Pergamon

Accid. Anal. and Prev., Vol. 28, No. 6, pp. 695–707, 1996

0001-4575/96 $15.00 + 0.00

PII: S0001-4575(96)00042-5

ACCIDENT PREDICTION MODELS FOR ROADS WITH MINOR JUNCTIONS

LINDA MOUNTAIN,[1,*] BACHIR FAWAZ[1] and DAVID JARRETT[2]

[1]Department of Civil Engineering, University of Liverpool, Brownlow Street, Liverpool, L69 3BX, U.K. and [2]School of Mathematics and Statistics, Middlesex University, The Burroughs, London, NW4 4BT, U.K.

(Received 27 November 1995; in revised form 16 April 1996)

Abstract—The purpose of this study was to develop and validate a method for predicting expected accidents on main roads with minor junctions where traffic counts on the minor approaches are not available. The study was based on data for some 3800 km of highway in the U.K. including more than 5000 minor junctions. The highways consisted of both single and dual-carriageway roads in urban and rural areas. Generalized linear modelling was used to develop regression estimates of expected accidents for six highway categories and an empirical Bayes procedure was used to improve these estimates by combining them with accident counts. Accidents on highway sections were shown to be a non-linear function of exposure and minor junction frequency. For the purposes of estimating expected accidents, while the regression model estimates were shown to be preferable to accident counts, the best results were obtained using the empirical Bayes method. The latter was the only method that produced unbiased estimates of expected accidents for high-risk sites. Copyright © 1996 Elsevier Science Ltd

Keywords—Road accidents, Highway safety, Prediction models, Junctions

INTRODUCTION

Estimates of expected accidents are required to assess the effects of accident remedial measures and to select highway locations that might benefit from remedial treatment. In the case of the effects of remedial measures, the objective is to compare the accident occurrence after treatment with that expected had the treatment not been implemented. In the case of site selection, it is necessary to assess whether a particular highway section or junction is unusually hazardous: whether more accidents occur than expected given the nature of the site and level of exposure (normally traffic flow for junctions, flow and length for links). Indeed when, as is common, road safety budgets are limited, it is also necessary to attempt to rank sites so that those which are either most hazardous or, ideally, where treatment would be most cost-effective, are investigated and treated first. The estimation of expected accidents, while fundamental to road safety assessments, is by no means straightforward.

The obvious approach is to use observed accidents. Accident totals or rates can be compared at treated sites in the periods before and after scheme implementation; sites can be ranked according to their observed accident totals and/or rates. While commonly used, accident counts are prone to numerous sources of error, most of which are well known and well documented: see, for example, Hauer (1980), Maher (1991), and Jarrett (1994). These include errors due to random variation, exaggerated estimates of treatment effects due to regression-to-mean effects, and trends due to external factors such as traffic growth or national and local safety policies and programmes. In the case of accident rates (e.g. accidents per million vehicles or per million vehicle-kilometres) there is an added problem if the relationship between accidents and exposure is non-linear as empirical evidence suggests (Satterthwaite 1981; Persaud and Dzbik 1993; Hauer et al. 1994; Fridstrøm et al. 1995). Accident rates are normally used to allow for variations in exposure to compare safety levels at different locations or times on the assumption that a lower accident rate indicates a safer site. If the relationship between accidents and exposure is non-linear, accident rates will vary with exposure so that a lower accident rate would not necessarily imply a safer site (Hauer 1995).

Alternatively expected accidents can be estimated using predictive models which relate accident occurrence to traffic volume and a range of attributes such as highway design features, traffic control features,

*Author for correspondence.

and site geometry. The quality of the estimates will depend on the quality of the models used. Appropriate, well-fitting models can give better estimates of expected accidents than counts (Persaud and Dzbik 1993). However, even these models will tend to underestimate the expected after accidents at sites selected on the basis of high accident frequencies in the before period and hence produce biased estimates of treatment effectiveness at high risk sites (Mountain et al. 1995b). In addition, there will inevitably be a degree of unexplained between site variation because all the factors which influence the accident potential at a site can never be included: sites which are similar in terms of the explanatory variables used in the regression model will still have different underlying mean accident frequencies due to site characteristics which remain unique and unmeasured (e.g. weather, quality of street lighting, and condition of the road surface).

Many models currently used have been developed using conventional regression modelling which assumes a normal error structure for the dependent variable (Satterthwaite 1981). However, since accident counts are non-negative and not normally distributed, such models are inappropriate: a generalized linear modelling approach which allows the use of a negative binomial error structure is preferable for accident models (e.g. Abbess et al. 1981; Maycock and Hall 1984; Miaou and Lum 1993). In the case of highway sections, in spite of evidence to the contrary, it is common to assume that accidents on links are proportional to traffic flow and link length, and to quote the coefficient of proportionality (the accident rate per million-vehicle-kilometres) for various carriageway types. In the U.K., for example, such accident rates are quoted for use in the cost benefit analysis (COBA) of alternative highway schemes (U.K. DoT 1981). Accidents at junctions are ideally modelled separately. Junction models do not normally assume a linear relationship between accidents and conflicting flows (e.g. U.K. DoT 1981; Maycock and Hall 1984; Pickering et al. 1986; Hall 1987; Hauer 1988). However, problems can arise because the application of such models requires, as a minimum, a knowledge of entry flows. Frequently, for junctions with minor roads (i.e. local distributors or access roads), suitable traffic counts are not available. Separate modelling of these junctions is not then possible unless counts are obtained specifically for the purposes of modelling. This could be an expensive procedure given the large number of traffic counts likely to be required, particularly at the site selection stage, and would in any case only provide data concerning current flows. A possible solution is to simply include minor junctions as part of a link, basing estimates of expected accidents on link and junction combined accident rates (e.g. U.K. DoT 1981). However, given the high risk of accidents at junctions relative to other locations on the highway network, variations in the frequency of minor junctions could have a considerable influence on accident frequencies and the estimates obtained are unlikely to be reliable. This is a particular problem when potential accident remedial sites are assessed because, for the relatively short sections of highway normally considered for treatment, minor junctions can be particularly important in determining the expected accident frequency (Mountain et al. 1995b).

Research has shown that the accuracy of regression model predictions can be improved using an empirical Bayes (EB) approach (see for example Hauer et al. 1988; Pendleton 1991; Mountain et al. 1992; Persaud and Dzbik 1993 and Kulmala 1994). In this, account is taken of the fact that sites which are identical in terms of the independent variables included in a regression model (i.e. have the same predictive model estimate of the expected accident frequency (μ)) will, in practice, have different expected accident frequencies (m) due to unexplained between site variation. In the EB method m is estimated as a weighted combination of the observed accidents at the site prior to any treatment (x_b) and a predictive model estimate of the expected accident frequency (μ): the inclusion of x_b takes into account specific site characteristics not included in the regression model while μ smooths out random variation. The EB estimate of the mean accident frequency is given by

$$\hat{m} = \alpha\mu + (1-\alpha)x_b \quad (1)$$

where the size of the weight α depends on the variance of m about its mean μ (i.e. on the accuracy of μ as a predictor of m). If the between site variation in m is assumed to be gamma distributed with shape parameter γ then

$$\alpha = [1 + \mu/\gamma]^{-1} \quad (2)$$

The limitations of the EB approach lie primarily in the availability of suitable predictive equations and of estimates of the shape parameter, γ.

