ELSEVIER

Advances in Engineering Software 32 (2001) 859–869

ADVANCES IN ENGINEERING SOFTWARE

www.elsevier.com/locate/advengsoft

Identification of sites for road accident remedial work by Bayesian statistical methods: an example of uncertain inference

B.G. Heydecker*, J. Wu

Centre for Transport Studies, University College London, Gower Street, London WC1E 6BT, UK

Received 3 November 1999; accepted 19 April 2001

Abstract

Road accident remedial work can be effected in many cases by implementing low cost engineering measures. These are intended to restore the safety performance of sites to which they are applied to prevailing levels according to their kind, traffic flows and design. This approach is, therefore, appropriate to sites where the mean accident frequency exceeds a standard value for that site. In this paper, we present several analyses that provide quantitative but uncertain indications of which sites should be considered for this kind of treatment according to their accident record. We describe four measures that are informative in this respect — the count of accidents, an estimate of the mean accident frequency, the amount by which this estimate exceeds a standard value, and the probability that it does so. We illustrate the use of these analyses by applying them to a dataset of accident records of priority controlled road junctions and compare the results of using the different measures.

Keywords: Bayesian; Log-linear model; Mean frequency; Road accidents; Road safety

1. Introduction

Reducing road accident casualty frequency is an endeavour of national importance in most countries, including Britain. This can be achieved in several ways, notable amongst which are engineering, enforcement, education and encouragement. Of these, road engineering improvements have been found to be highly cost-effective ways of improving road safety when they are targeted on sites at which the risk of accident occurrence is unusually high. As progress is made towards the elimination of high risk accident sites at which low cost remedial measures will be especially effective, identification of further sites for treatment of this kind becomes more difficult. In the present paper, we consider methods for the automatic identification of sites for detailed investigation on the basis of their accident record and their characteristics, including traffic flows.

Observations of accident occurrence at a site can be used to provide an estimate of the mean frequency, and this can be compared with appropriate standard values to indicate the scope for safety improvements. However, road accidents occur as rare random events, so the numbers observed are small and the presence of stochastic effects should be respected. Because of the random nature of accident occurrence, the mean frequency can never be known but rather can only be estimated. Whenever an observation of accident occurrence is used directly to estimate the mean frequency, further observations will tend to occur at a frequency that is closer to the true, but unknown, mean for that site. This effect, which occurs generally, is known as *regression to the mean*. An immediate consequence of this is that sites that are selected for investigation because of an observed accident frequency that is excessively high will tend to have lower accident frequencies after selection even if no remedial work is undertaken. This directional effect that arises when high observations are used as the basis for identification is known as *bias by selection*.

In the present paper, we consider how the mean accident frequency of a site can be estimated using various kinds of information including the accident record of that site which is directly relevant but if used on its own will give a biased estimate. We adopt a Bayesian statistical mechanism to eliminate this bias by estimating the size of the regression to the mean effect and then anticipating it. The mean value towards which regression takes place is estimated using a log-linear statistical model based upon various explanatory covariates. The strength of the regression effect is estimated using a separate log-linear model of the same kind. These models are estimated using the accident records of each of a population of similar sites according to the empirical

* Corresponding author. Tel.: +44-20-7679-1553; fax: +44-20-7679-1567.

E-mail address: ben@transport.ucl.ac.uk (B.G. Heydecker).

0965-9978/01/$ - see front matter
PII: S0965-9978(01)00037-0

Bayesian approach. We then proceed to show how the resulting qualified estimates of mean accident frequency for a site can be used as the basis for identification of sites at which the mean accident frequency is excessive and which, therefore, merit detailed investigation.

2. Bayesian analysis

2.1. Introduction

Bayesian statistical methods enable analysts to combine evidence from observations of stochastic systems with that from other sources, such as models that describe standard operation. The resulting estimates combine these different sources of information using weights that implicitly respect the amount of information in each of these sources and leads to estimates that are as accurate as is possible on the basis of the information that is used in them. This provides a facility to estimate the amount by which observations of accident frequencies should be adjusted in respect of regression to the mean: in the present context, this achieves correction for bias by selection. Beyond this, the Bayesian approach provides an indication of the accuracy to which mean accident frequencies are estimated and hence of their likely range of values: we develop this aspect to estimate the probability that the mean accident frequency at a site is greater than a standard value for sites of that kind and use this as the basis of a criterion for selection.

2.2. A Bayesian statistical model

A de facto standard form of Bayesian statistical technique has been developed [1] and is now widely applied to the analysis of road accident data. This approach treats the mean accident frequency at a site formally as an unknown quantity. In the absence of site-specific observations, this is described by a Bayesian prior distribution, the dispersion of which represents the uncertainty in the value of the mean frequency. Bayes' theorem provides a calculus to update this prior distribution in light of site-specific observations of accident occurrence: the resultant of this is the Bayesian posterior distribution. This represents the knowledge of the mean accident frequency by combining the information from each of the prior distribution and the observations in an appropriate way. Thus, the posterior distribution represents at once a modification to observations that will correct them for regression to the mean and a refinement of the prior one in light of the site-specific observations.

An appropriate choice of probability distribution for the number of accidents that occur during an interval of duration t at a site at which the mean frequency is μ per unit time is the Poisson:

$$P(n|\mu,t) = \frac{e^{-\mu t}(\mu t)^n}{n!} \quad (n \geq 0) \tag{1}$$

which has mean and variance given respectively by $E(n) = \mu t$ and $\mathrm{Var}(n) = \mu t$.

