april 26 - 28, 1988
amsterdam
the netherlands

traffic safety theory & research methods

Session 4: Statistical analysis and models

SWOV

This book was made possible through the support of 3M Europe

SESSION 4: STATISTICAL ANALYSIS AND MODELS

Summaries of the papers presented by the additional speakers

Geoff MAYCOCK & Mike MAHER, Transport and Road Research Laboratory, Crowthorne, United Kingdom
Generalized linear models in the analysis of road accidents; Some methological issues

Heinz HAUTZINGER, Institut für angewandte Verkehrs- und Tourismusforschung e.V. ITV, Heilbronn, Federal Republic of Germany
Statistical superpopulation models in traffic safety research

Full papers of other contributors

R.M. KIMBER & J.V. KENNEDY, Transport and Road Research Laboratory, Crowthorne, United Kingdom
Accident predictive relations and traffic safety

Snehamay KHASNABIS & Ramiz AL-ASSAR, Wayne State University, Detroit, Michigan, U.S.A.
An exposure-based technique for analyzing heavy truck accident data

KUO-LIANG Ting & CHIN-LUNG Yang, National Cheng Kung University, Tainan, Taiwan, Republic of China
A predictive accident model for two-lane rural highways in Taiwan

Risto KULMALA & Matti ROINE, Technical Research Centre of Finland
Accident prediction models for two-lane roads in Finland

G. TSOHOS, Aristotle's University of Thessaloniki, Greece & A. KOKKALIS, University of Birmingham, United Kingdom
Determination of black spots; A comparitive and correlation study of existing methods

Paul JOVANIS & HSIN-LI Chang, North Western University, Evanston, Ill. U.S.A.
Some observations on theory and methodology in safety research

Summaries of the papers presented by the additional speakers

Geoff MAYCOCK & Mike MAHER, Transport and Road Research Laboratory, Crowthorne, United Kingdom
Generalized linear models in the analysis of road accidents; Some methological issues

Heinz HAUTZINGER, Institut für angewandte Verkehrs- und Tourismusforschung e.V. ITV, Heilbronn, Federal Republic of Germany
Statistical superpopulation models in traffic safety research

GENERALIZED LINEAR MODELS IN THE ANALYSIS OF ROAD ACCIDENTS - SOME METHODOLOGICAL ISSUES

by
G. MAYCOCK and M. J. MAHER
Transport and Road Research Laboratory

1. INTRODUCTION

In recent years, generalised linear modelling has become a popular tool for the analysis of road accident data. This summary paper briefly presents the application of this technique to the analysis of data assembled during a study of accident-involved drivers at the Transport and Road Research Laboratory as a means of illustrating some of the methodological issues which have arisen during the modelling process. The final paper will include examples taken from recent analyses of junction accidents (see for example, Kimber and Kennedy, 1988).

2. THE 'ACCIDENT-INVOLVED' DRIVERS STUDY

In order to explore the relationship between the road accident frequencies of drivers and relevant individual characteristics, 229 car drivers who had been interviewed during the course of an 'on-the-spot' accident study, were invited to take part in further tests at the Laboratory. The visual, perceptual and performance abilities of these drivers were measured. They also completed a 'cognitive failure' questionnaire - to assess how forgetful or indecisive they were - and underwent hazard perception tests in a simulator to measure how long it took them to recognise hazards on the road. Basic information on age, estimated miles driven per year (exposure) and the number of accidents the subjects had experienced in the last 3 or 5 years of driving, were obtained by interview.

Details of the study and of the various statistical investigations carried out are reported elsewhere (Quimby, et al, 1986). The Generalized Linear Modelling analysis presented briefly here takes the frequency (accidents per year) of the self-reported accidents obtained by interview as the dependent variable, and relates this to other potential 'explanatory' variables measured in the study. The analysis relates to 145 drivers for which full data was available, and to the accidents they reported as experiencing in the last 3 years (excluding the 'on-the-spot' accident by which they were sampled). The form of the systematic component of the model fitted was:

$$E[A_i] = K\, T_i\, M_i^{\alpha} \exp \sum [b_j F_{ij}] \qquad (1)$$

where, A_i is the number of accidents reported by the ith individual in T_i years (in this case 3), M_i is the estimated annual average mileage relevant to the T_i years, and F_{ij} are j other explanatory variables; K, α and the b_j's are to be determined.

Equation (1) was fitted using GLIM (Baker and Nelder, 1978) with a LOG link and an OFFSET equal to the natural logarithm of the number of years (T_i) of accident data. The number of accidents is assumed to be a Poisson variable. The results are shown in Table 1, which includes a measure of the sensitivity of the various components, and an analysis of deviance.

The average frequency of accidents reported by the subjects in this study was 0.14 per year. Table 1 shows that age is an important determinant of accident frequency - accidents per year fall by about a factor of 2.8 over the 20-60 year age range. More interestingly accident frequency appears to be relatively insensitive to annual mileage travelled (exposure) - indeed in this small sample, the exponent of mileage is not statistically significant. (Mileage travelled proved however, to be significant in larger samples, though the exponent was still very much less than 1.0; the term is included here for completeness).

The remaining variables in the lower half of Table 1 are the laboratory measures which proved to be significant correlates of an individual's accident liability. The movement in depth test is a test of decision making ability. The sign of its coefficient is however noteworthy; it implies that the safer drivers took longer to respond to this particular test - a result which may be explained in terms of caution in decision making style. Median latency is a measure of the time it takes a driver to respond to a hazard in the simulator, and subjects reporting fewer accidents proved quicker at recognising hazards. The positive correlation shown between accident frequency and cognitive failure is also intuitively reasonable - though this may have something to do with the fact that the accidents were self-reported. The practical significance of these findings are discussed elsewhere (Quimby, et al, 1986); here we are concerned with the statistical methodology.

The figures shown in the upper half of Table 1 illustrate the kind of results to be expected from the analysis of a survey of self-reported accidents for which the measures of performance included in the lower half of the table are not available. (They could also - with different variables - represent a model relating accidents per year at a range of junctions to site specific variables). In the present example, after fitting a model which includes age and exposure, Table 1 shows that the residual deviance (139.6) is reasonably close to the number of degrees of freedom (142). Of course with a sample size of only 145 these statistics are not well defined, but this is a result which taken at face value, would suggest that the fitted model has accounted for all the systematic variation in the data leaving only a random Poisson error component (see 3.1 on goodness of fit statistics). We know in this case however, that significant systematic components are omitted from the model. The conclusion that the model 'fits well' is thus incorrect. Moreover, even though in general we may not have direct measures of all the explanatory variables likely to be useful model predictors, we might still like to obtain an estimate of the residual between-individual (or between-site) variation in accident frequency which could potentially arise from such unobtainable variables. The following section suggests a strategy for dealing with this situation.

3. MODEL FITTING

3.1 Goodness of fit statistics.

The principal statistic calculated by GLIM for the purpose of testing significance and goodness of fit is deviance. Deviance is a likelihood ratio statistic and is asymptotically distributed like χ^2. It has additive properties enabling an analysis of deviance to be presented analagously to analysis of variance. In general, the calculation of deviance from observed and estimated data values involves a scale factor which is dependent on the error distribution from which the data is assumed to be drawn.

In the case of Poisson errors the scale factor is 1, and in models where a constant term is fitted the scaled deviance is $y[\ln(y/\hat{\mu})]$ where y are the observed values and $\hat{\mu}$ are the model 'fitted values'. If this error distribution is correct, and providing the fitted values ($\hat{\mu}$) are generally greater than 1.0, the differences in scaled deviance obtained by fitting null terms to the model should be distributed like χ_1^2. This fact can be used directly as a test of the statistical significance of added terms. Moreover an overall 'goodness of fit' assessment can be made by reason of the fact that for a well-fitting model with an appropriate link function, error distribution and functional form, the expected value of the residual scaled deviance should approximately equal the number of degrees of freedom. (Appendix A of McCullagh and Nelder, 1983, provides a correction to deviance which seems useful for values of $\hat{\mu}$ lying between 1 and 20; this correction should not however be used when values of $\hat{\mu}$ in the vector of fitted values fall below 1).

Although the expected value of deviance is approximately 1 per degree of freedom whilst the model fitted values are greater than 1.0, it falls dramatically (at least for Poisson and Negative Binomial data) as μ falls below 1.0. Fig. 1 shows how the expected value of scaled deviance for Poisson and Negative Binomial distributions varies with μ. Thus a data set which has a high proportion of estimated accident frequencies less than 0.5, will have an expected value of the scaled deviance for the data set as a whole considerably less than the number of degrees of freedom. This is the case shown in Table 1. The expected value of deviance (calculated from the fitted values) is 129 - considerably less than the number of degrees of freedom (142).

An alternative test of overall goodness of fit is provided in GLIM by means of the 'generalised Pearson' χ^2 statistic. Assuming each data point to be unit weighted, this statistic (X^2) is:

$$X^2 = \sum \frac{(y - \mu)^2}{\text{(Variance function)}},$$ where the 'variance function' is the

variance of the assumed error distribution expressed as a function of the mean. In the case of a Poisson errors X^2 is: $\sum (y - \hat{\mu})^2/\hat{\mu}$. Differences in X^2 as between nested models are not χ_1^2 variables, so that this statistic cannot be used for testing the significance of adding terms to a model - note for example, the increase in X^2 as the movement in depth term is added. Moreover the variance of X^2 is a function of μ for small values so that difficulties arise in using this statistic for overall goodness of fit. By definition however, for a well fitting model with the appropriate error distribution (and variance function), the actual value X^2 should equal the number of degrees of freedom irrespective of the value of μ. In the case of the accident involved driver data presented in the upper part of Table 1, it will be seen that the value of X^2 for the simple model is 163.2 - considerably exceeding the number of degrees of freedom and indicating over dispersion in the residuals compared to Poisson errors.

It will be seen therefore that the agreement between the final model deviance and the number of degrees of freedom for the simple model (upper part of Table 1) is coincidental. It arises from over dispersion (which inflates the deviance) in combination with low values of accident frequency (less than 1.0) in the vector of fitted values (which reduces the deviance).

3.2 Over dispersion

The existence of over dispersion in real data is well known and the simplest technique for dealing with it is the use of 'quasi-likelihood' (McCullagh and Nelder, 1983). Such methods assume a common dispersion parameter which is independent of μ - rather like the residual variance in a least squares fit. In the present context an alternative treatment may be preferred. Over dispersion can arise in three ways:

(i) the systematic component of the model may be incorrect - available variables have not been included, or have not been included in the most appropriate form,

(ii) significant variables have had to be omitted from the model

(iii) the assumed error structure is inappropriate.

Normally, we would have hoped to eliminate the first as far as possible by attention to the range and the form of the explanatory variables used, and by experimenting with alternative model specifications. The most appropriate representation of the structure of the residual variation will be one which handles the combination of (ii) and (iii) sensibly.

As was suggested earlier, in analysing the accident data, we may be interested in estimating not only the effects of measured variables (eg. age and exposure in the case of drivers, or traffic flow and layout features in the case of junctions), but also the magnitude of the residual variability arising from other factors. The question here is - what sort of distribution of residual between-individual or between-site effects are we dealing with? Fig. 2 shows a histogram of the between-individual variation in accident frequency arising from the three factors represented in the lower half of Table 1. As expected, the distribution is positively skewed, and a Gamma distribution has been superimposed to represent the between-individual component of the accident variability corrected for age and exposure.

The Gamma assumption is a very convenient one, since it means that providing the within-individual accident generating process can be assumed to be Poisson, the sampling distribution of accidents is Negative Binomial - a distribution traditionally used to represent between-individual variations in observed accidents (Arbous and Kerrich, 1951). The variance of the Negative Binomial distribution is $\mu(\mu + k)/k$, where is the mean and k is the parameter of the underlying Gamma distribution. (Note: as k tends to infinity, the Negative Binomial distribution approximates to the Poisson). The value of k in the Gamma distribution can be regarded as a measure of the potential unexplained between-individual variation in accident liability once known variables and factors have been allowed for. It is a convenient representation as it implies that the unexplained variation has a constant coefficient of variation (equal to $1/\sqrt{k}$) which can in principle, be calculated as a function of sub-sets of the data.

The Gamma-Poisson model needs to be checked. The crucial test would be to check that the relationship between the variance and the mean within the data, corresponded to the Negative Binomial variance function given above. Some evidence on this point will be presented in the final paper.

The OWN fit facility in GLIM allows the Negative Binomial error distribution to be fitted directly. The scale factor for this distribution is 1, and the simplest estimator of k is that value which when a Negative Binomial fit is carried out makes the generalised Chi-square statistic (X^2) equal to the number of degrees of freedom. This is equivalent to determining k by the method of moments, and since the expected value of X^2 is independent of μ, the value of k so determined is not affected by low mean values. There are however other methods of estimating k which might be preferred. If e is the residual $(y - \hat{\mu})$, then $E[e^2] = \mu + \mu^2/k$ and an estimate of k is given by $\sum\hat{\mu}^2/\sum(e^2 - \hat{\mu})$; a plot of e^2 against μ should look like a quadratic passing through the origin. k may also be estimated by maximum likelihood methods.

These alternatives will be discussed in the final paper.

Clearly, determining k by equating deviance to the number of degrees of freedom as has been done previously (Maycock and Hall, 1984) is only satisfactory if low mean values (see 3.3 below) are not a problem. The use of Mean Deviance Ratio as an F statistic can also be misleading in these circumstances.

3.3 The low mean value problem

Once the problem of over dispersion has been satisfactorily resolved by either a quasi-likelihood method or the use of a Negative Binomial fit, a satisfactory method is required for testing the significance of extra terms in a model in the presence of low fitted values. We know in this situation that even if the Negative Binomial model is satisfactory, the calculated deviances will not be χ^2 (degrees of freedom) variables. There is however some evidence that the deviance differences are χ_1^2 variables, and this property of deviance difference is currently being studied in greater detail.

As a alternative to the use of deviance difference, significance of extra terms may be assessed by means of estimates of standard errors obtained either from the Negative Binomial model, or from the Poisson model using the 'jacknifing' technique. It is hoped to be able to incorporate an assessment of the relative usefulness of these alternatives in the final paper.

4. IN CONCLUSION

Some methodological issues which arise in the application of the Generalized Linear Modelling methodology to the analysis of between-individual accident liabilities of drivers or to the between-site variations in junction accident rates have been discussed. The issues have been illustrated by means of an analysis of the accident histories of accident involved drivers.

Two problems relating to the use of deviance as a test of significance and goodness of fit have been raised: the presence of over dispersion in the data due to between-individual systematic effects omitted from the model, and the reduction in the expected value of deviance when there is a predominance of fitted values less than 1.0 in the data set (or a high proportion of zeros in the observed accident frequencies).

Quasi-likelihood methods provide a simple method of dealing with over dispersion. The use of the Negative Binomial distribution for residuals may however be preferred, although further checking of this model is required. Work is in hand to investigate alternative methods of estimating the parameter k of the Negative Binomial model, and for judging the significance of extra terms in a model in the presence of both over dispersion and low fitted values.

5. ACKNOWLEDGEMENTS

The work described in this paper forms part of the programme of the Transport and Road Research Laboratory and the paper is published by permission of the Director.

Crown copyright. The views expressed in this Paper are not necessarily those of the Department of Transport. Extracts from the text may be reproduced, except for commercial purposes, provided the source is acknowledged.

6. REFERENCES

Arbous, A. G. and Kerrich, J. E. (1951) Accident statistics and the concept of accident-proneness. Biometrics, 7 (4), pp 341-432.

Baker, R. J. and Nelder, J. A. (1978) Generalised linear interactive modelling. The Glim system. Release 3. Rothamstead Experimental Station. Harpenden.

Kimber, R. M. and Kennedy, J. V. (1988) Accident predictive relations and traffic safety. Conference: Traffic Safety Theory and Research Methods, April 26-28, 1988, Amsterdam.

Maycock, G. and Hall R . D. (1984) Accidents at 4-arm roundabouts. Department of Transport, TRRL Report LR 1120: Crowthorne, (Transport and Road Research Laboratory).

McCullagh, P. and Nelder, J. A. (1983) Generalised linear models. Monographs on statistics and applied probability. Chapman and Hall.

Quimby, A. R., Maycock, G., Carter, I. D., Dixon, Rachel and Wall, J. G. (1985) Perceptual abilities of accident-involved drivers. Department of Transport, TRRL Report RR 27. Crowthorne (Transport and Road Research Laboratory).

TABLE 1

'Accident-involved' drivers
Model for individual accident frequency (accidents per year)
145 drivers - Poisson errors

Explanatory Variables	Regression Coefficients (S.E.) (1)		Sensitivity (2)	S Deviance /degrees of freedom (3)	Expected deviance	X^2 (3)
Constant (ln K)	-1.7			148.1/144		168.8
Miles per year (1000's)	0.11	(0.23)	1.4	147.4/143		166.9
Age (years)	-0.026	(0.013)	2.8	139.6/142	129.0	163.2
Movement in depth	-2.10	(0.84)	4.1	132.5/141		166.1
Median latency in the driving simulator	0.009	(0.004)	2.2	126.7/140		156.0
Cognitive failure questionnaire	0.030	(0.014)	2.7	122.2/139	118.3	141.2

(1) The regression coefficients and standard errors relate to the full model.
(2) Sensitivity is the ratio of the predicted accident frequencies at the 5 and 95 percentile points of the distribution of the relevant variable.
(3) Scaled deviance, degrees of freedom and X^2 relate to models containing terms up to and including the term on the current line of the table.

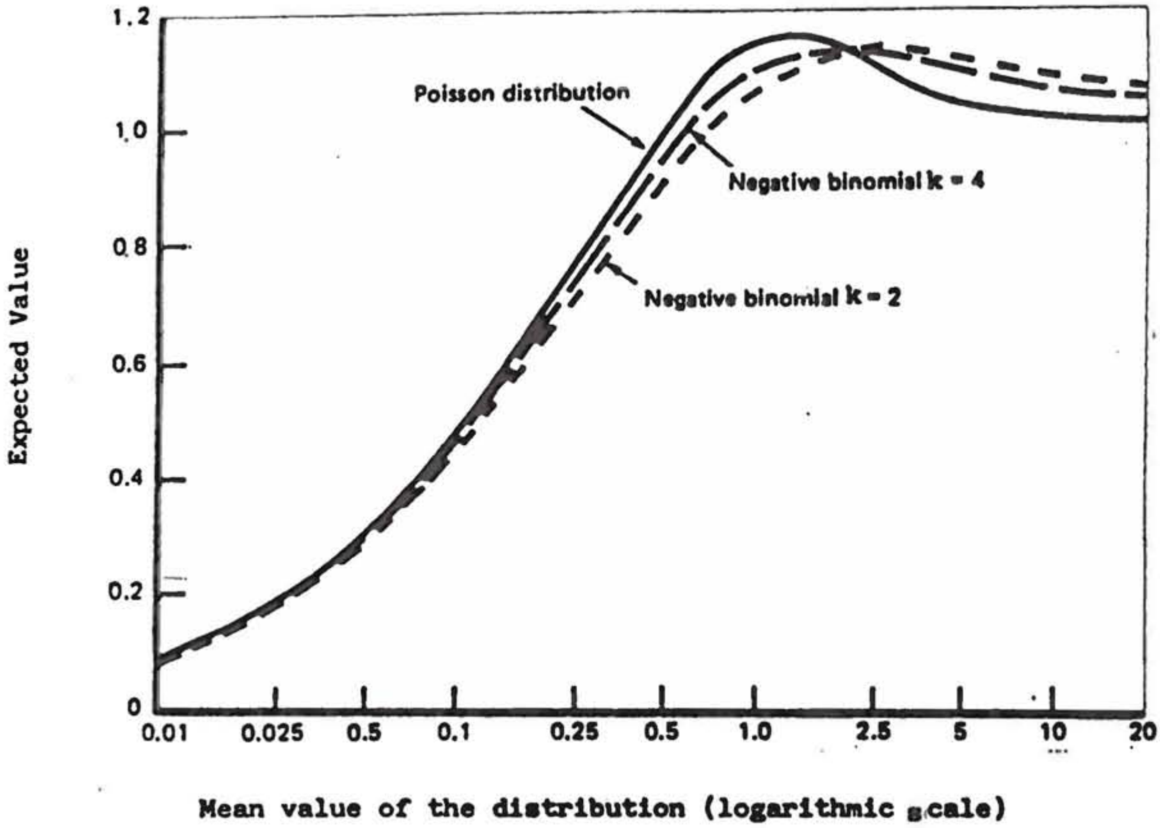


Fig 1. Expected values of scaled deviance for Poisson and negative binomial distribution as a function of the mean value.

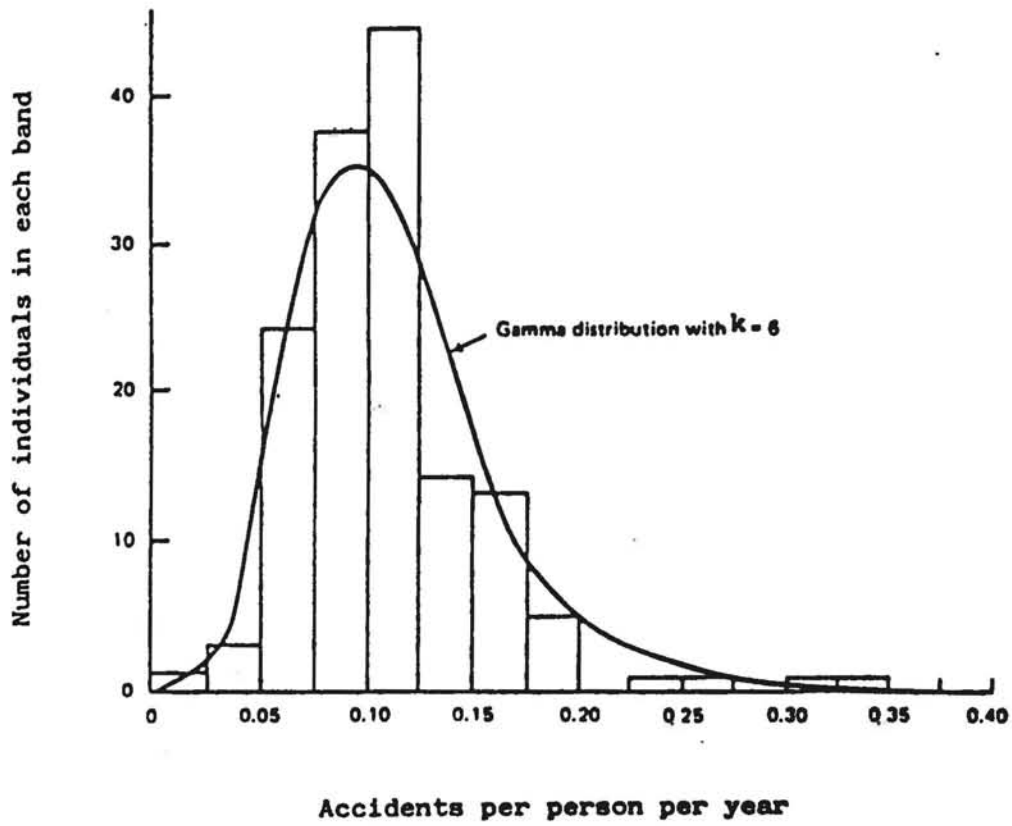


Fig 2. Between-individual distribution of accident liability implied by the 'model' of Table 5 once age and exposure have been allowed for.

STATISTICAL SUPERPOPULATION MODELS IN TRAFFIC SAFETY RESEARCH

Heinz Hautzinger

1. Statistical Concept

In classical sampling theory the population values $y_1, \ldots, y_N$ of the characteristic under study are considered as fixed. Consequently, the population total Y and mean $\bar{Y}$ are also fixed quantities. Stochastic elements are introduced into the analysis by randomly selecting n out of N elements and using the sample mean $\bar{y}$ as an estimator of $\bar{Y}$.

In traffic safety studies this concept is often not really adequate since the population values $y_1, \ldots, y_N$ are properly to be regarded as realizations of certain random variables $Y_1, \ldots, Y_N$. As a simple example consider the case where the population consists of all road crossing in a certain region and where y_i is the number of accidents at the i-th crossing during a specified period of time.

The distribution of $Y_1, \ldots, Y_N$ is usually called a "superpopulation" and in practice this distribution can often be specified up to some parameters. In our example, a simple specification would be to assume $Y_1, \ldots, Y_N$ to be independent Poisson variables with expectation $\mu > 0$. It depends on the research aim whether we are interested in the parameters of the superpopulation model (which in our example is the "accident rate" μ) or in the population mean $\bar{Y} = \Sigma Y_i / N$, which is of course, a random variable.

In both cases we shall select n units from the population and observe the realisations y^*_i of the corresponding random variables Y^*_i $(i=1,\ldots,n)$. The mean

$$\bar{y}^* = \Sigma y^*_i / n \tag{1}$$

of these realisations can then either be interpreted as an unbiased estimate (in the usual sense) of the fixed model parameter μ or as a "model-unbiased" prediction of the realisation of the population mean $\bar{Y}$ in the sense that $E(\bar{Y}^*) = E(\bar{Y})$, where the operator E refers to the superpopulation (and not to the sampling procedure).

Two results are of importance: If our superpopulation model is valid

1. the prediction interval for $\bar{Y}$ is narrower than the confidence interval for μ , and

2. unbiased estimation and prediction does not necessarily require random selection of units.

Superpopulation models are especially useful, if in addition to y_i the values x_i of an auxiliary (or explanatory) variable are available. The following rather general superpopulation model is of special importance:

$$(2) \qquad Y_i = \beta x_i + \delta(x_i)\, U_i \qquad (i=1,...,N)$$

where the U_i are independent identicially distributed random variables with $E(U_i) = 0$ and $var(U_i) = \sigma^2$ for $i=1,...,N$. The parameters β and $\sigma > 0$ need not to be known. Moreover, the x_i are assumed to be positive and known. The function $\delta(x)$ is also assumed to be positive for positive x-values and must be chosen according to the structure of the data. Typical examples are

$$(3) \qquad \delta(x) = 1 , \qquad \delta(x) = \sqrt{x} , \text{ and } \qquad \delta(x) = x .$$

Which functional form is to be preferred can be decided on the basis of a scattergram of (x_i, y_i)-values. CASSEL/SÄRNDAL/WRETMANN (1977) give a simple procedure how to construct a best linear unbiased prediction of the population mean $\bar{Y}$.

It has been mentioned that the above results are independent of the way the sample units have been selected. Actually, under the superpopulation model certain (non random) systematic or purposive sampling procedures are suggested by statistical theory in order to minimize the expected squared prediction error. Obviously, non random sampling bears the risk that our prediction is biased if the assumptions of the superpopulation model are not valid in reality. Therefore, robust random sampling strategies are recommended such that with probability close to 1 the eventual bias is small.

The concept of a superpopulation is a flexible way to incorporate a-priori-information into the estimation procedure. As such it is an ideal combination of theoretical and statistical considerations (accident model and sampling model). Actually, the concept has been developed in the context of ratio estimation. See BREWER (1963) and ROYALL (1970) . The assumption of a certain type of superpopulation model yields an unbiased ratio estimator and variance formula which are both simple and exact for any $n > 1$.

2. Superpopulation Models and Mixtures of Poisson Distributions: a Comparison

By the notion "superpopulation" we mean the joint distribution of $Y_1, \ldots, Y_N$, where Y_i is a random variable associated with the i-th element ("entity") of a population of size N . Thus far, this concept is related to the concept of "mixtures" of Poisson distributions developed by GREENWOOD/YULE (1920). There are, however, important differences between superpopulation and mixture models:

(a) In the case of a superpopulation model the population is assumed to be finite ($N < \infty$) and existent, whereas in the mixture model we often assume that the population is hypothetical and not finite.

(b) The expected value $E(Y_i)$ is in the superpopulation model thought to be a fixed but unknown quantity, which might, of course, vary from one unit to the other. In contrast to this, $E(Y_i)$ is treated in the mixture model as a random variable following a Gamma distribution.

