of overdispersion distribution has virtually no effect on the fit. Comparison of the results from Tables 3 and 4 shows that the variable shape model fit significantly better than the fixed shape models.

The reason why the spread of the overdispersion distribution reduces as the predicted value increases is probably because of aggregation effects. Sites with higher μ are generally sites that have longer length, or have long observation period, or have more traffic. A longer link, for example, might be thought of as being an aggregate of several shorter sub-links, with similar but not necessary identical overdispersion errors. The aggregation of these sub-links with partially correlated errors leads to a reduction in the relative amount of overdispersion.

Overall, then, we conclude that basing the choice of overdispersion distribution on the DIC values from MCMC fits can be very misleading, that the choice of overdispersion distribution has virtually no effect on the goodness-of-fit but that the freedom of allowing the C_v to vary with the mean has a more significant effect, and that this beneficial effect is of similar magnitude for each of the three distributions.

Estimation of current trend: a multinomial model

The links data consists of annual accident frequencies y_{it} and annual flows Q_{it} over the period 2005-2009 ($i = 1, \ldots, 334$; $t = 0, \ldots, 4$), along with link lengths L_i and other design variables that do not change over time. The disaggregate nature of the data (in the sense of disaggregation by year and not by accident type) allows the estimation of a current trend term of the form $\exp(\beta t)$ that multiplies the base year (2005) prediction so that we have:

$$\mu_{it} = a\, Q_{it}^{\alpha} L_i \exp\left(\frac{2b}{L_i}\right) \exp(\beta t) \qquad (3)$$

as the predictive model.

Whilst we could use the annual data as the observational unit, this involves the assumption of complete independence between all observations. This issue will be discussed in the next section. An alternative approach that avoids this assumption is to consider the model in two linked parts. In the first we model the aggregate data: using the total accidents at a site y_i and relating it to the average flow $\bar{Q}_i$ to obtain estimates of the parameters a, α and b. In the second part, we model the distribution of the total accidents at each site across the five years via a multinomial distribution. Using the estimate of α, the probability of observing the distribution of y_i (total accidents at site i) is given by

$$P(y_{i1}, \dots y_{i5}) = \frac{y_i!}{y_{i1}! \dots y_{i5}!} p_{i1}^{y_{i1}} \dots p_{i5}^{y_{i5}} \quad (4)$$

where the probabilities are given by:

$$p_{it} = \frac{Q_{it}^{\alpha} \exp(\beta t)}{\sum_s Q_{is}^{\alpha} \exp(\beta s)} \quad (5)$$

Then, to find the maximum likelihood estimate of β, we need to maximise

$$z = \sum_i \sum_t y_{it} \log(p_{it}) \quad (6)$$

with respect to the single parameter β using, for example, the `maxLik` function in R on the log likelihood function in (6) . To complete the loop between parts 1 and 2, we iterate between estimation of the trend and the aggregate fitting, with a revised set of weighted average flows:

$$\bar{Q}_i = \sum_t w_t Q_{it} \quad \text{where } w_t = \frac{\exp(\beta t)}{\sum \exp(\beta s)} \quad (7)$$

and repeat until convergence. R code for this is in section A5 of the Appendix.

To illustrate, consider the re-fitting of the model (1) using annual data on accidents and flows, with an exponential trend term. In the first iteration we assume $\beta = 0$, and an NB model is fitted to the aggregate data to give estimates for a, α and b. In the second half of the iteration, these estimates are used to fit the multinomial model in (5) and (6) to re-estimate β. Table 5 shows the estimates at the end of successive iterations. The rapid convergence is clear, even with the excessive level of precision quoted here.

Table 5: Convergence of the multinomial model

Iteration	a	α	b	β
1	0.08872734	0.7301594	0.03350563	-0.06004986
2	0.08848451	0.7310552	0.03347281	-0.06004527
3	0.08848452	0.7310551	0.03347281	-0.06004527
4	0.08848452	0.7310551	0.03347281	-0.06004527

The advantage of this approach is its simplicity: it uses the aggregate accident data in fitting the predictive model, and the multinomial model to estimate the trend (which here is a $100 * (1 - e^{\beta}) \approx 6\%$ per year reduction in accident risk). Crucially in this method, no assumptions are made about the independence of observations from difference years at the same site: a topic which will be discussed in the next section.