A convenient choice of prior distribution for the mean frequency μ is then the gamma, which is known as the conjugate distribution for the Poisson (see, for example, Robert [2] p. 147). Here, we adopt the parameterisation for the gamma distribution of n_b for the shape and t_b for the reciprocal of scale which gives the probability density function

$$p(\mu) = \frac{e^{-\mu t_\mathrm{b}}\mu^{n_\mathrm{b}-1}t_\mathrm{b}^{n_\mathrm{b}}}{\Gamma(n_\mathrm{b})} \quad (\mu > 0). \tag{2}$$

The mean and variance of μ are given, respectively, by $\mu_\mathrm{b} = n_\mathrm{b}/t_\mathrm{b}$ and $\sigma_\mathrm{b}^2 = n_\mathrm{b}/t_\mathrm{b}^2$. We note that the variance of this distribution varies with the square of the mean according to the reciprocal of the shape parameter n_b. Thus

$$\sigma_\mathrm{b}^2 = \mu_\mathrm{b}^2/n_\mathrm{b} \tag{3}$$

The marginal distribution $P(n|t)$ of observations of accident numbers when the mean frequency μ is distributed as $p(\mu)$ is given formally as

$$P(n|t) = \int_\mu P(n|\mu,t)p(\mu)\mathrm{d}\mu \quad (n \geq 0). \tag{4}$$

With the present choice of Poisson (1) and gamma (2) distributions, this gives rise to the negative binomial distribution for observations:

$$P(n|t) = \frac{\Gamma(n+n_\mathrm{b})}{n!\Gamma(n_\mathrm{b})}\left(\frac{t^n t_\mathrm{b}^{n_\mathrm{b}}}{(t+t_\mathrm{b})^{n+n_\mathrm{b}}}\right) \quad (n \geq 0), \tag{5}$$

which has moments

$$E(n|t) = \frac{n_\mathrm{b}}{t_\mathrm{b}}t = \mu_\mathrm{b}t$$

and

$$\mathrm{Var}(n|t) = \frac{n_\mathrm{b}}{t_\mathrm{b}}t\left(1+\frac{t}{t_\mathrm{b}}\right) = \mu_\mathrm{b}t + \sigma_\mathrm{b}^2t^2.$$

Thus the mean frequency of the observation distribution corresponds to that of the prior distribution, whilst the variance is the sum of that of a Poisson distribution with this mean and an element determined by the variance of the prior distribution. The observation process is therefore more dispersed than would be a Poisson process with the same mean, with the degree of overdispersion determined by the parameter t_b of the prior, which corresponds to the reciprocal of the scale parameter of the gamma distribution. This overdispersion of the observation process is a reflection of the uncertainty in the true mean frequency of accident occurrence.

According to Bayes' theorem [2], observation of n_o accidents in a period of duration t_o will cause the prior distribution $p(\mu)$ to be updated to the posterior distribution $p(\mu|n_\mathrm{o},t_\mathrm{o})$ according to

$$p(\mu|n_\mathrm{o},t_\mathrm{o}) \propto L(\mu|n_\mathrm{o},t_\mathrm{o})p(\mu) \tag{6}$$

where $L(\mu|n_o, t_o)$ is the likelihood of μ given the observation of n_o accidents during t_o, and is calculated as $L(\mu|n_o, t_o) = P(n_o|\mu, t_o)$. With the present choice of observation and prior distributions given by Eqs. (1) and (2), respectively, the posterior distribution also has the gamma form but with parameters $n_a = n_b + n_o$ and $t_a = t_b + t_o$. We see from this that the parameters n_b and t_b of the prior distribution bear interpretation as being equivalent to (respectively) a number of accidents that have been observed and a period of time over which the observations took place. The posterior mean μ_a is then

$$\mu_a = \frac{n_b + n_o}{t_b + t_o} \tag{7}$$

This can be rearranged conveniently using the mean frequency $\mu_o = n_o/t_o$ at which accidents have been observed at the site. Thus

$$\mu_a = \left(\frac{t_b}{t_b + t_o}\right)\mu_b + \left(\frac{t_o}{t_b + t_o}\right)\mu_o$$
$$= \mu_o + (1 + t_o/t_b)^{-1}(\mu_b - \mu_o) \tag{8}$$

This shows that the strength of the regression towards the prior mean μ_b is determined by the relative sizes of the prior parameter t_b and the duration of the observation period t_o through the Bayesian *shrinkage* term $(1 + t_o/t_b)^{-1}$. In the case that no observations are available, $n_o = t_o = 0$, $\mu_a = \mu_b$ so that the prior distribution prevails. On the other hand, with increasing amounts of observed information, corresponding to the limit $t_o \to \infty$, we have $\mu_a \to \mu_o$ so that the observations prevail and the prior distribution has diminishing influence.

2.3. Modelling of accident frequencies

The Bayesian combination of prior knowledge with that from observations has the effect of using prior information to moderate specific observations by furnishing estimates with values that are closer to expected value for a site of that kind. Once the principle of this adjustment has been accepted, two practical matters arise: what should be the value towards which the adjustment is made, and how strong should that adjustment be. These depend on the prior distribution of the mean accident frequency at the site in question. In the present approach, these matters are resolved according to values of the parameters n_b and t_b, or equivalently the mean and variance μ_b and σ_b^2 of the Bayesian prior distribution.

An appropriate choice for the prior mean μ_b is a standard value for the site, which could reasonably be expected to depend on flows, geometry and design features. An estimate of this is often [3–6] furnished by a log-linear model of a vector $\underline{\mathbf{x}} = (x_0, x_1, \ldots, x_m)^T$ of explanatory covariates for the site. Thus, we estimate the mean as

$$\mu_b = \exp(\underline{\boldsymbol{\beta}}^T \cdot \underline{\mathbf{x}}) \tag{9}$$

where $\underline{\boldsymbol{\beta}} = (\beta_0, \beta_1, \ldots, \beta_m)^T$ is a vector of parameters of the model. By convention, $x_0 = 1$ so that β_0 corresponds to the constant term of the model giving a base value $\bar{\mu}_b = \exp(\beta_0)$. Furthermore, x_1 is usually taken to be the logarithm of some measure Q of traffic flow so that $\mu_b \propto Q^{\beta_1}$. Models of this kind appear widely in the literature and their intended uses include appraisal of proposed designs.

There remains the matter of the dispersion of the prior distribution which determines the strength of the Bayesian correction. Hauer [6] and many others since suppose that the shape of the prior distribution is invariant between sites so that $n_b = k$ for some scalar k. Accordingly, the relationship between the variance and the mean of the prior distribution is $\sigma_b^2 = \mu_b^2/k$. In the present approach, we propose instead a log-linear model

$$\sigma_b^2 = \exp(\underline{\boldsymbol{\gamma}}^T \cdot \underline{\mathbf{x}}) \tag{10}$$

with a vector of parameters $\underline{\boldsymbol{\gamma}}$ for the variance of the prior distribution (2). Together, the models (9) and (10) are equivalent to the log-linear models

$$n_b = \exp((2\underline{\boldsymbol{\beta}} - \underline{\boldsymbol{\gamma}})^T \cdot \underline{\mathbf{x}}), \text{ and } t_b = \exp((\underline{\boldsymbol{\beta}} - \underline{\boldsymbol{\gamma}})^T \cdot \underline{\mathbf{x}}) \tag{11}$$

for the parameters n_b and t_b of the prior distribution. This represents a flexible extension to Hauer's [6] choice of $n_b = k$.