(c) Within the superpopulation concept we imagine our finite population to be a random sample of size N from a superpopulation and, additionally, we assume that a sample of n $(n < N)$ units has been selected from the population. In the mixture model on the other hand we only have an infinite hypothetical population and from this population a sample of size n .

In Section 1 the assumption was made that $Y_1, \ldots, Y_N$ are independent identical Poisson distributed random variables. This is, of course, one of the most simple superpopulation models. It can be generalised in a variety of ways. One possible modification would be, for instance, the assumption that the Y_i are Poisson distributed with expectation

(4) $\mu_i = \exp(\beta x_i)$ $(i=1,\ldots,N)$

where x_i is the value of an explanatory variable observed at the i-th unit and β is a parameter to be estimated. If the units were, for instance, crossings, the explanatory variable might be the volume of traffic flow at the crossing. Sampling theory under generalised linear models of the type described above is, however, just developing.

From (4) another difference between superpopulation models and mixtures of Poisson distributions becomes evident, namely, that the superpopulation model contains an <u>explicite</u> hypothesis on $E(Y_i)$. For instance, this expectation can either be regarded as

(I) being identical for all units in the population or

(II) being identical for all units belonging to a certain stratum of the population (but differing between the strata) or

(III) being a function of a certain explanatory variable (analogous to a regression model).

In contrast to this, the concept of a mixture of Poisson distributions does <u>not</u> contain such a hypothesis on the expected value of accident frequency of a <u>specific</u> unit. It merely contains an assumption on the distribution of the expected value in the population of units. From this point of view, the superpopulation model has the potential of being an explanatory model, whereas the mixture model is merely descriptive.

Of course, under the superpopulation model each of the three alternative assumptions (I),(II),(III) also generates a specific frequency distribution (not a probability distribution) of the expected values in the finite population of units:

Case (I) One-point distribution (degenerate distribution)

Case (II) Discrete distribution with relative frequencies equal to N_j/N, where N_j denotes the number of units in the j-th stratum.

Case (III) Distribution of the expected value depends upon the distribution of the x-variable.

There is a further difference in the two concepts as far as statistical inference is concerned. Under the superpopulation model we may on one hand <u>forecast</u> the total number

$$Y = Y_1 + \ldots + Y_N$$

of accidents in the population or the mean number of accidents per unit, i.e. the quantity

$$\bar{Y} = Y/N$$

(both Y and $\bar{Y}$ are random variables). On the other hand, we may estimate the expected value

$$E(Y) = E(Y_1) + \ldots + E(Y_N)$$

of the total number of accidents or the expected value

$$E(\bar{Y}) = E(Y)/N$$

of the mean number of accidents per unit. Both forecasting and estimation is based on a sample of n units ($n < N$). Under the mixture concept we do not have this distinction between forecasting and estimation.

Of course, we can think also of other forecasting or estimation problems. For instance, we could forecast the number $N(z)$ of units with exactly z accidents. Obviously, $N(z)$ is to be regarded as realisation of a random variable. The proportions

$$f(z) = N(z)/N \qquad (z=0,1,2,\ldots)$$

describe the distribution of the variable "number of accidents" in our population of N units. Under the superpopulation model the frequency distribution $f(z)$ of the characteristic "number of accidents per unit" in the population of size N is, of course, a stochastic quantity. Compared with this, within the framework of a mixture model $f(z)$ is a probability distribution in the usual sense (in the mixture model mentioned above $f(z)$ is a negative is a negative binomial distribution) and statistical analysis concentrates on estimation of the parameters of this distribution.

3. Applications of Superpopulation Models in Traffic Safety Research

In traffic safety research various types of populations are encountered: populations of individuals, vehicles, road sections, crossings, residental areas and so forth. Among the characteristics we observe at the single units of a population there is nearly always the number of accidents or some related veriable. Since the number of accidents of an individual, a road section or crossing and so forth is a random variable, the superpopulation model is a quite natural concept for traffic safety studies. It allows for a clear distinction between the fixed parameters of an underlying theoretical accident model and the random average number of accidents occuring under this model. This is of special importance for group comparisons which are frequently to be conducted in empirical traffic safety research.

Superpopulation models are also useful, if risk exposure quantities are to be estimated, e. g., from household travel surveys. For instance the total length of all car trips made by a population of individuals during a certain year may properly be regarded as a random variable. If we draw a random sample of households and ask for their travel behaviour on a specific day of the year (also randomly assigned to the houshold) we have to deal with two sources of random fluctuation: One due to sampling and the other due to the stochastic nature of the phenomenon under consideration.

A variety of other applications of superpopulation models exist. For instance, the author has based a large scale empirial survey, which was designed to quantify the accuray of official road traffic accident statistics on a superpopulation model for response errors. See HAUTZINGER et al. (1985). The basic idea was as follows: If we define the variable Y_i to be one and zero if an error occurs at the i-th accident or not, repectively, the total number Y of errors in the population of all accidents recorded by police is a random variable.

On the one hand, we are interested to estimate the probability that an error arises (which is a fixed model parameter) and on the other hand we would like to have a prediction of the random proportion of accidents which are affected by an error. It is shown in the full paper how traffic safety related surveys can be designed to be robust and efficient within the superpopulation framework.

References

BREWER,K.W.R. (1963). Ratio estimation in finite populations: Some results deducible from the assumption of an underlying stochastic process. Australian Journ. Stat., 5, 93-105

CASSEL,C.M., SÄRNDAL,C.E. and WRETMANN,J.H. (1977). Foundations of inference in survey sampling. John Wiley & Sons, New York

GREENWOOD,M., and YULE,G.U. (1920). An inquiry into the nature of frequency-distributions of multiple happenings, etc. J. R. Statist. Soc., 83, 255

HAUTZINGER,H., et al. (1985). Genauigkeit der amtlichen Straßenverkehrsunfallstatistik. Forschungsberichte der BASt, Heft 111, Bergisch Gladbach

ROYALL,R.M., (1970). On finite population sampling theory under certain linear regression models. Biometrika, 57, 377-387

Full papers of other contributors

R.M. KIMBER & J.V. KENNEDY, Transport and Road Research Laboratory, Crowthorne, United Kingdom
Accident predictive relations and traffic safety

Snehamay KHASNABIS & Ramiz AL-ASSAR, Wayne State University, Detroit, Michigan, U.S.A.
An exposure-based technique for analyzing heavy truck accident data

KUO-LIANG Ting & CHIN-LUNG Yang, National Cheng Kung University, Tainan, Taiwan, Republic of China
A predictive accident model for two-lane rural highways in Taiwan

Risto KULMALA & Matti ROINE, Technical Research Centre of Finland
Accident prediction models for two-lane roads in Finland

G. TSOHOS, Aristotle's University of Thessaloniki, Greece & A. KOKKALIS, University of Birmingham, United Kingdom
Determination of black spots; A comparitive and correlation study of existing methods

Paul JOVANIS & HSIN-LI Chang, North Western University, Evanston, Ill. U.S.A.
Some observations on theory and methodology in safety research

ACCIDENT PREDICTIVE RELATIONS AND TRAFFIC SAFETY

P M Kimber and J V Kennedy

Transport and Road Research Laboratory, UK

1. INTRODUCTION

1.1 This paper is concerned with the development and use of accident predictive relations. Such relations enable the annual frequency of accidents at a road junction, for example, to be predicted from the road layout (widths, markings and so on), the traffic and pedestrian flows, and a range of other factors.* They can be used

- to identify potential design improvements,
- to provide accident estimates for economic appraisal of road improvements;

and, in conjunction with traffic assignment models,

- to enable the effects on accidents of traffic management schemes to be predicted, and to identify casualty-reducing schemes.

1.2 The cost of accidents in Great Britain is about £2850m per annum; 80 per cent or so, some £2400m, is in built-up areas. A recent Government review of road safety[2] concluded that substantial savings could come from major new research in two areas: traffic management for safety, and behavioural research. Maycock[3] takes up some issues in behavioural research in another paper. Accident predictive relations are crucial to traffic management for safety, since they allow the accident consequences of measures to redistribute traffic and pedestrian flows to be estimated quantitatively. They can also point to behavioural issues, by focussing attention on the traffic manoeuvres at junctions which emerge as particularly accident prone.

1.3 The methods described here have been developed by the Transport and Road Research Laboratory in a series of cross-sectional studies to establish accident predictive relations for roundabouts, rural major/minor T-junctions and urban traffic signal junctions. Each of these junction types was tackled because of particular interest in design improvements to reduce casualties. Their places within the national accident picture are outlined later, in Section 4.

---

*By "accidents" we mean accidents involving death or personal injury; formal definitions are given for Great Britain in Reference 1.

1.4 This paper essentially sets out a broad methodology for such studies and examines their role in future applications. It is structured as follows. Section 2 sets out the methodological basis of the cross-sectional studies, and Section 3 gives illustrations from the results of the three studies that have been completed. Section 4 discusses future needs in the national accident context and work in progress. Section 5 summarises.

2. METHODOLOGY

2.1 Cross-sectional accident studies consider many junctions under a particular form of control. They provide a powerful means for identifying accident determinants by drawing together the accident types and numbers, the junction layout and control characteristics, and the traffic and pedestrian flows as they vary from one junction to another across the sample. The methods we describe here come from the TRRL studies; they were formulated first by Maycock and Hall[4], and expanded and developed by Pickering _et al_[5], and Hall[6]. Analytically, they draw heavily on generalised linear modelling techniques[7,8,9,10]. They allow the development of relations of the general form:

$$A = F(\underline{q},\underline{p},\underline{g},\underline{c}) \qquad \ldots (1),$$

where A is the frequency of injury accidents per year within 20m of the junction, and $\underline{q},\underline{p},\underline{g},\underline{c}$ are respectively the relevant sets of traffic flows (24 hour flows, expressed in thousands of vehicles), pedestrian flows, geometric layout variables (road widths etc), and, at traffic signal junctions, control variables (timings, stage sequences etc). F is a function to be determined.

Structure of studies; samples

2.2 The studies each divide into three main phases: (a) drawing a sample of junctions of a given type, stratified by traffic flow within the main movements (for example, on the major and minor arms of a T-junction), and by main junction features, so as to ensure a wide range in the important variables; (b) conducting a detailed survey of: accidents over the previous several years, junction layout and control variables, and traffic flow; and (c) statistical analysis of these data, and development of accident relations.

2.3 The sample has to be constructed carefully, and extensive prior reconnaissance is necessary before the first phase, (a), so as to ensure freedom from bias. Within each of the sample strata junctions are selected randomly, taking no account of accident numbers. A minimum of three years of accident data are needed - more if the accident frequency is low - but there should have been no major layout changes during the period. However, the sample is necessarily limited in size by constraints in data collection, since the requirements are extensive for each junction. Table 1 shows the main features of the TRRL samples.

TABLE 1: Accident statistics by junction type within the samples

	Rural T-junctions	Signals	Roundabouts		
			Small	Conventional	All
Number of sites	302	177	36	48	84
Period studied (months)	58	48	72	72	72
Junction years	1392	670	166	265	431
Number of accidents	674	1772	647	780	1427
Accidents per year	0.48	2.65	3.89	2.94	3.31
Severity (% fatal or serious)	36	20	17	16	16
Accident rate (per 10^8 total vehicle inflow)*	17.0	34.4	34.8	23.5	27.5

*But see Section 3.3

Analytic methods

2.4 The methodology is based on the usual generalised linear form,[7,8,9,10] consisting of: (i) a systematic component $\eta = a_o + \sum a_i x_i$. where η is a linear predictor variable, x_i are explanatory variables (i = 1, 2, ...), and a_i are regression coefficients; (ii) a random component representing the distribution of data about the regression line, which may come from a family of exponential functions; and (iii) a link function, f, $\eta = f(\mu)$ specifying the link between η and the mean values, μ, of the dependent variable. In 'classical' linear regression an identity link, $\eta = \mu$, is used and the random component taken as Gaussian with variance independent of μ. But in modelling accidents it is usual to assume Poisson errors and a log link function, $\eta = \ell n \mu$.

2.5 The most rudimentary models for the accident frequency contain flow variables only, in some simple algebraic combination - for example, as the total junction inflow Q. Allowing that without flow there would be no accidents, the power function

$$A = kQ^{\alpha} \qquad \text{... (2)}$$

is about the simplest logically consistent form, where k and α are to be determined.

2.6 Observations are of the numbers of accidents (AT) in a period of several years, T. Although such numbers are commonly regarded as Poisson variables, the frequencies, A, obtained from them by division (AT/T) are not. As it stands, therefore, equation (2) would have a non-Poisson error structure if the sample values of A were obtained in this way. It is easy to restore a Poisson structure by multiplying both sides of the equation by T:

$$AT = T.kQ^{\alpha} \qquad \ldots (3).$$

Then, taking a log link function

$$\eta = \ell nAT \qquad \ldots (4),$$

the coefficients α and k can be estimated from

$$\ell nAT = \eta = \ell nT + \ell nk + \alpha \ell nQ \qquad \ldots (5).$$

ℓnT is an 'offset' variable whose coefficient is constrained to unity.

2.7 More elaborate flow models, $A = k'Q_{\ell}^{\alpha} \; Q_{m}^{\beta}$, involving products of flows can be set up similarly. Q_{ℓ} and Q_m can either be sums of component flows, as in a 'cross-product' model where each represents the sum of inflows on opposite arms of a junction, or individual crossing movements, in which case A becomes the frequency of those accidents directly associated with the particular movements.

2.8 With a log link function, the simplest form of general relation incorporating geometric layout variables and junction control variables as well as flows is:

$$AT = T.kQ_{\ell}^{\alpha} \; Q_{m}^{\beta} \quad \exp \sum_i b_i g_i \qquad \ldots (6),$$

where the g_i, i=1,2, ..., represent layout and control variables, and b_i are coefficients to be determined. g_i can be of two types: continuous variables (eg road width) or discrete variables (usually 2-level) denoting the presence or absence of a feature (eg a road island). The effects of the latter can be put in a somewhat clearer form when their coefficients have been determined, by writing $\exp b_j g_j = (1 - c_j g_j)$ where $c_j = (1 - \exp b_j)$ and g_j is the variable, taking the value 0 or 1. This shows directly the percentage reduction ($100c_j$) when the feature is installed.

2.9 For clarity we have omitted pedestrian flows from equation (6), and do so for the remainder of the paper. The principles applying to them are essentially similar, and though they are a very important part of the accident picture, in methodological terms they would over-complicate the outline analysis we present here.

2.10 Maximum likelihood estimates of the coefficients in these models can be determined by means of the programs GLIM[9] or GENSTAT[10], given the link function and error structure. For relations of the type in equation (6), the method employed has been first to enter the flow variables alone; then to enter the geometric and control variables one at a time, taking first those which produce the largest reduction in the discrepancies between the fitted and observed values of AT. To explore the whole of the sample space means examining the effects of many variables. The most appropriate functions in the TRRL studies were chosen as those which combined simplicity, functional appropriateness, and statistical validity. Maycock and Hall examined in some detail the robustness of the functional form of equation (6) and found it superior to the alternative forms tried. Readers are referred to the TRRL Reports[4,5,6] for a full discussion.

Significance testing; goodness of fit

2.11 Significance testing is based on scaled deviance, a generalised goodness-of-fit statistic D defined by

$$D = -2\left\{\ln(\max L_c) - \ln(\max L_f)\right\} \qquad \ldots (7),$$

where $\ln(\max L_c)$ and $\ln(\max L_f)$ are respectively the log likelihood of the current model and of a 'full' model which fits all of the data points exactly. For Poisson distributed data

$$D = 2\sum_i (y_i \ln(y_i/\mu_i) + \mu_i - y_i) \qquad \ldots (8),$$

where i = 1,2, ... n runs over the n data points. For pure Posson errors and $\mu>0.5$ accidents per year, D is asymptotically distributed like χ^2 with n-p-1 degrees of freedom for a model with p parameters. For a well fitting model with such errors, the expected value $\xi(D)$ is approximately equal to the number of degrees of freedom[4]. For two nested models with df_1 and df_2 degrees of freedom respectively, the difference in D is distributed like χ^2 with $(df_1 - df_2)$ degrees of freedom. In principle this provides a basis for significance testing. However, the data do not always conform to the assumption of pure Poisson errors and $\mu>0.5$, and other strategies have then to be employed. Consider first deviations from Poisson errors, which arise from unexplained between-site variations in the accident frequency.

2.12 Extra-Poisson variation. Residual between-site error is conveniently represented by a probability density of Γ-form. Taken with the within-site Poisson errors, the sampling distribution over all sites can be shown correspondingly to be negative binomial[12]. D calculated from equation (8) is then no longer distributed like χ^2. In these circumstances the mean deviance ratio, MDR, can be used[9] instead of D:

$$\text{MDR} = \frac{\text{Deviance difference}/(df_1 - df_2)}{\text{Residual deviance}/df} \qquad \ldots (9)$$

where the residual deviance and df correspond to the best fitting model. MDR is distributed approximately as an F-statistic. An alternative is to specify negative binomial errors directly in GLIM; since the negative binomial distribution has two parameters, μ and S:

$$P(y) = \frac{\Gamma(S+y)}{\Gamma(S)y!} \left(\frac{S}{\mu+S}\right)^S \left(\frac{\mu}{\mu+S}\right)^y \qquad \ldots (10)$$

and S is unknown, the process requires some assumption about S. Maycock and Hall assumed all unexplained between-site error belonged to a single Γ-distribution and adjusted S progressively until, for the best models, the deviance, D', became equal to the number of degreees of freedom, the condition for a well-fitting model with negative binomial errors;[4] D' is given by

$$D' = 2\sum_i \left\{ y_i \ln(y_i/\mu_i) - (y_i + S)\ln((y_i + S)/(\mu_i + S)) \right\} \qquad \ldots (11),$$

and is distributed like χ^2. The coefficient estimates derived in this way for roundabout accident models were almost identical to those using a Poisson structure and the MDR statistic; estimates of the standard errors were about 25% greater. When S is determined in this way the within-site and between-site components of error can be separated in the models.

2.13 Cases when $\mu < 0.5$. Here, values of D fall below those expected for χ^2. Maycock[3] takes up this issue in another paper. Maher[11] has shown that for such cases the quantity

$$(D - \xi(D))/\left\{\text{Var}(D)\right\}^{1/2} \qquad \ldots (12)$$

may be used as a t-statistic, where D is as before and $\xi(D)$ and Var(D) are calculated using the fitted estimates of μ, $\hat{\mu}_i$ for sites i:

$$\xi(D) = \sum_i \sum_{y=0}^{N} d_i(y, \hat{\mu}_i).P(y|\hat{\mu}_i) \qquad \ldots (13)$$

$$\text{Var}(D) = \xi(D^2) + [\xi(D)]^2 \qquad \ldots (14)$$

and

$$\xi(D^2) = \sum_i \sum_{y=0}^{N} d_i^2(y, \hat{\mu}_i).P(y|\hat{\mu}_i)$$

$$d_i(y, \hat{\mu}_i) = 2\left\{ y\ln(y/\hat{\mu}_i) + \hat{\mu}_i - y \right\}$$

$$P(y|\hat{\mu}_i) = \hat{\mu}_i^y e^{-\hat{\mu}_i} / y!$$

It is usually sufficient to take N=20 for computational purposes.

3. SOME RESULTS FROM THE THREE STUDIES

3.1 The three TRRL studies completed over the past several years each produced extensive and detailed results for a wide range of accident types and vehicle manoeuvres, and it is possible only to give some brief illustrative examples here. The full results are given in detail in the original Reports.

3.2 Traffic flows and turning products proved fundamental, and in all cases they were very significantly associated with the accident frequency. Their effects can be represented within a hierarchy of models from 'global' total inflow models, equation (2), to disaggregate flow/geometry models, equation (6). However, it is only when accidents are brought into association with the relevant manoeuvres and intersecting flows that any lasting insight begins to emerge. Figures 1, 3 and 4 illustrate the many interactions involved. Moreover, though they are useful in some applications, the coarser flow models inevitably subsume correlations between flows and junction design features within the sample - for example higher flows tend to be associated with wider roads in the population, and a properly representative sample will reflect that. It means the flow dependence in such 'flow-only' models will continue to hold only so long as these correlations are maintained in future design practice, and this in part circumvents the objective, which is to discover potential improvements in design. Such implicit constraints are not obvious unless the effects of geometric variation are separated. The separation of geometric variation in the 'flow-geometry' models is thus of fundamental importance.

3.3 Both total inflow models and cross-product models suffer from these drawbacks. For total inflow models, the interpretation is further complicated by the different priority status of the inflows on different roads - for example at a T-junction where accident numbers will depend strongly on the distribution of flows between non-turning major road traffic and minor road traffic. A total inflow model for a roundabout with balanced inflows between arms is therefore not comparable with one for a T-junction with very heavy major road flows. Total inflow models are not given here mainly for these reasons, and cross-product models are given as the coarsest level of modelling. For the models described in the following Sections, all terms and coefficients are significant at the 5% level or better.

Four-arm roundabouts

3.4 Figure 1 shows the primary accident types and traffic flows at roundabouts.

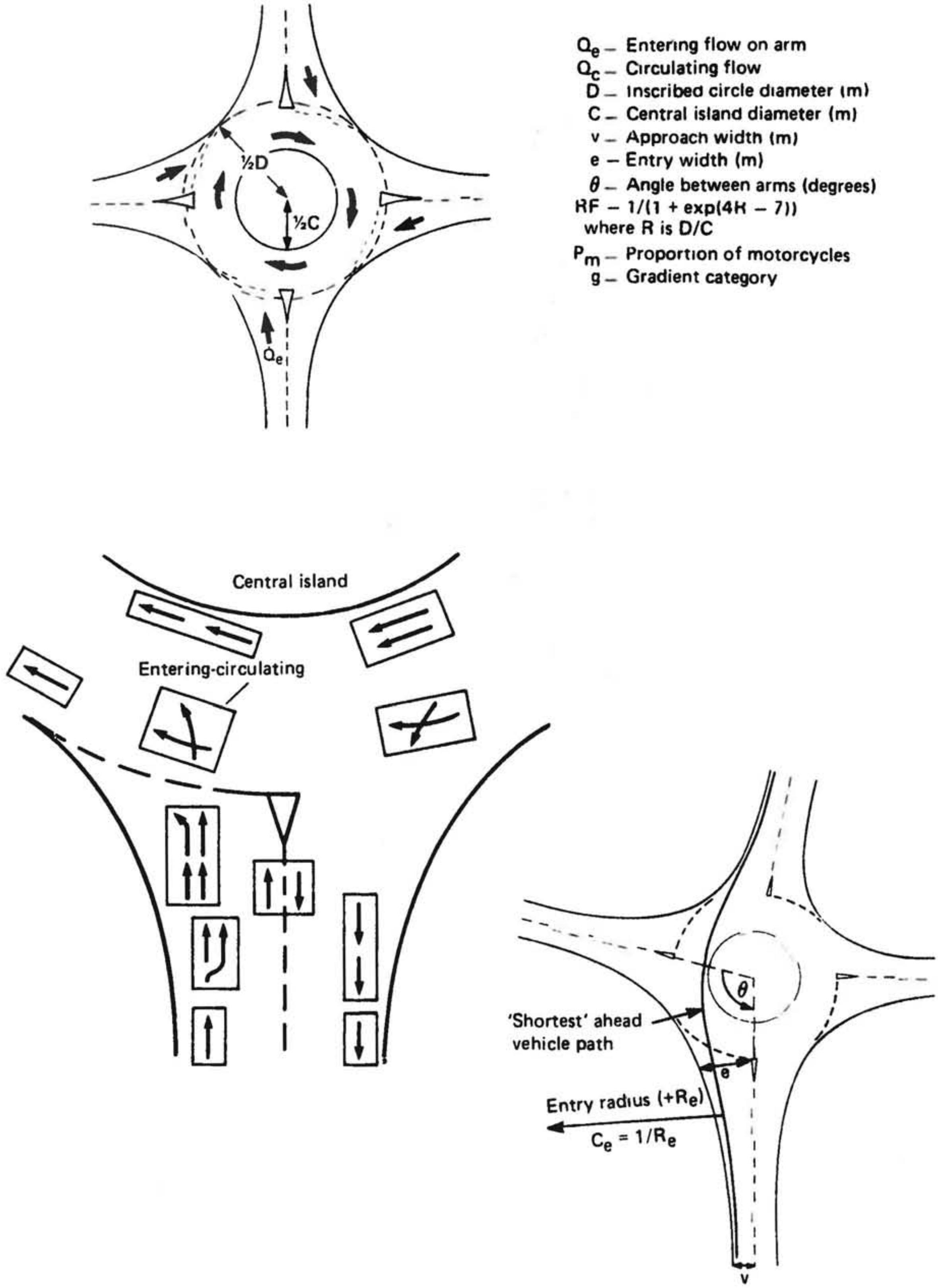


Fig. 1 Entering-circulating accidents at roundabouts showing the important safety parameters and defining the vehicle path curvature C_e (right)

Because of the symmetry of the priority system the problem of accident and flow classification by manoeuvre reduces essentially to that for a single entry arm. Table 2 gives percentages of accidents by type. It shows a very clear difference in accident patterns between small island roundabouts and conventional roundabouts (ie those with a large central island). At small island roundabouts 71% of accidents were of the entry-circulating type whereas only 20% were at conventional roundabouts, where single vehicle accidents (30%) and approaching accidents (25%) were relatively more important.

TABLE 2: Percentage of accidents in the samples by accident type and junction category

Rural T-junctions		Traffic Signals		Roundabouts		
					Small	Conventional
Rear shunt	19.7	Approaching	8.7	Approaching	7.0	25.3
Right turn from major	22.1	Principal right turn	26.5	Entering-circulating	71.1	20.3
Right turn minor	27.4	Other right turn	6.5			
		'Right angle'	13.2			
Left turn	3.4	Left turn	3.2			
Single vehicle	14.4	Single vehicle	8.7	Single vehicle	8.2	30.0
Pedestrian	1.8	Pedestrian	28.8	Pedestrian	3.5	6.4
'Other'	11.2	'Other'	4.3	'Other'	10.2	18.0

3.5 Total accident frequencies for the whole roundabout could be predicted by the simple cross-product model

$$A = K_1(QP)^{0.68} \qquad \ldots (15)$$

where Q and P are the sums of inflows on opposite arms. The constant K_1 was determined separately for small-island roundabouts and conventional roundabouts, and differed between them: $K_1 = 0.095$ for the first, $K = 0.062$ for the second.

3.6 As an example of a particular accident type, we consider entry-circulating accidents. These were associated with the intersecting flows Q_e and Q_c (Figure 1) and could be predicted by

$$A_{ec} = K_2 Q_e^{0.68} Q_c^{0.36} \qquad \ldots (16)$$

Again the constant was determined separately for the two classes of roundabouts with the result $K_2 = 0.088$ for small-island roundabouts, $K_2 = 0.017$ for conventional roundabouts. The difference arose from characteristic differences in geometric layout between the two classes, whose effects were resolved by the full model where the layout parameters defined in Figure 1 are represented explicitly:

$$A_{ec} = 0.046\, Q_e^{0.65} Q_c^{0.36}\, \kappa \exp(-40.3C_e + 0.16e(1 - v/18) - 1.0(RF)) \quad \ldots (17)$$

This expression consists essentially of three parts. The first is the flow function; the second, $\kappa = \exp(0.21P_m - 0.008\theta + 0.09g)$ is a multiplier representing the effect of layout and traffic parameters in effect 'fixed' from the designer's point of view; and the third - the remainder of the expression - is a multiplier determined by the parameters C_e, e, v, and RF which can be adjusted by the designer. The most important of the adjustable parameters to emerge was the minimum vehicle path curvature on entry C_e: increases in C_e produce marked reductions in the accident frequency.

3.7 Expressions of similar general form were derived for the other accident types. A common feature to emerge from this study, and the others, was that some geometric parameters influenced several different accident types in different ways, producing a compound effect depending on flow. Figure 2 summarises the results for the effect of C_e on all accident types at one arm of a roundabout. It can be seen that although its effect is slightly to increase single-vehicle accidents and approaching accidents, the reduction in entry-circulating accidents dominates, and overall accidents are reduced very significantly.