The purpose of this study was to develop and validate a method for predicting expected accidents on main roads with minor junctions where traffic counts on the minor arms are not available. The approach did not assume a linear relationship between accidents and exposure and was based on developing regression models using the generalized linear modelling package, GLIM, which allows the specification of a negative binomial error structure for the dependent variable. The estimates of expected accidents based on the regression models developed were compared with models where accidents are assumed to be proportional to exposure, with accident counts, and with EB estimates.

THE DATA

The data used in the study comprised details of highway characteristics, accidents and traffic flows on networks of main roads in seven UK counties for periods of between 5 and 15 years. The networks represented a total of some 3800 km of highway. The road networks were restricted to UK A- and B-roads outside major conurbations: motorways were excluded on the basis that they were not the direct responsibility of the county councils, C-roads and unclassified roads on the basis that traffic counts were not generally available for minor roads[1]. In addition to road type, the network roads were categorized according to carriageway type (single or dual) and speed limit ($\leqslant$40 mph (urban) and $>$40 mph (rural)). Records of all personal injury accidents occurring on the network and reported to the police during the study period were obtained. Damage only accident data is not normally collected in the UK and thus this study was restricted to injury accidents only. For the purposes of this study the relevant information for each accident was the date and location, although other information was included in the database. In the case of flows, records were obtained of all counts carried out on the networks during the study period. These were converted to annual average daily traffic flows (AADT) using appropriate factors and flows for any years without counts obtained by linear interpolation. Data concerning the composition of the traffic (e.g. the percentage of heavy goods vehicles), its speed and levels of pedestrian flow, while potentially relevant to the prediction of accident frequencies, were not generally available and thus not used in this analysis.

Junction accidents were defined as accidents occurring within 20 m of the extended kerblines of the junction. Major junctions were classified as junctions between network roads (i.e. between A- and/or B-roads); minor junctions as junctions between a network road and minor roads (i.e. C-roads or unclassified roads). The modelling of major junction accidents will be the subject of a separate paper.

[1] The prefixes A, B or C given to classified highways in the UK are part of a numbering system rather than a classification system but the prefixes do generally coincide with the importance of the highway in the national road network: A-roads are generally roads of national or regional importance; B-roads typically connect smaller centres of population; C-roads and unclassified roads form the local distributors and access roads.

In order to allocate accidents to minor junctions, locational information was required. The database was set up using the FoxPro relational database package. Within the database each county highway network was represented by a node and link system. In this, a node was either a major junction, or a point on a highway where there was a change of carriageway type or speed limit; links were sections of highway between nodes. Locations of nodes, treatments, accidents and minor junctions were then represented in terms of a kilometerage along a specified road number from a specified node. Within the database, accidents at minor junctions were allocated by comparison of the kilometerages of accidents and minor junctions. The determination of whether a study section included any minor junctions was achieved by comparing the start and end kilometerages of the study section with those of the minor junctions.

Some of the characteristics of the highway networks used to develop the regression models are summarized in Table 1. Of the 3800 km of highway included, some 75% were subject to a speed limit of more than 40 mph. However, of the 26,000 accidents occurring on these roads, more than 50% occurred on roads subject to speed limits of 40 mph or less. An average of about one third of the total link accidents on urban roads occurred at minor junctions, as compared with one fifth on rural roads. The frequency of minor junctions on urban roads was approximately three times that for rural roads. The observed total link annual accident rates averaged over the seven counties corresponds fairly closely with the COBA link and junction combined accident rate for rural roads but for urban roads the rates tend to be rather lower than the COBA values. This may be due in part to the exclusion of major conurbations from the sample data and in part due to local variations: there were substantial between county variations in accident rates.

Figure 1 illustrates the influence of minor junctions on the observed total link accident frequencies on the networks. The links in the study networks were grouped according to the number of minor junctions per kilometre and the average annual accident frequency for each group plotted against the mean number of minor junctions per kilometre. The plots show an increase in accident frequency with minor junction frequency for all six link categories.

In order to compare the estimates of expected accidents obtained using the alternative estimation methods, data ideally would be required for sites where the expected accident frequency was known. While expected accidents cannot be observed, locations which are unaffected by any form of engineering

Table 1. Network data for seven counties

	Urban			Rural		
	A single	A dual	B single	A single	A dual	B single
Length of network (km)	731.1	61.1	213.9	1 914.7	242.1	621.2
Number of minor junctions	1 798	122	502	2 036	260	641
Average number of minor junctions/km	3.0	3.2	2.4	1.1	1.2	1.2
Total link accidents	10 980	1 184	1 926	9 234	1 358	1 301
Percentage of total link accidents at minor junctions	32%	36%	30%	21%	26%	25%
Observed annual accident rate (per million vehicle-km)						
link & minor junction	0.75	0.69	0.69	0.38	0.21	0.51
COBA* annual accident rates (per million vehicle-km)						
link & junction	0.946	1.022	0.997	0.355	0.233	0.457
link only	0.319	0.335	0.402	0.244	0.165	0.344

*Values supplied by UK Department of Transport to be published for use in COBA10.

scheme likely to affect accidents could be an appropriate group provided that the expected accidents at these untreated sites are not subject to changes over time arising from general accident trends. Accident trends may arise due to various national or local safety policies and programmes including enforcement, education and legislation. The existence of trend at untreated sites was investigated using a time series of the annual accident totals at these sites (Mountain et al. 1995a). No trend in total injury accidents was detected although there was a decreasing trend in fatal and serious accidents. Since this study is concerned with total injury accidents it was thus assumed that trend effects could be ignored: untreated sites could be used to compare the estimates of expected accidents obtained using the alternative estimation methods.

For untreated sites studied over two time periods (notional before and after periods), the expected accidents in the after period will be m even when the sites are selected on the basis of the accident frequency in the before period. In order to generate a sample of sites similar in character to typical treated sites (in terms of location, length, type, accident frequency etc.), the untreated sites were sections of highway which had been subject to some form of engineering scheme during the study period but studied during a 2 year period when they were not treated. Accidents in the first year without treatment were taken as the notional before accidents, x_b, and those in the second year as the after accidents, x_a.