This approach, then, envisages mean accident frequencies μ_b that vary systematically according to the explanatory covariates $\underline{\mathbf{x}}$. It also envisages dispersion between the mean accident frequencies of sites that have identical values for their explanatory covariates that also varies systematically according to the explanatory covariates but in a way that need not depend on the mean.

2.4. The empirical Bayes approach

The present approach allows for three distinct kinds of information to be used in conjunction to estimate the mean accident frequency at a site. These are: the characteristic elements such as geometry and design features at the site; variables such as the traffic flows; and the specific accident record of the site. The dependence of the accident frequency on the geometry and the design features can be estimated as a matter of statistical modelling from the accident records of a suitable population of sites: this can be achieved by maximising the likelihood of the parameters $\underline{\boldsymbol{\beta}}$ and $\underline{\boldsymbol{\gamma}}$ of the log-linear models (11) for the parameters n_b and t_b of the prior distribution (2) of the mean accident frequency by considering the negative binomial distribution (5) of observations.

The model fitting can be achieved conveniently by maximising the sum over all sites of the logarithm $\mathcal{L}$ of the likelihood of the parameter values given the accident records of that site. Thus, for a set of observations of $n_o^{(i)}$ accidents occurring in a period of duration $t_o^{(i)}$ at site i, $(1 \le i \le I)$ where the explanatory covariates have the values $\underline{\mathbf{x}}^{(i)}$,

we solve

$$\underset{\underline{\beta},\underline{\gamma}}{\text{Max}} \sum_{i=1}^{I} \mathcal{L}(\underline{\beta},\underline{\gamma}|n_o^{(i)}, t_o^{(i)}) \tag{12}$$

where the log-likelihood is calculated according to $\mathcal{L}(\underline{\beta},\underline{\gamma}|n,t) = \log_e P(n|t)$ using the observation distribution (5) with parameter values n_b and t_b calculated according to Eq. (11) with $\underline{x} = \underline{x}^{(i)}$ for site i. With the present choice of Poisson and gamma distributions, there is no reason to suppose that Eq. (12) corresponds to a convex programming problem, so multiple local optima and unbounded solutions are possible.

Although the likelihood that is maximised corresponds to the observation distribution (5), the parameters can be identified with those of the prior distribution (2) of the mean frequency. Thus the effect of solving this statistical model-fitting problem is to establish a prior distribution of the form (2) for sites of the kind that have been observed. The present formulation allows this prior distribution to have both shape and scale that depend on the observable characteristics $\underline{x}$ of the site. Because this prior distribution (2) is established by fitting the distribution (5) to observations, this is known as an *empirical Bayes* approach.

3. Identification of sites

3.1. Introduction

In road accident remedial work, the issue arises of which sites are to be investigated in detail with a view to accident remedial treatment. Some consideration of the accident record of a site is entirely appropriate in this process, but a crucial quantity in this identification is not the historical but rather the hypothetical future expected frequency of accident occurrence in the absence of any treatment. According to the Bayesian paradigm, this information is encapsulated in the posterior distribution, because it represents the full state of knowledge concerning the site. In particular, the mean of the posterior distribution provides the most accurate available estimate of the accident frequency for a site in future years if the site remains unchanged (Hauer [1] p. 190). Furthermore, the dispersion of the posterior distribution indicates the degree of uncertainty in that estimate.

3.2. Statistical methods

Several criteria are available for the selection of road sites for initial investigation that can be developed within analyses of the present form. These include measures of the absolute frequency of accident occurrence per unit of time, measures of risk of accident involvement per unit of traffic, measures of excess of either of these quantities by comparison with appropriate standard values, and measures of belief that the accident frequency is excessively high. We consider several of these possibilities in turn.

3.2.1. Accident count

In Britain, a criterion in current use [7] for initial investigation of a site is that n_o, the number of personal injury accidents recorded during the most recent 3 years (t_o) for which data are available, is at least 12. This corresponds to an unadjusted measure of absolute frequency of occurrence and has the merit of simplicity. In cases where the standard frequency of accident occurrence is comparably high, a possible reasonable conclusion from the investigation is that no appropriate accident remedial treatment is available for the site so that despite its record, it might not be considered suitable for treatment. Against this is the problem that the unadjusted accident frequency $\mu_o = n_o/t_o$ implicit in this will be subject to regression to the mean and, hence, will not be a reliable estimate of future accident frequencies in the absence of treatment. Because the critical value in this selection criterion is deliberately set above the population mean, this regression will be predominantly downwards, thus leading to bias that can be quantified using Eq. (8) as $(1 + t_o/t_b)^{-1}(\mu_b - \mu_o)$.

3.2.2. Bayesian estimate of accident frequency

To overcome the problem of bias by selection that arises from regression to the mean, the Bayesian posterior expectation μ_a of the mean accident frequency given by Eq. (8) can be used instead of the observed mean μ_o. This provides an unbiased estimate of the absolute frequency of accident occurrence, and hence indicates sites at which a high frequency of accident occurrence is expected in the future. Use of this as a criterion to indicate further investigation would prioritise sites with a high accident frequency whether or not it is excessive for a site of that kind, or corresponds to a high risk of accident involvement for individual vehicles.