Rural T-junctions

3.8 These lack the symmetry of the priority system at roundabouts and accident types and flow interactions are rather more complex. Figure 3 shows the main classes. From Table 2, right-turning accidents form the largest accident category, accounting for almost half the accidents. Layouts with painted areas on the major road to separate turning traffic ("ghost islands", see Figure 3) were associated with 35% fewer accidents overall at the high flow sites. Table 1 shows the accident rate to be much lower than at the other junction types, but this reflects mainly the relatively high proportion of non-turning major road flows compared to the minor flows (see 3.3 above). Accident severities were substantially higher than at the other junction types. The simple cross-product model for total accident frequency took the form

$$A = 0.24(QP)^{0.49} \quad \ldots (18)$$

where Q is the sum of the flows into the junction from the major road arms and P is the inflow from the minor arm.

3.9 We use two main accident types to illustrate the disaggregation into components - simple rear end shunts in the major road stream approaching from

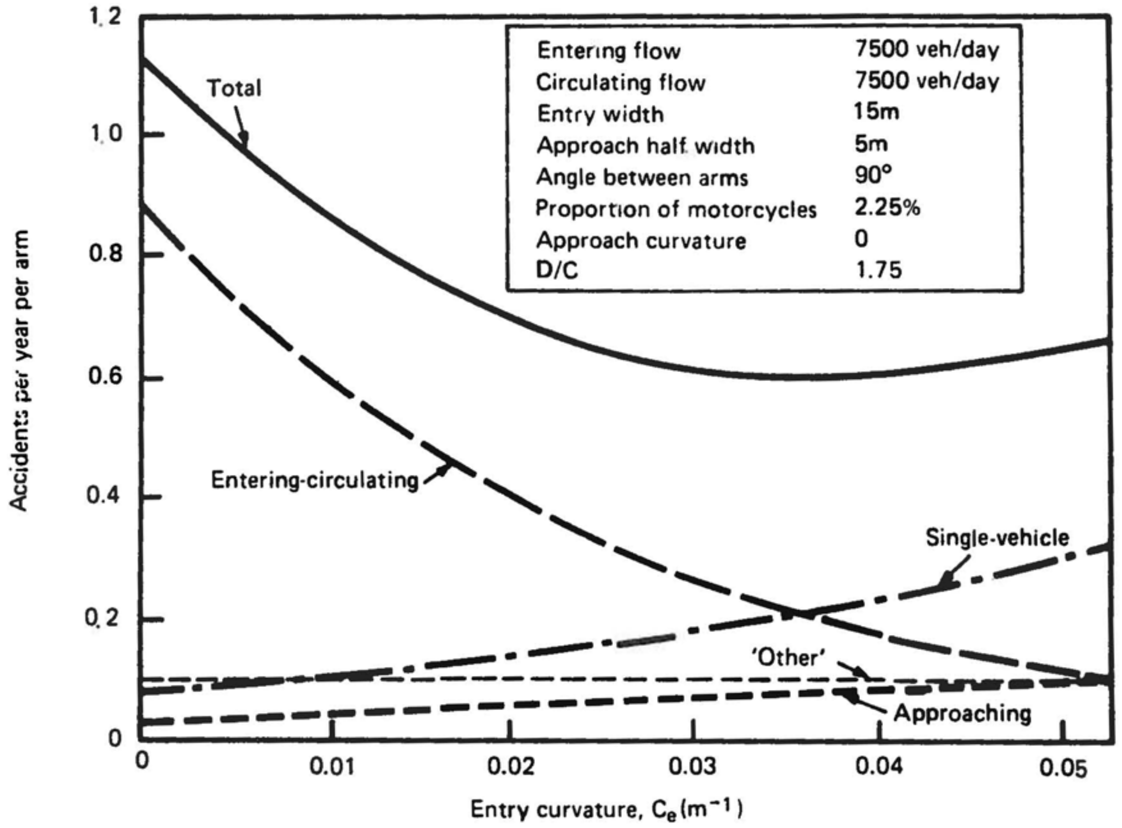


Fig. 2 The predicted effect of entry curvature on roundabout accidents (from Maycock and Hall[4])

left to right on Figure 3, and right-turning accidents from the minor road. For the first, the frequency A_s was strongly associated with the flows Q_1 and Q_2 and could be predicted by

$$A_s = 0.026\, Q_1^{1.14}\, Q_2^{0.33} \qquad \ldots (19)$$

Submodels of this form developed for two classes of junction, one with ghost islands on the major road and the other without, indicated lower frequencies with ghost islands. The full analysis also showed that the accident frequency decreased as the width of the major road, v_1, increased. These effects are represented in the flow-geometry relation:

$$A_s = 0.18(1 - 0.71\delta_G)\, Q_1^{1.39}\, Q_2^{0.46}\, \exp(-0.48v_1) \qquad \ldots (20)$$

where $\delta_G = 1$ for sites with a ghost island and zero for those without. A_s is thus less by 71% at sites with ghost islands. The interaction between flow and geometric variables is illustrated by equations (19) and (20): in equation (19) correlations between flows and geometry are subsumed within the indices; in equation (20) the indices represent the dependence of A_s on flow at constant geometry. The statistical separation of the two types of variation, with flow and with geometry, is described fully by Pickering et al[5].

3.10 The second example is the right-turning manoeuvre out of the minor road. The accident frequency A_r, was associated with the flows Q_3 and Q_6, and the simple flow model took the form:

$$A_r = 0.215\, Q_3^{0.32}\, Q_6^{0.82} \qquad \ldots (21)$$

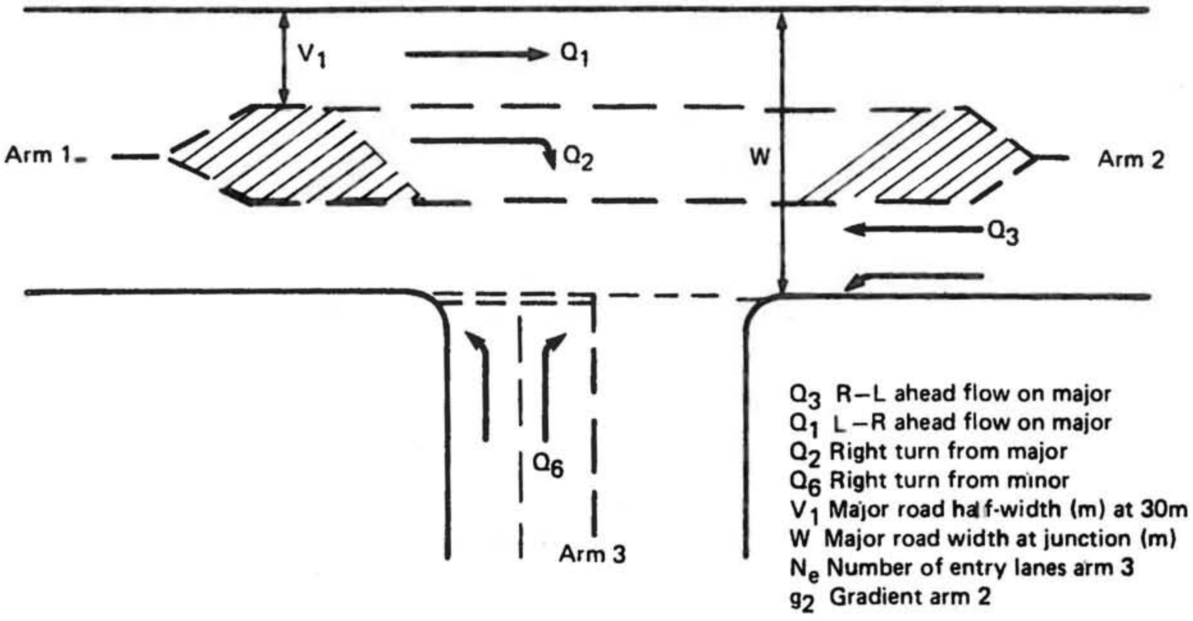
and the flow-geometry model:

$$A_r = 0.038\, Q_3^{0.21}\, Q_6^{0.72}\, \kappa'\, \exp(0.14W + 0.37N_e) \qquad \ldots (22)$$

where the symbols are as in Figure 3. κ' is a 'fixed' term determined by the gradient g_2: $\kappa' = \exp 0.075g_2$, and is unity at flat sites. The accident frequency is higher at the larger junctions where W and N_e are larger.

Four-arm urban traffic signal junctions

3.11 These are more complicated still: the symmetry of priorities of the roundabout case is again missing, and there is now a wide range of signal control variables to add to the basic geometric variables. Moreover, pedestrian activity is very significant, though we do not take that up here. The accident types and flow interactions are many, and accidents have to be carefully grouped to provide a basic structure. Jerry et al[13] discuss this problem and provide an analysis of accidents at Canadian junctions. Figure 4 shows the main accident groupings adopted by Hall[6] in the TRRL study, and the corresponding geometric and flow variables. We can only present a small fraction of the full results here.



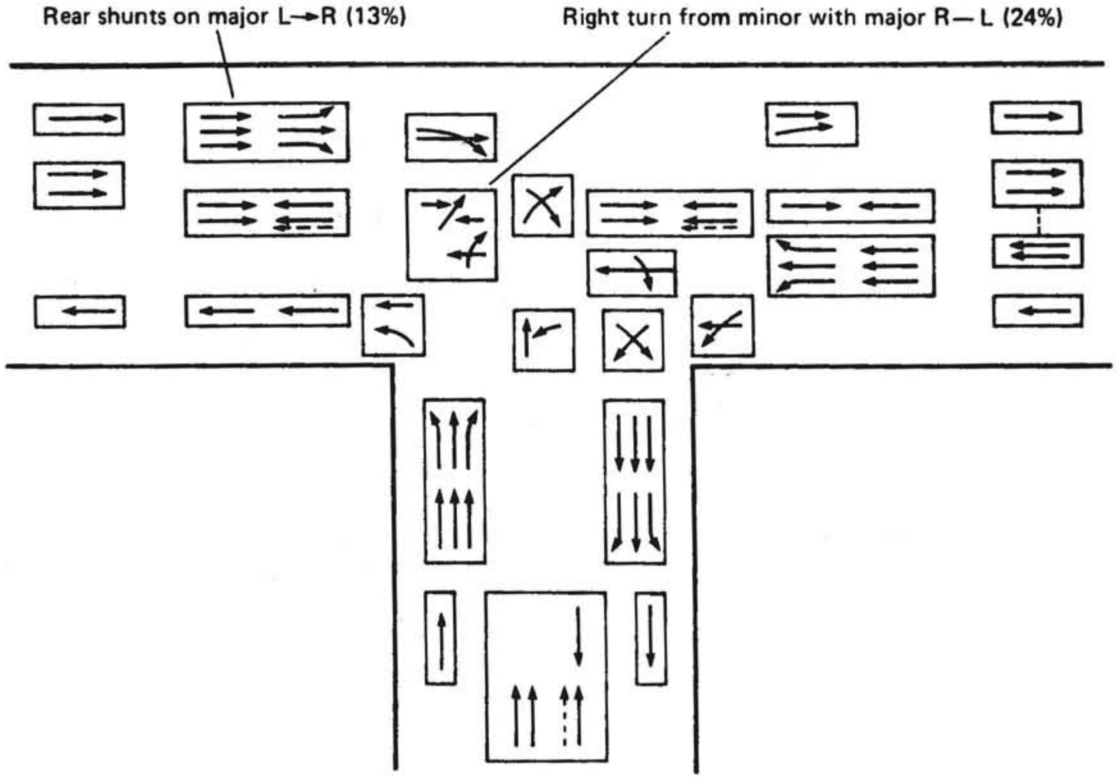
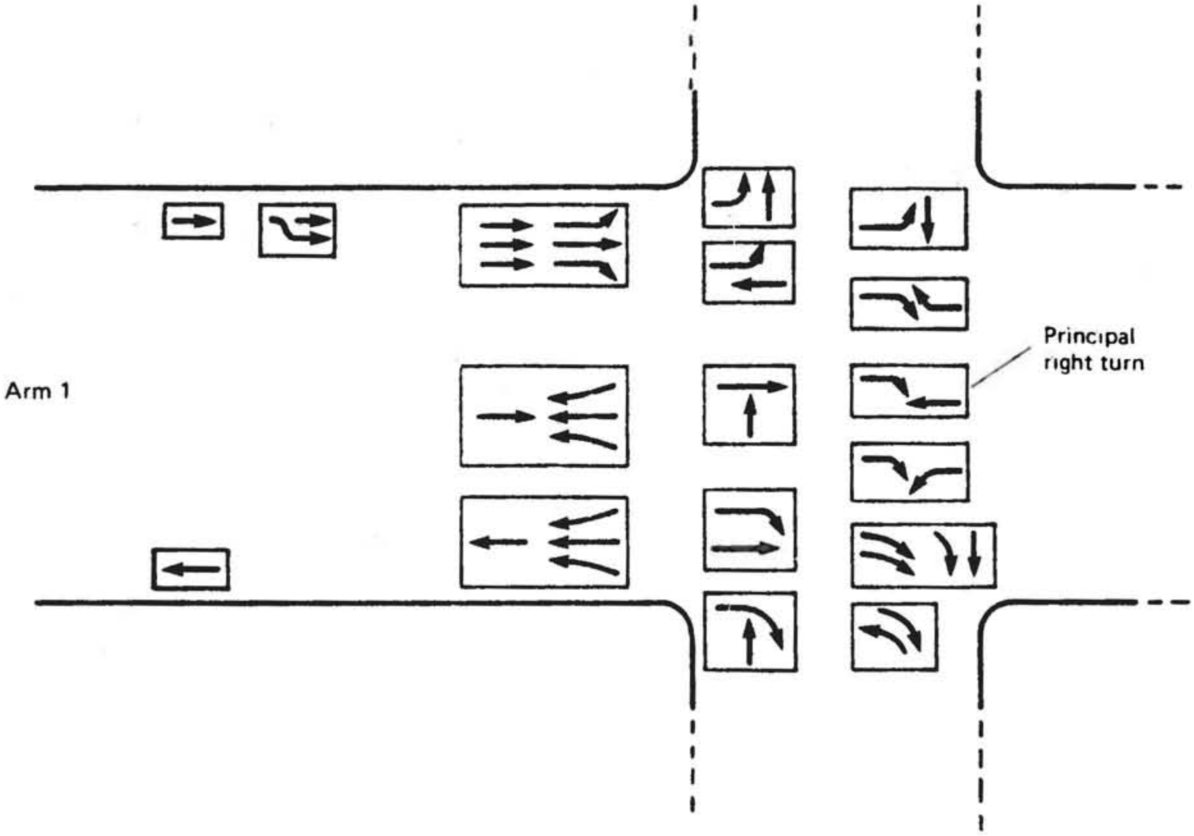


Fig. 3 Accidents at rural T junctions showing: rear shunts on the major road (left to right) and accidents between right turners from the minor road with vehicles travelling from right to left on the major road



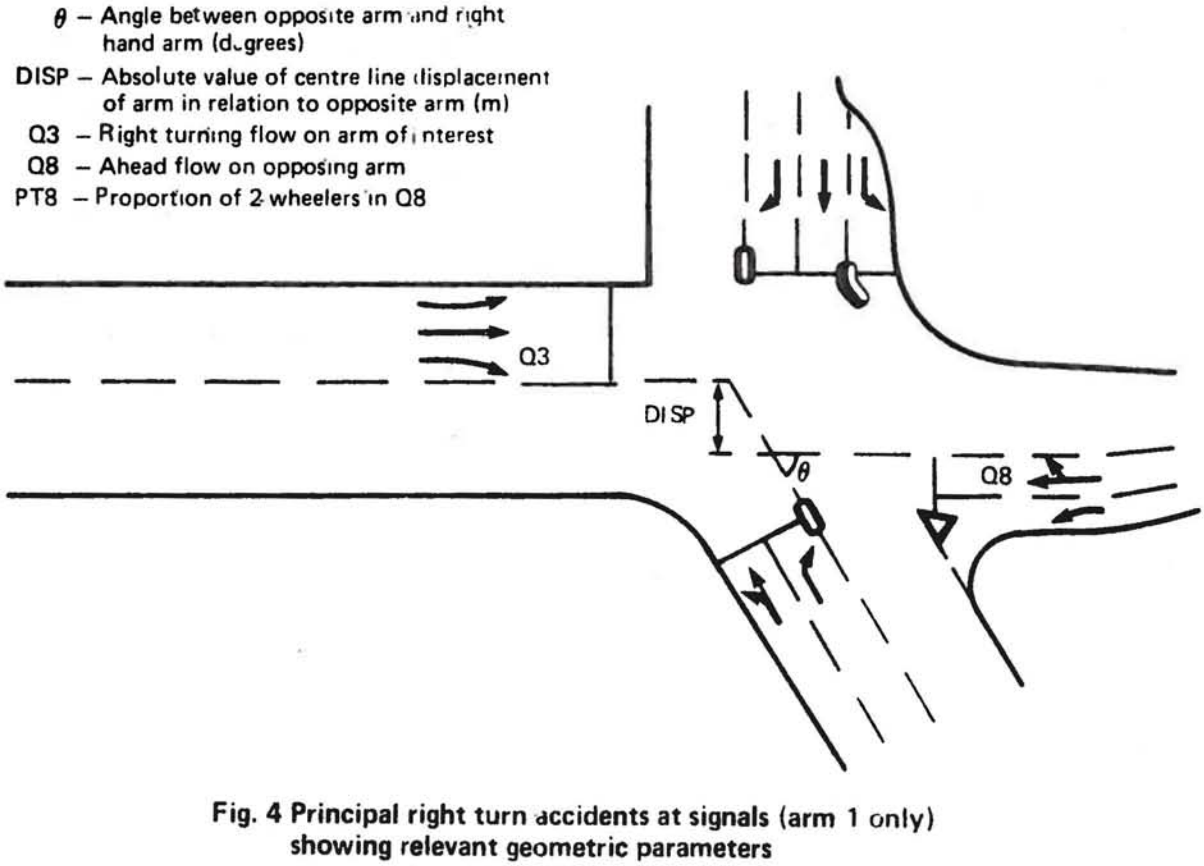


Fig. 4 Principal right turn accidents at signals (arm 1 only) showing relevant geometric parameters

3.12 The simple cross-product flow model for total junction accidents gave:

$$A = 0.152(QP)^{0.63} \quad \text{... (23)}$$

where Q and P are as in the roundabout case.

3.13 As an example of one accident type of many, we take the principal right-turn accidents, this is the largest single group, accounting for about a quarter of all accidents, and has preoccupied designers for many years in trying to achieve safe and efficient designs. The accident frequency A_{pr} per arm was associated with the flows Q_3 and Q_8; the simple flow model gave:

$$A_{pr} = 0.103\ Q_3^{0.5}\ Q_8^{0.61} \quad \text{... (24)}$$

and the model with all significant layout and control variables gave the relation:

$$A_{pr} = 0.179\ Q_3^{0.59}\ Q_8^{0.48}\ \kappa''(1 + 0.32\delta_c)(1 - 0.9\delta_s)\exp(0.85C_{18} + 0.12C_{12}) \quad \text{... (25)}$$

This relation is essentially in four parts: the first is the flow function; the second $\kappa'' = \exp(-0.017\theta - 0.1DISP + 2.76PT8)$ is a multiplier representing the effect of 'fixed' layout parameters; the third $(1 + 0.32\delta_c)$ and the fourth $(1 - 0.9\delta_s)\exp(0.85C_{18} + 0.13C_{12})$ are multipliers representing respectively the effects of a central island (an 'adjustable' layout parameter), and of the signal control variables. Accidents are higher by 32% with a central island ($\delta_c = 1$) than without ($\delta_c = 0$), and lower by 90% with a separate right turn stage ($\delta_s = 1$) than without ($\delta_s = 0$). They increase as C_{18}, the arrival rate per second of green, increases, ie if the proportion of green time is decreased, and as the intergreen C_{12} increases.

3.14 In all some fourteen predictor relations of this general form were developed, according to accident type, and are expounded in detail by Hall. The balance between the accident changes they produce as a function of the design variables, within the total frequency, has yet to be fully explored, as has the trade-off between accidents and vehicle delays. Taken together, they provide considerable insight into the accident risks, and how they might be reduced by design changes.

4. FUTURE NEEDS

The role of accident predictive relations

4.1 In the Introduction, we gave three important uses of accident predictive relations. The first, to identify potential design improvements, is fairly self-evident. As to the second and third, to allow economic appraisal of road improvements and to investigate traffic management strategies, it is not obvious *a priori* that they might not be achieved at a rather more modest level by the use of "aggregate" rates (by road or junction type). In fact, the evidence points to the conclusion that for urban traffic safety appraisal they cannot. The reasons are these.

4.2 It is quite clear that traffic flow variables are crucial in determining accident frequencies. The relations are non-linear in the flows; simple rates per unit of traffic are not therefore sufficient to define accident numbers independently of flow. Moreover, the functional dependence on traffic flow is different for accidents associated with the various different traffic manoeuvres at a junction. This means that the accident consequences of traffic redistribution within a network of roads can only be satisfactorily predicted by means of accident-flow relations which apply to the relevant intersecting flows themselves. For example, a simple total inflow model cannot predict the accident reductions from banning right turns at a series of traffic signal junctions. Neither can a cross-product model. Similarly, the effects of changes in junction layouts on accidents, which depend upon the flows, will simply not appear in an appraisal unless sufficiently discriminating accident predictive relations are used.

4.3 The same will apply for pedestrian activity. Pedestrian accidents are very significant: one-third of fatalities in GB are pedestrians, and 95% of pedestrian fatalities are in built-up areas. The provision and siting of crossing facilities will influence the patterns of intersecting vehicular and pedestrian flows, and hence the accident totals; but unless the accident predictive relations treat these interactions explicitly, the appraisal of traffic management schemes will appear neutral to such things, and possible casualty reductions will be lost.

Traffic management for safety

4.4 These arguments point towards two needs:

(a) a need for methods to predict traffic and pedestrian redistribution effects in road networks following traffic management changes, and

(b) a need for sufficiently discriminating accident predictive models for the major components of road networks - the main types of junction and road link.

4.5 Redistributional effects. Traffic assignment models already allow the effects of traffic assignments to be predicted, given a matrix of origin-destination demand flows. Pedestrian activity is more difficult to cope with, because of the adaptability of pedestrian travel patterns, and it is unlikely that a directly equivalent form of modelling will prove feasible; but because of the extent of the pedestrian accident problem in built-up areas, and the potential effects of junctions, crossing facilities, and strategic traffic re-routing on it, it will nonetheless be necessary to make explicit allowance for changes in pedestrian activity following traffic management changes.

4.6 Junction types and accidents. The distribution of accidents between the main junction types and road links within built-up areas is shown in Table 3. The data are for 1983; equivalent figures have not been collated for later years, but although absolute costs have risen by about 50% since 1983, the distribution of costs between categories can be expected to be similar. Junctions generate nearly two-thirds of all accident costs, and links just over a third. Most accident costs are on single-carriageway roads, primarily at T-junctions (34%) and on the links themselves (34%). Four-arm traffic signal junctions and roundabouts together account for 7%. But in formulating research programmes, these percentages only provide broad indicators. They say nothing

TABLE 3: Analysis of accidents in built-up areas in GB. The data are for 1983; broadly similar distributions of costs can be expected for current conditions; absolute costs have increased by about 50%

Road feature (junction unless otherwise stated)	Accident cost £m	Percentage total cost	Personal injury accidents	% involving pedestrians	% involving cyclists
On single carriageway roads					
Major minor T	514	34	65,286	30	17
" " cross-roads	111	7	14,734	23	14
" " Y	21	1	2,960	28	15
Private drives	51	3	7,449	9	22
Signal cross-roads	62	4	8,501	30	11
Signal Ts	24	1.5	3,371	36	12
Roundabouts	42	3	7,499	12	23
Other junctions	48	3	6,110	34	15
Links	516	34	55,383	39	12
On dual-carriageway roads					
Major minor T	40	3	3,970	30	14
Signal cross-roads	20	1	2,305	22	8
Other	26	1.5	3,208	23	17
Links	55	4	5,085	51	9

about the susceptibility of the figures in the different categories to possible accident reduction measures. Such susceptibilities are by their nature difficult to estimate until the accident risks associated with particular vehicle and pedestrian manoeuvres, flows, and road layout and land-use characteristics have been established. The studies outlined in Section 3, which were mounted primarily to investigate potential design improvements also go some way to providing accident predictors for urban accident appraisal; but the analysis of Table 3 shows that some 87% of the urban junction accident bill remains uncharted in these terms. The largest costs, £514m pa come from urban T-junctions. Urban road links generate a further £516m pa. So between them these two features alone account for more than £1000m pa in 1983 costs.

4.7 Studies of urban T-junctions and road links. Urban T-junctions differ substantially from rural ones in a wide range of factors, including vehicle speeds, on-street parking, pedestrian densities, layout features, and land-use characteristics. It is not feasible therefore to translate the results of the rural T-junction study into the urban context. Neither do any link accident models exist at the appropriate level of discrimination. We have therefore embarked on a major study of urban T-junctions and road links. The boundary between the two is a fine one, because of the multiplicity of minor access points along any urban link, ranging from very lightly trafficked junctions to private drives and retail access points. The sample will encompass about 300 stretches of urban road links totalling around 150 km in length overall. Within this length we expect around 3600 very lightly trafficked minor priority junctions. Stratification will be primarily by traffic flow and pedestrian flows across the road, but will take account of land-use type and parking activity. This sample will be complemented by another comprising a further 300 busy T-junctions stratified by major road flow, minor road flow, and pedestrian crossing flows. Accident data will be collected for the last five years (personal injury accidents), and a comprehensive set of measurements made of flow (by turning movement at junctions), pedestrian flows, layout, and land-use variables and traffic behaviour variables (speeds, parking practices). The study will take about two years.

5. CONCLUSION

5.1 This paper has outlined new methodologies which can be used to develop relations between accidents, traffic and pedestrian flows, and road layout features by means of cross-sectional studies. Although the minimum data requirement is quite large, the yield, in terms of clearly differentiated results for a range of important traffic and pedestrian conflicts, is high. Past studies have pointed to positive design improvements encapsulated at several points in British Department of Transport Advice and Standards. The accident predictive relations are, or will be, incorporated in the widely used DTp computer programs for junction design ARCADY2 (Assessment of Roundabout Capacity and DelaY)[14], PICADY2 (Priority Intersection Capacity and DelaY)[15], and OSCADY (Optimised Signal Capacity and DelaY)[16]. Traffic management appraisal calls for relations, of the type developed, to be used in conjunction with traffic assignment models. There is substantial work to be done to establish a satisfactory basis for appraising the safety aspects of traffic management in built-up areas, and developing casualty reduction strategies. A major study is now in progress to investigate urban T-junctions and road links.

5.2 Whilst there are substantial international differences in road user behaviour and in accident numbers and patterns, much of the basic methodology of the studies described here could be applied elsewhere. Studies conducted on a similar basis in different countries could not only bring out similarities in accident causative processes but also provide valuable indications of which successful national practices could be tried elsewhere. It is planned to explore some of these issues in a short Workshop at the end of this Conference.

6. ACKNOWLEDGEMENTS

The work described in this paper forms part of the programme of the Transport and Road Research Laboratory and the paper is published by permission of the Director.

7. REFERENCES

1. DEPARTMENT OF TRANSPORT. Road Accidents Great Britain, 1986. The Casualty Report. HMSO London (1987).

2. DEPARTMENT OF TRANSPORT. Road Safety: the Next Steps. Report of the Inter-Departmental Review of Road Safety Policy (1987).

3. MAYCOCK G. Modelling individual differences in accident liability - some methodological issues. Int. Symposium on Traffic Safety Theory and Research Methods. Inst. Road Safety Research SWOV, The Netherlands (1988).

4. MAYCOCK G and HALL R D. Accidents at four-arm roundabouts. TRRL Report LR 1120 (1984).

5. PICKERING D, HALL R D and GRIMMER M. Accidents at rural T-junctions. TRRL Report RR65 (1986).

6. HALL R D. Accidents at four-arm single carriageway urban traffic signals. TRRL Contractors Report CR 65 (1986).

7. NELDER, J A and WEDDERBURN, R W M. Generalised linear models. J. R. Statist. Soc. A, 135, Part 3, p370 (1972).

8. McCULLAGH, P and NELDER, J A. Generalised linear models. Chapman and Hall (1985).

9. BAKER, R J and NELDER, J A. Generalised linear interactive modelling (GLIM) Rothamsted Experimental Station, Harpenden (1978).