Where untreated sites spanned a node it was necessary to separate the site into its component parts for the purposes of modelling. For example, if the site included a change of speed limit from say 40 mph to derestricted, then the urban and rural components of the site would have different predictive equations and, for the EB method, the components would require separate $\hat{m}$ estimates, based on the μ and x_b values for the appropriate segment. Where the node was a major junction a similar procedure was required but the major junction itself formed an additional component part of the site. Since this paper is concerned only with modelling link and minor junction

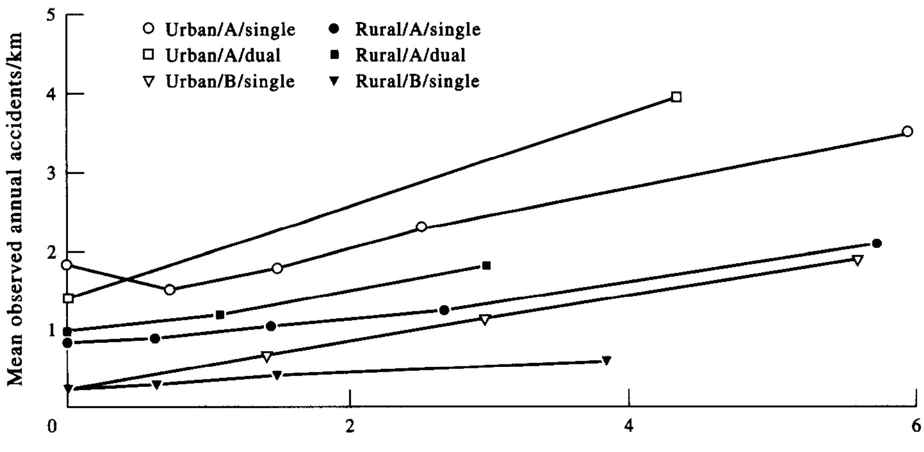

Fig. 1. Variation of annual accidents with the number of minor junctions per kilometre.

accidents, the major junction components were not included in the analysis.

THE REGRESSION MODELS

The regression models were developed using a generalized linear modelling approach based on the GLIM statistical package. An excellent description of this approach is given by Maycock and Hall (1984). Initially, the most appropriate form of the model was investigated using a Poisson error structure. The results are summarized in Table 2. The basic model, in which μ is assumed to be proportional to the link length (l) and traffic flow (t) (i.e. $\mu = \alpha t l$) as in COBA (UK DoT 1981), is compared with three alternative model forms as indicated in the table. These are all log-linear models with link length and traffic flow included as explanatory variables. The second and third alternative models also include minor junctions, in the form of total number of minor junctions and number of minor junctions per kilometre respectively. The form of these latter models is the simplest given that the regression analysis involves a logarithmic transformation and Fig. 1 would seem to confirm that this form is not unreasonable.

Since the deviance for each model is much larger than its degrees of freedom, the models are over-dispersed relative to the Poisson distribution. The main reason for this is that, although accidents at a site can be reasonably assumed to be Poisson distributed, the overall error structure of the model will also include between site variations in accident rates and will not therefore be Poisson. To take this into account, the significance tests for comparing each of the three alternative models with the basic model are based on the mean deviance ratio (MDR) calculated as:

$$\text{MDR} = \frac{\text{Deviance difference}/(df_1 - df_2)}{\text{Residual deviance}/df_2}$$

In this case, df_1 is the degrees of freedom of the basic linear model and df_2 of the alternative model. The residual deviance is the deviance of the alternative model. The MDR can be compared with the critical points of the F distribution in the usual way (Maycock and Hall 1984).

The results given in Table 2 show that all of the alternative models give a significantly ($p < 0.01$) better fit to the data than the linear model and confirm that minor junctions are an important explanatory variable in the prediction of link accident frequencies. The best fit model includes minor junctions in terms of the number per kilometre.

If the between site variation is assumed to be gamma distributed, the resulting mixed distribution is negative binomial and thus the detailed development of the model proceeded on the basis of a negative binomial error structure. This error structure is not a standard distribution within GLIM and must thus be fitted using GLIM's user-defined model facility. In order to determine the shape parameter, γ, of the between sites distribution an iterative procedure was used. The basis of this approach was the assumption that the variance of the observed accident frequency (i.e. $\mu + \mu^2/\gamma$) is estimated by the squared residuals of the regression model. An initial estimate of γ was obtained using the residuals from a model with a Poisson error structure. This estimate was then used in the negative binomial model and new values of γ estimated until satisfactory convergence was achieved (e.g. Hauer et al. 1988; Persaud and Dzbik 1993).

The results of the analysis are shown in Tables 3 and 4. The model referred to as LOGLIN1 was developed from the best fitting model in Table 2. A second model (referred to as LOGLIN2) was developed in which link accidents and minor junction accidents were modelled separately. The model form is summarized in Table 3. For the minor junction accidents the explanatory variable is the major road flow.

Both the deviance and Pearson's chi-squared statistic were used to evaluate the goodness of fit of the models. The deviance is asymptotically distributed like chi-squared so that, for a well fitting model, the expected value of the deviance will approximately equal the number of degrees freedom (i.e. the expected value of the scale parameter is approximately 1) provided that the fitted values are greater than about 0.5. However, for smaller values of the

Table 2. Comparison of alternative model forms based on the Poisson error structure

Model form*	Deviance	Degrees of freedom	Mean deviance ratio	Significance
$\mu = \alpha t l$	14,176	1455	—	—
$\mu = \alpha t^{\beta 1} l^{\beta 2}$	12,805	1453	77.8	$p < 0.01$
$\mu = \alpha t^{\beta 1} l^{\beta 2} \exp(bn)$	11,795	1452	97.7	$p < 0.01$
$\mu = \alpha t^{\beta 1} l^{\beta 2} \exp(bn/l)$	11,364	1452	119.8	$p < 0.01$

*Where μ = predicted annual total link accidents
t = total two-way annual link traffic flow (million vehicles per year)
l = length of link (km)
n = number of minor junctions within the link and b, α, β_1, and β_2 are coefficients to be estimated.