Assumption of independence

When using disaggregate data for the accident frequencies y_{it} and flows Q_{it} the question arises as to whether it is safe to assume independence between years at each site. Given that the overdispersion error represents the effect of the unobserved design variables of a site, it would seem likely that this effect remains the same from one year to another at any site, so

that although the Poisson errors are independent, the overdispersion errors are common (or at least highly correlated) across different years at any site. Therefore if the number of accidents at site i in year t is denoted by y_{it} it may be better to assume that these are drawn randomly and independently from Poisson distributions with means m_{it} and that $m_{it} = f_i\,\mu_{it}$ where the factor f_i is randomly drawn from the overdispersion distribution independently for each site i but is constant for all years t at any site. The question arises as to whether the fitted model from this formulation is much different from that from a model in which independence between the y_{it} is assumed for all sites and all years.

We therefore have two alternative versions of the model:

M1: $y_{it} \sim Poisson(m_{it})$ where $m_{it} = f_{it}\mu_{it}$ and the f_{it} are independent for all i, t.

M2: $y_{it} \sim Poisson(m_{it})$ where $m_{it} = f_i\mu_{it}$ and the f_i are independent for all i.

These models were fitted to the links data using WinBUGS for the four overdispersion distributions (gamma, lognormal, Weibull and VSG) and for the two versions of each model. The results are shown in Table 6. In the VS-PG, the coefficient of variation $C_v = c\,\mu^n$ where μ is the per-year average prediction for a site. The WinBUGS code for version M2 is given in section A6 of the Appendix.

Table 6: Results from models with alternative independence assumptions

Model	a	β	dispersion	C_v	DIC
M1: PG	0.0733	-0.063	r = 3.475	0.5364	3065.8
M2: PG	0.0714	-0.060	r = 2.934	0.5839	2969.2
M1: PLN	0.0732	-0.062	σ = 0.5144	0.5504	3088.3
M2: PLN	0.0716	-0.060	σ = 0.5502	0.5946	2980.2
M1: PW	0.0732	-0.062	v = 2.047	0.5120	3052.2
M2: PW	0.0716	-0.061	v = 1.799	0.5752	2963.0
M1:VS-PG	0.0730	-0.059	c = 0.519 n = -0.264	0.623 (μ = 0.5) 0.466 (μ = 1.5)	3049.2
M2: VS-PG	0.0730	-0.062	c =0.529 n = -0.298	0.650 (μ = 0.5) 0.469 (μ = 1.5)	2957.6

The results show the superiority of the VS-PG over the three other models but more strongly show the superiority of version M2 in each case. This confirms that the overdispersion error is, for any site, constant from year to year. The DIC for model (M1) indicates this assumption of independence (in both i and t) is not valid.

If it is not safe to assume independence between years at the same site, it might also be the case that it is not safe to assume independence between all links on the same scheme. Miaou and Song (2005), Wang *et al* (2009), and Noland and Quddus (2004) have fitted models that include a spatial autocorrelation effect: in the first case, for ranking sites (intersections in a city and rural links across a state), allowing for the distance between sites; in the second case, between neighbouring segments of a motorway when investigating the possible effect of congestion on accident frequency; and in the latter, between neighbouring wards when

considering the effect of deprivation on road casualties. Recall that in the rural links data there are 334 links drawn from 73 schemes: some schemes contain just a single link, whilst others comprise up to ten links. Therefore if link i is part of scheme n, it could be that there is correlation between the overdispersion errors for links on the same scheme, on the grounds that the design of one the link within a scheme will be similar to that of another link within the same scheme. In an extreme case, with perfect correlation, the y_i may be drawn randomly and independently from Poisson distributions with means $m_i = g_n\mu_i$ where the overdispersion factor g_n is randomly and independently drawn from the overdispersion distribution for each scheme n but is the same for all links i within that scheme.

Here we can formulate three alternative versions of the model:

M1: $y_{it} \sim Poisson(m_{it})$ where $m_{it} = f_i\mu_{it}$ and the f_i are independent for all i.

M2: $y_{it} \sim Poisson(m_{it})$ where $m_{it} = g_n\mu_{it}$ and the g_n are independent for all n.

M3: $y_{it} \sim Poisson(m_{it})$ where $m_{it} = f_i g_n\mu_{it}$ and the f_i are independent for all i, and the g_n are independent for all n.

These were fitted for the gamma model, with results as in Table 7. In each of M1 and M2 there is just one source of overdispersion error: from the link and the scheme respectively. In the case of M3 there are two separate sources: from the link (with shape r^i) and from the scheme (with shape r^n). It can be seen from the results that the best fitting model is M1, as it clearly has the lowest DIC value. The very large shape value for the scheme errors in M3 confirms that there is no virtually no correlation between the errors for different links within the same scheme. So, the conclusion is that it is safe to assume independence between all links, including those on the same scheme: that is there is no discernible scheme effect.