10. ALVEY, N G et al. GENSTAT. A General Statistical Program, Rothampsted Experimental Station, Harpenden (1977).

11. See Appendix B of Reference 5.

12. ABBESS, C, JARRETT, D and WRIGHT, C C. Accidents at blackspots: estimating the effectiveness of remedial treatment, with special reference to the 'regression-to-mean' effect. Traff. Engng. Control, 19, 447-50 (1978).

13. JERRY, C N, HAUER, E, and LOVELL, J. Are longer signal cycles safer? Proc. Canadian Multidisciplinary Road Safety Conference V Calgary, Alberta (1987).

14. DEPARTMENT OF TRANSPORT. ARCADY 2 User Manual HECD/R/30 (1985).

15. DEPARTMENT OF TRANSPORT. PICADY 2 User Manual HECB/R/31 (1985).

16. DEPARTMENT OF TRANSPORT. OSCADY User Manual HCSL/R/42 (1987).

Crown Copyright. The views expressed in this Paper are not necessarily those of the Department of Transport. Extracts from the text may be reproduced, except for commercial purposes, provided the source is acknowledged.

AN EXPOSURE-BASED TECHNIQUE FOR ANALYZING
HEAVY TRUCK ACCIDENT DATA

By
Snehamay Khasnabis[1]
and
Ramiz Al-Assar[2]

INTRODUCTION

Measures of exposures used in accident analysis are complex and not well understood (3,6). In accident studies one must establish at the outset an appropriate exposure measure to compute accident rates (8,10). The development of such a measure might appear to be a simple task; however, certain conceptual problems must be resolved when the objective is to separate accident data into two or more vehicle categories (e.g., trucks, passenger cars, etc.). The problem arises from a lack of agreement among traffic experts as to what constitutes exposure to accident, particularly when a comparison of accident data by different vehicle categories is the object of the analysis. Current literature on accident exposure indicates little agreement among experts on how to incorporate exposure factors in accident analysis (9,15).

Exposure in accident analysis can be regarded as "opportunity or risk of accident involvement," and can, in its simplest form, be measured by Vehicle Miles of Travel (VMT) generated on a given facility over a specified period of time, usually one year. Implicit in the designation of VMT as exposure is the premise that increased travel generated on a given facility results in greater accident risks. Therefore, the measure of performance or the accident rate must reflect the effect of varying amounts of travel.

[1] Professor, Department of Civil Engineering, Wayne State University, Detroit, MI 48202.

[2] Graduate Assistant, Department of Civil Engineering, Wayne State University, Detroit, MI 48202.

The above rate is appropriate in comparing temporal or spatial trends in accident experience. However, certain methodological problems would arise if one were to use the same measure in comparing accident data for different vehicle categories, e.g., heavy trucks vs. light trucks. The object of this paper is twofold: first to address this methodological issue, and second to present a procedure for analyzing accident data involving trucks of varying sizes, along with a casestudy application.

BACKGROUND INFORMATION

By extrapolating the definition of exposure for the purpose of analyzing truck accident data, one could compute the following:

$$\text{Truck Accident Rate} = \frac{\text{Number of Accidents Involving Trucks}}{\text{VMT Generated by Trucks}} \qquad \text{(A)}$$

The use of the above measure implies that for a specific vehicular category, exposure to accidents is caused by travel generated only by that type of vehicle. It can, however, be argued that exposure to accident for a particular vehicle type i is caused not only by travel generated by type i itself, but also by travel generated, in part, by all other types of vehicles present in the traffic stream. For example, a total of 70,000 truck accidents was recorded in Michigan in 1982, where a truck accident is defined as one that involves at least one truck. Note that these accidents involved approximately 76,000 trucks and 48,000 non-trucks (mostly passenger cars). An argument could be made that truck accidents are, at least in part, the result of conflicts between trucks and nontrucks, as evidenced by the involvement of 48,000 non-trucks. Thus, the measure used to compare accident data should reflect the exposure effect of these non-trucks or, alternatively, the rate should have in the numerator those accidents that involved only trucks.

Khasnabis, et al., in their earlier research, discussed the above methodological issue, and presented three possible approaches for analyzing accident data involving specific vehicular categories (11). In the above study, the authors used "trucks" and "passenger cars" as the specific vehicular categories and demonstrated the application of these approaches using an accident data base for the state of Michigan. The three approaches, presented briefly, are as follows:

Approach 1

Approach 1 requires the categorization of the accident data into truck accidents (accidents involving at least one truck) and passenger car accidents (accidents involving at least one passenger car). Next, the percentage of passenger cars in truck accidents is computed, and the VMT attributable to passenger cars is included in the denominator along with the VMT for trucks. A similar procedure is followed to include the truck VMT in the compilation of the passenger-car accident rate. Thus, by the above definition:

$$\text{Truck Accident Rate} = \frac{\text{Accidents Involving at Least One Truck}}{(\text{Truck VMT} + \text{Contribution of VMT by Pass. Cars})} \qquad \text{(B)}$$

Note that the purpose of including the contribution of VMT by passenger cars in equation (B) is to account for the increased opportunity of interaction resulting from the presence of other vehicles in the traffic stream. In computing the accident rate for passenger cars by this method, a similar contribution by trucks in the VMT attributable to truck-car accidents must to be added in the denominator.

Approach 1 has one inherent deficiency. Comparison of the accident rates for the two vehicle categories by this method does not ensure the use of mutually exclusive data bases. Specifically, an accident between a truck and a passenger car (which is considered a typical truck accident) would be accounted for in both categoreis by this method, thus leaving the analysis open for interpretation.

Approach 2

Approach 2 requires the development of a rate based on a numerator containing the number of vehicles involved in accidents rather than the number of accidents. This approach represents a significant departure from the traditional practice used in most accident analyses, where the number of accidents (as opposed to the number of vehicles) has been used in the numerator. Thus, according to this approach:

$$\text{Truck Involvement Rate} = \frac{\text{Trucks Involved in Accidents}}{\text{Total Truck VMT}} \qquad \text{(C)}$$

Note that equation (C) ensures the use of mutually exclusive data bases with no overlap of sample space in the two rates to be compared. However, the method totally disregards the concept of "opportunity for interaction" between different vehicles by separating trucks and passenger cars into the two distinct categories. Also, the use of number of vehicles in the numerator may unrealistically "inflate" the rate for passenger cars due to the fact that most multi-vehicle truck accidents involve a passenger car as the second vehicle, while most multi-vehicle passenger car accidents do not involve a truck as the second vehicle.

Approach 3

Approach 3, an outgrowth of approach 1, attempts to incorporate into the analysis the use of mutually exclusive data bases, ensuring that a given accident is considered only once as an entity in a comparison pair. The procedure requires the computation of three sets of accident rates, as follows, even though the objective is to compare accident involvement by two types of vehicles.

$$\text{Truck-Only Accident (TOA) rate} = \frac{\text{Number of Accidents Involving Trucks Only}}{(F_t \times \text{Truck VMT})} \tag{D}$$

$$\text{Passenger Car Only Accident (POA) Rate} = \frac{\text{Accidents Involving Passenger Cars Only}}{(F_p \times \text{Passenger Car VMT})} \tag{E}$$

$$\text{Combined Accident (CA) Rate} = \frac{\text{Accidents Involving All other Vehicles}}{(\text{VMT Attributable to All Other Vehicles})} \tag{F}$$

where

$$F_t = \frac{\text{Number of Trucks Involved in All Truck Accidents}}{\text{Number of All Vehicles Involved in All Truck Accidents}} \tag{G}$$

and

$$F_p = \frac{\text{Number of Passenger Cars Involved in All Non-truck Accidents}}{\text{Number of All Vehicles Invovled in All Non-truck Accidents}} \tag{H}$$

Note that in equations (D) and (E) the numerator is the number of accidents in which all of the vehicles involved (as opposed to at least one vehicle, as used in equation C) are vehicles of a given category, i.e., truck or passenger car. The numerator and the denominator in equation (F) are the complements of the accidents and exposures, respectively, considered together in equations (D) and (E). Thus, all accident and exposure data not considered in the previous two equations are contained in the last equation. Further, in equation (G), a truck accident is one that involved at least one truck. Similarly, in equation (H), a non-truck accident is one that does not involve any truck at all. The advantage of using equations (D), (E) and (F) is that each of the three categories represents mutually exclusive and homogeneous subsets of the data base, with no overlap in the sample space. Note also that the limiting value of F_t and F_p is between 0 and 1. In reality, however, F_t is likely to be within a range of 0.6 and 0.7 and F_p between 0.85 to 0.95, with very little year-to-year variation.

Scope of This Paper

The procedure developed by Khasnabis, et al. was used to analyze truck and passenger car accidents in Michigan (11). However, it can be used to study any two or three accident categories, where the assessment of the relative role of these vehicular groups is the object. In Michigan, trucks have historically accounted for only 15% of all travel expressed in VMT, and yet at least one truck is involved in 25% of all accidents (10,14). The increasing number of highway fatalities in recent years has caused researchers to question the relative role of trucks (particularly heavy trucks) in the incidence of traffic accidents (4, 5, 17). Additionally, the passage of the 1982 Surface Transportation Assistance Act, which made it possible for heavier, longer and wider trucks to operate on selected national highways, has raised concerns in the minds of many safety experts (12, 16).

The purpose of the research from which the paper is developed was to adapt one of the three procedures to gain an understanding of the phenomenon of heavy truck accidents in Michigan, by analyzing the historical accident and exposure data. The following definitions have been used in this study:

Accident: An incident for which an official accident report was filed. In Michigan, all accidents involving personal injury or property damage exceeding $200 require an official report.

Truck Accident: An accident for which at least one vehicle was coded as being either a straight truck (single unit) or a semi-tractor.

Double-Bottom (DB): A combination of a truck or truck-tractor and two trailers, with an overall length exceeding 55 feet (up to a maximum of 65 feet).

Single-Bottom (SB): A combination of a truck or truck-tractor and one trailer.

The specific objectives of this paper are as follows:

1. To present a procedure for analyzing heavy truck accident data by proper incorporation of exposure factors involving vehicles of different categories.
2. To determine if there is any significant difference in the accident experiences of the three truck categories, Double-Bottom Trucks, Single-Bottom Trucks, and all other trucks, as reflected by the 13-year data base (1971-83) in the state of Michigan.

METHODOLOGY

A modified form of approach 3 was used to gain an understanding of heavy truck accident phenomena. In equations (D) and (E) the factors F_t and F_p were introduced partially to discount the effect of other vehicles in the exposure estimation. Using the same approach, the following rates can be derived:

$$\text{Double-Bottom Only (DBO) Rate} = \frac{\text{Number of Accidents Involving DB's only}}{F_D \times \text{DBO VMT}} \qquad \text{(I)}$$

$$\text{Single-Bottom Only (SBO) Rate} = \frac{\text{Number of Accidents Involving SB's only}}{F_S \times \text{SBO VMT}} \qquad \text{(J)}$$

$$\text{All Other Trucks (AOT) Rate} = \frac{\text{Number of Accidents Involving AOT's}}{\text{AOT VMT}} \qquad \text{(K)}$$

$$\text{where } F_D = \frac{\text{Number of DB's Involved in all DB Accidents}}{\text{Number of All Vehicles Involved in DB Accidents}} \qquad \text{(K)}$$

$$\text{and } F_S = \frac{\text{Number of SB's Involved in all SB Accidents}}{\text{Number of All Vehicles Involved in SB Accidents}} \qquad \text{(L)}$$

Note that in equation (K), a DB accident is one that involved at least one DB truck and similarly in equation (L), an SB accident is one that involved at least one SB truck. Unfortunately, during the study of the heavy truck accident data, relevant information to compute the parameter F_D and F_s was not available. Hence the numerical values of F_D and F_S were assumed unity. The authors recognize that the validity of this assumption is questionable, because it partially ignores the "opportunity for interaction" concept associated with measurement of exposure. However, since the emphasis of this paper is on methodological aspects and the case study is for demonstration of the proposed approach only, the above assumption appears acceptable. It was felt intuitively that numerical values of F_D and F_S would not be drastically different from each other; hence the conclusions of the case study are likely to remain unchanged, even though there could be some changes in the accident rates computed, if realistic values of F_D and F_S were used.

A two-stage analytic procedure was used to conduct the study:

a) In stage I, an overall statistical analysis of the truck accident data was performed for the analysis period 1971-1983. A two-way analysis of variance was performed to obtain a broad understanding of the most significant factors contributing to truck accidents.

b) In stage II, accident data were categorized into three groups: Class of Trafficway, Severity of Accidents, and Type of Vehicle. The purpose of this categorization was to create a more uniform data matrix to permit a better comparison of the accident data.

Development of Database

Two major databases were developed on an annual basis for each of the 13 years of accident and exposure data. These are briefly discussed below.

Accident Data: Accident data were collected for three different categories, namely, Double-Bottom truck accidents, Single-Bottom truck accidents, and All Other truck accidents. This data was divided into three categories according to severity: Fatal, Personal Injury, and Property Damage. The accident data were furthered categorized into 3 classes of trafficways.

VMT Data: There were two primary sources for calculating truck VMT data: The Highway Statistics (7) and the American Trucking Trends (1). For each of these two sources, total VMT was calculated by multiplying the number of trucks registered in the State of Michigan by the average travel rate in miles per truck, computed from nationwide data. The implicit assumption was that there is no significant difference in the nationwide and statewide travel rates. No information on travel rate for trucks for the State of Michigan was available. An assumption was necessary.

The VMT data generated were compared with a third independent data source, namely, the five-year census data based on information collected through the "Truck Use and Inventory" survey, available for the years 1972, 1977 and 1982 (2). The relative closeness of the data from these three independent sources indicated that the information generated was realistic. It was also assumed that the travel generated by out-of-state trucks was balanced by travel generated outside the State by vehicles registered within Michigan. No effort was thus made to account for truck travel generated in the State by out-of-state trucks, or to discount travel generated by Michigan trucks outside the State boundaries.

Truck VMT data thus obtained was divided into two categories, Double-Bottom Trucks and Single-Bottom Trucks, with the assumption that the travel generated by these two vehicular categories is portortional to their corresponding registration. Lastly, the VMT data compiled for each of the three vehicular groups was further categorized into three class of trafficway follwing a similar estimation procedure. In the absence of any information on truck VMT by functional classification of highways, the only way to derive estimates was to use the classes of trafficway (CTW) used in the census data; these were:

Long range:	[Those traveling more than 200 miles.]
Short range:	[Those traveling less than 200 miles.]
Local:	[Short distances.]

It was assumed that long-range trafficways are facilities with the highest design standard (i.e., interstates and expressways), while those in the shorter range categories are major and minor arterials and/or collectors.

Data Analysis

A three-step process was followed to compute the accident rates. First, information on the number of annual accidents was classified into a three-dimensional matrix, "TOV" (Type of Vehicle), "CTW" (Class of Traffic Way), and "SOA" (Severity of Accident) (27 cells, with three levels for each dimension). Next, VMT data was categorized into three classes of Trafficway (CTW), following the procedure described above. Finally, accident rates were compiled according to equations (I), (J), and (K), with data obtained from the first two sets.

Two types of statistical tests were performed. In stage I, a two-way Analysis of Variance (ANOVA) was conducted following the principles of factorial design, using the Statistical Package SPSS. Standard t-test were conducted in Stage II, which compared the differences betweens the mean accident rates of the two vehicular groups, categorized by the class of trafficway and severity of accident. A null hypothesis was set up and tested for the accident rates as follows:

NULL HYPOTHESIS (H_0:): There is no significant difference between the mean accident rates of a specific severity group and class of trafficway of the compared types of vehicles.

A 5 percent level of significance ($\alpha = .05$) was used for these statistical test. The analysis of variance and "t" - tests required the assumption of the normality of the distribution of accident data. The authors recognize that the validity of this assumption is questionable and suggest either a pre-testing of normality of distribution or logarithmic transformation of the variables to ensure normality in future studies.

RESULTS

The results of the statistical analysis are presented here for each of the two stages:

Stage I: An analysis of variance (ANOVA) was performed following the "Factorial Design" type of statistical experiment, as follows.

Factor	Level
1. Type of Vehicle (TOV)	3 levels - Single Bottom (SB), Double Bottom (DB), and All Other Trucks (AOT)
2. Class of Trafficway (CTW)	3 levels - Long Range, Short Range, and Local

The ANOVA performed for total accidents and fatal accidents are reported separately in Tables 1 and 2. A total of 119 observations is included in each of these ANOVA tables, being the result of three TOV levels, three CTW levels, and thirteen years of data; the measure of performance is the number of annual accidents per vehicle miles of travel, computed according to equations I, J, and K.

Table 1 shows that for total accidents, both the main effects (CTW and TOV) and their two-factor interaction (CTW x TOV) are statistically significant at the 5 percent level. To provide a more direct interpretation:

(1) Accident experience changes significantly with changes in the three vehicular categories for the same class of trafficway (TOV main effect).

(2) Accident experience changes significantly with changes in the classes of trafficway for the same type of vehicle (CTW main effect).

(3) Accident experience changes significantly with changes in the vehicular categories as the class of trafficway changes, or vice versa (TOV x CTW interaction).

Table 2 shows similar data for fatal accidents. Contrary to popular belief, neither the type of vehicle, nor the class of trafficway, nor their interaction appear to have any statistical significance. The lack of significance here, the authors feel, should not be used to infer that the variables are not important. Perhaps ANOVA is a crude tool used for a delicate operation, when the data base suffered from low frequencies. The test presented in Stage II addresses this question in greater detail.

Table 1

ANOVA Results: Effect of Class Trafficway (CTW) and Type of Vehicle (TOV) on Total Accident Rate

Source of Variation	Sum of Squares	DF	Mean Square	F
Explained	1.026	8	0.128	11.919*
-Main Effect	0.507	4	0.127	11.782*
-CTV	0.218	2	0.109	10.145*
-TOV	0.289	2	0.144	13.420*
-Interaction	0.519	4	0.130	12.056*
Residual	1.162	108	0.011	
Total	2.188	116	0.019	

Table 2

ANOVA Results: Effect of Class of Trafficway (CTW) and Type of Vehicle (TOV) on Fatal Accident Rate

Source of Variation	Sum of Squares	DF	Mean Square	F
A. Explained	0.048	8	0.006	1.013
- Main effects	0.023	4	0.006	0.962
- CTW	0.011	2	0.006	0.959
- TOV	0.011	2	0.006	0.964
- Interaction	0.025	4	0.006	1.065
B. Residual	0.638	108	0.006	
Total	0.686	116	0.006	

Stage II: In this set of analyses, statistical comparisons of annual accident rates in various severity groups between DBO's and SBO's and between DBO's and AOT's for long-range and for short-range type facilities are presented. Tables 3 and 4 show that for the long-range facilities the DBO's have experienced significantly higher accident rates than SBO's and AOT's respectively. The above conclusion is borne out by the rejection of the Null Hypothesis in all the tests.

Results of similar analysis with short-range types of facilities are presented in Tables 5 and 6. In all the cases analyzed, the DBO's have experienced higher accident rates than SBO's or AOT's. From an inspection of the data presented, it is also clear that the accident rates for compatible cells are much higher for short-range facilities than for long-range ones. This finding supports an earlier finding in Stage 1, that class of trafficway is an important variable in explaining changes in accident rates.

CONCLUSIONS

This study was conducted as part of an unsponsored research project in the Department of Civil Engineering, Wayne State University, during the period 1985-86. The objective of the study was to develop a procedure for evaluating the relative role of heavy trucks in highway accidents, to demonstrate the feasibility of the approach by applying it to an actual case study, and to assess whether the type of facility has any effect on heavy truck accident experience.

The procedure used is a modified version of an exposure-based method used by the principal author in an earlier study in conjunction with factorial design techniques, to compare truck accidents with passenger car accidents. Analysis of variance and ttests of means were used to examine the accident data for the State of Michigan, and conclusions are as follows:

(1) The procedure developed is a viable approach for analyzing heavy truck accident data and, for the most part, lends itself to the use of data commonly available from state transportation agencies, the U.S. Department of Transportation, and the U.S. Bureau of Census.

Table 3

Comparison of Mean Accident Rates
Between DBO's and SBO's at Long Range Facilities

Accident Type	Mean Rate*	Test	$t_{calculated}$	$t_{critical}$	DF	Conclusion
Fatal	0.0001 0.0005	SBO's vs. DBO's	5.04	1.782	12	(Reject H_o)
P.I.	0.0005 0.0071	SBO's vs. DBO's	8.99	1.782	12	(Reject H_0)
P.D.	0.019 0.0160	SBO's vs. DBO's	9.19	1.782	12	(Reject H_0)
Total	0.0017 0.0242	SBO's vs. DBO's	10.04	1.782	12	(Reject H_0)

H_0: No difference between accident rates of compared class

* Expressed as Number of Accidents Per Million VMT.

Table 4

Comparison of Mean Accident Rates Between DBO's and AOT's at Long Range Facilities

Accident Type	Mean Rate*	Test	$t_{calculated}$	$t_{critical}$	DF	Conclusion
Fatal	0.0005 0.0001	DBO's vs. AOT's	4.57	1.782	12	(Reject H_o)
P.I.	0.0071 0.0032	DBO's vs. AOT's	4.93	1.753	15	(Reject H_0)
P.D.	0.0160 0.0070	DBO's vs. AOT's	5.99	1.734	18	(Reject H_0)
Total	0.0242 0.0096	DBO's vs. AOT's	5.80	1.734	18	(Reject H_0)

H_0: No difference between accident rates of compared class

* Expressed as Number of Accidents Per Million VMT.

Table 5

Comparison of Mean Accident Rates Between DBO's and SBO's at Short Range Facilities

Accident Type	Mean Rate*	Test	$t_{calculated}$	$t_{critical}$	DF	Conclusion
Fatal	0.0001 0.0049	SBO's vs. DBO's	2.89	1.782	12	(Reject H_0)
P.I.	0.0003 0.0721	SBO's vs. DBO's	2.92	1.782	12	(Reject H_0)
P.D.	0.0008 0.1770	SBO's vs. DBO's	3.01	1.782	12	(Reject H_0)
Total	0.0011 0.2542	SBO's vs. DBO's	3.00	1.782	12	(Reject H_0)

H_0: No difference between accident rates of compared class

* Expressed as Number of Accidents Per Million VMT.

Table 6

Comparison of Mean Accident Rates
Between DBO's and AOT's at Short Range Facilities

Accident Type	Mean Rate*	Test	$t_{calculated}$	$t_{critical}$	DF	Conclusion
Fatal	0.0049 0.0005	DBO's vs. AOT's	2.61	1.782	12	(Reject H_o)
P.I.	0.0721 0.0225	DBO's vs. AOT's	2.0	1.782	12	(Reject H_0)
P.D.	0.1770 0.0605	DBO's vs. AOT's	1.97	1.782	12	(Reject H_0)
Total	0.2542 0.0844	DBO's vs. AOT's	1.99	1.782	12	(Reject H_0)

H_0: No difference between accident rates of compared class

* Expressed as Number of Accidents Per Million VMT.

(2) Both type of truck and type of facility as individual factors, as well as their interaction, appear to have significant effects upon truck accident experience in Michigan.

(3) For all severity categories of accidents considered (Total, Fatal, Personal Injury and Property Damage), DBO's appear to have experienced higher accident rates than SBO's or AOT's.

(4) Generally, truck accident rates on long-range facilties appear to be lower than those on short-range facilities. This trend is expected because of the better design standards associated with long-range facilities.

(5) Because of problems associated with the availability of truck accident data, it was not possible fully to incorporate the concept of "opportunity for interaction" in exposure measurement in the case study analysis. The proposed procedure, however, allows for incorporating this effect if appropriate data is available.

(6) Further studies are recommended to refine the procedure to include the contributions to exposure by other vehicles involved in heavy truck accidents in a manner compatible with the available data base. Also, in future studies effort should be made to pre-test the normality of distribution of accident data, before ANOVA and t-test are used. If necessary, operations such as log-transformation of accident rates should be conducted to ensure normality. Lastly, the "t"tests conducted on DBO's vs SBO's, are equivalent to performing multiple contrasts. Future research should use multiple range tests (e.g., Duncan's LSD) for such purposes.

REFERENCES

1. American Trucking Trends, Department of Research and Transport American Trucking Association, Inc., Washington D.C. annual.

2. Census of Transportation, Truck Use and Inventory Survey, Bureau of Census, U.S. Department of Commerce 1972, 1977, 1982.

3. Chapman, R., The Concept of Exposure. Accident Analysis and Prevention, Vol. 5, 1973, pp. 95-110.

4. Chira-Chavla,T., Cleveland, D.E., and Kostyniuk, L.P., Severity of Large Truck and Combination Vehicle Accidents in Over the Road Service: A Discrete Multivariate Model. Transportation Research Record 975, TRB, National Research Council, Washington, D.C., 1984, pp. 23-36.

5. Chira-Chavala, T. and Cleveland, D.E., Causal Analysis of Accident Involvements for the Nation's Large Trucks and Combination Vehicles. Transportation Research Record 1047, TRB, National Research Council, Washington, D.C., 1985, pp. 56-64.

6. Greene, D.L. and Loeble, A.S., Vehicle Miles of Travel Statistics, Lifetime Vehicles Miles of Travel and Current Methods of Estimating Vehicle Miles of Travel, Oak Ridge National Laboratory, Oak Ridge, TN, ORNL/TM-6327, Feb. 1979.

7. Highway Statistics, Office of Highway Planning, Federal Highway Administration; annual.

8. Jovanis, P. and Dellear, J., Exposure-Based Analysis of Motor Vehicle Accidents, Transportation Research Record 910.

9. Jovanis, P. and Chang, H. Modelling the Relationship of Accidents to Miles Travelled, Transportation Research Record 1068, National Research Council, Washington, D.C., 1986, pp. 85-89.

10. Khasnabis, S. and Atabak, A., A Comparison of Accident Data for Trucks and for All Other Motorized Vehicles in Michigan, Transportation Research Record 753, TRB, National Research Council, Washington, D.C., 1980, pp. 9-14.

11. Khasnabis, S. and Reddy, T.R., Systematic Procedure for Incorporating Exposure Factors in Truck Accident Analysis. Transportation Research Record 910, TRB, National Research Council, Washington, D.C., 1983, pp. 36-43.

12. Khasnabis, S., "Operational and Safety Problems of Trucks in No-passing Zones on Two-lane Rural Highways", Transportation Research Record #1052, National Research Council, pp. 36-44, 1986.

13. McGee, H.W., Synthesis of Large Truck Safety Research, Final Report, NHTSA, U.S. Department of Transportation, 1981.

14. Michigan Traffic Accident Facts. Michigan Department of State Police, Lansing, (annual).

15. Scott, R.E. and O'Day, J., Statistical Analysis of Truck Accident Involvements. Highway Safety Research Institute, Univ. of Michigan, Ann Arbor, December 1971.

16. Twin Trailer Trucks, Transportation Research Board Special Report 211, National Research Council, Washington, D.C., 1986.

17. Vallette, G.R., et al. The Effect of Truck Size and Weight on Accident Experience and Traffic Operations. Final Report. Biotechnology,Inc., Falls Church, Va.; FHWA, U.S. Department of Transportation, 1980.

A Predictive Accident Model for
Two-Lane Rural Highways in Taiwan

by

Kuo-Liang Ting
and
Chin-Lung Yang

Department of Communication and
Transportation Management Science
National Cheng Kung University
Tainan, Taiwan 70101
Republic of China

ABSTRACT

This study is concerned with identification and quantification of the complex relationships among geometric design elements and accidents, and with the construction of a predictive model of traffic accidents based on these physical factors and other operational characteristics. A complete data set covering 2-year period of accidents occurred on major two-lane rural highways in Taiwan is used for the analysis. To relax the more strict assumptions of normality and linearity, it begins by creating categorical variables through a series of statistical procedures. Several intercorrelated variables are either grouped into new variables to conform with design practice, or represented by single variables to produce meaningful results. Automatic Interaction Detection (AID) technique is then used to explore the structure of the refined data and to reveal interactions between variables. Prior to the construction of Multiple Classification Analysis (MCA) model, the interactions have to be identified and their significance tested. A graphic method accompanied by statistical tests has been developed in this study, which uses information directly obtained from the AID analysis. Consequently, the

interactive terms are introduced in the MCA model to replace the corresponding raw variables. The model thus formulated performs reasonably well on the data set in spite of its inherent imperfections.