Table 3. Comparison of the best fit regression models

Model*	Deviance	Pearson χ^2	Degrees of freedom	γ	Systemmatic variation explained by the model estimated using: Elvik index	Freeman–Tukey R^2
LOGLIN1 $\mu=\alpha_1 t^{\beta 1} l^{\beta 2}\exp(\mathrm{bn}/l)$	1420.1	1480	1452	1.80	60%	68%
LOGLIN2						
$\mu_L=\alpha_2 t^{\beta 3} l^{\beta 4}$	1410.4	1495	1452	1.63	55%	66% } 71%
$\mu_J=\alpha_3 t^{\beta 5}$	3631.4	4605	4932	0.51	33%	43% } 71%

*For LOGLIN2 $\mu=\mu_L+n\mu_J$
where μ_L = predicted annual link only accidents
μ_J = predicted annual minor junction accidents

Table 4. Model coefficients

Model	Model parameter	Urban: A single	Urban: A dual	Urban: B single	Rural: A single	Rural: A dual	Rural: B single
LOGLIN1	α_1	0.808	0.808	0.608	0.481	0.481	0.362
	β_1	0.614	0.614	0.614	0.614	0.614	0.614
	β_2	0.986	0.630	0.986	0.986	0.630	0.986
	b	0.104	0.104	0.104	0.104	0.104	0.104
LOGLIN2	α_2	0.784	0.628	0.603	0.453	0.363	0.348
	α_3	0.131	0.131	0.131	0.088	0.088	0.088
	β_3	0.644	0.644	0.644	0.644	0.644	0.644
	β_4	0.957	0.658	0.957	0.957	0.658	0.957
	β_5	0.648	0.648	0.427	0.648	0.648	0.427

mean, the expected values of the scale parameter fall rapidly (Maycock and Hall 1984). In the data used here, fitted values of less than 0.5 were encountered, most notably for the minor junctions. It has been suggested that, for overdispersed Poisson distributions, the scale parameter should be estimated using Pearson's chi-squared rather than the deviance (Aitkin et al. 1989). Therefore this statistic was also calculated and its values are given in Table 3. It will be noted that the Pearson's chi-squared values are consistently higher than the deviance although both give similar estimates of the scale parameter except for the minor junction part of the LOGLIN2 model. For this the Pearson's chi-squared value is closer to the degrees of freedom than the deviance. Overall the results show there is little to choose between the LOGLIN1 and LOGLIN2 models in terms of the fit to the network data.

Recently Fridstrøm et al. (1995) have developed various goodness-of-fit measures for generalized Poisson regression models which give a measure of the percentage variation explained by the models in a similar way to the coefficient of determination (R^2) as used in ordinary least squares regression. These goodness-of-fit measures, however, use the properties of the Poisson distribution (i.e. the variance of a Poisson variable is equal to its mean) to take account of the inevitable random variation in accident counts, measuring the goodness-of-fit in terms of the percentage of systematic (i.e. explicable) variation explained by the model. This is important because the scope for random variation is not constant but rather is larger when the expected accidents are smaller. Of the goodness-of-fit measures developed, Fridstrøm et al. (1995) find that three are almost equivalent: the "Freeman–Tukey R^2" (based on the Freeman–Tukey transformation residuals); the "Elvik index" (based on the overdispersion parameter); and the "scaled deviance R^2" (based on the log-likelihood ratio). As the last of these measures is not straightforward to compute, only the first two have been used here and are presented in Table 3. While the Freeman–Tukey R^2 produces estimates of the explained variation some 10% larger than the Elvik index, both measures confirm that there is little to choose between the LOGLIN1 and LOGLIN2 models in terms of the fit to the total link accidents, the models explaining some 60–70% of the systematic variation. These results imply that additional explanatory variables (such as traffic composition, pedestrian flows and minor road entry flows) would improve the fit of the models. However, it was estimated that the exposure measure used in the LOGLIN1 ($t^{\beta 1} l^{\beta 2}$) model explained some 10–15% more of the systematic variation than the linear exposure measure (tl) confirming the inadequacy of the proportional model.

By adding variables in turn to the LOGLIN1 model it was found that exposure explained some 46–56% of the systematic variation (as estimated by the Elvik index and the Freeman–Tukey R^2 respectively), the category variables (road type, carriageway type and speed limit) some 10% and minor junction frequency 4%. It is also worth noting that, for total link accidents, it was estimated that only some 2–4% of the sample variation can be attributed to random factors whereas for minor junctions (where there are fewer accidents) random variation accounts for as much as 40% of the sample variation.

The estimated model coefficients are summarized in Table 4. The estimated powers of traffic flow ($\beta 1$, $\beta 3$ and $\beta 5$) are significantly less than 1 for all six highway categories indicating a less than proportional relationship between traffic volume and reported injury accidents. The models show that reported injury accidents increase by about 6% when the traffic flow increases by 10% and thus imply that link accident rates are potentially misleading: the accident rate per million-vehicle-kilometres on otherwise identical links (i.e. equally hazardous from an engineering point of view) will be lower for links with higher flows. This is consistent with the findings of previous research (see, for example, Satterthwaite 1981; Persaud and Dzbik 1993; Hauer et al. 1994 and Fridstrøm et al. 1995). Available empirical evidence infact suggests that not only is the relationship between traffic volume and the total number of injury accidents non-linear, but also that it varies according to accident type and severity. In the case of accident type there is, for example, a tendency for single-vehicle accident rates to decline with flow while collision rates increase initially and then decrease at higher levels of flow. In the case of severity, the damage-only accident rate appears to increase with traffic volume (Satterthwaite 1981) while the fatality rate decreases (Fridstrøm et al. 1995). Fridstrøm et al. (1995) suggest that the decrease in the average accident severity with increasing volume could be due to reductions in speed arising from increased traffic density. The authors are currently developing models disaggregated according to accident severity and would hope to report on this work in the near future.

The estimated powers of link length ($\beta 2$ and $\beta 4$) for all single-carriageway roads are not significantly different ($p<0.05$) from 1 while for dual-carriageways they are significantly less than 1. Thus, whereas accidents on single-carriageways are proportional to link length, on dual-carriageways they are less than proportional. The accident rate per million-vehicle-kilometres on otherwise identical sections of dual-carriageway (i.e. equally hazardous from an engineering point of view) will be lower for longer sections; doubling the length of the study section will increase the expected accident frequency by only 55%. The non-linear relationship with link length for dual-carriageways was not expected. A possible explanation is that the influence of junctions (major and/or minor) on accident frequencies extends beyond the area specified by the U.K. standard definition of a junction accident: an accident occurring at or within 20 m of a junction. There is evidence that higher accident frequencies can be observed in the band 20–100 m from a junction (Pickering et al. 1986) than elsewhere on a link and this would have a greater influence on shorter links. In addition, shorter links tend to have more junctions per kilometer than longer links (in part due to the effect of division by small values of l). It might be that a higher junction frequency generates interaction effects that further increase the length of road over which the presence of a junction affects traffic operations and this effect may not be fully accounted for by the models.