Table 7: comparison of three model versions, investigating scheme effects

Model	a	β	dispersion	C_v	DIC
M1: PG	0.0714 (0.0045)	-0.059	$r^i = 2.925$	0.5364	2969.1
M2: PG	0.0781 (0.0058)	-0.058	$r^n = 6.300$	0.5839	3073.9
M3: PG	0.0721 (0.0049)	-0.061	$r^i = 3.053$ $r^n = 259.6$	0.5504	2992.4

Similar fits were carried out for other distributional forms (Weibull and lognormal), and the conclusions were the same.

Prediction uncertainty

The purpose of fitting models is generally to use them for prediction. Whilst the fitted model, such as that in (1) or (2), provides a point estimate for μ, it is usually desirable to have an idea of the uncertainty that should be attached to this prediction of the number of accidents to be expected at a new site: either in the form of a standard error or as a confidence interval.

Wood (2005) showed how these may be derived for the standard negative binomial regression model. Here we provide a more general treatment for a variety of models and fitting methods. Since $m = f\mu$, with $E(f) = 1$, we have $\log(m) = \log(f) + \log(\mu)$, so that because of the independence of the $\log(f)$ and $\log(\mu)$variables, we can write:

$$Var(\log(m)) = Var(\log(f)) + Var(\log(\mu)) \quad (8)$$

Because of this additive form, and the fact that both μ and m are non-negative, it is best to obtain confidence intervals on the log scale, and then transform. The second term on the right hand side is the variance of the linear predictor: $\eta = \log(\mu)$ and, following a fit in R, $Var(\eta)$ is given by the function `predict` for any specified new sites with given values of Q, L and T. Since $E(f) = 1$, $Var(\log(f))$ is given approximately by C_v^2 which, for the gamma case, is equal to $1/r$ where r is the shape estimated in the fitting process.

For example, consider the model in (1) with gamma overdispersion, the estimates from which are given in the first row of Table 3, fitted using R. Details of the calculations, and the R code, for the calculations and plotting of 95% CIs for both μ and m can be found in section A7 of the Appendix. Figure 8(a) shows a plot of the 95% confidence intervals for μ and m, for sites that have values of Q ranging from 0 to 60 (the range in the fitting data set is from 2 to 42), whilst holding T at 5 years and the link length L at 1km (close to the mean for the data set). It is noticeable how the CI widens appreciably as Q increases, especially beyond the range in the original data. It can also be noted how much wider the confidence intervals are for m compared with those for μ.

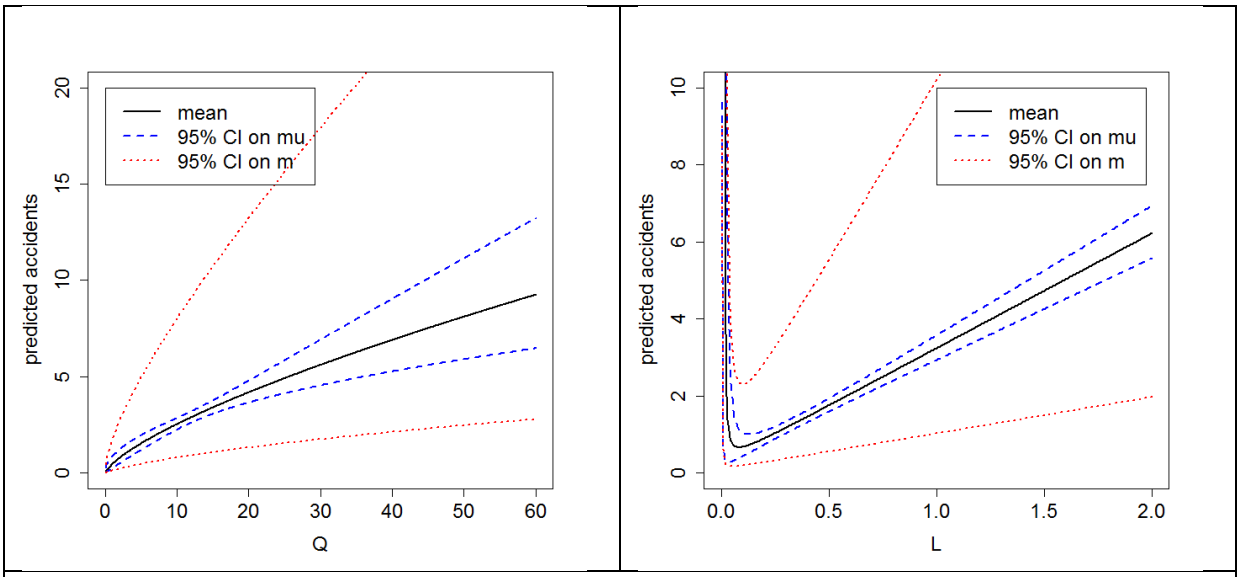

Figure 8: plots showing the 95% confidence intervals on predicted values of μ and m as (a) Q varies, and (b) as L varies (gamma case, using R).