A PREDICTIVE ACCIDENT MODEL FOR

TWO-LANE RURAL HIGHWAYS IN TAIWAN

INTRODUCTION

Accidents on two-lane rural roads have been examined by many researchers and are of great concern to highway engineers of many countries in the world. These roads constitute a large portion of highway facilities and involve relatively high accident rates. Geometric design, traffic use, frequency and charactar of intersectional and access conflict points, and physical condition on these routes vary widely. Thus, without some understanding of their interactive effects on safety on these roads, choices from among many possible improvements and locations are particularly difficult, to achieve the greatest safety benefit from investments in highway modernization.

Despite many studies, the understanding of the effects of geometic design on safety has not been adequate to predict the accident response to individual geometric design element changes. The effects of a few dominant elements have been identified; however, the obviously complex interactions among geometric elements and characteristics on accidents are neither well known nor adequately understood.

The objectives of this research were to explore the interactive effect of geometric design elements and traffic characteristics on accidents on two-land rural roads, and to identify some promising prediction models useful in engineering decisions. Attention is limited to the provincial highways in Taiwan with average daily traffic (ADT) values of 2,000 p.c.u.'s or greater. A procedure of joint use of two multivariate techniques, AID (Automatic Interaction Detection) and MCA (Multiple Classification Analysis), was applied in the modeling phase.

In the following section, previous studies and recent methodological developments in the area are reviewed. The proposed method is then explained, followed by the analysis procedure and model results based on real-life data. This paper is concluded with a summary of the major findings and extensions of the research.

BACKGROUND

Among many variables associated with accident analysis, traffic volume is usually considered the most important explanatory variable. Its effect on road accidents is somewhat better understood and it is generally accepted that there is a positive relationship between VMER (vehicle-mile exposure rate) and ADT (Kihlberg and Tharp, 1968; Shannon and Stanley, 1978). However, different relationship has been reported for tangent sections of road (Baldwin, 1946), or for single vehicle accidents (Zegeer and Mayes, 1979). When accident measures other than VMER are used, such as accidents per mile-year (MYER), the effects of traffic volumes are even stronger (Zegeer and Mayes, 1979; Billion and Stopher, 1957; Versace, 1960; Cleveland and Kitamura, 1978; Cleveland, et. al., 1984 and 1985). The effect of access point density and its interactive effect with ADT were also important, especially in predicting multi-vehicle accidents (Cleveland, et. al., 1984 and 1985).

The discussion of other variables, such as geometric design elements, speed limit, etc., in the explanation of different type of accidents are enormous (for example: Gupta & Jain, 1973; Polus, 1980; Cleveland, et. al. 1984 and 1985). The findings from these studies about the effects of geometric design elements on safety are mixed and conflicting, especially for lower range of ADT (Schoppert, 1957; Perkins, 1956; Rinde, 1977). The effect of a single geometric element is difficult to identify because of the mixing or confounding of these elements in actual highway installations (Rinde, 1977; McBean, 1982). This probably results in overestimating the positive effect of better individual geometric improvements because higher-quality alignments are found more frequently with better cross-section geometric elements on high ADT facilities (Zegeer, et. al., 1981). The interacting effects of the individual elements and the high correlations among these elements were clearly shown in an early study using factor analysis (Versace, 1960).

Mathematical models relating accidents to geometric design elements have been constructed by several researchers (Gupta and Jain, 1973; Roy Jorgensen and Associates, 1978; Blackburn, et. al., 1978; Graham and Harwood, 1982). The functional specifications of these models are generally of linear form; the model fit in terms of variance explained has been relatively poor. Exceptions to this can be found in the multiple linear interactive model developed by Dart and Mann (1970) and the flexible models using second derivatives suggested by Jara-Diaz and Gonzalez (1986).

In contrast to these models using continuous explanatory variables, a descriptive model rather than an explanatory one, has been constructed by Cleveland and Kitamura (1978) to predict off-road accidents. Same type of analysis using AID technique for exploration appeared in later versions of the model, with an attempt to fit simple categorical or mathematical models (Cleveland, et. al., 1984 and 1985). The grouping of design elements frequently used together as a result of design policies into so-called bundles has been recommended for effective modeling. In an earlier application of AID technique, Snyder (1974) used a broader, but less-detailed set of explanatory variables which include the adjacent land use and physical and social characteristics of the region, as well as physical characteristics of the roads. With separate analysis applied to different type of facilities, no interaction terms are found in the additive MCA or regression model.

METHODOLOGY

A complex set of relationships exist involving travelers, vehicles, roadways and environments in a transport system for making trips, and thus in each accident occurrence resulting from occasional system failures that are not compensated for. Because of the complexity of the relationships as well as the large number of characteristics associated with accident occurrence, the traffic safety profession has discovered that direct theoretical analysis is of limited value. Hence, data developed from accidents themselves are analyzed to search for these characteristics and relationships, called inductive modeling. The effort has been directed toward identifying the relationships between accident occurrence and geometric and traffic characteristics. The sample studied will be of site rather than accidents for obtaining the likelihood of accident occurrence under certain conditions. The data file is thus road-segment based, which contains the accident history as well as the physical descriptors of the site. This data is to be analyzed by appropriate multivariate techniques.

A model should be formulated to include the most significant explanatory variables or predictors and to combine them in an accurate structural form, sometimes called a construct. The selections of variables and the fucntional form are generally guided by prior knowledge or based on theoretical considerations. To construct an inductive model based on a large number of predictors, an analyst always faces with problems such as: mixing of continuous and categorical variables; non-linearities in relationships; intercorrlations between the predictors; the interaction effects, etc. Nevertheless, the nonlinear effects and interactions among predictors are more difficult to deal with. The use of cross-classification tables (contingeny tables) can relax some of the more restricted assumptions imposed by many other multivariate techniques. Despite its general simplicity and thus wide use, the method of cross-classification tables presents a serious problems in the analysis with a large number of predictors, each having several categories. The sample is soon segmented to subgroups characterized by sparse observations.

The approach proposed here is to use AID as a preliminary search tool, followed by MCA for model parameterization, each compensating for other's limitations. Both techniques have advantages over conventional analysis of variance or multiple regression technique in that the programs can accept predictor variables in form as weak as nominal scales, do not require linearity or somewhat restricted assumptions, and accept unequal number of observations in cells.

The AID Technique

Since its introduction in the mid-1960's (Morgan and Sonquist, 1963; Sonquist and Morgan, 1964), the AID technique has been widely adopted by marketing researchers (for example: Assael, 1970; Armstrong and Andress, 1970; Green, 1978). Besides its limitations and inexpert use in the area being criticized by Doyle and Fenwick (1975), the technique draws on no sample theory; thus no information can be obtained on the relative importance of the statistical significance of the predictors.

The basic concept of the AID method is to partition the total sample into the most homogeneous groupings in terms of the variance in the dependent variable. All independent variables are categorical. The algorithm considers each variable in turn as the possible basis for splitting the sample into two subgroups. Thus for each variable that partition is found which maximizes between group sum of squares, defined as:

$$BSS = N1\bar{Y}1^2 + N2\bar{Y}2^2 - N\bar{Y}^2$$

where N and $\bar{Y}$ are the sample size and mean of the dependent variable in the parent group.

N1 and $\bar{Y}1$ are the sample size and mean of the dependent variable in split group 1.

N2 and $\bar{Y}2$ are the sample size and mean of the dependent variable in split group 2.

The program then splits the sample on that variable which affords the largest such between sum of squares. The two groups so found then become candidates for splitting. The process continues until terminated by one of the three stopping rules: a group becomes too small; the variance in a group is too small; or no possible split can significantly reduce BSS.

The MCA Technique

The MCA technique examines the interrelationships between several predictor variables and a dependent variable within the content of an additive model (Andrews, Morgan, and Sonquist,1967). MCA is directly related to analysis of variance in its more complex form; it can also be viewed as the dummy variable multiple regression, but with easier interpretation of the model coefficients. Mathematically, the model specifies that a coefficient be assigned to each category of each predictor; thus the score on the dependent variable for each unit can be calculated as:

$$Y_{ij...n} = \bar{Y} + A_i + B_j + ... + E_{ij...n}$$

where $Y_{ij...n}$ = the score of unit n who falls in category i of predictor A, category j of predictor B, etc.

$\bar{Y}$ = grand mean of the dependent variable

A_i = the effect of membership in the ith category of predictor A

B_j = the effect of membership in the jth category of predictor B

.
.
.

$E_{ij...n}$ = error term for this unit

This set of coefficients can be obtained by solving a set of normal equations so that the sum of the squared errors is minimized. The normal equations can be solved by matrix inversion or by a series of successive approximation in an iterative procedure, which are available in most statistical analysis packages. The method assumes that the data being examined can be understood in terms of an additive model. When interactions are known to be present, one can use a combined variable, sometimes called a pattern variable, to replace individual variables.

The Proposed AID/MCA Approach

The basic concepts of using AID and MCA jointly are derived from the work by Cleveland, et.al. (1981), based on the search strategy suggested by Sonquist (1970) and Sonquist, et. al. (1971).

It has been applied in the area of marketing research by Newman and Staelin (1971). A similar approach of using AID as the preliminary search tool, but followed by a logit model, was used for the analysis of dichotomous dependent variables (Green, 1978). The basic concept of the joint use of two techniques is for them to serve complementary functions. The former technique provides guidance on which predictors, which categories within predictor, and which types of interactions to be included in the second-stage analysis. The latter provides an explicit parameterization of the model and appropriate significance tests. The approach proposed entails the following steps:

1. All the predictor variables are expressed categorically. The continuous ones have to be transformed by the least singificant difference method, one of several methods available today. The number of categories within various predictors should be as large as limited by the AID program.

2. AID is applied as a screening procedure prior to the second stage of MCA. The results will suggest the existence and general pattern of interactions.

3. The interactions are located by a graphic method and tested for significance by ANOVA. Only significant interaction terms are to be considered.

4. The variables having strong interactive effects are grouped, becoming a pattern variable to be included in the

MCA analysis.

5. After making sure the problem of extreme multicollinearity is not present, the MCA program is used to estimate the additive model.

ANALYSIS AND MODELS

A data set containing information on traffic, geometric and environmental conditions, and accident experience on two-lane rural roads within the jurisdiction of Taiwan Provincial Government was analyzed. The accident data covering a 2-year period, over 393 sections of major provincial highways, each 3 kilometers long, were acquired from the official source; however, only those accidents involving deaths and injuries were available for the analysis. The entire sample has not been further classified by accident type, such as single-vehicle or off-road, because it would result in extreme skewness in the dependent variable. The data describing the physical and operational characteristics of these roads were immediately available through the inventory files maintained and periodically revised by the Bureau of Public Roads, Taiwan. The information on traffic flow along each road section should be noticed. The range of ADT selected is between 2,000 and 15,000 passenger car units (p.c.u.'s) per day, characterizing high-volume two-lane, rural highways. Due to the mixing of motorcycles in the traffic stream, it is believed that number of vehicles is not a good measure of traffic conditions. Vehicular counts of different types were thus transformed into p.c.u.'s by their passenger car equivalents (p.c.e.). The percentages of motorcycles and trucks and buses, respectively, were retained as other variables to measure the extent of flow nonhomogeneity. These variables and others related to geometric designs are listed in Table 1.

Table 1 - The Description of Data File

Variable Name	Description	Unit
VY1	Accidents per section	No.
V1	Roadbase width	m
V2	Pavement width	m
V3	Length of bridge w/ width <= pavement width	m
V4	Culverts w/ length <= pavement width	No.
V5	Pipes w/ length <= pavement width	No.
V6	Intersections	No.
V7	Guardrail	m
V8	Ditch	m
V9	Signs	No.
V10	Lightings	No.
V11	Length w/ grade 5-7%	m
V12	Length w/ grade 5-8%	m
V13	Length w/ grade 5-9%	m
V14	Length w/ grade 5-10%	m
V15	Length w/ grade 5-11%	m
V16	Length w/ grade 5-12%	m
V17	Length w/ radius <= 15m	m
V18	Length w/ radius <= 30m	m
V19	Length w/ radius <= 45m	m
V20	Length w/ radius <= 60m	m
V21	A.D.T.	p.c.u.'s/day
V22	Motorcycles	%
V23	Trucks & buses	%
V24	A.D.T.	vehicles/day
V25	Terrain	-
V26	Speed limit	kph

Data Transformation

Prior to AID/MCA analysis all the continuous explanatory variables have to be transformed into categorical ones. This was carried out by some statistical methods of making no overlaps of averages between groups, subject to the criterion of least-significance difference set at a certain level. The number of categories within each predictor was arbitrarily set to six, which was automatically reduced, if necessary, by the merging feature of the program. The correlation between roadbase width and pavement width exists in the sample, resulting from the design practices, but can be remedied by using a new definition of so-called bundles. Other variables that are highly correlated in their own nature and make up a factor in the factor analysis were investigated, e.g., Variables 11 thru 16, 17 thru 20, and 21 and 24. Only one variable was chosen from each factor and was eligible for entering the model later. Finally, some variables that are of similar nature and measuring the same effect, i.e., culverts and pipes shorter than the pavement width and signs and lightings, respectively, were grouped together. The definition of roadway width bundles and the resulting categories in the explanatory variables are shown in Tables 2 and 3, respectively.

Table 2 - Definition of Roadway Width Bundles

Roadway Width Bundle (NEW1) Category	Roadbase Width (V1)	Pavement Width (V2)
1	6.4- 9.0m	6.4- 8.0m
2	9.0-10.5m	8.0-10.5m
3	10.5-12.5m	10.5-12.5m
4	12.5-15.0m	12.5-15.0m
5	9.0-12.5m	6.4-10.5m
6	12.5-15.0m	8.0-12.5m

The AID Analysis

AID was first applied to the data using the variable codes of Table 3. Because fourteen potential variables, each ranging from 2

Table 3 - Definition of Categorzied Variables

Variable	Definition if New	No. of Categories	Range Coding
NEW1	V1 & V2	6	See Table 2
V3	–	6	1=0-5 2=6-10 3=11-20 4=21-25 5=26-40 6=41-1000
NEW4	V4+V5	3	1=0-1 2=2-3 3=4-7
V6	–	2	1=1 2=2-26
V7	–	6	1=0-25 2=26-90 3=91-130 4=131-450 5=451-700 6=701-2798
V8	–	6	1=0-280 2=281-800 3=801-1300 4=1301-1850 5=1851-2800 6=2801-4869
NEW9	V9+V10	4	1=0-20 2=21-30 3=31-50 4=51-238
V16	–	2	1=0-30 2=31-1310
V20	–	3	1=0-50 2=51-100 3=101-691
V21	–	3	1=2054-5400 2=5401-11100 3=11101-14729
V22	–	6	1=14-20 2=21-30 3=31-40 4=41-50 5=51-60 6=61-79
V23	–	6	1=3-5 2=6-10 3=11-15 4=16-20 5=21-30 6=31-45
V25	–	3	1=level 2=rolling 3=mountainous
V26	–	4	1=30 2=40 3=50 4=60

to 6 categories, were involved, the output would become voluminous. The AID branching was truncated at the point where the minimum subgroup size of 5 or the reducibility criterion of 0.6% (in BSS/TSS) was not met. This AID run explains 40.86% of the variance. Figure 1 shows a partial description of the AID tree diagram that emerged from this stage of the analysis.

Note that the sample is first split (at level 1) on the variable of roadway width bundles. In the category of the worst design standards of 2-lane rural roads, pavement width between 6.4-8 meters and no lateral clearance, the major contribution to variance explanation is from splits based on the length of roadside ditches and on the terrain. In general, the worst-designed roads on level terrain (Group 22) have significantly more accidents while those in rolling or mountainous terrain (Group 23) or those with longer roadside ditches (Group 21) experienced fewer accidents.

In the category of better-designed roads, most of them on level terrain having wider pavement with/without lateral clearance, the important explanatory variables are traffic related, ADT (in p.c.u.'s) and % of motorcycles. In the category with mortorcycles consisting 50% or more of the traffic, the road sections with fewer signs and lightings (Group 10) or those on level terrain with longer guardrails as well as more signs and lightings (Group 19) are less accident-prone. On the other hand, those sections with shorter guardrails and lots of signs and lightings (Gourp 18) are more accident-prone. In the category with fewer motorcycles (less than 50%), the low ADT group (Group 6) and the middle ADT group on rolling terrain (Group 15) have fewer

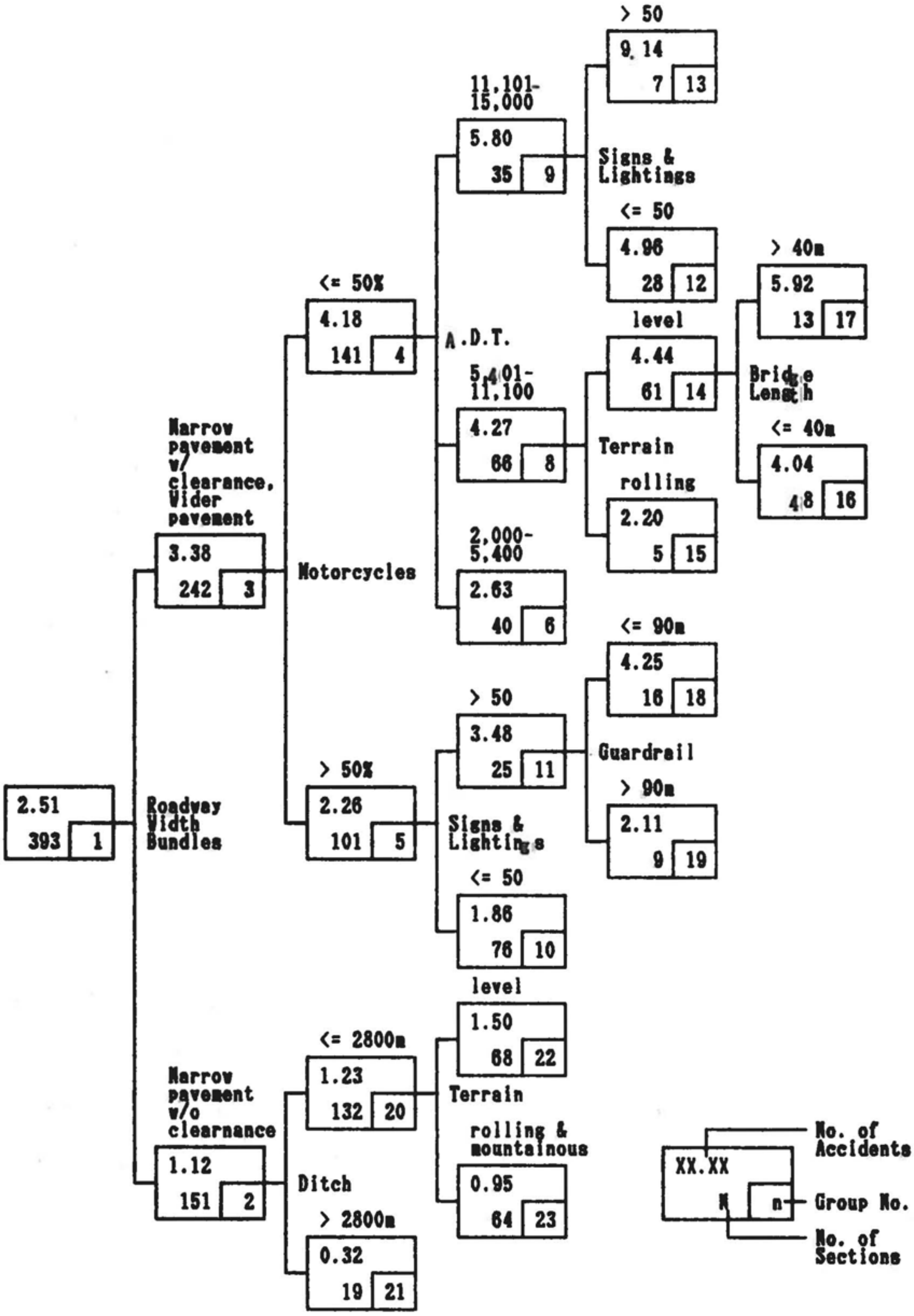


Figure 1 - AID Diagram of Total Number of Accidents

accidents. The middle ADT group on level terrain but with shorter bridge length (Group 16) are less accident-prone than those with longer bridge length (Group 17). In the high ADT groups, the sections with more signs and lightings (Group 13) are more accident-prone than those with fewer signs and lightings (Group 12).

From the above discussion, and the asymmetry in the tree diagram itself, it is obvious that complex interactive effects exist among several road and traffic descriptors on the accident occurrence. Other strong predictors, although failing to appear in the AID splits because of their strong correlation with others, were also retained in the data set for further analysis.

The Interaction Terms

Besides the tree itself, commonly used methods for displaying the AID results include tables showing the proportion of variation explanable by each predictor, tables of effect profiles, and the graph of effect profiles. The means profile chart is most useful for revealing the differential effects of a variable in various subgroups. If there appear to be major differences between profile lines, then the variable can be considered a candidate for inclusion in an interactive term.

The concept of congruence was applied in the analysis for locating the interactive variables and finding out the form the interaction takes. Variables were ordered in sequence by their explanatory power or theoretical importance, and the differential effect profiles of each variable in various subgroups formed by major AID plits as well as in the total sample were plotted. Figure 2, as well as Table 4, shows the effect of variable NEW1 (roadway width bundles) in groups 4, 5, 8 and 9 and in the total sample. The lines associated with subgroups 8 and 9 (also subgroup 6 not shown) and their parent group 4 are not parallel. The major split variable was ADT, which could be susceptibe to the effect of variable NEW1. The interactive effect between these two was then tested using an ANOVA and turned out significant at 0.005 level. Other similar, statistically significant 2-way interactive effects include those between ADT and number of intersections and between ADT and length of bridges. Having the largest explanltory power among the three, the interaction between ADT and roadway width bundles alone was considered for constructing a new term, to avoid too complex higher-order interaction terms.

The process of combining the variables of ADT and roadway width bundles was aided by the AID splits and the cross-classification means table so that it would not result in too many empty cells. Category 1 (narrow pavement with no lateral clearance) and Categories 5 and 6 (wide pavement with sufficient lateral clearance) of the roadway width bundles, respectively, are somewhat homogeneous and were considered independently with the ADT. The rest of the categories (medium or wide pavement with no lateral clearance) was classified by low ADT and medium and high ADT's. The definition of the resulting categories of the combined variable or interaction term is shown in Table 5:

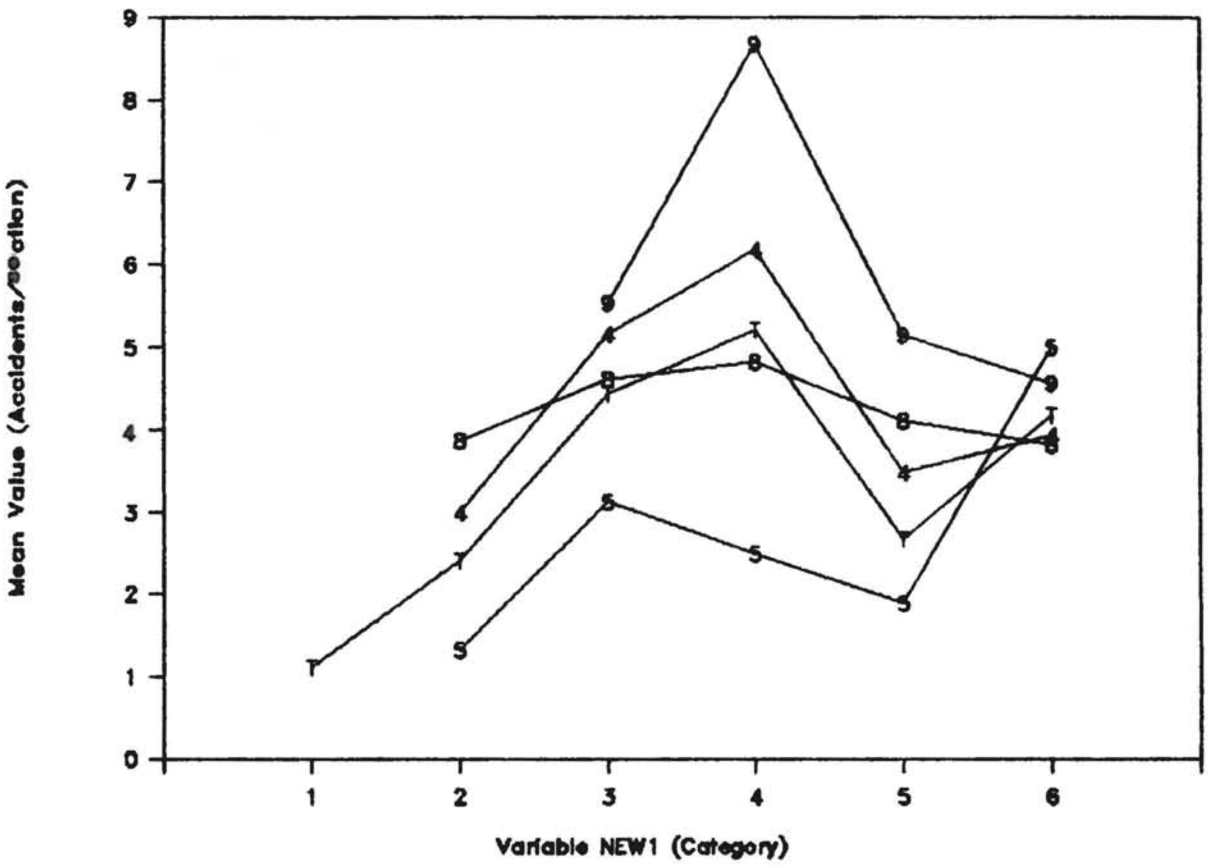


Figure 2 - Plot of the Effect of Variable NEW1 in Groups 4,5,8,9 and in the Total Sample

Table 4 - Mean Effect of Variable NEW1 in Groups 4,5,8,9 and in the Total Sample

Variable NEW1	Total		Group 4		Group 5		Group 8		Group 9	
Category	Size	Mean	Size	Mean	Size	Mean	Size	Mean	Size	Mean
1	151	1.12	-	-	-	-	-	-	-	-
2	17	2.41	11	3.00	6	1.33	8	3.88	-	-
3	47	4.43	30	5.17	17	3.12	13	4.62	15	5.53
4	23	5.22	17	6.18	6	2.50	11	4.82	6	8.67
5	132	2.67	65	3.48	67	1.90	28	4.11	7	5.15
6	23	4.17	18	3.94	5	5.00	6	3.83	7	4.57

Table 5 - Definition of Interaction Term Between Roadway Width Bundles and ADT

Interaction Term Category	Roadway Width Bundles Category	ADT Category
1	1	All
2	5,6	All
3	2,3,4	2,3
4	2,3,4	1

The MCA Model

The final stage of the analysis is to estimate the model using MCA. The model is of additive form with interactive variables of interest being replaced by combined variables (pattern variables). The data set manipulated previously was used as input to the statistical analysis package SAS for solving the normal equations used by MCA. The summary statistics printed by the program including the etas, betas, unadjusted and adjusted coefficients are listed in Table 6.