However it is not clear why these effects should only be apparent on dual-carriageways. Overall, the minor junction frequency for single- and dual-carriageway links is similar (Table 5). Shorter links do have higher numbers of minor junctions per kilometer than longer links: overall the minor junction accident frequency for links less than 1 km in length is twice that for links over 3 km and dual-carriageway links shorter than 1 km have a minor junction accident frequency some 20% larger than single-carriageways of similar length (Table 5). It is also possible that the effects of junctions persist over longer distances on dual-carriageways because of higher speeds. In addition it may be that junction related accidents are more influential on dual-carriageways due the absence of other hazards. Dual-carrigeways, because of their higher design standards and absence of opposing traffic, are likely to have fewer accidents due to other hazzards such as skidding on sharp bends, overtaking on bends and crests, or pedestrian accidents at crossings. Thus, whereas on well-designed dual-carriageways intersections are likely to be the only major hazard, on single-carriageways intersections will be one of many hazards. The results of the modelling would tend to support this hypothesis. For links of equal length, minor junction frequency and flow level, the ratio of single- to dual-carriageway accidents is estimated to be $l^{0.36}$ on the basis of the LOGLIN1 model. This implies that for links of less than 1 km the accident frequency on single-carriageways is less than on dual-carriageways (potentially due to the extended influence of the dual-carriageway major junctions) while for longer links the single- carriageway accident frequency is larger than for dual-carriageways (potentially due to the influence of other

Table 5. Minor junction frequencies on single and dual carrigeway roads

	Number of links:		Average number of minor junctions per kilometer on:		
Link length	Single	Dual	All carriageway types	Single carriageways	Dual carriageways
All lengths	1324	135	1.9	1.9	1.8
<1 km	408	46	2.5	2.4	2.9
>3 km	364	33	1.2	1.3	0.9

single-carriageway hazards). Further investigation is, however, needed to test this hypothesis.

The non-linear model form not only implies that link accident rates are potentially misleading (the accident rate per million-vehicle-kilometres on otherwise identical links will be lower for links with higher flows and, in the case of dual-carriageways, for longer link sections), but also implies that the linear model, as used in COBA, is inappropriate, even for sections without minor junctions

This point is emphasized in Fig. 2 in which, for links grouped by traffic volume, the mean observed and predicted annual accidents are plotted against the mean annual traffic volume. The predicted annual accidents were obtained using LOGLIN1, LOGLIN2 and the COBA accident rates. The COBA estimates were obtained in two ways, referred to as COBA1 and COBA2. The COBA1 estimates use the link and junction combined accident rates for all link segments irrespective of the presence or absence of minor junctions. These are the estimates which would be obtained in the absence of information concerning

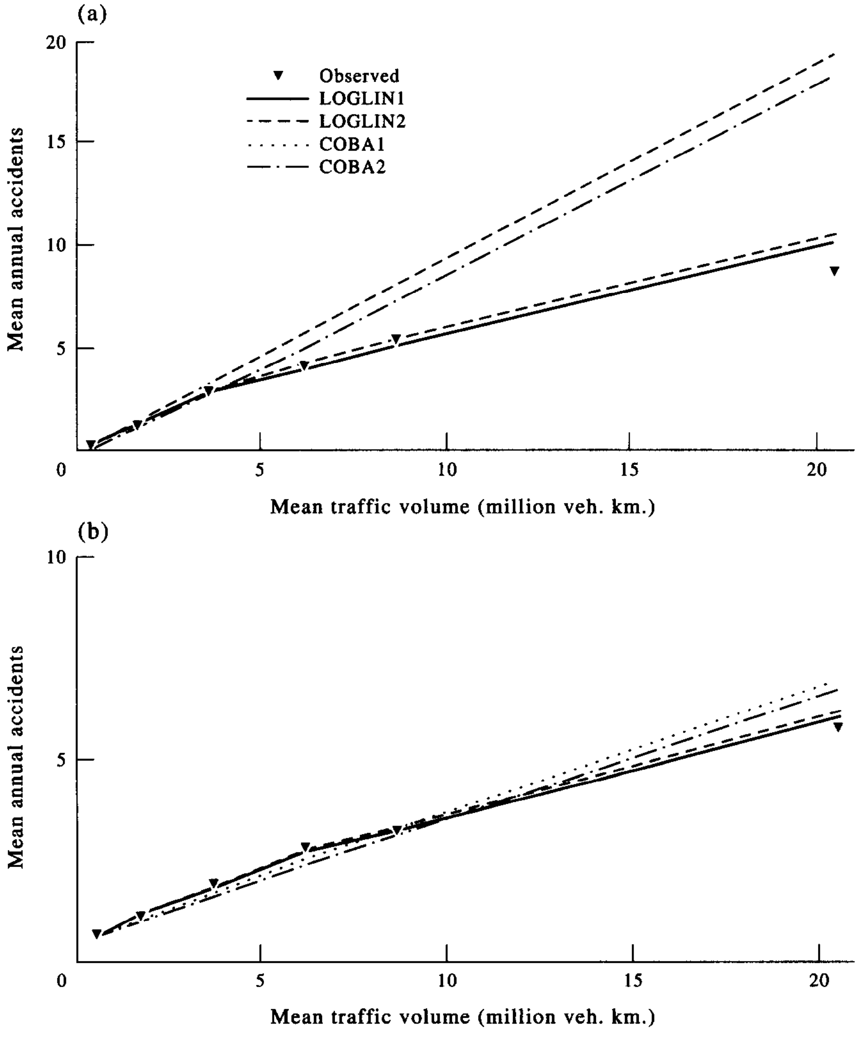

Fig. 2. Regression model predictions. (a) Urban roads; (b) rural roads.

the locations of minor junctions. The COBA2 estimates use the link only accident rates for link segments without minor junctions and the link and junction combined rates for segments with one or more minor junctions. It is clear that the proportional models as used in COBA increasingly tend to overestimate the annual accidents as traffic volumes increase, particularly on urban links. This figure also shows the close correspondence between the LOGLIN1 and LOGLIN2 estimates and that the COBA2 estimates are marginally preferable to COBA1.

THE QUALITY OF THE ESTIMATES

The next stage in the analysis was to apply the models to the sample of untreated link segments. For each of these link segments the expected value of x_a is m. Here the values of m are estimated using three methods. With the before and after comparison of accident counts, m is estimated by x_b; with the predictive models the estimate is μ; while with the EB method the estimate is $\hat{m}$. Let $\tilde{m}$ denote any one of these three estimates. The estimated treatment effect, T, for a group of sites is then defined as

$$T=\Sigma x_a/\Sigma\tilde{m} \quad (3)$$

If, for each site, $\tilde{m}$ was equal to the true value m, then the expected value of T would be 1. The actual value of T will differ from 1 because of discrepancies between x_a and $\tilde{m}$. These discrepancies will arise both because of the errors in estimating m and also because of random variation in x_a about m. Thus we can write

$$\begin{aligned} x_a-\tilde{m} &= (x_a-m)+(m-\tilde{m}) \\ &= \text{random error} + \text{estimation error} \quad (4) \end{aligned}$$

It is the estimation errors which are of interest here. Although the mean squared error provides a useful summary measure it does not distinguish between the random and estimation errors. The separation of these two sources of error is in practice difficult. The expected value of the random error in eqn (4) will always be equal to zero. If the EB estimate $\hat{m}$ is used then the expected value of the estimation error will also be zero. However, because of selection bias, it need not be zero if either of the other two estimates of m are used. It follows that, where sites have been selected on the basis of their before accident frequencies, the estimate T will be biased unless the EB estimate is used. In all cases the variance of T will depend on the between sites variation (i.e. the extent to which m differs from its predicted value μ). However, because the variance of m depends on μ an exact analysis is fairly complicated. For the purpose of the present analysis a simpler but approximate approach was adopted. GLIM was used to regress x_a on each estimate $\tilde{m}$ (i.e. x_b, μ, and $\hat{m}$). The regression model was linear, with a Poisson error structure and constrained to pass through the origin. The model was suggested by the fact that x_a has a Poisson distribution with mean m. Where m is replaced by $\tilde{m}$ the quality of the fit to the data will depend on which estimate of m is used.