Similarly if we investigate the prediction uncertainty when we vary L over the range 0 to 2 kms, whilst keeping Q = 14 (the approximate mean of the data), and T = 5 years, we obtain the plot in Figure 8(b), where the rapid rise in the predicted number of accidents is evident as L becomes small. The CI width also increases rapidly. This effect was discussed earlier and

as a consequence seven short links (of less than 50m) were omitted from the fitting process. Altogether these plots illustrate the potential dangers in extrapolating beyond the extent of the fitting data set.

For models fitted using WinBUGS, similar information can be obtained rather simply by adding a set of dummy links for which predictions, and their uncertainties, are to be calculated. These extra links do not affect the model fitting (they have no values for the observed numbers of accidents) but use the fitted model. In Figure 9, for example, a lognormal distribution was assumed for the overdispersion (the model whose results were given in the third row of Table 3). In addition to the 334 real links, 60 dummy links were included. These had no observed accident frequencies, but all had T = 5 years, and L = 1km, and values of flow Q ranging from 1 to 60. At the end of the model run, using 25,000 iterations following 5000 iterations for burn in, the output showed not only the statistics for the model parameters a, b, α and σ but also for the μ_i and m_i (i = 1, ... 60), including the mean, 2.5% percentile and 97.5% percentile. From these the graphs in Figure 9(a) can be drawn. Figure 9(b) shows the same sort of results for variation in the link length L which was allowed to vary over the range 0 to 2 kms. The slight wobbles in the upper limit for m are due to the process being Monte Carlo. The overall nature of these plots for this lognormal case, fitted in WinBUGS, is very similar to those produced for the gamma case, fitted in R, in Figure 8, thereby showing a reasonable degree of robustness of the uncertainty estimation to the distributional assumptions.

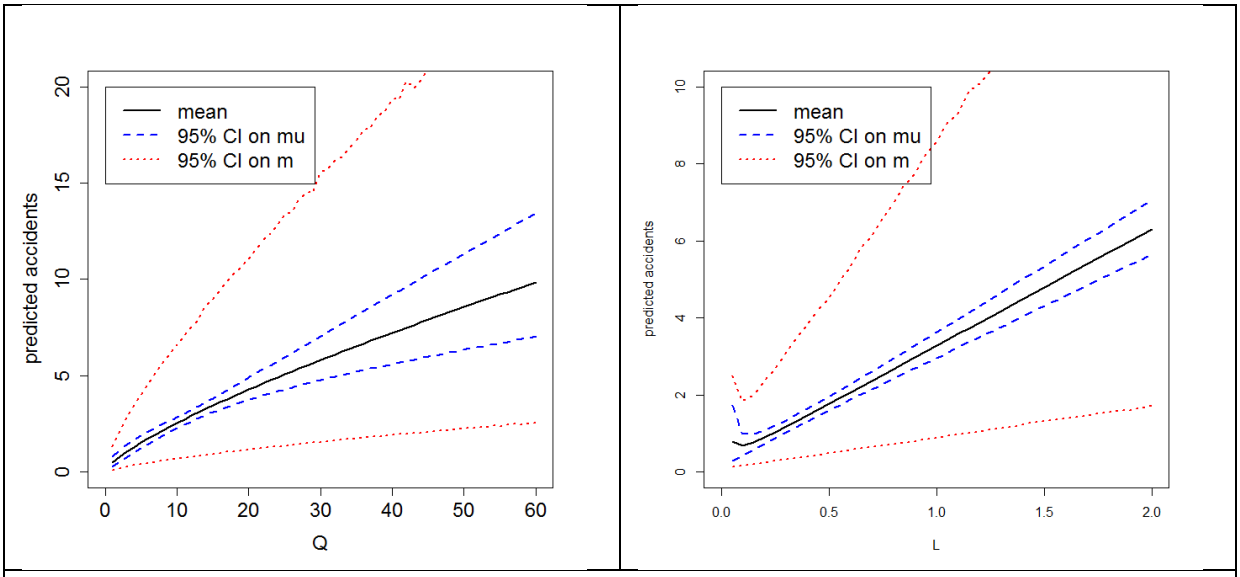

Figure 9: plots showing the 95% confidence intervals on predicted values of μ and m as (a) Q varies, and (b) as L varies (lognormal case, using WinBUGS).