The MCA model thus constructed explains approximately 30% of the total variance, a moderately predictive system. The interaction term involving roadway width bundles and ADT's explains almost half, 15%, followed by percentage of motorcycles, 8%. Other significant variables, e.g., signs and lightings, terrain, and guardrail length explain between 1% and 3% of the variance. The variables insignificant by the F-test at 0.05 level, but having strong correlation with the significant ones, are retained in the model. The adjusted coefficients measure the predictive power of one variable by holding all other predictors, i.e., all other

Table 6 - Summary Statistics of the MCA Model

Variable/Category	N	Unadjusted Deviation	eta	Adjusted Deviation	beta
NEW1V21 (Roadway Width & ADT Bundles)					
1	151	-1.35		-1.01	
2	155	0.39		0.26	
3	70	2.12		1.68	
4	17	-1.93		-1.70	
			0.50		0.39
V22 (% Motorcycles)					
1	16	0.80		0.62	
2	31	-0.64		-0.38	
3	69	1.07		1.18	
4	123	0.24		0.28	
5	100	-0.70		-0.98	
6	54	-0.47		-0.29	
			0.25		0.28
NEW9 (Signs & Lightings)					
1	197	-0.53		-0.19	
2	69	0.30		0.14	
3	72	0.02		-0.39	
4	55	1.51		1.01	
			0.26		0.16
V7 (Guardrail Length)					
1	138	0.44		0.14	
2	37	0.38		0.37	
3	23	0.52		0.48	
4	74	0.02		0.05	
5	40	-0.91		-0.66	
6	81	-0.63		-0.26	
			0.19		0.11
V25 (Terrain)					
1	279	0.47		0.20	
2	77	-0.91		-0.38	
3	37	-1.82		-0.71	
			0.28		0.12
V8 (Ditch Length)**					
1	58	0.20		-0.14	
2	79	-0.18		0.16	
3	60	-0.01		0.13	
4	59	0.10		-0.14	
5	82	-0.04		0.14	
6	55	0.02		-0.29	
			0.04		0.06
V3 (Bridge Length)**					
1	256	-0.19		-0.06	
2	20	0.94		0.75	
3	32	0.08		0.08	
4	10	-0.11		-0.24	
5	15	0.29		-0.57	
6	60	0.39		0.15	
			0.12		0.08

Table 6 - Continued

Variable/Category	N	Unadjusted Deviation	eta	Adjusted Deviation	beta
V20 (Length w/ Radius <= 60m)**					
1	355	0.10		-0.03	
2	15	-1.24		0.12	
3	23	-0.77		0.43	
			0.12		0.04
V16 (Length w/ Grade 5-12%)**					
1	377	0.23		0.06	
2	56	-1.37		-0.38	
			0.21		0.06

Grand mean = 2.51 accidents/section

$R^2 = 0.33$; R^2 adj = 0.27

$F = 5.74$; $F^*(41,361,0.05) = 1.35$

"**" - Nonsignificant by approximate F-test at 0.05 level

predictors are assumed distributed as they are in population at large. To obtain the average number of accidents on a particular road segment, one simply add the adjusted coefficients of membership in certain categories to the grand mean. The main effects of individual categories within each variable are summarized as follows:

1. The interactive effects of roadbase width, pavement width, and ADT are quite complex. The segments with narrow pavement and no lateral clearance and those with wider pavement and no lateral clearance but having lower ADT's have the lowest accident counts. The segments with wider pavement and no lateral clearance but having higher ADT's have the highest accident counts. Obviously, ADT is still the most dominant factor in accident occurrence.

2. For the effect of motorcycles, more accidents occurred in the range of 31-40% while fewer in the range of 51-60%. Besides other traffic and road conditions, this may well be explained by the degree of disturbance versus the degree of homogeneity in the traffic stream.

3. The effect of total number of signs and lightings seems somewhat contradictory. The sections with more signs and lightings have more accidents. The existence of these devices may imply somewhat complex traffic and environmental conditions, their effects not being captured by other variables.

4. The effect of guardrail length may seems contradictory as well. The sections with shorter guardrails have experienced more accidents. This may better be explained by relating guardrails with terrain. The sections on level terrain are less guardrail-dependent; they are characterized by more accidents associated with wider pavement having higher ADT's.

5. The effect of ditch length should also be investigated along with terrain. The sections on rolling or mountainous terrain accompanied by longer ditches are generally associated with lower design standards and lower ADT's. Fewer accidents occurred on these sections.

6. The effect of bridge length is not montonic. More accidents occurred on sections in the middle range of bridge length while fewer in the high range.

7. The effects of curve length and grade length are somewhat different. The sections with more length on curves are more accident-prone while those with more length on steep grades are less accident-prone.

CONCLUSIONS AND RECOMMENDATIONS

This study was concerned with accident occurrence on two-lane rural highways and its relationship to traffic and road and environmental conditions. A national data set of two-lane rural accident experience, involving 393 three-kilometer road sections with ADT between 2,000 and 15,000 p.c.u.'s which recorded 987 accidents in 2-year period, was studied. Within the data set, the continuous variables were first categorized, followed by the grouping of intercorrelated geometric or operational variables into bundles, or into factors to be represented by single variables. A descriptive model was then constructed by AID technique for revealing the general pattern of interactions. With the aid of the AID analysis, a series of means profile charts were generated; the variables showing significant interactive effects by the ANOVA were candidates for combination. Finally, an explanatory MCA model was constructed with parameters to show the importance of individual variables, including the interaction terms which have replaced the raw variables.

The most important findings from this research are viewed as follows:

1. Strong interactive effects exist among the road and traffic descriptors that simple models based on original variables will not suffice for the accident prediction. This necessitates the use of many combinations of variables, as bundles or interaction terms, in effective modeling.

2. The joint use of AID/MCA techniques allows each to supplement the other's limitations. The AID provides some insight into the relative importance of individual variables and their complex interactive effects. The information on which predictors, and which categories within predictor, to include in the MCA analysis is also very useful. The MCA model having explicit parameterization and appropriate significance tests should check with the AID results. Nevertheless, some important variables not appearing in the AID splits should not be ignored in the MCA analysis; failing to include correlated variables generally leads to less predictive power for those included.

3. The analysis uses section-length exposure rate rather than the conventional vehicle-mile exposure rate to permit the ADT to be treated as a classification or an independent variable. The results show that for the worst-designed sections, frequently associated with lower ADT's, the terrain-associated variables serve as a proxy for the ADT.

For better-designed sections, the traffic-related variables show much stronger effects; the terrain related variables are not as strong as previously. The variable of signs and lightings seems to be a proxy for the complexity of road and environmental conditons not captured by other variables.

4. The constructed MCA model explains about 30% of the total variance in the dependent variable, having moderately predictive power. The adjusted coefficients show that the interaction term of the roadway width-ADT bundles has the strongest effect on accident occurrence, followed by % motorcycles, signs and lightings, terrain, and guardrail length. By adding the effects of membership in certain categories to the grand mean, one can predict the number of accidents on a road section of interest. Such a simple additive model can be very useful for engineers in determining the location and magnitude of safety improvements.

5. The analysis has illustrated the danger in basing decisions to improve a given element on simple comparisons when it really is the joint effect of the differences in several such elements that is responsible for observed accident differences.

Beyond the procedures and findings summarized, several recommendations are made for further studies. As high-quality data files with many more accidents become available, this study should be repeated to test and refine the conclusions that were found in this research. To reduce the skewness in the dependent variable (accidents/section) for more effective modeling, it is usually achieved by increasing the length of sections or study period. Both suffer the problem of changes in traffic and/or road and environmental conditions; the optimum combination of length and perid should be studied. An option is to vary the length of sections having homogeneous physical and operational characteristics. Attention should also be paid to the development of more concrete, theoretically sound procedures for categorizing continuous variables. For the search procedure of identifying interactions, alternative approaches such as using the creterion of dependency between dependent variable and each of the predictors, rather than variance explanation of predictors, suggested by Perreault and Barksdale (1980) should be implemented. Their procedure also has the feature of pairwise merging, and then separating, of the response levels on each of the predictors to determine the smallest number of groupings. As for the final explanatory model, several alternatives are available, including log-linear models. In all, furthering the knowledge in the construct of accident occurrences and models would significantly improve the evaluation process of the highway safety programs.

ACKNOWLEDGEMENT

The work reported here was conducted as a university research project under the sponsorship of National Science Council, Republic of China.

REFERENCES

1. Andrews, F.J., J.N. Morgan, and J.A. Sonquist (1967), Multiple Classification Analysis, Survey Research Center, Institute for Social Research, University of Michigan, Ann Arbor, Michigan.
2. Armstrong, J.S. and J.G. Andress (1970), "Exploratory Analysis of Marketing Data: Trees vs. Regression", Journal of Marketing Research 7, pp. 487-492.
3. Assael, H. (1970), "Segmenting Market by Group Purchasing Behavior: An Application of the AID Techniques", Journal of Marketing Research 7, pp. 153-158.
4. Baldwin, D.M. (1946), "The Relationship of Highway Design to Traffic Accident Experience", Papers and Discussions, AASHO 32nd Annual Convention, pp. 103-109.
5. Billion, C.E. and W.R. Stopher (1957), "A Detailed Study of Accidents as Related to Highway Shoulders in New York State", Proceedings 36, HRB, National Research Council, Washington, D.C., pp. 497-508.
6. Blackburn, R.R., D.W. Harwood, A.D. St. John, and M.C. Sharp (1978), Effectiveness of Alternative Skid Reduction Measures - Volume 1: Evaluation of Accident Rate-Skid Number Relationships, Midwest Research Institute, Kansas City, Missouri.

7. Cleveland, D.E. and R. Kitamura (1978), "Macroscopic Modeling of Two-Lane Rural Roadside Accidents", Transportation Research Record 681, TRB, National Research Council, Washington, D.C., pp. 53-62.

8. Cleveland, D.E., L.P. Kostyniuk, K.L. Ting, C.P. Green, T. Chirachavala, P. Landau, K. Riegel, and M. Pohlman (1981), Advanced Accident Causal Models, Department of Civil Engineering, University of Michigan, Ann Arbor, Michigan.

9. Cleveland, D.E., L.P. Kostyniuk, and K.L. Ting (1984), "Geometric Design Element Groups and High-Volume Two-Lane Rural Highway Safety", Transportation Research Record 960, TRB, National Research Council, Washington, D.C., pp. 1-13.

10. Cleveland, D.E., L.P. Kostyniuk, and K.L. Ting (1985), "Design and Safety on Moderate-Volume Two-Lane Roads", Transportation Research Record 1026, TRB, National Research Council, Washington, D.C., pp. 51-61.

11. Dart, O.K., Jr. and L. Mann, Jr. (1970), "Relationship of Rural Highway Geometry to Accident Rates in Louisiana", Highway Research Record 312, HRB, National Research Council, Washington, D.C., pp. 1-16.

12. Doyle, P. and I. Fenwick (1975), "The Pitfalls of AID Analysis", Journal of Marketing Research 12, pp. 408-413.

13. Graham, J.L. and D.W. Harwood (1982), "Effectiveness of Clear Recovery Zones", NCHRP Report 247, TRB, National Research Council, Washington, D.C., 68 pp.

14. Green, P.E. (1978), "An AID/Logit Procedure for Analyzing Large Multiway Contingency Tables", Journal of Marketing Research 15, pp. 132-136.

15. Gupta, R.C. and R. Jain (1973), Effect of Certain Geometric Design Characteristics of Highways on Accident Rates for Two-Lane Roads in Connecticut, Department of Civil Engineering, University of Connecticut, Storrs, Connecticut.

16. Jara-Diaz and Gonzalez (1986), "Flexible Models for Accidents on Chilean Roads", Accident Analysis and Prevention 18, No. 2, pp. 103-108.

17. Kihlberg, J.K. and K.J. Tharp (1968), Accident Rates as Related to Design Elements of Rural Highways, NCHRP Report 47, HRB, National Research Council, Washington, D.C., 173 pp.

18. McBean, P.A. (1982), The Influence of Road Geometry at a Sample of Accident Sites, TRRL Report 1063, Transport and Road Research Laboratory, Crowthorne, Berkshire, England.

19. Morgan, J.N. and J.A. Sonquist (1963), "Problems in the Analysis of Survey Data and a Proposal", Journal of the American Statistical Association 58, pp. 415-434.

20. Newman, J.W. and R. Staelin (1971), "Multivariate Analysis of Differences in Buyer Decision Time", Journal of Marketing Research 8, pp. 192-198.

21. Perkins, E.T. (1956), "Relationship of Accident Rate to Highway Shoulder Width", Bulletin 151, HRB, National Research Council, Washington, D.C., pp. 13-14.

22. Perreault, W.R., Jr. and H.C. Barksdale, Jr. (1980), "A Model-Free Approach for Analysis of Complex Contingency Data in Survey Research", Journal of Marketing Research 17, pp. 503-515.

23. Polus, A. (1980), "The Relationship of Overall Geometric Characteristics to the Safety Level of Rural Highways", Traffic Quarterly 34, No. 4, pp. 575-585.

24. Rinde, E.A. (1977), Accident Rate Versus Shoulder Width, California Department of Transportation, Sacramento, California.

25. Roy Jorgensen and Associates, Inc. (1978), Cost and Safety Effectiveness of Highway Design Elements, NCHRP Report 197, TRB, National Research Council, Washington, D.C., 237 pp.

26. Schoppert, D.W. (1957), "Predicting Traffic Accidents from Roadway Elements of Rural Two-Lane Highways with Gravel Shoulders", Bulleting 158, HRB, National Research Council, Washington, D.C., pp. 4-18.

27. Shannon, P. and A. Stanley (1978), "Pavement Width Standards for Rural Two-Lane Highways", Transportation Research Record 685, TRB, National Research Council, Washington, D.C., pp. 20-23.

28. Snyder, J.C. (1974), "Environmental Determinants of Traffic Accidents: An Alternate Model", Tranportation Research Record 486, TRB, National Research Council, Washington, D.C., pp. 11-18.

29. Sonquist, J.A. (1970), Multivariate Model Building - The Validation of a Search Strategy, Survey Research Center, Institute for Social Research, University of Michigan, Ann Arbor, Michigan.

30. Sonquist, J.A. and J.N. Morgan (1964), The Detection of Interaction Effects, Survey Research Center Monograph No. 35, Institute for Social Research, University of Michigan, Ann Arbor, Michigan.

31. Sonquist, J.A., E.L. Baker, and J.N. Morgan (1971), Searching for Structures, Survey Research Center, Institute for Social Research, University of Michigan, Ann Arbor, Michigan.

32. Versace, J. (1960), "Factor Analysis of Roadway and Accident Data", Bulletin 240, HRB, National Reserach Council, Washington, D.C., pp. 24-32.

33. Zegeer, C.V. and J.G. Mayes (1979), Cost Effectiveness of Lane and Shoulder Widening of Rural Two-Lane Roads in Kentucky, Research Report 524, Division of Research, Bureau of Highways, Kentucky Department of Transportation, Lexington, Kentucky.

34. Zegeer, C.V., R.C. Deen, and J.G. Mayes (1981), "Effect of Lane and Shoulder Widths on Accident Reductions on Rural, Two-Lane Roads", Transportation Research Record 806, TRB, National Research Council, Washington, D.C., pp. 33-43.

Risto Kulmala & Matti Roine
Technical Research Centre of Finland

ACCIDENT PREDICTION MODELS FOR TWO-LANE ROADS IN FINLAND

BACKGROUND

We developed models for describing the safety of two-lane main highways outside urban areas mainly for prediction purposes and to locate hazardous road sections. In the second phase we developed the model further to enable us to evaluate the safety effects of different road characteristics, and to provide the road authorities with a tool for road planning. The model applied to road sections outside junctions. The work was commissioned by the Roads and Waterways Administration.

Our study material consisted of 4857 road sections on two-lane highways outside urban areas with a total of 15492 police-reported accidents in the years 1981 - 1986. 4208 of these accidents resulted in death or injury. These sections were formed so that certain road characteristics, such as road width, speed limit etc., remained constant throughout the section.

As the coverage of accident statistics varies between 10 and 90 % depending on the type of accidents and the part of the country, we decided to concentrate on fatal and injury accidents. The coverage of accident statistics based on police reports is ca. 70 % for these accidents, and it does not vary considerably according to accident types.

ACCIDENT MODELS

Basic models

Our goal was to produce accident models that would explain the accident occurrence on two-lane highways outside built-up areas. These models should be based on the statistical data available for the road authorities. The data consisted of road sections with speed limit 80 and 100 km/h. Only road sections on paved main roads and sections with no major road improvements during 1978 - 1983 were included in the analysis.

Homogeneous road sections were formed. A new road section was introduced when speed limit, the width of pavement and pavement material changed, pedestrian and bicycle way started or ended and road lighting started. Each homogeneous road section formed a record with data on accidents, traffic and road geometry.

Models were based on the theory of generalized linear models. These models are extensions of classical linear models and consist of tripartite form: random component, systematic component and the link between the random and systematic component. We regarded that the error distribution was Poisson because our purpose was to explain accident occurrence. The systematic component of the models was to describe the way that the expected count of accidents were related to the independent explanatory variables. With Poisson error distribution we used the log link function.

Our models consisted of six different models, two speed limit classes (80 and 100 km/h) and three traffic volume classes (ADT: under 1500, 1500 - 3000, above 3000 motor-vehicles/day). The standard model formula was:

$$A = k * S^a * \exp(\Sigma b_i * x_i)$$

, where

A = fatal and injury accidents in 1978 - 1983
S = mileage
x_i = variables
κ, a, b are coefficients to be estimated.

Models were estimated with the GLIM-package and we used the Scaled Deviance (SD) for significance testing. The Scaled Deviance is:

$$SD = -2 * (\log(\max L) - \log(\max L_f))$$

where

$\log(\max L)$ = maximized log-likelihood for the current model
$\log(\max L_f)$ = maximized log-likelihood for the full model

The best explanatory variables were taken into the models after fitting mileage as the measure of exposure (table below). We noticed that all the models included both the width of pavement and the passing sight distance > 300 meters (%) describing the effect of road geometry. The percentage of heavy vehicles, lorries and buses, turned out to be an important additional traffic variable on road sections with ADT less than 3000.

Speed limit	Average Daily Traffic (ADT)	Modeltype	SD /d.f
80	< 1500	S + L + N + K.RP	1.122
	1500 - 3000	S + L + N + RP	1.177
	> 3000	S + L + N	1.302
100	< 1500	S + L + N + R + L.K	1.167
	1500 - 3000	S + L + N + K.RP	1.137
	> 3000	S + L + N + K	1.567

The modeltype in the table above describes the variables that were included in the models. The K.RP formula if not preceded by K and R is interpretated as "RP. within K" and means nesting. The variables in the models are:

S = mileage (continuous)
L = pavement width (< 7,5, 7,5 - 8,5, 8,6 - 9,5, >9,6 m)
N = passing sight distance > 300 m (%)
K = average curvature (different classes)
R = percentage of heavy vehicles (continuous)
RP = percentage of heavy vehicles (classified)

Development of basic models

The concept of basic models was to indicate the best explanatory variables and dependencies between accident frequency and variables. The models were not aimed at countermeasure effect analysis. Therefore, we made some further analysis to get accident models for prediction of effects of safety. We used the latest accident and traffic data (period 1981 - 1986).

The data had to be homogenized so that there would be comparable data sets for most of the alterations of the variables. We left out all the road sections that were considered to be in built-up areas, all minor roads in the northermost part of Finland because of the under reporting of accidents and some very deviant road sections in southern Finland. After several analyses we ended up with a single model with the necessary variables that can be used in road and safety policy planning.

For the development of this model we used data from 2730 accidents. The model was in close agreement with the data, the Scaled Deviance is 3040 with 2720 d.f (degrees of freedom). The mean-squared-error of the new model is even smaller than the MSE of the six previous models.

The new model is:

$$A = 0{,}1377 * S^{0{,}9767} * \exp(\Sigma\, b_i * x_i)$$

where

A = fatal and injury accidents in 1981 - 1986
S = mileage

exp():

- 0,4581	* L2	(1, if pavement width 8,6-9,5 m, else 0)	
- 0,1555	* L2	(1, if pavement width >9,5 m, else 0	)
- 0,005455	* N	(passing sight distance >300 m (%)	)
+ 0,009096	* RP	(percentage of heavy vehicles	)
+ 0,001331	* K	(average curvature	)
+ 0,05874	* LR	(1, if pavement width < 8,6 m and speed limit 100 km/h)	
+ 0,3564	* LR	(1, if pavement width 8,6-9,5 m and speed limit 100 km/h)	
+ 0,2179	* LR	(1, if pavement width > 9,6 m and speed limit 100 km/h)	

In the model, the expected number of accidents depends on mileage, pavement width, passing sight distance, percentage of heavy vehicles, curvature and speed limit. The expected number of accidents on the road sections is directly proportional to mileage (exposure), power of mileage is almost 1,00.

When the effect of the other variables is omitted the accident risk is lowest if pavement width is 8,6 - 9,5 m. Passing sight distance percentage has a remarkable effect on accident risk, risk decreases with improving road geometry. Heavy vehicles affect overtaking and seems to increase accident risk on road sections.

The model predicts that higher speed limit raises the accident risk. The effect of speed limit depends on pavement width. When speed limit is changed from 80 to 100 kmph, the risk increases 6 % if the pavement width is < 8,6 m, 42 % if the pavement width is 8,6 - 9,5 m, and 24 % if the pavement width is > 9,6 m.

This model can be used for evaluation of effects of road improvements if the effect on variables in the model is calculated. We have also an interactive PC-program based on the model above that predicts safety effects of designed road improvements.

THE STABILITY OF ACCIDENT COUNTS

Various methods to estimate the expected number of accidents were tested. The accident data of of the road sections was divided into two populations, the first period 1978 - 1981 and the second 1981 - 1983. Road sections longer than 10 kilometers were excluded so that the study material consisted of 3696 road sections on two lane highways outside urban areas. The data contained 1951 fatal and injury accidents in the first period and 1834 in the second period. The reported number of accidents was thus ca. 6 % lower during the second period.

We used the Poisson probability function for the accident frequency of a single entity and the Gamma function for the populations of studied entities (see later: comparison of models). If the assumptions are reliable the negative binomial distribution reflects the number of accidents on entities of a real population. The results are presented in the table below. We concluded that the model describes very well the occurrence of accidents in the two populations of entities. Because of the definition of an entity it is natural that there exists variation and the expected number of accidents differs between the populations. Later on we made some further analysis of this variation.

Accidents per Section (x)	Number of entities having x accidents Actual 78-81	Neg.Bin. 78-80	Actual 81-83	Neg.Bin. 81-83
0	2605	2631	2598	2618
1	644	609	682	652
2	245	242	245	244
3	107	109	100	101
4	44	52	34	44
5	25	26	19	20
6	11	13	7	9
7	6	7	4	4
8	3	4	3	2
9	1	2	3	1
10	0	1	1	0
11	0	1	0	0

Our problem is usually two-fold. We do not know exactly the expected number of accidents on entities in the past without analyzing accident data. The accident history of entities has very often been used as a direct estimate for future counts of accidents. Latest research results indicate that this belief may also be erroneous.

We have used the Poisson and Gamma function assumptions when producing estimates for the expected count of accidents on entities. As Hauer et.al have shown, the Gamma distribution can be estimated as follows:

$$a = x / (s^2 - x)$$
$$b = x^2 / (s^2 - x)$$

Where x (mean) and s^2 (variance) depend on n(x), the number of entities with x accidents:

$$x = \Sigma\, x * n(x) / n$$
$$s^2 = \Sigma\, (x - x)^2 * n(x) / n$$

The variance of the expected number of accidents (m) depends on the reported accidents and is smaller than Var(x), if the m's are not equal in the population:

$$Var(m) = Var(x) - E(m) = s^2 - x$$

It has been shown that the estimator T_1 minimizes $E((T - m)^2$. We assume that $p(x) = n(x) / n$ where n is the total amount of entities.

$$T_1 = (x + 1) * p(x + 1) / p(x)$$

The variance of T_1 can be estimated by:

$$Var(T_1) = T_1^2 * ((1 / n(x+1) + (1 / n(x))$$

The variance of estimates depends on the number of entities and accidents. Smoothened estimates are produced by fitting a weighted regression curve through the points of estimates T_1.

We can get the third estimate for the expected amount of accidents in the populatition of entities using the equation /Hauer/:

$$T_2 = x + (E(x) / Var(x)) * (E(x) - x)$$

The average number of accidents in 78 - 80 was 0,528, variance 1,204, estimated a = 0,781 , b = 0,412. The estimates T_2 can be calculated by the model:

$$T_2 = x + 0,4385 * (0,5279 - x)$$

The weight in the curve fitting was inversely proportional to the points variation with the largest point having a weight of 1. We got the model:

$$T_3 = 0,24 + 0,514 * T_1$$

The R^2 of the model is about 0,99, so the fit is good. We concluded that the estimates T_3 are not much better than T_1:s (table below). All the calculated estimates are undoubtedly better than the number of reported accidents on various entities, and quite free from the regression-to-the-mean effect.

Accidents per section 78 - 80	Average of accidents 81 - 83	Estimates T_1	$Var(T_1)$	T_2	T_3
0	0,29	0,25	0,0001	0,23	0,25
1	0,71	0,76	0,0033	0,79	0,76
2	0,94	1,31	0,0231	1,35	1,28
3	1,73	1,64	0,0868	1,92	1,79
4	1,45	2,84	0,5063	2,48	2,31
5	2,36	2,64	0,9124	3,04	2,82
7	3,83	4,00	8,0000	4,16	3,85

When studying the number of accidents during the time-periods, it seems that there exists a trend in the development of safety. This trend should also be considered, because it affects the m:s (safety). Firstly, we have assumed that the expected number of accidents per unit of exposure remains unchanged. An estimate for the expected amount of accidents and the variance per entity during the second period is then:

$$E(m_2) = (e2 / e1) * E(m_1)$$

$$Var(m_2) = (e2 / e1)^2 * E(m_1)$$

However, our data pointed out that this estimate for the reduction of the variance was not very accurate. It is possible that the safety improvement is more concentrated on the risky road sections. We assumed here that the reduction is proportional to the amount of accidents on entities and the average number of accidents equals the average during the second time period (est2). The calculated two estimates are presented in the next table.

Accidents per Section (x)	Number of Actual 78-81	entities Actual 81-83	having x Neg.Bin. est1	accidents Neg.Bin. est2
0	2605	2598	2668	2608
1	644	682	601	662
2	245	245	232	246
3	107	100	102	101
4	44	34	47	44
5	25	19	23	19
6	11	7	11	9
7	6	4	6	4
8	3	3	3	2
9	1	3	2	1
10	0	1	1	0
11	0	0	0	0

It is possible to estimate the distributions of m:s within the groups of entities. The Gamma probability function is then:

$$f(m/x) = (1+a)^{(x+b)} * m^{(x+b-1)} * e^{-m(1+a)} / g(b)$$

The expected number of accidents on entities in 1981-1983 can then be calculated using the data from the first period and the conditional Gamma distribution. We have presented both estimates (est1 and est2) in the table below.

The calculations indicate that the marginal estimates (totals 1981-1983) are slightly better if the additional safety benefit is estimated. However, the differences according to the conditional Gamma distributions are insignificant.

Accidents per section 78-80		Number of entities having x accidents during 81-83									
		0	1	2	3	4	5	6	7	8	9
0	est1	2187	312	76	21	6	2	1	0	0	0
	est2	2118	361	90	25	7	2	1	0	0	0
	data	2019	448	102	32	4	0	0	0	0	0
1	est1	354	173	72	28	11	4	2	1	0	0
	est2	352	176	73	28	10	4	1	1	0	0
	data	422	104	66	37	14	8	3	0	0	0
2	est1	88	73	43	22	10	5	2	1	0	0
	est2	90	74	43	21	10	4	2	1	0	0
	data	112	73	37	18	2	0	1	1	0	1
3	est1	25	30	23	14	8	4	2	1	0	0
	est2	26	31	23	14	7	4	2	1	0	0
	data	25	34	25	8	3	1	1	1	0	0
4	est1	7	10	10	7	5	3	1	1	0	0
	est2	7	11	10	7	4	2	1	1	0	0
	data	12	17	8	2	2	2	0	1	0	0
5	est1	3	5	5	5	3	2	1	1	0	0
	est2	3	5	5	4	3	2	1	1	0	0
	data	5	3	4	8	2	2	1	0	0	0

A COMPARISON BETWEEN DIFFERENT PROBABILITY MODELS

Assumptions

We studied the accidents by using the following assumptions /Hauer/:

The PDF (probability density function) of accidents for a single entity (junction, road section etc. in a specified period) follows the Poisson distribution if the expected number of accidents m is fixed. If the m's of the population of entities varies with a PDF of G(m), where G(m) is assumed to be of a two-parameter Gamma family, the PDF of accident counts in the population is the negative binomial distribution.