3.2.3. Potential for accident reduction

The excess of the accident frequency relative to a standard value for that site can be used to indicate sites at which there is potential for reduction in accident frequency by improving the safety performance of the site to an appropriate standard level. This measure of excess is known as the *potential for accident reduction* (PAR) [8] and is here denoted as R. The mean μ_b of the Bayesian prior distribution can be used as the standard value, so that if the mean μ_a of the Bayesian posterior distribution given by Eq. (8) is used as an estimate of the current accident frequency, we have

$$R = \mu_a - \mu_b = (1 + t_b/t_o)^{-1}(\mu_o - \mu_b) \tag{13}$$

This, then, corresponds to a partition of the difference $\mu_o - \mu_b$ between the observed accident frequency μ_o and the mean prior estimate μ_b into two: a transitory part that

Table 1
Summary accident statistics of the road junctions

	Sites	Accidents (n)	Mean $E(n)$	Variance Var(n)
All	389	524	1.35	8.31
$n = 0$	202	0	0	0
$n > 0$	187	524	2.80	13.21

arises from stochastic effects which will be eliminated by regression to the mean, and an enduring part that arises because the mean frequency for that site differs from the standard value. The latter part is identified as the potential for accident reduction.

3.2.4. Bayesian probability of excess

In the present work, we also develop use of the posterior probability P_a that the mean accident frequency exceeds a standard value μ_h for that site. Thus

$$P_a(\mu > \mu_h) = \int_{\mu=\mu_h}^{\infty} p(\mu|n_o, t_o)\mathrm{d}\mu \tag{14}$$

which depends on the detailed specification of the posterior distribution $p(\mu|n_o, t_o)$. Our choice of standard value μ_h for this purpose is the median of the prior distribution, so that in the case where no observations are available, $P_a(\mu > \mu_h) = 0.5$, providing a neutral value for this measure. Values that exceed 0.5 indicate evidence that the mean frequency at a site exceeds the appropriate standard value and hence that the site should be investigated; values close to 1 indicate a high degree of certainty in this.

4. Example calculations

4.1. Introduction

In order to illustrate the use of the present Bayesian statistical analysis, we show how it can be used for the indication of sites for road accident investigation work. For this illustration, we use a database consisting of a random sample of rural 3-arm junctions in the USA state of Minnesota: this is one of the datasets investigated by Vogt and Bared [9]. We model the mean and the variance of the number of accidents that were recorded at each site during the 5 year period 1985–1989 (inclusive) as depending on traffic flows, and characteristics of the sites including signed speed limit and certain geometric measurements. Using the analysis presented in Section 3, we identify sites at which the number of accidents recorded exceeds the value estimated from the fitted model, and quantify this excess according to the various statistical estimates of size and certainty. We then compare and discuss these approaches to site identification.

4.2. The accident data

The dataset that is used for this example analysis consists of a random sample of 389 three-arm junctions sites of two-lane two-way roads, with stop control on the minor road. The accidents include personal injury and also property damage only ones that occurred either in the intersection, or within 250 feet of it and having a cause pertaining to it [9]. Because the dataset is site-based rather than accident based, it includes sites at which no accidents are recorded: these number 202 and constitute slightly over half of the dataset. The mean number of accidents during the 5 year period occurring at sites in the whole dataset is 1.35 whilst the mean number at the 187 sites where any accidents occurred is 2.80. These statistics of the dataset are summarised in Table 1.

Several covariates are available in the dataset which are described in Table 2. All of these were considered as candidate explanatory variables in the statistical model, but some were eliminated during the modelling procedure, because there was insufficient statistical evidence that they were associated with the distribution of accident frequencies.

The covariates that are available include exogenous, circumstantial, and design ones. As is usual for models of this kind, the measures of traffic flow were transformed by taking the natural logarithm, as described in Section 2.3. The intersection angle φ describes the absolute deviation from a right-angle of the angle between the centre lines of the major and minor roads at the intersection. The categorical variable H that describes roadside hazard was treated as an ordinary numeric variable; it has a scale that ranges from 1 (no hazard) to 7 [10], though no site in this dataset had a value greater than 5. The driveway exits

Table 2
Summary of covariates

Variable	Symbol	Units	Minimum	Mean	Maximum	Proportion 0
Major road flow	q^M	$\log_e$ (vehicles/day)	5.30	7.80	9.87	0
Minor road flow	q^m	$\log_e$ (vehicles/day)	1.61	5.41	8.34	0
Intersection angle	φ	Degrees	0	13.4	90	0.51
Roadside hazard	H	(Categorical)	1	2.11	5	
Driveway exits	D	1/(152 m)	0	1.26	9	0.38
Vertical curvature	C^V	Percent/(30 m)	0	0.14	4.39	0.53
Horizontal curvature	C^H	Degrees/(30 m)	0	1.21	29	0.54
Speed limit	S	Miles/hour	22.5	52.7	55	
Turning lanes	T	(Categorical)				0.56

Table 3
Characteristics of models fitted to the full dataset (389 sites)

	Moments	Constants only	Constant shape ($n_b = 1.883$)	Full model
Number of parameters	2	2	7	16
Mean (accidents/5 years)	1.347	1.347	1.350	1.304
Variance (accidents/5 years)^2	8.310	5.324	9.503	8.740
Joint log-likelihood	− 614.32	− 603.48	− 498.18	− 483.35

variable D indicates the number of these that join the major road within 76 m (250′) of the of the intersection. Two measures of curvature were available — vertical curvature C^V measured as change in gradient per unit road length averaged over all crests on the major road within 76 m of the intersection, and horizontal curvature C^H measured as change in direction per unit length of the major road averaged over all curves on the major road within 76 m of the intersection. The design variables include the speed limit S that was signed on the major road, and a categorical indicator T for the presence of dedicated nearside and offside turning lanes on the major road. Note that when the speed limit varied within a site, the mean value was used.

4.3. The statistical models

The statistical model specified by Eqs. (9) and (10) was fitted according to the empirical Bayes procedure described in Section 2.4 by maximising the likelihood according to Eq. (12). This was conducted in three distinct steps starting from a constants-only model specified by the mean and variance of the data as presented in Table 1. The first step was to maximise the likelihood of the constants β_0 and γ_0, using the moments estimators as initial values. The second step was to introduce as many of the covariates described in Table 2 as could be supported for the mean accident frequency by the data, allowing a single parameter to be fitted for the shape of the prior distribution; in this case the variable corresponding to intersection angle φ, the driveway exit density D, and the presence of turning lanes T were excluded. The third step was to model systematic variations in the shape of the prior distribution according to values of the covariates: in this case, the indicator T for the presence of turning lanes was included in the model equations. The results of this process, including model mean and variance, and the resulting maximised joint log-likelihood, are summarised in Table 3.