Conclusions

The paper has reported on a number of methodological issues that arise in the re-calibration and the re-fitting of existing predictive accident models, here illustrated by application to data on the accidents occurring over a five-year period on 334 rural single-carriageway links in the UK. These issues are not specific to that particular type of data but may arise in a wide

variety of circumstances. Therefore the intention is to contribute to the steadily-widening literature on the subject of predictive accident modelling.

In the context of scaling up an out-of-date model to enable it to fit well to current data, the question of goodness-of-fit measures arises as there are a number of seemingly sensible criteria by which to assess the effect of applying any given scaling factor; these include the unbiassedness, minimising the RMS error or the RMS relative error, and minimising the mean absolute deviation or scaled deviance. The paper highlights the fact that these different criteria generally will result in quite different values of the scaling factor and that there is, therefore, no unique or correct criterion. Others have proposed graphical approaches, such the CURE plot, in which the pattern of the cumulative residuals is inspected, to provide insights into where the model does not fit well, or to confirm its satisfactory behaviour. The paper has suggested the use of "binned" residual plots as a way of overcoming the inevitable variability associated with individual observations and their residuals, as it is easier to detect any systematic pattern without the eye being distracted by a small number of outliers.

The widely-used negative binomial regression model implies that the overdispersion distribution is gamma. Developments in statistical modelling software (especially the wider use of MCMC methods) have demonstrated that other distributions such as the lognormal and Weibull are practicable alternatives to the gamma. The paper has shown that when these alternative models are fitted, their resulting C_v values are very close, and that more substantial improvements in goodness-of-fit come from allowing this coefficient of variation to depend on the predicted mean μ in a power function. The form and effect of this additional freedom is somewhat similar to that adopted by Lord and Park (2008), but uses fewer parameters. This variable shape device $C_v = c\mu^n$ may be equally-well applied to other distributions, and not just to the gamma, it has been shown. Of some importance is the finding that the DIC values from MCMC fits are not a reliable way to compare the goodness-of-fit for models with different overdispersion distributions. The numerical integration maximum likelihood fits showed that there was hardly any difference between the log likelihood values from the different models. This suggests that the choice of overdispersion distribution is not particularly important.

When using disaggregate (eg annual) flow and accident count data in order to fit a model with current trend estimated, it is important to consider the way in which the overdispersion errors are modelled. Since the overdispersion error (the difference between a site mean $m_i$and the predicted value μ_i) is due to site factors that are not included in the predictive model, it seems likely that these site factors will largely persist at a site from year to year. Therefore it is important to recognise this in the model formulation and not treat the observations from all links and all years as if they were independent. It is relatively easy, using MCMC methods, to formulate a model either with independent overdispersion errors f_{it} or with constant (year-to-year) errors f_i . Comparisons between these alternative formulations showed, with the links data set, that the latter model gave a far superior DIC value, confirming that the overdispersion effect does remain constant at any site from year to year, and that the independence model is incorrect. A simple alternative way to fit such a

model, without the need for MCMC methods was shown to be the use of the multinomial model to estimate the trend effect.

Finally, whilst in principle almost any model form for the distribution of observations can be used and the likelihood function written as a numerical integral, and maximised using some general function minimisation routine, users generally are influenced in the choice of model formulation by the availability of software and algorithms to solve the model. Therefore, the negative binomial model, with fixed shape, has dominated in predictive accident modelling work in recent years. But the availability of MCMC methods, and other algorithms in open source software such as R now means that the typical user has a wider choice of model and solution methods than was previously the case. By providing details of algorithms, and code for R and WinBUGS, we hope that this encourages others to try alternative model formulations and thereby improve the fit to their data.

Acknowledgements
The authors wish to express their thanks to the UK Engineering and Physical Sciences Research Council for their funding to the universities of Leeds and Liverpool that enabled this project to proceed, and to Lancashire County Council, and many other local authorities, for their help in providing accident and flow data.

References

AASHTO (2010). *Highway Safety Manual – 1st Edition*.