The mean and variance of the fatal and injury accidents were in our material: mean $X = 0.866$
variance $S = 1.470$

If the PDF of accident counts in the population would be Poisson, the mean would equal the variance. This is clearly not the case. If this is a result of varying expected accident counts in the population i.e. varying m:s and the Gamma assumption above is correct, the probability density of m:s is:

$f(m) = a^b m^{b-1} e^{-am}/g(b)$,

where g(b) is the value of the one-parameter Gamma function at point b.

The parameters a and b can be estimated from the data /Hauer/:

$a = X / (S - X)$
$b = X^2 / (S - X)$

The probability of an entity in the population to have x accidents is:

$P(x) = (a/(a+1))^b \ (b(b+1)\ldots(b+x-1))/((a+1)^x x!)$,

which is the negative binomial distribution.

<u>Comparison</u>

In our data, a = 1.435 and b = 1.243. The table below lists the actual accident counts, and the expected counts on the basis of negative binomial distribution, and Poisson distribution (m = 0.866).

Also this table shows that the Poisson model does not correspond to the data very well. This is not very surprising as the Poisson model assumes each section to have the same expected number of accidents. The negative binomial model, however, is in close agreement with the data.

Accidents per Section (x)	Number of entities having x accidents Actual Data	Neg. Binomial Model	Poisson Model
0	2528	2517	2042
1	1268	1285	1769
2	592	592	766
3	265	263	221
4	114	114	48
5	56	49	8
6	23	21	1
7	7	9	0
8	2	4	0
9	0	2	0
10	0	1	0
11	2	0	0

This shows that the m's really vary in the population. But is it also a question of varying safety from the point of view of e.g. a single road user or a traffic engineer?

Accident risks and risk exposure

Accident risks are usually used as a measure of traffic safety, and expressed in the form number of accidents/ exposure. For road sections accident risk is traditionally calculated as the ratio between the number of accidents and vehicle mileage, and called accident rate. The expected number of accidents (m) can thus be expressed as a product between the expected accident rate (R) and vehicle mileage: m = R x mileage. The accident models presented elsewhere in this paper show that the number of accidents is indeed approximately proportional to vehicle mileage.

To estimate the effects of different road characteristics on safety, or to predict the number of accidents, we are always interested in the accident rates, as we usually have reasonably accurate information on vehicle mileage, and its changes. The question is now: in which way do the R's vary in the population of road sections? To study this we divided the data in different categories on the basis of vehicle mileage. The classification interval was 1 million vehicle kilometers, and the mileage as well as accident data were from a period of 6 years. The mean and variance of the accident counts for each mileage class are shown below.

Mileage class (million veh.km)	Accidents on road sections		Number of road sections
	Mean	Variance	
1-2	0.1658	0.1604	550
2-3	0.2509	0.2810	562
3-4	0.3255	0.3489	513
4-5	0.4869	0.5655	382
5-6	0.6062	0.7508	353
6-7	0.7560	0.9241	250
7-8	0.7837	0.8424	245
8-9	0.7940	0.9017	199
9-10	0.9305	1.2692	187
10-11	0.9226	0.8900	155
11-12	1.1159	1.7472	164
12-13	1.1927	1.4904	109
13-14	1.2692	1.6246	130
14-15	1.5429	1.4813	105
15-16	1.3786	1.5122	103
16-17	1.6477	1.5643	88
17-18	1.5632	1.8303	87
18-19	1.7683	1.7852	82
19-20	1.9706	3.1336	68
20-21	1.4310	2.1092	58
21-22	1.9818	1.9441	55
22-23	1.9273	2.4762	55
23-24	2.5135	2.2011	37
24-25	2.3529	1.8731	51
25-26	2.3902	3.0440	41
26-27	2.2973	3.1036	37
27-28	2.4737	3.4451	38
28-29	2.8837	2.9149	43
29-30	2.8333	3.2472	30
30-31	2.6857	3.8691	35
31-32	2.7000	3.5276	30
32-33	2.6000	1.9715	15

The close connection between the number of accidents and mileage is evident in the table. The mean accident count approximately equals its variance in many mileage classes, and closer inspection of the accident data shows that the accident counts within these mileage classes follow the Poisson distribution. In the classes, where the variance is clearly larger than the mean, the negative binomial model fits better with the data than the Poisson model. Still, in most of these cases, the Poisson model does not differ significantly from the actual accident data.

The conclusion to be drawn from the table above is that the variance of the expected number of accidents in the total population is mainly due to the variance of mileage i.e. exposure instead of "safety" expressed as accident risk or rate. The accident rates seem also to vary, but in a smaller scale. A part of the variance of accident counts within mileage classes is naturally due to the variance of mileage, too. Still it is evident that there exist real safety differences in the population of Finnish road sections. Some of the differences were explained by our accident models as shown elsewhere in this paper.

On the basis of the study we stress the importance of accounting for the effect of exposure on accident counts. Otherwise conclusions drawn from the available accident data can often be misleading.

REFERENCES

Hauer, E., Lecture Notes, Workshop given at Highway Safety Research Center, University of North Carolina, Summer 1986.

Kallberg, V-P., Kulmala, R. & Roine, M., Pääteiden onnettomuuksien riippuvuus tie- ja liikenneteknisistä tekijöistä (Accident prediction models for two-lane roads in Finland). Espoo 1987. Technical Research Centre of Finland, Research Reports 488. 60 p. + app. 14 p.

DETERMINATION OF BLACK SPOTS. A COMPERATIVE AND CORRELATION STUDY OF EXISTING METHODS.

Tsohos G., Dr. Engineer, Associate Professor, University of Thessaloniki, Greece

Kokkalis A., Transportation Engineer, University of Birmingham, England.

ABSTRACT

It is well known, that the determination of black spots on a road network is of great importance for the optimization of traffic safety performance. Since a long time, various methods based on statistical theory have been presented to permit the engineers to locate hazardous sections on road networks. This paper evaluates the rationale of the most common existing methods, which can be used to ensure the identification of black spots. Comparison and correlation of the results each method yields, is also attempted.

Traffic accident data have been obtained from a research project on traffic safety held by Thessaloniki University. Concerning accident analysis on the national road network in Northern Greece. Four methods of black spot identification have been used :

a. Absolute number of traffic accidents
b. Use of Poisson's distribution
c. Traffic accident indices
d. Accident severity indices.

After the statistical analysis of approximately 2000 accidents, it has been concluded that :

a. Important differences exist on identifying black spots according to the above mentioned methods.
b. Poisson's distribution gives more optimistic results in comparison to traffic accident and accident severity indices.
c. Lamm's absolute number of accidents method correlates better with all other methods.
d. A combination of methods must be used to confirm the existence of a black spot.

DETERMINATION OF BLACK SPOTS. A COMPERATIVE AND CORRELATION STUDY OF EXISTING METHODS.

Tsohos G., Dr. Engineer, Associate Professor, University of Thessaloniki, Greece
Kokkalis A., Transportation Engineer, University of Birmingham, England.

INTRODUCTION

It is well known that traffic accidents consist a significant problem in modern societies with many social and economic consequences in either personal or national scale. The advent of motorvehicles, apart from its obvious numerous advantages, produced many serious problems, the most important of which is the road accidents. Throughout the world a significant number of people fall victims of road accidents creating serious personal or even social distress. Furthermore, the national economy of a country suffers consinderable losses as a result of accidents causing the killing of or injuries to people and the damaging of property.

The problem of traffic safety is very keen in Greece. Proportionally to the number of vehicles, in Greece occured twice as many road accidents as those occured in other Western European countries, during the last decade. However, highway accident statistics indicate that the annual number and rate of accidents is declined[5]. This, along with the fact that the annual vehicle-kilometers of travel have consinderably increased throughout the same period, gives an indication that positive gains are being achieved from recent safety efforts.

Generally, highway safety programs are aimed at reducing traffic accident fatalities, injuries and property damages attributable to highway system failures, as opposed to those attributed to vehicle or driver failures. An analysis of accidents on a road shows that in addition to a comperatively uniform distribution of accidents over the whole road's lenght, a considerable portion of them occur on relatively short sections , generally known as black spots or black kilometers. (depending on their length). The identification of these hazardous spots or sections in the road network, where traffic accidents tend to cluster and the proposal of certain remedial measures, is the most fruitful way of preventing accidents and enhancing roadway safety.

Quite a lot of methods exist for the identification of black spots, most of them based on statistical theory. The results that they yield vary considerably depending on the rationale and the methodology each one follows. The evaluation of the most well known of the existing methods as well as the comparison and correlation of the results they yield, is the subject of this report.

METHODS OF BLACK SPOT IDENTIFICATION

The four most commonly in use methods of black spot identification are :

a. The absolute number of traffic accidents
b. The use of Poisson's distribution
c. The traffic accidents indices
d. The accident severity indices.

A brief outline of these methods follows.

Absolute Number of Traffic Accidents

Using the absolute number of traffic accidents, an accident risk level can be assigned in each section of the road network in proportion to the actual number of accidents occuring there in each year. Then, the level which corresponds to a hazardous road section can be determined and subsequently each road section can be classified in relation to its accident risk level.

Babkov(2) considers a road section as a hazardous one, when 3 at least road accidents occur there every year, whilst Benner et al(3) consider this number to be 4. Lamm et al(4) divide the specific road in one kilometer long sections and classify them in order of increasing traffic accidents. The sections belonging to the upper 15% of the above series are considered as hazardous ones and treated as black sections.

Use of Poisson's Distribution

It is generally accepted that road accidents are accidental events and therefore the probability of an accident to occur in a road section during a specific time period follows the distribution of accidental events known as the Poisson's distribution. However, in certain section or spots of the road network traffic accidents occur in considerably higher frequences which by no means can be accepted as accidental and is indisputably attributed to the specific road characteristics prevailing there. Thus, with the aid of Poisson's distribution black sections on aroad can be identified.

The first step is to separate the road network into sections with similar geometric and traffic characteristics. In these sections the average number of accidents per kilometer represents the mean of the Poisson's distribution, i.e.

the number of accidents expected to occur in each one kilometer long subsection, if only accidental factors govern the occurrence of an accident. In sections with higher frequency of accidents their causes can be attributed with a certain level of confidence to other than accidental facts. When this level of confidence exceeds 90% the researcher is quite convinced that other than accidental events govern the high frequency of accidents in this specific subsection, which therefore is identified as a black subsection.

Traffic Accident Indices

Traffic accident indices are widely used for the estimation of the accident risk in specific road sections. Quite a lot of indices have been proposed. In the most commonly used ones the number of accidents is given in relation to the population of the area, or to the traffic volume of the road section, or to the number of vehicle-kilometers travelled or even to the length of the road network. Black sections are considered those, where the above indices take higher than the average values.

Accident Severity Indices

In all methods described till now the seriousness of the accidents has not been taken into account. However, the quantitative assessment of traffic accidents is quite necessary for a rational classification of road sections in relation to their accident risk. This quantitative assessment can be achieved by the introduction of certain factors and coefficients, which take into consideration the severity of the accident and the amount of property losses occured. For this purpose the following formula have been proposed.

Severity Index = $P_1 \cdot n_1 + P_2 \cdot n_2$

where : n_1, n_2 = number of accidents resulted in injuries or fatalities respectively

and P_1, P_2 = corresponding severity factors for each type of accident.

The formula can be easily extended to include more types of traffic accidents, if the relative data are available.

The values of these severity factors are determined according to the losses to the national economy due to the specific type of road accident. Typical values of these factors are given in Table 1. The inevitable differences in assessing the cost of accidents existing in various countries result in the differences in the values of the severity factors appeared in this table.

Critical Evaluation of the Methods of Black Spot Identification

Traffic safety on road sections should be assessed according to the number, the

Type of accident	Severity factor according to				
	Reinhold	Bitzl	Fisher	U.S.A.	U.S.S.R.
Unregistered	-	-	-	-	1
Damage only	1	1	1	1	3
Light injury	5	30	2	5	0.5
Heavy injury	70	30	8	5	8
Fatality	130	100	40	23	135

Table 1. Values of accident severity factors proposed by various authors (source : ref. 2)

frequency and the seriousness of accidents occuring there. An integrated method of black spot identification must tak into account all the above factors. Thus, simply the number of accidents occuring on a road section irrespectively from the traffic volume is an imperfect criterion for black spot determination. Furthermore, even if two road sections have the same traffic volume levels and number of accidents, but they markedly differ in the severity of the casualties, it is not again acceptable to be considered as similarly hazardous.

Taking these principles into account the absolute number of traffic accident method of black spot determination, apart from its simplicity, has the serious disadvantages of not considering the traffic volume and the severity of the accidents. The same critisism applies to the use of Poisson's distribution for the identification of black spots. This method however, has the advantage of providing a sound statistical basis. The use of traffic indices to locate road black sections takes into consideration various parameters which reflect traffic conditions, i.e. traffic volume, number of vehicle-kilometers travelled etc. The disadvantage of the ignorance of the severity of the accidents still exists. Finally, the use of various severity indices reflecting the seriousness of the casualties is the most advanced method for black spot identification. However, the discrepancies existing in the values of the severity factors proposed by various authors, is a certain weakness of the method.

DETERMINATION OF THE STUDY AREA

The Traffic and Road Research Laboratory of the University of Thessaloniki has recently completed a research project concerning the traffic accident analysis in the national road network in Northern Greece, during a 5 year period (1979-

1983). Six of the most important national roads (Fig. 1) have been selected for a comperative and correlation study of the various black spot identification methods.

Figure 1. National roads of the Northern Greece consindered in this study (Scale : 1:2.000.000)

All the roads are single carriageways and have been separated into sections with similar geometric and traffic characteristics. According to Greek normal practice as fatal accidents are determind those in which death occured on the spot or during the transfer of the victim to the hospital and as injury accidents are determined those in which the sufferer has been transfered to the hospital for treatment. Due to incomplete data it was impossible to distinct between light and serious injuries. Furthermore, damage only accidents are totaly ignored. Table 2 shows the traffic accidents occured during this 5 year period in the 6 national roads in Northern Greece. To achieve a sound basis for comparison it was considered better to divide each road section in uniform, one kilometer long, subsections from which the most hazardous ones would be probably identified as black subsections.

APPLICATION OF THE VARIOUS BLACK SPOT IDENTIFICATION METHODS

The absolute number of traffic accidents method has been applied as it is described in the relative paragraph.

In the identification of black subsections by using the Poisson's distribution three level of confidence 90,95 and 98% are applied. In these level of confidence accidental factors are correspondingly unlikely to be the unique causes of

NATIONAL ROAD		THESSALONIKI SERRES				SERRES DRAMA			THESSALONIKI KAVALA					KAVALA XANTHI			XANTHI KOMOTINI			KOMOT. ALEX.	
ROAD SECTION	FROM (km)	9	17	50	73	1	30	40	11	35	70	101	129	4	9	27	1	10	51	0	51
	TO (km)	17	50	73	94	30	40	66	35	70	101	129	156	9	27	53	10	51	57	51	65
AVERAGE DAILY TRAFFIC VOLUME		19[illegible]0	13000	3360	7200	2665	1718	1807	6515	5250	5809	2400	4150	8369	4881	3288	2980	3084	6013	1853	4100
NO. OF INJURY ACCID.	NO. OF FATAL ACCID.	NUMBER OF 1 Km LONG SECTIONS, WITH A CERTAIN NUMBER OF ADDIDENTS																			
0	0		8		2	3	1	10			4	3	2	1	1		2	7	1	9	
1	0	1	6	5	2	5	4	8	3	8	2	2	2			2	1	7	1	12	1
0	1		1									1						1		5	
2	0	2	7	5	2	5	3	2	4	1	3	3	3		1	6		10		8	1
1	1		1			1		2		2	1			1							
3	0		3	4	2	3	1		2	6	1	2	4		2	6		4		3	
2	1		2			2		1	1	3	2	3	1		1		1	1		1	
1	2				1				1	2		1		1							
4	0	1	1	3	4	4	1	1		1	2	2	1		1	2			1	6	3
3	1					1			3	3		2	2			1					1
2	2					1						1	1		2	1					
5	0	2		1		1		2	1	1	1	1	2	1	1	2		1	1	2	3
4	1		1		3				2	1	1	1	1					1		2	
3	2		1								1	1			1	1			1		
6	0		2			1			1	1	2	1	2		1		1	2			2
5	1	1		3	1											1					1
4	2									1	2										
7	0			1					1		2		1		1			1			
6	1									1		2			1			1			
5	2					1					1				1						
8	0									2			1		1	1		2		1	1
7	1				1	1						1				1	1	2			
9	0										1		1						1		
8	1																1				1
7	2								1						1		1				
9	1								1							1				2	
8	2									1	1					1					
11	0	1			1										1						
9	2				1								1								
9	3																1				
12	1										2		1	1							
14	0										1	1									
14	1			1					1				1			1					
12	3										1				1						
17	0				1																
14	3								1	1											
13	4								1												
16	3																	1			
TOTAL NUMBER OF ACCIDENTS		36	64	84	99	82	17	33	131	135	161	104	129	23	103	120	48	130	24	117	69
ACCIDENTS/ km		4.5	1.9	3.7	4.7	2.8	1.7	1.3	5.5	3.9	5.2	3.7	4.8	4.6	5.7	4.6	5.3	3.2	4.0	2.3	4.9

TABLE 2. Accidents in 6 National Roads in Northern Greece (Ref. 1)

the accidents.

The traffic accident index selected in this study is based on the traffic volume of the road section. On road sections which are homogeneous as regards their geometric elements and their traffic volumes, the accident rate is determined by the formula :

$$V_{Rt} = \frac{z \cdot 10^6}{365 \cdot Q \cdot L \cdot N}$$

where : z = is the total number of accidents

Q = is the traffic volume (vehicles per day)

L = is the length of the road section (km)

and N = is the time period (years).

Addionally the traffic index on each road subsection is calculated by the ratio:

$$V_{Rs} = \frac{z \cdot 10^6}{365 \cdot Q \cdot N}$$

where : z = is the number of accidents in each one kilometer long subsection and the rest variables as above.

In those subsections where $V_{Rs} > V_{Rt}$ the potential accident hazard is high so that the specific subsection is identified as a black one.

For the application of the accident severity index method, three sets of severity factors are used, which are : (8.50), (7.70), (12.100), the first number assessing the injuries and the second the fatalities. Applying these values the severity index for each kilometer of the road section, as well as the average severity index over the total length of the road section are calculated. This last value is multiplied by a coefficient, which takes successively the values 1.2 , 1.5 and 2.0 . The product is compared with the severity factors found for each one kilometer long road subsections. Obviously, as black subsections are identified those in which the severity factor exceeds the value of the product.

The number of black subsections identified by using each method are presented in table 3.

Critical Evaluation of the Results

Since the number of road sections examined, as well as their total length is quite high, arbitrary limits reflecting the average acceptable percentage of black subsections in relation to the total number of subsections, can be set. Thus, as acceptable percentage is considered every figure lying between 15% and 20%. Results found within these limits are obtained by Lamm's method of absolute number of traffic accidents, by using Poisson's distribution in practically

NATIONAL ROAD		THESSALONIKI SERRES				SERRES DRAMA			THESSALONIKI KAVALA					KAVALA XANTHI			XANTHI KOMOTINI			KOMOT. ALEX.		TOTAL NUMBER OF BLACK SECTIONS	PERCENTAGE
ROAD SECTION	FROM (km)	9	17	50	73	1	30	40	11	35	70	101	129	4	9	27	1	10	51	0	51		
	TO (km)	17	50	73	94	30	40	66	35	70	101	129	156	9	27	53	10	51	57	51	65		
	LENGTH	8	33	23	21	29	10	26	24	35	31	28	27	5	18	26	9	41	6	51	14		
AVERAGE DAILY TRAFFIC VOLUME		19000	13000	3600	7200	2665	1718	1807	6515	5250	5809	2400	4150	8369	4881	3288	2980	3084	6013	1853	4100		
METHODS		NUMBER OF BLACK SECTIONS DETERMINED BY THE VARIOUS METHODS																					
ABSOLUTE NUMBER OF TRAFFIC ACCID.	BABKOV [2]	0	0	1	1	0	0	0	3	1	1	0	1	0	1	1	0	1	0	0	0	11	2
	BENNET [3]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	LAMM [4]	1	5	3	3	4	2	4	4	5	5	4	4	1	3	4	1	6	1	8	2	70	15
USE OF POISSON'S DISTRIBUTION	90% level of confid.	1	5	5	4	4	2	4	5	7	6	5	6	1	3	5	4	11	1	13	2	94	20
	95% level of confid.	1	5	2	4	3	1	4	4	5	6	4	5	1	2	5	2	9	1	7	1	72	16
	98% level of confid.	1	4	1	3	2	0	3	3	4	5	2	4	1	2	3	1	7	1	3	0	50	11
TRAFFIC ACCIDENT INDEX		4	18	9	8	15	5	8	7	13	13	13	11	2	8	9	5	11	3	17	8	187	40
ACCIDENT SEVERITY INDICES	INJURY 8 FATAL 50 COEFF. 1.2	2	8	5	9	8	2	6	6	10	10	13	9	2	7	9	4	12	2	19	7	150	32
	INJ. 8 FATAL 50 COEFF. 1.5	2	8	5	6	7	2	6	5	6	9	8	5	2	4	7	2	9	2	13	6	114	25
	INJ. 8 FATAL 50 COEFF. 2.0	2	5	4	2	4	1	6	4	3	5	4	4	1	1	5	1	5	1	6	1	65	1
	INJ. 7 FATAL 70 COEFF. 1.2	2	8	5	9	7	2	6	6	13	10	13	8	2	8	8	4	10	2	13	3	139	30
	INJ. 7 FATAL 70 COEFF. 1.5	2	8	4	8	7	2	6	6	7	8	8	7	2	7	7	2	9	1	11	3	120	25
	INJ. 7 FATAL 70 COEFF. 2.0	1	6	4	2	6	1	5	3	5	8	3	4	0	3	5	2	6	1	10	2	77	17
	INJ. 12 FATAL 100 COEFF. 1.2	2	8	5	9	7	2	6	6	13	10	13	9	2	7	8	4	12	2	13	4	142	30
	INJ. 12 FATAL 100 COEFF. 1.5	2	8	4	8	7	2	6	6	6	8	8	7	2	6	8	2	9	1	11	3	120	25
	INJ. 12 FATAL 100 COEFF. 2.0	2	6	4	3	6	1	5	3	5	8	3	4	1	2	6	2	6	1	11	2	80	17

TABLE 3. Number of black sections (1 km in length) on the 6 National Roads in Northern Greece.

all three levels of confidence and by using the accident severity indices with the multiplying coefficient having the value of 2, irrespectively from the values of the indices themselves. Benner's and Babkov's proposals for black spot determination identify unacceptably low number of black subsections, obviously because the criterion set (3 and 4 at least traffic accidents annualy) is difficult to be met. On the other extreme, traffic accident index method gives a very high percentage of black subsections (40.2%). Also, Poisson's distribution method yields more optimistic results than those obtained by the traffic index method and by the severity index method. Finally, inspection of Table 3 shows that the influence of the coefficients used in the accident severity index method in the determination of the number of black subsections, is considerably stronger than the influence the values of the severity factors have.

Correlation of the Results

An attempt to correlate the results, the four methods of black spot identification yield, is made by the calculation of the correlation coefficients (r) between all pairs of the different methods. The results are presented in Table 4. In cases where the value of r exceeds 0.85 , the correlation is considered to be high. On the other hand, where r is less than 0.70 the correlation is considered as poor.

Inspection of Table 4 shows that the use of the Poisson's distribution at the 98% level of confidence yields the lowest correlation with every other method, whereas Lamm's method of absolute number of traffic accidents has the highest correlation with all other methods. Poisson's distribution method at the 90% and 95% level of confidence correlates fairly well with the rest of the methods. The same applies to the traffic index method and the accident severity index method. The values of the severity factors which presents the better correlation with other methods are (7.70) and (12.100), the second being slightly better. Finally, the value of the coefficient which enhances the correlation of the severity index method, is 1.5 .

CONCLUSIONS

This study confirmed the important differences existing in black spot identification according to the various methods in use. Thus, it is the authors' opinion that a combination of two methods of black spot identification should be always made. The methods proposed for this combination are the Poisson's distribution at the 95% level of confidence and the accident severity index method. The most appropriate values of the severity factors determined here are 12 for

METHODS		LAMM'S ABSOLUTE NUMBER OF TRAFFIC ACCIDENTS	USE OF POISSON'S DISTRIBUTION: 90% level of confidence	USE OF POISSON'S DISTRIBUTION: 95% level of confidence	USE OF POISSON'S DISTRIBUTION: 98% level of confidence	TRAFFIC ACCIDENT INDEX	ACCIDENT SEVERITY INDICES: INJURY: 8, FATAL.:50 COEFFICIENT : 1.2	INJURY: 8, FATAL.:50 COEFFICIENT : 1.5	INJURY: 8, FATAL.:50 COEFFICIENT : 2.0	INJURY: 7, FATAL.:70 COEFFICIENT : 1.2	INJURY: 7, FATAL.:70 COEFFICIENT : 1.5	INJURY: 7, FATAL.:70 COEFFICIENT : 2.0	INJURY:12, FATAL.:100 COEFFICIENT : 1.2	INJURY:12, FATAL.:100 COEFFICIENT : 1.5	INJURY:12, FATAL.:100 COEFFICIENT :2.0
LAMM'S ABSOLUTE NUMBER OF TRAFFIC ACCIDENTS			0.911	0.895	0.743	0.851	0.892	0.925	0.831	0.842	0.919	0.917	0.859	0.913	0.901
USE OF POISSON'S DISTRIBUTION	90% level of confid.					0.692	0.869	0.831	0.708	0.776	0.801	0.826	0.819	0.798	0.827
	95% level of confid.					0.688	0.808	0.819	0.789	0.797	0.863	0.800	0.849	0.875	0.800
	98% level of confid.	0.743				0.561	0.587	0.624	0.677	0.684	0.756	0.729	0.768	0.750	0.636
TRAFFIC ACCIDENT INDEX		0.851	0.692	0.688	0.561		0.801	0.857	0.723	0.787	0.833	0.850	0.787	0.821	0.811
ACCIDENT SEVERITY INDICES	INJ. 8 FATAL 50 COEFF. 1.2	0.892	0.869	0.808	0.587	0.801									
	INJ. 8 FATAL 50 COEFF. 1.5	0.925	0.831	0.819	0.624	0.857									
	INJ. 8 FATAL 50 COEFF. 2.0	0.831	0.708	0.789	0.677	0.723									
	INJ. 7 FATAL 70 COEFF. 1.2	0.842	0.776	0.797	0.684	0.787									
	INJ. 7 FATAL 70 COEFF. 1.5	0.919	0.801	0.863	0.756	0.833									
	INJ. 7 FATAL 70 COEFF. 2.0	0.917	0.826	0.800	0.729	0.850									
	INJ. 12 FATAL 100 COEFF. 1.2	0.859	0.819	0.849	0.768	0.787									
	INJ. 12 FATAL 100 COEFF. 1.5	0.913	0.798	0.875	0.750	0.821									
	INJ. 12 FATAL 100 COEFF. 2.0	0.901	0.827	0.800	0.636	0.811									

TABLE 4. Correlation coefficients between numbers of black spots on each road section determined by various methods.

injury accidents and 100 for fatal accidents, as well as, the more appropriate value for the multiplying coefficient is 1.5 . Finally, Lamm's proposal of the characterization as "blacks" of the 15% of the most hazardous road subsections appears to provide a sound initial estimation for a black spot identification study.

References

1. Iktinos Consultant Engineers "Traffic Accident Analysis and Remedial Measures for the Promotion of Traffic Safety in the National Road Network". Ministry of Public Works, Thessaloniki 1985, Greece.