Table 4
Summary of accident statistics of the main dataset

	Sites	Accidents (n)	Mean $E(n)$	Variance Var(n)
All	370	437	1.18	7.73
$n = 0$	202	0	0	0
$n > 0$	168	437	2.60	13.33

Substantial improvement in fit was achieved in each of the three steps of this procedure, as indicated by increase in the joint log-likelihood. The most pronounced improvement in fit was achieved with the introduction of the explanatory model for the mean of the accident frequency in the constant shape model, but according to the likelihood ratio test [11, p. 86] the additional improvement in log-likelihood of over 14 that was achieved by the extension to the full model justifies the use of the additional nine parameters (this corresponds to a test value $X^2 = 28$ compared with the critical value of $\chi^2_{0.05} = 16.9$ with 9 degrees of freedom). The stability of the mean number of accidents, especially during the likelihood maximisation of the constants only model, justifies the choice of parameterisation of the log-linear model as ($\underline{\boldsymbol{\beta}}$, $\underline{\boldsymbol{\gamma}}$) in Eqs. (9) and (10) in preference to the corresponding coefficients in the log-linear models for n_b and t_b in Eq. (11).

However, the absolute value of the joint log-likelihood of 483.35 that resulted from this fitting process is rather larger than the number of residual degrees of freedom 373, being the number of observations (389) less the number of fitted parameters (16) whereas for a well-fitting model, the former should not exceed the latter [3, p. 53]. Inspection of the fit of the model to the data revealed that for several of the sites, the accident record was inconsistent with the corresponding model estimate. This was judged by comparing the model estimate $\mu_b t_o$ of the expected number of accidents during the observation period with the observed number n_o using the statistic:

$$X^2 = \frac{(n_o - \mu_b t_o)^2}{\mu_b t_o}. \tag{15}$$

Table 5
Characteristics of models fitted to the main dataset (370 sites)

	Moments	Constants only	Constant shape ($n_b = 3.210$)	Full model
Number of parameters	2	2	7	13
Mean (accidents/5 years)	1.181	1.181	1.186	1.151
Variance (accidents/5 years)^2	7.731	4.350	7.488	4.222
Joint log-likelihood	− 554.27	− 538.16	− 411.01	− 393.56

Table 6
Fitted parameter values and their 0.95 confidence limits for the log-linear models of mean and variance (370 observations)

Variable	Coefficient (full model)	0.95 confidence limits		Coefficient (constant shape model)
		Lower	Upper	
Mean				
Constant	− 13.45	− 13.55	− 13.36	− 14.71
Major road flow	0.876	0.864	0.887	0.959
Minor road flow	0.618	0.603	0.633	0.605
Speed limit	0.0146	0.0126	0.0165	0.0246
Roadside hazard	0.132	0.0913	0.171	0.181
Horizontal curvature	0.0351	0.00931	0.0571	0.0413
Variance				*Shape*
Constant	− 1290	− 1291	− 1289	3.210
Major road flow	4.293	4.185	4.391	
Minor road flow	1.075	0.919	1.217	
Speed limit	22.61	22.60	22.63	
Roadside hazard	− 2.083	− 2.548	− 1.637	
Horizontal curvature	0.705	0.245	1.144	
Vertical curvature	10.31	7.541	12.85	

A heuristic sequential procedure was followed of removing observations for those sites with the greatest values of this test statistic and refitting the statistical model to the remainder. This procedure stabilised after some 19 observations had been removed, corresponding to a critical value of 7 for the test statistic in Eq. (15). The effect of this was that the dataset was partitioned into a main part consisting of 370 observations to which the empirical Bayes modelling procedure was applied, and a complementary part of 19 outlying observations, which were retained separately for subsequent analysis. Summary statistics of the main dataset are given in Table 4.

The results of the model fitting procedure with the reduced dataset are summarised in Table 5: in this case three of the candidate explanatory covariates were removed from the model as there was insufficient evidence for their correlation with variations in the observed accident frequencies. The resulting model fit was improved substantially by the removal of the outliers, so that the absolute value of the maximised joint log-likelihood of 394 was comparable with the residual freedom of 357 (= 370 − 13). As was the case with the statistical model of the full dataset, the mean accident frequency remained remarkably stable through the fitting procedure. Similarly, the benefit in terms of improvement in model fit with increasing model flexibility remained substantial.

The statistically significant model parameters, together with their 0.95 confidence intervals estimated according to the likelihood ratio principle [11, p. 112] are given in Table 6 for the full model, together with the corresponding estimates for the constant shape model for comparison. The variables describing each of the number of driveway exits and presence of turning lanes were found not to contribute significantly to the modelling process and so were excluded from it. The vertical curvature was found to contribute significantly to the model for variance but not to that for mean accident frequency. The other variables shown in Table 6 contribute, in some measure, to each of the log-linear models.

According to the full model, variations in the mean accident frequency are proportional to those in major road flow raised to the power 0.876, and with those in minor road flow raised to the power 0.618; these sub-linear associations are consistent with those found in other studies [3–5]. As would be expected, the coefficients of each of speed limit, roadside hazard, and horizontal curvature are all positive, indicating an increase in mean accident frequency with each of them. The resulting log-linear model represents the best available joint representation of the systematic variations in the mean and variance of the accident frequencies at the sites in the main part of the dataset.

The parameter values fitted for the log-linear model of mean accident frequency in the constant shape model are generally similar to those of the full model, though in four of the six cases the estimates do not lie within the 0.95 likelihood-based confidence interval for those of the full model. The coefficient of the logarithm of major road flow in this case is close to unity, indicating an approximately linear relationship between this flow and mean accident frequency. The value of 3.210 for the shape parameter n_b that was fitted to this reduced dataset of 370 sites is substantially greater than that of 1.883 for the full dataset of 389 sites: this indicates that the prior distribution is more informative in this case.

4.4. Results

4.4.1. Introduction

We now consider the results of applying in turn each of the criteria that were discussed in Section 3.2 for the

Table 7
Notation used in describing the results of the statistical analysis

Symbol	Description	Equation	Section
n_o	Number of accidents observed		3.2.1
μ_a	Bayesian posterior mean accident frequency	(8)	3.2.2
R	Potential for accident reduction	(13)	3.2.3
P_a	Bayesian probability of excessive frequency	(14)	3.2.4

indication of sites that have unusually high accident frequencies. For each of these criteria, we identify those sites that are indicated by it, and present their values according to all of the criteria and their associated standard values as estimated by the log-linear model (9) for the mean of the prior distribution. We discuss the results of this in terms of both indications of sites for further investigation with a view to accident remedial treatment, and of the likely effect on evaluation of any such treatment. Finally, we comment on the 19 outlying observations that were not used in the modelling process but were retained separately.