In all three cases the slope of the fitted regression model will be equal to T. The estimated standard error of the slope coefficient then allows an assessment of whether T differs significantly from 1. In making this assessment it is important to take account of overdispersion in the model. With random errors alone the regression model would give a scale parameter close to 1; the deviance should be distributed like chi-squared with $(N-1)$ degrees of freedom, where N is the number of sites. The contribution of the estimation error to the variance of x_a about the regression line will lead to overdispersion and hence increase the deviance of the fitted model. The exact form of this overdispersion will again depend on which estimate is used. It can be shown that the variance of x_a will generally increase faster than linearly with $\tilde{m}$, so that there will be greater overdispersion for sites with high accident frequencies. Thus the estimate of the scale parameter obtained from the Poisson model will give an indication of only the average level of overdispersion. Furthermore, if sites are selected on the basis of having a high accident frequency in the before period, larger values of the scale parameter are likely to be obtained. It can also be shown that the degree of overdispersion will be less when the EB estimate $\hat{m}$ is used than when m is estimated by the predicted value μ. Thus the scale parameter can be expected to be greater for the regression of x_a on μ than for the regression of x_a on $\hat{m}$. Finally, when x_a is regressed on x_b, larger values of the scale parameter can again be expected; these reflect the inappropriateness of the model as well as overdispersion.

The results obtained were broadly as expected and are summarized in Tables 6 and 7. In Table 6 the results are presented both for all untreated link segments and for a sub-set of untreated blackspots which were arbitrarily defined for the purposes of this study as untreated link segments with 2 or more accidents in the before period. The untreated blackspots were used to illustrate the errors in estimates which might occur at sites selected for remedial treatment: sites selected, at least in part, on the basis of a high accident frequency in the before period. In Table 7 the untreated link segments are sub-divided into two groups according to whether or not the segment includes any minor junctions.

Table 6. An assessment of the errors in the estimates obtained for untreated links

Method		All untreated link segments *Number of segments = 1006* $\sum x_b = 685$ $\sum x_a = 727$ T	Scale parameter	Mean squared error	Untreated "blackspots"[a] *Number of segments = 168* $\sum x_b = 488$ $\sum x_a = 323$ T	Scale parameter	Mean squared error
Accident count		1.06 (0.07)	3.09	1.63	0.66** (0.05)	1.81	5.33
Predictive model	COBA1	0.97 (0.04)	1.31	1.28	1.37** (0.10)	1.65	3.04
	COBA2	1.15** (0.05)	1.31	1.27	1.51** (0.11)	1.76	3.19
	LOGLIN1	1.04 (0.04)	1.27	1.23	1.50** (0.10)	1.52	3.18
	LOGLIN2	1.01 (0.04)	1.25	1.20	1.43** (0.10)	1.52	3.06
EB method	COBA1	1.04 (0.04)	1.14	1.14	1.03 (0.07)	1.50	2.98
	COBA2	1.17** (0.05)	1.14	1.15	1.11 (0.08)	1.54	3.02
	LOGLIN1	1.06 (0.04)	1.13	1.11	1.06 (0.07)	1.46	2.81
	LOGLIN2	1.04 (0.04)	1.12	1.11	1.01 (0.07)	1.46	2.87

[a]"Blackspots" are defined here as link segments with 2 or more observed accidents in the before period.
*Significantly different from 1 ($p \leqslant 0.05$); **significantly different from 1 ($p \leqslant 0.01$).

Table 7. A seperate assessment of the errors in the estimates obtained for untreated links with and without minor junctions

Method		All untreated link segments *without* minor junctions *Number of segments = 571* $\sum x_b = 224$ $\sum x_a = 234$ T	Scale parameter	Mean squared error	All untreated link segments *with* minor junctions *Number of segments = 435* $\sum x_b = 461$ $\sum x_a = 493$ T	Scale parameter	Mean squared error
Accident count		1.04 (0.11)	2.48	0.75	1.07 (0.10)	3.89	2.79
Predictive model	COBA1	0.81** (0.05)	1.00	0.56	1.08 (0.06)	1.68	2.22
	COBA2	1.32** (0.09)	1.02	0.55			
	LOGLIN1	0.97 (0.06)	1.04	0.55	1.07 (0.06)	1.58	2.12
	LOGLIN2	1.01 (0.07)	1.01	0.54	1.00 (0.06)	1.57	2.07
EB method	COBA1	0.89* (0.06)	0.91	0.48	1.13* (0.06)	1.41	2.00
	COBA2	1.27** (0.08)	0.94	0.50			
	LOGLIN1	1.00 (0.06)	0.94	0.49	1.08 (0.06)	1.38	1.91
	LOGLIN2	1.03 (0.06)	0.92	0.49	1.04 (0.06)	1.38	1.93

*Significantly different from 1 ($p \leqslant 0.05$); **Signficantly different from 1 ($p \leqslant 0.01$).

Consider first the estimates of T. The values for all untreated links are given in column 1 of Table 6. Only the COBA2 predictive model estimates, and the EB estimates based on these, are significantly different from 1. This is a somewhat unexpected result since, in the COBA2 estimates, a distinction is made between links with and without minor junctions and it might have been expected that the estimates would have been better than the COBA1 estimates. The explanation of this result can be found in Table 7. The estimates of T for segments without minor junctions are given in column 1. The COBA1 estimates use the COBA link and junction combined accident rates while the COBA2 estimates use the link only rates. As would be expected, the COBA1 predictive model gives an estimate of T significantly less than 1 (i.e. the link and junction combined accident rate overestimates accidents on segments without minor junctions) but, rather unexpectedly, the COBA2 predictive model estimate is significantly greater than 1. For segments with minor junctions both COBA1 and COBA2 estimates use the COBA link and junction combined accident rates. Since the estimate of T based on these rates (Table 7, column 4) is greater than 1 there is some compensation in the COBA1 estimates when segments with and without minor junctions are taken together as in Table 6. Even when used in the EB method, Table 6 indicates that neither of the COBA predictive models give unbiased estimates of T when segments with and without minor junctions are considered separately. Accident counts, both of the LOGLIN models, and the EB method based on these models give unbiased estimates of T for both groups.