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

Cheng L., Geedipally S.R. and Lord D. (2012). The Poisson-Weibull generalized linear model for analyzing motor vehicle crash data. Transportation Research Board 91st annual meeting, Washington DC, January 22-26, 2012.

DfT (2010a) *Road Casualties Great Britain 2009*. Department for Transport, London, UK.

DfT (2010b) *Transport Trends 2009*. Department for Transport, London.

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

Harwood, D.W., Council, F.M., Hauer, E., Hughes W.E., and Vogt A. (2000). *Prediction of the Expected Safety Performance of Rural Two-Lane Highways*. FHWA-RD-99-207, U.S. Department of Transportation..

Hauer, E. and Bamfo, J. (1997). Two tools for finding what function links the dependent variable to the explanatory variables. In: Proceedings of the ICTCT 1997 Conference, Lund, Sweden.

Lord D. and Geedipally S.R. (2011). The negative binomial–Lindley distribution as a tool for analyzing crash data characterized by a large amount of zeros. *Accident Analysis and Prevention* 43, 1738 – 1742.

Lord D. and Mannering F.L. (2010). The statistical analysis of crash-frequency data: a review and assessment of methodological alternatives. *Transportation Research Part A: Policy and Practice* 44 (5), 291–305.

Lord D. and Park P. Y-J. (2008). Investigating the effects of the fixed and varying dispersion parameters of Poisson-gamma models on empirical Bayes estimates. *Accident Analysis and Prevention*, 40, 1441-1457.

Lunn, D.J., Thomas, A., Best, N., and Spiegelhalter, D. (2000) WinBUGS - a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing*, 10, 325--337. http://www.mrc-bsu.cam.ac.uk/bugs/

Maher M.J., and Mountain L.J. (2009) The sensitivity of estimates of regression to the mean. *Accident Analysis and Prevention* 41, 861-868.

Maher M.J. and I. Summersgill (1996). A comprehensive methodology for the fitting of predictive accident models. *Accident Analysis and Prevention* 28(3), 281-296.

Maycock G. and Hall R.D. (1984) Accidents at four-arm roundabouts. Report LR1120, Transport Research Laboratory, Crowthorne, UK.

McCullagh P. and Nelder J.A. (1983). *Generalized Linear Models*. Chapman and Hall, London.

Miaou S-P. and Song J.J. (2005). Bayesian ranking of sites for engineering safety improvements: decision parameter, treatability concept, statistical criterion, and spatial dependence. *Accident Analysis and Prevention* 37, 699–720.

Mountain L.J., Hirst W.M. and Maher M.J. (2005). Are speed enforcement cameras more effective than other speed management measures? The impact of speed management schemes on 30mph roads. *Accident Analysis and Prevention* 37, 742-752.

Noland R.B. and Quddus M.A. (2004). A spatially disaggregate analysis of road casualties in England. *Accident Analysis and Prevention* 36 (6), 973–984.

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

Persaud B. and Lyon C. (2007) Empirical Bayes before–after safety studies: Lessons learned from two decades of experience and future directions. *Accident Analysis and Prevention* 39, 546–555.

R Development Core Team (2006). R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org.

Sawalha Z. and Sayed T. (2006). Transferability of accident prediction models, *Safety Science,* 44(3), 209-219.

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

Summersgill I., Kennedy J. and Barnes D. (1996) Accidents at 3-arm priority junctions on urban single-carriageway roads. Report 184, Transport Research Laboratory, Crowthorne, UK.

Walmsley D.A., Summersgill I. and Binch C. (1998) Accidents on modern rural single-carriageway trunk roads. Report 336, Transport Research Laboratory, Crowthorne, UK.

Wang C., Quddus M.A. and Ison S.G. (2009). Impact of traffic congestion on road accidents: a spatial analysis of the M25 motorway in England. *Accident Analysis & Prevention,* 41(4), 798-808.

Wood A., Mountain L.J., Connors. R. and Maher M.J. (2012). Updating predictive accident models of modern rural single carriageway A-roads. To appear in *Transportation Planning and Technology.*

Wood G.R. (2005). Confidence and prediction intervals for generalised linear accident models. *Accident Analysis and Prevention* 37, 267-273.

Appendix A

Here we give sample code (for R and WinBUGS) for some of the model fitting and other calculations carried out in the paper. In each case, only the core part of the code is shown: that is, excluding data entry and manipulation. The versions used were R 2.14.1 and WinBUGS 14.