The notation that is used in this section is indicated in Table 7, together with references to the equations that specify them.

4.4.2. Accident count

When the number of accidents that have been recorded at a site is used as the sole criterion for selection of a site for investigation, this will tend to lead to over-estimation of the future accident frequency if the site remains unchanged. In the present case, a total of nine sites had a record of eight or more accidents during the 5 year observation period. Of these, only three had 12 or more accidents and could possibly have been indicated for investigation according to the current British guidelines of 12 or more personal injury accidents in three consecutive years. These nine sites are detailed in Table 8 together with the various estimates resulting from the empirical Bayesian analysis presented here.

When the null model in which the only parameters correspond to constants to represent the moments is used for the Bayesian prior distribution, the prior estimate of the mean accident frequency does not vary between sites: the mean and variance of the prior are 0.2362 (accidents/year) and 0.1268 (accidents/year)^2, respectively. Using this prior distribution, the posterior distribution, and hence the posterior mean accident frequency μ_a and the probability of excess P_a vary only with n_o, the number of accidents observed at a site. Accordingly, although the posterior mean accident frequency for these sites is generally quite low when no explanatory variables are used in estimation of the prior distribution, the posterior probability that they have excessive accident frequency is uniformly high.

Table 8
Analysis of sites with greatest recorded numbers of accidents

		Null model		Full model			
Site	n_o	μ_a	P_a	R	μ_b	μ_a	P_a
63	39	5.7464	1.000	6.492	1.199	7.691	1.000
64	16	2.3953	1.000	0.000	3.118	3.118	0.5000
24	14	2.1039	1.000	0.000	1.728	1.728	0.5000
100	9	1.3754	1.000	0.000	1.795	1.795	0.5000
3	9	1.3754	1.000	0.6715	1.052	1.723	0.9978
130	8	1.2297	1.000	0.000	1.874	1.874	0.5000
259	8	1.2297	1.000	0.000	0.8931	0.8931	0.5000
260	8	1.2297	1.000	0.000	0.6765	0.6765	0.5000
14	8	1.2297	1.000	0.0232	0.8442	0.8674	0.6264

On the other hand, when the full log-linear modelling approach (9) and (10) is used to estimate the prior distribution for these sites, the strength of the prior distribution is so great that in only three cases (sites 63, 3 and 14) is the posterior distribution substantially different. In these cases, some potential is identified for accident reduction, and the probability of the posterior mean accident frequency being excessive ranges from 0.6264 (site 14), which is only slightly greater than the neutral value of 0.5, to values that are close to unity (sites 63 and 3). In light of the results of applying the full log-linear modelling approach, we conclude that notwithstanding the high frequency of occurrence of accidents at these sites, there is in the most part only weak evidence that their future mean accident frequency will be higher than usual. Only three of the nine sites that would be identified using this approach show any appreciable potential for accident reduction, and only in two of those cases is there strong evidence for this in the form of a high posterior probability P_a of excessive mean accident frequency.

4.4.3. Bayesian estimate of accident frequency

The Bayesian posterior estimate of mean accident frequency could be used as a criterion for selection of sites for investigation. This represents the most accurate available estimate of the future mean accident frequency if the site remains untreated, and is free from bias caused by regression to the mean. In any case, this estimate should be used to provide the value against which future observations are compared to evaluate any accident remedial treatment that is applied. Because of this, use of this value to select sites for investigation could be supported by the case that it will indicate sites at which the frequency of accident occurrence is greatest.

The Bayesian prior distribution that is used in this analysis will influence the posterior estimate and hence the identification of sites for investigation. The two choices for estimation of the prior distribution that are investigated here are the null (moments only) model and the log-linear models (9) and (10). When the null model is used for the prior distribution, the parameters n_b and t_b are invariant between sites so that according to Eq. (8), for the constant period of $t_o = 5$ years the posterior estimate of the mean varies linearly with the number of accidents observed

Table 9
Analysis of sites with high posterior mean accident frequencies when the full log-linear models (9) and (10) are used for the prior distribution

		Null model		Full model			
Site	n_o	μ_a	P_a	R	μ_b	μ_a	P_a
63	39	5.7464	1.000	6.492	1.1985	7.6909	1.0000
64	16	2.3953	1.000	0.000	3.1178	3.1178	0.5000
130	8	1.2297	1.000	0.000	1.8736	1.8736	0.5000
72	7	1.0840	1.000	0.000	1.8150	1.8150	0.5000
100	9	1.3754	1.000	0.000	1.7950	1.7950	0.5000
24	14	2.1039	1.000	0.000	1.7280	1.7280	0.5000
3	9	1.3754	1.000	0.672	1.0517	1.7233	0.9978
129	7	1.0840	1.000	−0.372	2.0728	1.7005	0.2362
89	7	1.0840	1.000	0.000	1.6229	1.6229	0.5000
131	5	0.7926	1.000	0.000	1.3525	1.3525	0.5000
12	3	0.5012	0.987	0.000	1.3040	1.3040	0.5000
9	5	0.7926	1.000	0.000	1.0732	1.0732	0.5000
172	7	1.0840	1.000	0.000	1.0184	1.0184	0.5000

from 0.0641 (accidents per year) if no accidents have been observed to 5.892 if 40 accidents have been observed. Analysis of the sites that have the greatest number of accidents observed (being eight or more during the 5 year period, corresponding to posterior mean frequencies of 1.23 or more accidents per year) is presented in Table 8, the content of which has been discussed in Section 4.4.2 above. Thus if no allowance is made for traffic flows and other influences on accident frequencies, these sites would be expected to have about four or more accidents during a future 3 year period if they were not treated in any way.

Results for the 13 sites that have the highest values of the posterior mean μ_a (all those in excess of one accident per year) are shown in Table 9 for the case that the log-linear models (9) and (10) are used to estimate the Bayesian prior distribution.