The results for the untreated blackspots (Table 6, column 4) indicate that only the EB method produces estimates of T which are not significantly different from 1. As expected, the accident counts result in an estimate of T which is significantly less than 1 due to the regression-to-mean effect. The magnitude of the regression-to-mean effect was 34% ($T=0.66$). Thus, on the basis of a before and after comparison, the effects of any remedial treatment at these sites would have been exaggerated by some 34%. The predictive models, on the other hand, all give T values significantly greater than 1 (in the range 1.37 to 1.51) i.e. an apparent increase in accidents in the absence of treatment. The effects of any remedial treatment would have been underestimated by between about 40 and 50% using this approach. This again is as would be expected. When sites are selected on the basis of a high observed accident frequency in the before period, although random fluctuations will in part account for the high accident count, the selected sites will never-the-less tend to include sites with higher than average long term mean accident frequencies for their type and exposure level (i.e. genuinely hazardous sites). Thus we would expect Σx_a to be larger than $\Sigma \mu$ and hence T to be greater than 1. Only the EB method results in unbiased estimates of T when sites are selected on the basis of a high accident frequency. This is an important result since these are the sites which are of most interest to road safety engineers.

The EB method also performs better than the accident counts and predictive models in terms of the other measures used in Tables 6 and 7. The EB method results in mean squared errors which are substantially less than those associated with accident counts: some 30% less for all sites; 45% less for blackspots. Relative to the predictive models, the EB method gives a reduction in mean squared errors of about 10%. As expected, the scale parameter values for the EB method are smaller than those for any other method.

CONCLUSIONS

Based on data for highway networks in seven UK counties, accident prediction models have been developed for six types of highway. The estimated coefficients are summarized in Table 4. The results of the modelling lead to the conclusion that accidents on highway links are not proportional to exposure (i.e. traffic flow and link length) as is commonly assumed. For all six carriageway types, accident frequencies are shown to be a non-linear function of annual traffic flows and, for dual-carriageways, a non-linear function of link length. This implies that link accident rates can be misleading when comparing safety levels at different locations or times. In addition it has been shown that the presence of minor junctions has an important influence on link accident frequencies. Account may be taken of these either by including the number of minor junctions per kilometre as an explanatory variable (as in LOGLIN1), or by modelling minor junction accidents separately (as in LOGLIN2): there is nothing to choose between these approaches in terms of the quality of the estimates obtained. Both of the LOGLIN models are preferable to the linear models used in COBA since the latter result in biased estimates both for higher levels of traffic flow and for highway sections without minor junctions.

Of the three methods used to estimate expected link accidents, the EB was found to be best followed by the predictive models. The predictive model estimates were preferable to estimates based on counts because they resulted in much smaller mean squared errors in the estimates. Although the EB estimates only gave a small reduction in the mean squared errors relative to the predictive models, the EB method is superior because it was the only method which produced unbiased estimates of the treatment effect at high-risk sites.

Acknowledgements—This paper is based on the results of research funded by the UK Engineering and Physical Sciences Research Council. The data for the project were obtained from the county councils of Cheshire, Cornwall, Cumbria, Devon, Kent, Lancashire and Warwickshire. We would like to thank these authorities for their help in providing the data and Ken Lupton for his assistance in data collection and computer programming.

REFERENCES

Abbess C.; Jarrett D.; Wright C. Accidents at blackspots: estimating the effectiveness of remedial treatment with special reference to the "regression-to-mean" effect. Traffic Engng and Control 22 (10):535–542; 1981.

Aitkin M.; Anderson D.; Francis B.; Hindle J. Statistical modelling in GLIM. Oxford: Clarendon Press; 1989.

Fridstrøm L.; Ifver J.; Ingebrigtsen S.; Kulmala R.; Thomsen L. Measuring the contribution of randomness, exposure, weather, and daylight to the variation in road accident counts. Accid. Anal. and Prev. 27 (1):1–20; 1995.

Hall R. D. Accidents at four-arm single carriageway urban traffic signals. , Crowthorne, UK: Transport Research Laboratory; CR 6: 1987.

Hauer E. Bias-by-selection: overestimation of the effectiveness of safety countermeasures caused by the process of selection for treatment. Accid. Anal. and Prev. 12 (2):113–117; 1980.

Hauer E.; Ng J.; Lovell J. Estimation of safety at signalised intersections. Transportation Research Record 1185:48–58; 1988.

Hauer E.; Terry D.; Griffith M. Effects of resurfacing on safety of two-lane rural roads in New York state. Transportation Research Record 1467:30–37; 1994.

Hauer E. On exposure and accident rate. Traffic Engng and Control 36 (3):134–138; 1995.

Jarrett, D.F. Statistical analysis of road accident frequencies. In Proc. Symposium on Safety in the Road Environment, LNEC, Lisbon; 1994.
Kulmala R. Measuring the safety effect of road accident measures at junctions. Accid. Anal. and Prev. 26 (6):781–794; 1994.
Maher M. J. The evaluation of road safety schemes: problems with statistical appraisal. In Proc. County Surveyors' Society/University Transport Studies Group Joint Seminar on The Evaluation of Road Safety Schemes. London; 1991
Maycock G.; Hall R. D. Accidents at four-arm roundabouts. Crowthorne, UK: Transport Research Laboratory; LR 1120; 1984.
Miaou, S-P; Lum, H. Modeling vehicle accidents and highway geometric design relationships. Accid. Anal. Prev. 25: 689–709; 1993.
Mountain L.; Fawaz B.; Sineng L. The assessment of changes in accident frequencies on link segments: a comparison of four methods. Traffic Engng and Control 33 (2):429–431; 1992.
Mountain L.; Jarrett D.; Fawaz B. The safety effects of highway engineering schemes. Proc. Institution of Civil Engineers. Transport 111:298–309; 1995.
Mountain L.; Fawaz B.; Jarrett D. The influence of minor junctions on link accident frequencies.. Traffic Engineering and Control 36 (12):673–679; 1995.
Pendleton, O. J. Application of new accident analysis methodologies. Vol.1: New methodologies. US Department of Transportation. Report FHWA-RD-90-091; 1991.
Persaud B.; Dzbik L. Accident prediction models for freeways. Transportation Research Record 1401:55–60; 1993.
Pickering, D.; Hall, R. D.; Grimmer, M. Accidents at rural T-junctions. Crowthorne, UK: Transport Research Laboratory; RR 65: 1986.
Satterthwaite S. P. A survey of research into relationships between traffic accidents and traffic volumes. Crowthorne, UK: Transport Research Laboratory; SR 692: 1981.
UK DoT. COBA manual. Department of Transport, London: Assessment Policy and Methods Division; 1981.

APPENDIX

The aim of this Appendix is to provide further mathematical details of the methods used in the paper, and in particular to prove some of the assertions made in the section on the quality of the estimates.