A1. Fitting a Poisson-lognormal model in WinBUGS

```
# lognormal model for SC links data
# y = accidents; Q = flow; L = link length
model{a <- exp(k)
  sigma <- 1/sqrt(tau); av <- -0.5/tau  # to make E(f)=1
  for(j in 1:NLinks){
    y[j] ~ dpois(m[j])
    m[j] <- f[j]*mu[j]
    log(mu[j]) <- k+log(L[j])+alpha*log(Q[j])+b*(2/L[j])
    f[j] ~ dlnorm(av, tau)
   } # end Links loop
 k ~ dnorm(0,0.0001)
 alpha ~ dnorm(0,0.0001)
 b ~ dnorm(0,0.0001)
 tau ~ dgamma(0.001,0.001)
}
```

A2. Fitting the Lord and Park generalised negative binomial model in R

```
require(MASS);require(maxLik); require(miscTools)
# uses max likelihood to fit variable-shaped NB
# y = vector of accidents; Qav = vector of flows;
```

```
# L = vector of link lengths; T = no. years data
nrate <- 2/L

# fit standard (fixed-shape) gamma first
fit <- glm.nb(y ~ log(Qav)+nrate+offset(log(T*L)))
k0 <- summary(fit)$coef[1,1]; alpha0 <- summary(fit)$coef[2,1];
b0 <- summary(fit)$coef[3,1]; shape0 <- theta.ml(fit);
param0 <- c(k0,alpha0,b0,shape0,0,0)
# extracted coeffs to act as initial values in VS fit

# set up log likelihood function
logL <- function(param){
k <- param[1];alpha <- param[2];b <- param[3];
k1 <- param[4];alpha1 <- param[5]; b1 <- param[6]; s <- param[7]
mu <- exp(k)*T*(Qav^alpha)*L*exp(b*nrate)
shape <- exp(k1)*T*(Qav^alpha1)*(L^s)*exp(b1*nrate)
t1 <- log(gamma(y+shape))-log(gamma(y+1))-log(gamma(shape))
t2 <- shape*log(shape)+y*log(mu)-(shape+y)*log(mu+shape)
z <- sum(t1)+sum(t2)}

# now optimise log likelihood for variable-shaped case
optbeta <- maxLik(logL,start=param0)
summary(optbeta)
```

A3. Fitting a variable-shaped Poisson-Weibull model in WinBUGS

```
# variable shape Weibull model for SC links data
# y = accidents; Q = flow; L = link length
model{a <- exp(k)
  for(j in 1:NLinks){
    y[j] ~ dpois(m[j])
    m[j] <- f[j]*mu[j]
    log(mu[j]) <- k+log(L[j])+alpha*log(Q[j])+b*(2/L[j])
    v[j] <- 1/(c*pow(mu[j],n))  # 1/v is power function of mu
    lam[j] <- pow(exp(loggam(1 + 1/v[j])),v[j]) # to make E(f)=1
    f[j] ~ dweib(v[j], lam[j])
  }   # end Links loop
 k ~ dnorm(0,0.0001)
 alpha ~ dnorm(0,0.0001)
 b ~ dnorm(0,0.0001)
 n ~ dnorm(0, 0.001)
 c ~ dgamma(0.001,0.001)
}
```

A4. Drawing a CURE plot in R

```
nLinks = 334; nrate <- 2/L; T <- 5; alpha <- 0.831; b <- 0.0576
a <- 0.0552;
mu <- a*T*(Q^alpha)*L*exp(b*nrate)
res <- y - mu   # residuals
# order the data by increasing value of Q
o <- order(Q) ; rord <- res[o]; Q <- Q[o]; muord <- mu[o]
r <- cumsum(rord); vr <- cumsum(muord); vmax <- max(vr)
# r = cumulative resids; vr = var(cumulative resids)
se = sqrt(vr*(1-vr/vmax))  # se(cumulative residuals) (adjusted)
Qmax <- max(Q); rmin <- min(r); rmax <- max(r)
plot(c(0,Qmax),c(rmin,rmax),type="n",xlab="flow",ylab="cumulative
residuals"); lines(Q,r,lwd=0.25);
lines(c(0,Qmax),c(0,0),lwd=2.0); lines(c(0,0),c(rmin,rmax),lwd=2.0)
lines(Q,2*se,lwd=0.25,lty="dotted");lines(Q,-2*se,lwd=0.25,lty="dotted")
```