Inspection of these results shows that at 10 of these sites, the prior distribution of the mean accident frequency is so strong that it is affected little by the observed data: the posterior mean is therefore high mainly because the prior one is. The sites at which there is some potential for accident reduction are numbers 63 and 3, and at these two sites there is also a high degree of confidence that the mean accident frequency is excessive as indicated by values of P_a that are close to unity. Of the sites described in Table 9, number 129 has a negative value for R because the observed accident frequency is less than the mean of the prior estimated using the log-linear model (9): it is identified by use of the posterior mean accident frequency as an indicator because despite the relatively low observed accident frequency, the posterior estimate remains high.

Table 10
Analysis of sites with greatest potential for accident reduction R

		Null model		Full model			
Site	n_o	μ_a	P_a	R	μ_b	μ_a	P_a
63	39	5.7464	1.000	6.492	1.1985	7.6909	1.0000
3	9	1.3754	1.000	0.672	1.0517	1.7233	0.9978
110	5	0.7926	1.000	0.501	0.3602	0.8616	0.9982
65	6	0.9383	1.000	0.228	0.3685	0.5968	0.9408
354	3	0.5012	0.987	0.114	0.4204	0.5345	0.8011
338	4	0.6469	0.998	0.092	0.5600	0.6516	0.6990

4.4.4. Potential for accident reduction

The potential for accident reduction represents the Bayesian posterior estimate of the amount by which the accident frequency at a site would be reduced if performance were restored to an appropriate standard value after due allowance is made for characteristics of the site. This provides a criterion for selection that prioritises sites that have the greatest estimated excess accident frequency after allowing for regression to the mean and hence those at which there appears to be most to be gained by treatment.

Results for the six sites that have the highest values of the potential for accident reduction R are shown in Table 10 for the case that the log-linear models (9) and (10) are used to estimate the Bayesian prior distribution.

This shows that the sites with the greatest potential for accident reduction had fairly high, though not all of the highest, observed accident frequencies and hence posterior mean μ_a when the null (moments-only) model was used for the prior distribution of the mean accident frequency. In all cases, the posterior probability P_a of excessive mean accident frequency are close to unity, indicating strongly that these sites have greater than average accident frequencies for this sample of 370 sites. When the full log-linear models (9) and (10) are used to estimate the moments of the Bayesian prior distribution, all these sites are found to have prior mean accident frequencies that exceed the population value of 0.2362 per year. Furthermore, the resulting posterior estimates of this exceed those arising from use of the null prior model in five of the six cases. Of these sites, all had relatively high values of posterior probability P_a of excessive mean accident frequency, indicating that these sites have greater accident frequencies than the appropriate standard ones given their flows and design characteristics: the three sites with the greatest potential for accident reduction indicated this strongly with values of P_a that exceed 0.99.

4.4.5. Bayesian probability of excess

The Bayesian posterior probability P_a that a site has an excessive mean accident frequency provides an indication of sites at which the future accident record is expected to be worse than is usual. When the full log-linear models (9) and (10) are used to estimate the prior distribution of the mean frequency, due allowance is made in this for characteristics of the sites. This then provides a criterion for selection that prioritises sites that have the greatest probability of having excessive accident frequency but without including any weighting for the size of that excess.

Results for the six sites that have the greatest values of the

Table 11
Analysis of sites with high posterior probability of excessive mean accident frequencies when the full log-linear models (9) and (10) are used for the prior distribution

		Null model		Full model			
Site	n_o	μ_a	P_a	R	μ_b	μ_a	P_a
63	39	5.7464	1.000	6.492	1.1985	7.6909	1.0000
110	5	0.7926	1.000	0.501	0.3602	0.8616	0.9982
3	9	1.3754	1.000	0.672	1.0517	1.7233	0.9978
65	6	0.9383	1.000	0.228	0.3685	0.5968	0.9408
257	2	0.3555	0.927	0.061	0.1151	0.1763	0.8214
354	3	0.5012	0.986	0.114	0.4204	0.5345	0.8011

posterior probability P_a in the case that the log-linear models (9) and (10) are used to estimate the Bayesian prior distribution are shown in Table 11.

This shows that the sites with the greatest values of posterior probability of excessive accident frequency generally also have high potentials for accident reduction R: indeed, five of the six sites in each of Tables 10 and 11 are the same. The other site that was identified according to the present criterion was site 257 at which only two accidents were recorded over 5 years and has an estimated potential for accident reduction of only 0.061 per year. Of these six sites, three had values of posterior probability P_a greater than 0.99, indicating a strong belief that their future accident frequency will on average be worse than usual for sites of their characteristics, whilst the next site had a value close to 0.95. On the other hand, only two of these sites had posterior estimates of mean accident frequency that exceed one per year, compared with the 13 such sites in the 370 that are listed in Table 9. This shows that a high value of the posterior probability of excessive accident frequency when characteristics of the site are taken into account is a substantially different indicator to a high posterior estimate of mean accident frequency.

Table 12
Comparison of sites selected according to various criteria. (Values that indicate investigation are shown in bold)

	Selection criterion				
Site	n_o	μ_a	R	P_a	Count
63	**39**	**7.691**	**6.492**	**1.000**	4
3	**9**	**1.723**	**0.672**	**0.998**	4
24	**14**	**1.728**	0.000	0.500	2
64	**16**	**3.118**	0.000	0.500	2
100	**9**	**1.795**	0.000	0.500	2
130	**8**	**1.874**	0.000	0.500	2
65	6	0.5968	**0.228**	**0.941**	2
110	5	0.8616	**0.501**	**0.998**	2
354	3	0.5345	**0.114**	**0.801**	2
14	**8**	0.8674	0.023	0.626	1
259	**8**	0.8931	0.000	0.500	1
260	**8**	0.6765	0.000	0.500	1
72	7	**1.8150**	0.000	0.500	1
89	7	**1.6229**	0.000	0.500	1
129	7	**1.7005**	−0.372	0.236	1
9	5	**1.0732**	0.000	0.500	1
131	5	**1.3525**	0.000	0.500	1
12	3	**1.3040**	0.000	0.500	1
338	4	0.6516	**0.092**	0.699	1
257	2	0.1763	0.061	**0.821**	1

4.4.6. Sites not used in the modelling process

We now consider briefly the 19 sites that were removed from the full dataset of 389 observations in the process of fitting the log-linear models (9) and (10) by the empirical Bayes approach described in Section 2.4. These outlying sites have recorded accident frequencies that are consistently greater than the corresponding estimates from the log-linear prior model (9) for the mean. Despite that, the absolute accident frequencies are not high at many of these sites, so that few of them would be indicated for further investigation by any of the criteria developed here.