Empirical Bayes estimates. Consider first the EB method of estimating the mean accident frequency at a site; the results here are well known (see, for example, Hauer et al. 1988) but will be used in the remainder of this Appendix. Suppose that the accident frequency x at a site is Poisson distributed with mean m, where m is itself assumed to be a random variable having a gamma distribution with mean μ and shape parameter γ, so that the variance of m is μ^2/γ. It follows that the unconditional distribution of x is negative binomial, with mean μ and variance $\mu+\mu^2/\gamma$. The mean μ is assumed to depend on site characteristics, so will vary from site to site, while γ is assumed to be a constant. In practice, the predictive equation for μ and the value of γ are estimated by fitting a negative-binomial regression (predictive) model to an appropriate set of sites, such as all links.

Under these assumptions, the posterior distribution of m given x (the distribution of m amongst sites with a predicted accident frequency of μ and having a given observed number x of accidents) is gamma with shape parameter $\gamma+x$ and mean

$$\hat{m}\equiv E(m|x)=\frac{\gamma+x}{\gamma/\mu+1}$$

This is the EB estimate of m, and can be expressed in the form of equations (1) and (2) of the main part of the paper:

$$\hat{m}=\alpha\mu+(1-\alpha)x,$$

where

$$\alpha=\frac{1}{1+\gamma/\mu}=\frac{\mu}{\mu+\mathrm{var}(m)}.$$

Note that α is a function of μ as well as of γ, so its value will be different for different sites. The variance of the conditional distribution of m given x is

$$\mathrm{var}(m|x)=\frac{\hat{m}^2}{\gamma+x}=\frac{\gamma+x}{(\gamma/\mu+1)^2}=\frac{\hat{m}}{\gamma/\mu+1}.$$

Random and estimation errors. In practice, the EB estimate $\hat{m}$ will be calculated using the before accident frequency x_b at a site. If there is no treatment effect, the after accident frequency x_a will again be Poisson distributed with mean m, and any estimate of m can be regarded as a prediction of x_a. As in the main part of the paper, let $\tilde{m}$ denote any one of the three estimates $\hat{m}$, μ or x_b. To compare these estimates, consider the error in using $\tilde{m}$ as a prediction of x_a. We can write

$$x_a-\tilde{m}=(x_a-m)+(m-\tilde{m})$$

or

$$\begin{matrix}\text{prediction}\\ \text{error}\end{matrix}=\begin{matrix}\text{random}\\ \text{error}\end{matrix}+\begin{matrix}\text{estimation}\\ \text{error}\end{matrix}$$

Because of selection effects, in order to investigate the properties of these two sources of error, it is important that any expectations are taken conditional on x_b. First take expectations conditional on m also. Since $E(x_a|m)=m$, and $\tilde{m}$ is either equal to μ or a function of x_b, we have

$$E[x_a-\tilde{m}|x_b,m]=0+(m-\tilde{m}).$$

Hence, taking expectations again (but still conditional on x_b),

$$E[x_a-\tilde{m}|x_b]=0+\left(E(m|x_b)-\tilde{m}\right)$$
$$=\hat{m}-\tilde{m}.$$

Thus the conditional expectation of the random error will always be equal to zero. If the EB estimate $\hat{m}$ is used then the conditional expectation of the estimation error will also be zero. However, it will not be zero if either of the other two estimates of m are used.

This result suggests that the estimation error is split into two components, so that

$$x_a-\tilde{m}=(x_a-m)+(m-\hat{m})+(\hat{m}-\tilde{m})$$

The first two terms on the right-hand side have expectation zero (conditional on x_b), while the third term is just the difference between the estimate used $\tilde{m}$ and the EB estimate $\hat{m}$.

Mean square errors Squaring this three-term expression and taking expectations conditional on x_b and m gives

$$E[(x_a-\tilde{m})^2|x_b,m]=E[(x_a-m)^2|m]$$
$$+(m-\hat{m})^2+(\hat{m}-\tilde{m})^2+2(m-\hat{m})(\hat{m}-\tilde{m}).$$

(This again uses the fact that $\tilde{m}$ and $\hat{m}$ depend on μ and x_b only, and the assumption that x_a and x_b are independent given m; since $E(x_a|m)=m$, the other cross-product terms vanish.) The first term on the right-hand side equals $var(x_a|m)=m$. Taking expectations again, conditional on x_b, gives

$$E[(x_a-\tilde{m})^2|x_b]=E(m|x_b)+E[(m-\hat{m})^2|x_b]+(\hat{m}-\tilde{m})^2+0,$$

the last term being zero because $E(m|x_b)=\hat{m}$. Hence

$$E[(x_a-\tilde{m})^2|x_b]=E(m|x_b)+var(m|x_b)+(\hat{m}-\tilde{m})^2$$

$$=\hat{m}+\frac{\hat{m}}{\gamma/\mu+1}+(\hat{m}-\tilde{m})^2.$$

When applied to a single site, the first term here is the mean (or expected) square random error; the second term is the mean square estimation error when the EB estimate is used; and the third term represents the additional mean square estimation error when any other estimate of m is used. It follows that the mean square error for a group of n sites, $(1/n)\sum(x_a-\tilde{m})^2$, can be expected to be smaller when the EB estimate of m is used than if either of the other two estimates are used.

Regression on $\tilde{m}$. The different estimates of m were compared by regressing x_a on each, assuming a linear model with zero intercept and a Poisson error structure. For such a model, the variance is assumed to be equal to the expected value of x_a; this would be the case if the true values of m were known, since $E(x_a|m)=m$ and $\text{var}(x_a|m)=m$. Where m is replaced by the EB estimate $\hat{m}$, it is still the case that $E(x_a|\hat{m})=E(x_a|x_b)=\hat{m}$, but the results of the previous section show that the variance is now given by

$$\text{var}(x_a|\hat{m})=\hat{m}+\frac{\hat{m}}{\gamma/\mu+1}.$$

Since this is always greater than $\hat{m}$, the model is overdispersed, by a degree which depends on μ, and which will be greater for sites with larger values of μ.

If the estimate $\tilde{m}=N$ is used and there are no selection effects, then x_a will be distributed about μ according to a negative binomial distribution. The linear model is still appropriate, but the variance is a quadratic function of μ so the degree of overdispersion again increases with μ. The presence of selection effects complicates matters considerably. Since the expected value of the estimation error is non-zero, the linear model with zero intercept is not appropriate. Moreover, the dispersion of x_a about μ is increased by the term $(\hat{m}-\mu)^2$ relative to the variance when the EB estimate is used. Both of these features will worsen the fit of the model. The situation is similar when the estimate $\tilde{m}=x_b$ is used. Both the non-zero expectation of the estimation error, and the term $(\hat{m}-x_b)^2$ in the mean square error, will make the fit of the regression model worse than when $\tilde{m}$ is equal to the EB estimate $\hat{m}$.