A5. Multinomial model for fitting model with trend in R

```
require(MASS);require(maxLik); require(miscTools)
# y[i,t] = accidents on link i in year t
# Q[i,t] = flow on link i in year t
# L = vector of link lengths
# T = no. years; N = no. links
ytot <- dim(N); Qav <- dim(N);p <- dim(T);mu <- dim(T);sumL <- dim(N);w <-
dim(T)
nrate <- 2/L;
beta <- 0  # initial estimate of beta (current trend)
for(i in 1:N) {ytot[i] <- sum(y[i, ])}

#  Stage 1: fit using aggregate data, with current beta value, to estimate
k, alpha, b
for(t in 1:T){w[t] <- exp(beta*(t-1))}
for(i in 1:N){Qav[i] <- sum(w*Q[i, ])/sum(w)}  # weighted average of flow
fit <- glm.nb(ytot ~ log(Qav)+nrate+offset(log(T*L)))
k <- exp(fit$coef[1]); alpha <- fit$coef[2]; b <- fit$coef[3]

# Stage 2: fit multinomial model, with current k, alpha, b to get beta
estimate
beta0 <- beta
logL <- function(beta){
for(i in 1:N){
for(t in 1:T){mu[t] <- k*(Q[i,t]^alpha)*L[i]*exp(b*nrate[i])
p[t] <- mu[t]*exp(beta*(t-1))}
sump <- sum(p)
for(t in 1:T){p[t] <- p[t]/sump}
sumL[i] <- sum(y[i, ]*log(p))}
totL <- sum(sumL)}
optbeta <- maxLik(logL,start=beta0)  # find MLE of beta
beta <- optbeta$estimate  # new estimate of beta
k; alpha; b; beta   # print latest estimates.  Return to Stage 1.
```

A6. WinBUGS code for fitting VS-PG model with trend

```
# VSG model for SC links data with trend
# overdispersion error same in each year
# y[j,t], Q[j,t] = accidents/flows on link j in year t; L[j] = link length
model{a <- exp(k)
    for(j in 1:NLinks){
    f[j] ~ dgamma(shape[j], shape[j]) # ensures E(f) = 1
    for (t in 1:5){
      y[j,t] ~ dpois(m[j,t])
      m[j,t] <- f[j]*mu[j,t]
      log(mu[j,t]) <- k+log(L[j])+alpha*log(Q[j,t])+b*(2/L[j])+beta*(t-1)
    } # end t loop
  muav[j] <- mean(mu[j, ])  # average mu for this site
  cv[j] <- c*pow(muav[j],n)  # cv a function of av mu
  shape[j] <- 1/sqrt(cv[j])
  } # end Links loop
  k ~ dnorm(0,0.0001)
  alpha ~ dnorm(0,0.0001)
  b ~ dnorm(0,0.0001)
  beta ~ dnorm(0,0.0001)
  n ~ dnorm(0,0.001)
  c ~ dgamma(0.001,0.001)
}
```

A7. Prediction uncertainty in R for the Poisson-gamma model

```
# y = accidents; L = link length(km); T = time period; Q = flow (000s)
nrate <- 2/L  # junction rate (per km)
fit <- glm.nb(y ~ log(Q)+nrate+offset(log(T*L)))
# plot of prediction and uncertainty for range of values of Q
# keeping T = 5, and L = 1
Qnew <- 0.1*c(1:600)  # range of Q's for predictions: (0,60)
Tnew <- rep(5,600); Lnew <- rep(1,600); nratenew <- 2/Lnew
new <- data.frame(Qnew,nratenew,Tnew,Lnew)
colnames(new) <- c("Q","nrate","T","L")
pred <- predict(fit,new,se.fit=TRUE)
eta <- pred$fit; etase <- pred$se.fit; mu <- exp(eta)
mulow <- exp(eta - 1.96*etase); muhigh <- exp(eta + 1.96*etase)
lmse <- sqrt(etase^2+1/shape);
mlow <- exp(eta - 1.96*lmse); mhigh <- exp(eta + 1.96*lmse)
plot(c(0,60), c(0,20),type = "n",xlab="Q",ylab="predicted accidents",
cex.axis=1.5,cex.lab=1.5)
lines(Qnew,mu,lty="solid",col="black",lwd=2)
lines(Qnew,mlow,lty="dotted",col="red",lwd=2)
lines(Qnew,mhigh,lty="dotted",col="red",lwd=2)
lines(Qnew,mulow,lty="dashed",col="blue",lwd=2)
lines(Qnew,muhigh,lty="dashed",col="blue",lwd=2)
legend(0,20,c("mean","95% CI on mu","95% CI on
m"),lty=c("solid","dashed","dotted"),col=c("black",
"blue","red"),lwd=2,cex=1.5)
```