4.5. Discussion

A range of criteria based upon a Bayesian statistical analysis of accident data are available for the identification of road sites for detailed investigation with a view to low cost engineering works for accident prevention and reduction. In this paper, we have presented and discussed four different possibilities for this, namely n_o, the observed accident count, μ_a, the Bayesian posterior estimate of mean accident frequency, R, the potential for accident reduction, and P_a, the Bayesian posterior probability of excessive mean accident frequency. Each of these is an uncertain indicator for sites and provides a different view of the process of road accident reduction. As a consequence of this, use of different of these criteria will lead to the identification of different sites for investigation.

The 20 sites in the present main dataset of 370 observations that are identified for investigation by one or more of these criteria are listed in Table 12, together with a count of the number of criteria that indicate each of them. Two of the sites (63 and 3) are indicated by all four of the criteria: of these, the indication of site 63 is unsurprising in view of its record of 39 accidents, whilst site 3 has a record of nine accidents which is fewer than occurred at either of sites 24 and 64 which are each indicated by only two criteria. In all, seven sites are indicated by two criteria, but the pairs of criteria are either the observed accident count and the Bayesian posterior estimate of mean accident frequency (sites 24, 64, 100 and 130), or the potential for accident reduction and the Bayesian posterior probability of excess accident frequency (sites 65, 110 and 354). Eleven other sites are indicated by just one of the four criteria, with each criterion indicating at least one site without corroboration from other criteria.

Of the nine sites in the full dataset of 370 that have eight or more accidents recorded, six are identified by at least one other criterion: the remaining three sites (14, 259 and 260)

have strong prior distributions with relatively low mean values so that they are indicated neither on grounds of high posterior mean nor on ones based upon a difference between posterior and prior distributions of mean frequency.

The presence of several sites with null values for both the potential for accident reduction and the posterior probability of excess accident frequency arises because in these cases, the prior distribution has sufficient strength that it is not modified substantially by the information in the observations that are available in the generation of the posterior distribution: this will affect both of these criteria. In the whole of the dataset of 370 observations, 251 sites have values of posterior probability of excess accident frequency that are within 2×10^{-2} of the neutral value of 0.5 and absolute values of potential for accident reduction that are less than 5×10^{-4}: this substantial proportion of over 2/3 of the sites can be considered as ones at which there is little evidence that road accident remedial work in the form of low-cost engineering measures would be effective in reducing the frequency of accident occurrence.

5. Conclusions

The selection of sites for detailed investigation with a view to low-cost engineering treatment is an important part of road accident remedial work. However, the relatively low frequency of road accident occurrence makes an appropriate treatment of their stochastic nature important. This makes the identification of appropriate sites difficult: the outcome of this selection process will itself be subject to random variations according to the occurrence of accidents and hence will be inherently uncertain. The requirements of any method for this include qualified confidence that effort will be directed appropriately as a consequence of its use and a degree of robustness with respect to the quality of the data that are available.

Several distinct, though interrelated, criteria have been explored in this paper using a real road accident database. All of these are based upon a Bayesian statistical analysis, which has been developed here within an empirical Bayesian framework. This shows that use of the various criteria lead to the identification of different sets of sites for further investigation. In view of this, several criteria could be used in a complementary manner in order to build a broad informative description of the characteristics of the sites. Measures that can then be assessed jointly include the chance that treatment will be effective, the likely magnitude of the benefit if a site were improved to the prevailing safety level for one with its characteristics, and the most accurate available estimate of the future mean accident frequency if the site were not treated.

Application of this analysis using the accident record of a set of real road junctions shows that many sites conform closely to the general pattern whilst a few deviate from it and are indicated for further investigation. About half of those that are indicated by any of the methods are indicated by two or more, showing some degree of commonality between the various criteria. Beyond the matter of indicating sites for further investigation, this approach also provides estimates of quantities such as the likely mean accident frequency if the site remains untreated which will be useful in further analysis, including monitoring and evaluation following treatment.

Acknowledgements

The authors are grateful to Jeffrey Paniati and Michael Griffith of the US Department of Transportation Federal Highway Administration for providing the data that were used to illustrate the analysis presented here. The authors are also grateful to two anonymous referees for their helpful comments. This research was funded by the UK Engineering and Physical Sciences Research Council.

References

[1] Hauer E. Observational before–after studies in road safety: estimating the effect of highway and traffic engineering measures on road safety. Oxford: Pergamon, 1997.
[2] Robert CP. The Bayesian choice: a decision-theoretic motivation. London: Springer, 1994.
[3] Maycock G, Hall RD. Accidents at 4-arm roundabouts, Transport and Road Research Laboratory report LR 1120, Crowthorne: TRL 1984.
[4] Hall RD. Accidents at four-arm single carriageway traffic signals, Transport and Road Research Laboratory report CR 65. Crowthorne: TRL 1986.
[5] Pickering D, Hall RD, Grimmer M. Accidents at rural T-junctions, Transport and Road Research Laboratory report RR 65, Crowthorne: TRL 1986.
[6] Hauer E. Empirical Bayes approach to the estimation of unsafety: the multivariate regression method. Accident Anal Prevent 1992;24(5):457–77.
[7] Institution of Highways And Transportation. Highway safety guidelines: accident reduction and prevention. London: IHT, 1990.
[8] McGuigan DRD. The use of relationships between road accidents and traffic flow in black spot identification. Traffic Engng Control 1981;22(8/9):448–53.
[9] Vogt A, Bared J. Accident models for two-lane rural segments and intersections. McLean, Virginia: FHWA, 1997.
[10] Zegeer CV, Hummer J, Herf L, Reinfurt D, Hunter W. Safety effects of cross-section design for two-lane roads, Report FHWA-RD-87-008, Washington, DC: FHWA 1986.
[11] Aitkin M, Anderson D, Francis B, Hinde J. Statistical modelling in GLIM. Oxford: Clarendon, 1989.