2. Babkov V.F. : "Road Conditions and Traffic Safety" MIR Publishers, Moscow 1975.

3. Benner E. et al : "Beseitigung von Unfallstellen, Band 3, Identifikation von Unfallstellen" , Herausgeber, Köln, November 1978.

4. Lamm R., Klockner J. : "Untersuchung des Unfallgeschehens im Landkreis Leer unter Besonderer Berucksichtigung Strassenbaulicher und Verkehrstechnischer Gesichtspunkte" , Gutachten, 1977.

5. Tsohos G., Dalaveras A. : "The Problem of Road Safety in Greece. A Survey of Road Accidents", Technika Chronika Scientific Journal of the Technical Chamber of Greece, Vol. 5, No 1, 1985.

SOME OBSERVATIONS ON THEORY AND METHODOLOGY IN SAFETY RESEARCH

By

Paul Jovanis
Associate Professor of Civil Engineering
and Transportation

and

Hsin-Li Chang
Assistant Professor

I. INTRODUCTION

This paper argues in support of a structured method of conducting safety analyses that is directly related to the title of the conference. Specifically, we argue that safety theory should be explicitly considered during the development and application of statistical methods for safety analyses. This is more than simply a call for "correct" use of statistics. We believe that significant progress on contemporary safety issues can only be made if theory is consonent with statistical method. At initial research stages, the theory may evolve from a conceptual model; at subsequent stages it may be inferred from relevant disciplines such as psychology, physiology or economics for example.

In addition to closer connections between theory and statistical tests, there is a need, we believe, for greater fertilization across methodologies and disciplines. For example, findings obtained through laboratory experiments should be considered when formulating models of driver cognitive processes. Positive crossfertilization occurs all too infrequently. The second section of this paper discusses potential linkages between different safety methodologies.

Finally, we present an example of a statistical method, based upon survival analysis, that is at least consistent with conceptual models of exposure. We present the methodology and an example of a new technique that can be used to test important empirical questions, but in a way that is consistent with contemporary notions of exposure and other theory.

The occurrence of accidents, must be compared to the number of opportunities available to be involved in an accident. Some representation of these opportunities is commonly referred to as exposure to accident risk. Hauer develops a definition of a unit of exposure as a trial in which the outcomes are an accident (possibly of several types) or a non-accident (Hauer, 1982). Safety (as measured by accident occurrence) is the product of the probability of having an accident (also called risk) and the number of exposure units. Factors contributing to accident risk are thus conceptualized as affecting the probability of an accident.

A major problem in combining accident data with exposure is that accidents are discrete events. Data describing accidents routinely come from reports describing accident outcome and characteristics such as driver, vehicle, roadway and environment at the time of the accident. Exposure data are much more aggregate, typically based upon measured or estimated daily, weekly, monthly or often yearly travel. A fundamental dilemma in studies of accident occurrence is how to combine exposure and accident data in a meaningful and consistent way so that the contribution of individual factors to accident risk can be identified.

All accident prediction models in the previous literature have been developed using aggregate exposure data. The use of aggregate data to construct an accident analysis model results in the loss of individual information and a clouding of the relationship between risk components and accident occurrence. Disaggregate data have been commonly used in travel demand research due to their improved explanatory capabilities, but they have not been commonly used in safety research, particularly for exposure data.

A variety of research approaches have been used to explore the risk factors of highway operations. These include the laboratory driving simulator [e.g. Hulbert and Wojcik, 1971] inobtrusive observation of on-road operations, detailed multi-disciplinary assessment of accident causes [e.g. Treat et al. 1977] and a wide variety of statistical analyses. A shortcoming of these four approaches is the failure to relate their findings quantitatively to accident risk due to the lack of appropriate exposure data. These methodologies are reviewed in more detail in Section II of this paper.

One factor hindering resolution of these problems is failure to use a consistent explanatory framework for accident occurrence. This framework should clearly differentiate risk of accident involvement from accident occurrence which is the interaction of risk and exposure. Hauer provides an excellent discussion of these issues [Hauer, 1982]. It would be advantageous if one could utilize concepts from Hauer to develop a framework that could provide a bridge between the aggregate observation of accident data and the disaggregate results obtained from laboratory experiments and detailed causal assessments. This connection would be an advance over the way of in which accidents are thought of as the result from interactions of the driver, vehicle, roadway and environment [ITE, 1976] without careful consideration of how these interactions occur.

The remainder of the paper is divided into three sections. First, we discuss four methodologies commonly used to study accident occurrence and causes. The methodologies are compared along four dimensions with the objective of identifying opportunities for findings from one methodology to influence another. This is intended to meet the objective of idnetifying areas of crossfertilization across methodologies.

The following section develops a framework for the study of accident occurrence that we believe is consistent with theory and the concept of exposure. We believe that the framework can be used to guide statistical analyses that are more theoretically and conceptually consistent. The paper concludes with a summary description of a methodology based upon survival theory that offers significant advantages over many other statistical techniques.

II. A TYPOLOGY SAFETY RESEARCH METHODOLOGIES

A. Overview

We have constructed a typology of traffic safety research methodologies in Table 1. Four different methodologies are identified: laboratory experiments, on-the-road study, accident causal analysis and correlational analyses. For each of these categories, we denote whether data are collected at the aggregate or disaggregate level and also whether these methodolgies address 4 topics that, we believe, are important in the identification of accident causality. The four topics are defined as follows:

Driver actions - the ability of the methodology to identify specific driver actions (or lack of actions) that may contribute to a crash. This includes both studies of driver capabilities (through laboratory experiments) and studies of driver behavior during on-the-road studies.

Accident Occurrence Process - the ability of the methodology to identify the process of accident occurrence as a series of events or collisions.

Exposure - the ability of the methodology to explicitly include exposure to accident risk as well as accident data and characteristics.

Actual Accident Involvement - the ability of the methodology to analyze actual accident data.

B. Laboratory Experiment

Laboratory experiment or simulation can be used to study details of driver or vehicle actions which may be linked to accidents but are difficult to observe in the field. Laboratory experiments commonly study actions such as steering wheel movement [Crandall, Duggar and Fox, 1966], lateral and longitudinal position [Barrett, Kobayashl and Fox, 1968], velocity estimation [Salvatore, 1968], breathing rate [Beers, Case and Hulbert, 1970], and vigilance [Heimstra, 1970]. In those experiments or simulations, the relationship between independent variables and these intermediate measures is applied directly and then inferences are made about the effect of these independent variables on highway accident risk.

The Advantages of laboratory experiments include safety of the subjects, control of some confounding variables and possibly reduced costs compared to field observation. We also face several shortcomings, foremost among them is the questionable generalization of the laboratory findings to the actual highway environment [Shinar, 1978].

While laboratory experiments allow us to obtain individual disaggregate performance data they are limited in their ability to provide insight in the process of accident occurrence and, obviously, do not contain data on actual involvements. It is also difficult to generalize observations from the laboratory to a broad population to gain insight on exposure to risk. In the

Table 1: A Typology of Traffic Safety Research Methodologies

Methodology \ Involved Topic		Driver Actions	Accident Occurrence Processes	Exposure	Actual Accident Involvement
Laboratory Experiment	Aggre.				
	Disagg.				
On-The-Road Study	Aggre.				
	Disagg.				
Accident Causal Analysis	Aggre.				
	Disagg.				
Correlation Analysis	Aggre.				
	Disagg.				

Previous Research   Proposed Research

parlance of this conference, these studies can be thought of as testing cognitive models.

B. On-the-Road Studies

Studies of drivers in actual conditions include application of the traffic conflicts technique [e.g. Perkins, 1969, Older and Spicer, 1976], inobtrusive observation of individual drivers and vehicles [Shinar, Rockwell and Malecki, 1975] and on-road measurements of drivers in instrumented vehicles [e.g. Platt 1970; Hrelander, 1976 and Fuller, 1980].

The major advantage of on-the-road research is that results obtained from it may be immediately applicable to the highway environment. Its major disadvantage is that many variables are not under strict experimental control and the results may be due to uncontrolled variables, and/or limited to the specific location where the study was conducted. While individual drivers are studied, it is not possible to directly relate these studies to outcomes (accidents). Exposure to the risks understudy are also difficult to assess. Many of these studies can be thought of as addressing "behavioral" models.

C. Accident Causal Analysis

An accident results whenever one or more factors -- labeled as the accident cause or causes -- deviates from the norm to such an extent that the system cannot accommodate it [Shinar, 1978]. One of the most consistent findings in accident research is that accidents are typically caused by more than one factor. Each factor cited as causal may be a cause only in the context of the other causes.

The most prominent study for accident causal analysis is the Indiana University's Trilevel Study of The Causes of Traffic Accidents [Treat et al., 1977]. These three levels of accident investigation include: (1) routine police investigation, (2) "on-site" investigation by specially trained technicians who rushed to the accident site immediately after notification by the police, and (3) "in-depth" investigation by a multidisciplinary accident investigation team who examined and interviewed the driver, reconstructed a complete diagram of site and vehicles' paths, and examined the accident vehicle in a specially equipped garage. The study results show that human factors, identified as probably or definite causes, are related to approximately 91 percent of the traffic accidents.

This study has had a great influence on subsequent safety research so that it is obviously of major importance. It's major limitation is the luck of exposure data which does limit some interpretation of their results.

D. Correlational Analysis

A variety of statistical approaches have been applied to safety studies. Usually, analysts combine the accident data with controlled exposure and test the hypothesis of interest. The simplest type of study is the comparison of the mean and variance of the accident involvement rates, which is undertaken to test the equality of accident risks between different exposure groups. Examples of this technique include the work of Foldvary [1979], who explored accident involvement rates in terms of characteristics of driver, vehicle, road, and driving environment, Meyers [1981], who compared the accident rates of truck and passenger-cars on limited-access facilities and a comparison of weather effect on auto and truck accident involvement rates by Jovanis and Delleur [1982].

Linear regression models have been widely used in safety studies. Usually, the accident involvement rates are considered the dependent variables in most of linear regression analyses of safety study, and the risk components to be detected are assigned to the independent variables. Those risk components include travel speed [Hall and Dickinson, 1974; Lavette, 1977], traffic volume [Oppe, 1979; Ivey et al., 1981; Ceder and Livneh, 1982], as well as weather and vehicle [Jovanis and Delleur, 1982].

Three particular properties of accident occurrence argue against the application of linear regression analysis to highway safety studies. First, the discreteness of accident occurrence will cause the error terms to be heteroskedastic in the linear regression analysis [Ruijgrok and Van Essen, 1980; Montgomery and Peck, 1982], even if one uses accident rates instead of the number of accidents [Jovanis and Chang, 1985]. Second, the non-negativity of accident measure of the dependent variable also impose restrictions on the applicability of the linear regression techniques. Third, the error terms are not normally distributed due to the characteristics of non-negativity and small value of discrete dependent variable. This makes us unable to generate the correct confidence intervals for estimated parameters. In order to improve the shortcomings of linear regression analysis in safety study, one discrete model -- the Poisson Regression Model, has been applied in the study of accident occurrence. Hamerslag [1982] used it to detect the effects of road characteristics and traffic volume on the accident involvement rates. Jovanis and Chang [1985] described the accident occurrence on a closed highway system as a Poisson process in which the daily expected number of accidents is a function of daily traffic exposure and weather condition.

Some multivariate analysis techniques other than regression analysis are also used in safety study. The automatic interaction detection (AID) technique has been used to categorize the explanatory variables in order to discriminate the accident involvement rates for different exposure groups [Snyder, 1974; Cleveland and Kitamura, 1978]. Koornstra [1969] used one set of categorical data to detect the relationship between type of seat belt and location of injury. Hakkinen [1979] studied how professional drivers classified as safe drivers versus accident drivers differ in terms of driver's characteristics by discriminant analysis. He also reduced the original twenty-six driver characteristics to six factors by factor analysis to give a concise representation of risk components to accident involvement. An aggregate logit model of discrete multivariate analysis was applied to study the severity of large-truck and combination-vehicle accidents in over-the-road service by Chirachavala [1984].

The common denominator of all above statistical or correlation analyses for traffic safety study is the absence of an explicit explanatory framework for accident occurrence. That is, those efforts emphasized the estimation of statistical relationships in the available data and attempted to intepret those relationships. A preferred approach is the development of an understanding of the underlying process which determined those relationships, and the development of an analysis framework which can capture those relationships. Furthermore, all exposure-based accident prediction models in previous literature were developed with aggregate data. The use of aggregate data to construct an accident prediction model will cloud the relationships between risk components and accident occurrence.

E. The Relationship of The Proposed Survival Theory Model to Previous Methodologies.

A complete traffic safety research framework should combine the knowledge of the driver's behavior, accident occurrence process, exposure and accident involvement together. While Each research approach has its own advantages and disadvantages, it would be useful if we could evolve a set of statistical methods that have the capability to use knowledge gained from the other three types of methodologies.

If we can develop a method to capture disaggregate exposure, we may be able to connect the study findings regarding driver behavior with actual accident involvement. We all know it is hard to collect disaggregate exposure data, but it is harder to collect disaggregate exposure data without a research framework to guide us how to collect it. The survival theory model is proposed as an example of how to fill the theoretical gap between previous traffic safety studies. It is our main purpose to develop a research framework for disaggregate modeling on highway safety study by combining elements of driver behavior with a conceptualized model of the accident occurrence process, exposure data and data describing actual accident outcomes. The conceptualization of accident occurrence is described next.

III. A CONCEPTUAL FRAMEWORK FOR THE PROCESS OF ACCIDENT OCCURRENCE

A. The Driver As An Information Processor.

Though driving has been modeled as information processing for some time [Shinar, 1978], there have been no attempts so far to use these concepts to develop a feasible and quantitative model for highway safety research. In order to extend this conceptual idea, some effort needs to be placed on the detailed observation of how the information comes to a driver as well as how the driver responds to it and keeps his vehicle on the road.

Figure 1 shows us how the risk factors bring their information to driver through direct or indirect ways. This hypothetical information propagation structure offers a useful guideline to think about the risk potential of the driving task and helps us to realize the possible interactions between risk factors. We observe that there are three paths to bring the environment information to the driver. First, the environment can directly pass its information to the driver and affect driver's performance. The driver's vision, for an example, will be hurt when driving under the bad weather or poor lighting conditions. Second, environment can affect the roadway conditions and then indirectly deliver its information to the driver. One of these examples is that snow will make the roadway slippery and require much more driving effort of drivers. Third, environment also affects the vehicle and asks more careful driving of the drivers, e.g., strong wind will make small vehicles less unstable.

Roadway has two ways to transmit its information to drivers. Different roadway designs can bring different extents of driving difficulty directly to the driver, or indirectly to the driver through affecting vehicle's performance, e.g., a narrow mountain roadway might bring a lot of pressure to driver particularly for large vehicles.

The vehicle is the closest element of contact to the driver while driving. The vehicle passes its information directly to the driver. Though most of this information is coming from the environment and the roadway, there is still some information to the driver created by the vehicle itself, such as travel speed or mechanical defect problems.

A driver makes his decision based on the information he receives. Different drivers may make their decisions in different ways. These decisions then result in different drivers' performance. Driver's decisions control the vehicle performance and feedback to affect the driver's further decision again. They have no effect on altering the conditions of the roadway and the environment.

B. Conceptualized Accident Occurrence Process.

An attempt trying to conceptualize the accident occurrence process starts with a microscopic observation of individual vehicles, from the start to end of their movement. Interest of this observation centers on how an accident is initiated, what the contributing risk components are, and how those risk components work together. The knowledge received from this microscopic

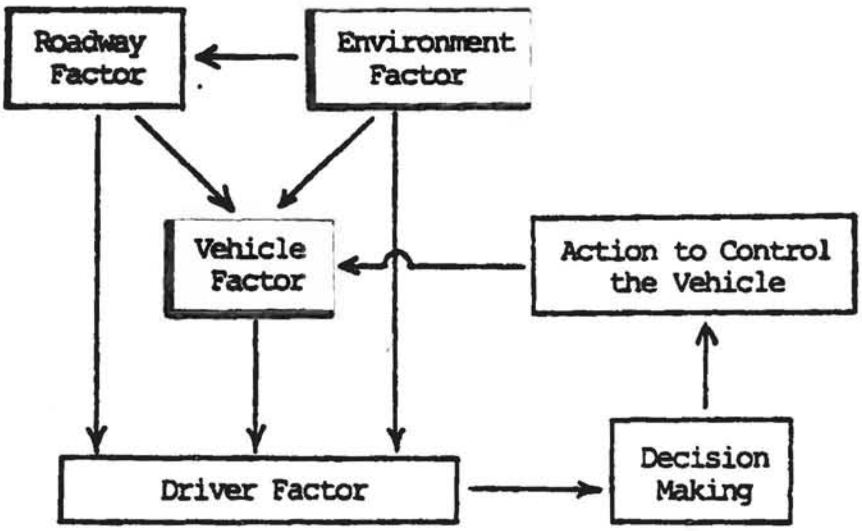


Figure 1: A Conceptual or Information
Processing by 'er (Modified from Shinar, 1978)

observation can then help us to develop the conceptual model we seek. Based on this conceptual relationship between accident occurrence and corresponding risk components, the risk to be involved in an accident can then be mathematically formulated in terms of those risk components.

The movement of a vehicle cannot continue infinitely due to the limitation of fuel tank capacity or fatigue of driver. Temporary stops may be necessary during traveling. Hence, travel to fulfill an activity may be finished either by only one continuous movement or by several segments of continuous movement. For different segments of continuous movement, the operating characteristics may or may not dramatically change. Furthermore, the time between two consecutive segments of continuous movement may affect the operating risk for the continuous movement following the stop, if the fatigue of driver is a factor affecting the highway operation risk. In order to capture the reality of highway operation risk, the selection of time frame to undertake observations and model formulation is a crucial issue. The time frame will vary, however, depending on the nature of the safety system to be investigated. For example, we may choose a twenty-four hour observation on auto traveling process due to the periodical characteristics of daily activity pattern. An origin-to-destination observation may be undertaken on truck traveling process. In general, a trip usually means a complete journey. It may consist of more than one segment of continuous movement, that starts after and ends with a long enough rest, in order to make the observed trips in our selected time frame reasonably independent from other trips not observed.

B.1 Accident Generating Process.

The traveling process for one vehicle trip is conceptually described in Figure 2. Essentially, the characteristics of driver, vehicle and trip (e.g. trip purpose) are given before the vehicle trip starts. We call those given characteristics the initial conditions of movement. In terms of accident risk, those initial conditions imply some risk potential for accident involvement. For example, the lack of enough rest prior to starting one trip will affect driver's alertness and increase the accident risk. With these initial conditions, the driver starts to undertake his information processing task and seeks to attain the required performance in order to maintain vehicle operation. Working along with the varying environment and roadway conditions, those initial conditions may or may not change as the vehicle proceeds to run.

The vehicle ends its exposure with a stop. Stops can be classified into two categories -- accident involvement and nonaccident stop, according to the definition of the chance set up. A nonaccident stop always results in a period of rest before the vehicle starts another continuous movement. Based on the criterion we have chosen to define the trip, we can assign the nonaccident stop to be the end of one trip or a temporary rest depending on how long the nonaccident stop lasts. A new continuous movement following the stop may come into the information processing system again with another set of initial conditions.

Our microscopic observation on individual vehicles terminates with the successful finish of one trip or being involved in an accident. We call the accident generating process the process that the driver experiences in seeking to survive in a risk system from the starting to the ending of one trip. For accident involved trips, our observation can measure the lifetimes of those

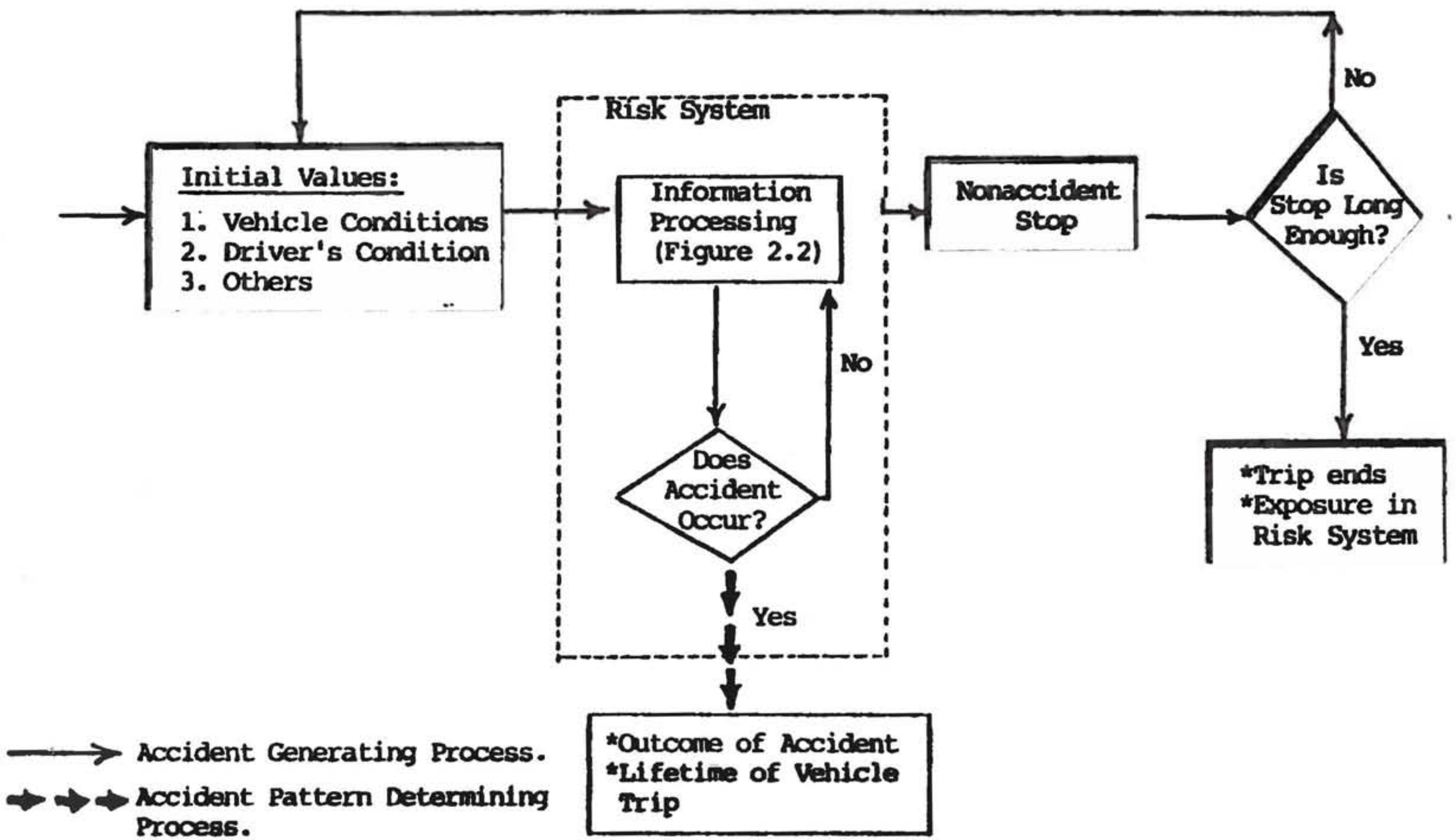


Figure 2: Conceptual Framework of Accident Occurrence

trips and outcomes of those accidents. However, for nonaccident trips, the only information we have is their survival after a given amount of exposure in the risk system. In terms of the survival analysis, models in the next section, these individual trips are censored (i.e. we do not observe the failure time).

B.2 Accident Patterning Process.

In the accident generating process, our interest is to figure out how the risk system determines whether or not an accident will occur. However, when an accident is initiated, the risk system will affect the outcome of accident again. This outcome includes the number of involved vehicles, type of collision, severity of injury and so on. We call the accident patterning process the process in which the risk system determines the outcome of an accident. Therefore, the risk system will not only dominate the accident generating process, but also control the accident pattern determining process. The risk components for the accident generating process operate during the whole vehicle trip, but only have an instantaneous effect on the accident patterning process.

Contrary to the accident generating process, the accident patterning process may have little to do with the travel exposure. Hence, the study associated with accident patterns can be easily undertaken through the data already in accident reports, obviating the most difficult issue in highway accident study -- exposure data. However, though studies of accident patterns can help us to find the strategies to reduce the severity of injury or property damage when a vehicle is involved in an accident, they are limited in how much they can contribute to identify how to avoid accident involvement.

B.3 Competing Accident Patterns.

In preceding sections, the accident generating process and accident patterning process are thought of as two sequential steps. However, if specific accident patterns are thought to have their own accident generating processes and compete with each other to stop the continuation of one vehicle trip, then the accident occurrence process may be constructed as a competing risk problem. Those specific accident patterns can be classified by accident causes, or accident outcomes. Whenever one of those two accident patterns appears first, the vehicle trip will be terminated, and the other will not occur.

It might be interesting to see the transition between accident patterns as some associated risk factor for one specific accident pattern has been reduced, if the accident occurrence process is formulated as competing accident patterns. For example, we might like to identify the reduction in right angle accidents along with possible increases in rear end accidents if skid resistance treatments are given to an intersection approach.

Several problems should be carefully considered before we formulate the accident occurrence process by competing accident patterns. First is the interdependency between different accident patterns. This is because different accident patterns might not be mutually exclusive. For example, an injury accident always comes with some property damage. Second, there are usually several common risk components between different accident patterns. The critical controversy is whether the accident generating processes for different accident patterns work independently. If they do not work independently, it will be very difficult to formulate the accident occurrence process by competing risk approach and further theoretical consideration will be required. At this initial stage of model development, we assume that accident patterns are independent.

We next procede to the mathematical formulation of an analysis approach based upon this conceptualization.

IV. MODELLING HIGHWAY ACCIDENT OCCURRENCE

A. Formulation of the Hazard Function

According to our conceptual structure, we can find that the accident generating process possesses some characteristics which will critically affect our consideration about what mathematical approach is appropriate to formulate this problem. First, system hazard is composed of all the risk components which may be constant, situational, and elapsed-time. Hence, the system hazard varies during the trip. Second, an accident is the only event that can occur during a vehicle trip other than to successfully complete the trip. These characteristics allow the accident generating process to be modelled as a survival process. Third, only a few trips among the observed trips will be involved in an accident. Using the concept of variable system hazard, our interest is to observe how long the vehicle can survive before an accident occurs.

Let T be a nonnegative random variable representing the lifetimes of individual trips in some population. Let f(t) denote the probability density function of t and let the distribution function be

$$F(t) = Pr\ (T < t) = \int_0^t f(x)\, dx \qquad (4\text{-}1)$$

The probability of an individual trip surviving till time t is given by the survival function

$$S(t) = Pr\ (T > t) = \int_t^\infty f(x)\, dx \qquad (4\text{-}2)$$

Note that S(t) is a monotone decreasing continuous function with $S(0) = 1$ and $S(\infty) = \lim_{t\to\infty} S(t) = 0$. The concept of hazard function h(t) is defined as

$$h(t) = \lim_{\Delta t \to 0} \frac{Pr\ (t < T < t + \Delta t \mid T > t)}{\Delta t} = \frac{f(t)}{S(t)} \qquad (4\text{-}3)$$

The hazard function specifies the probability density function of being involved in an accident at time t, given that the vehicle trip survives up until t.

The functions of f(t), F(t), S(t) and h(t) give mathematically equivalent specification of the distribution of T. It is easy to derive expressions for S(t) and f(t) in terms of h(t), since $f(t) = -S'(t)$. Eq. (4-3) implies that

$$h(x) = -\frac{d}{dx} \log S(x)$$

thus

$$\log S(x) \Big|_0^t = -\int_0^t h(x)\, dx \qquad (4\text{-}4)$$

and since $S(0) = 1$, we find that

$$S(t) = \exp\left[-\int_0^t h(x)\, dx\right]$$

For some purposes it is also useful to define the cumulative hazard function

$$H(t) = \int_0^t h(x)\,dx$$

which, by Eq. (4-4), is related to the survival function by $S(t) = \exp[-H(t)]$. It can be observed that since $S(\infty) = 0$, then $H(\infty) = \lim_{t\to\infty} H(t) = \infty$. Thus the hazard function $h(t)$ for a continuous lifetime distribution possesses the properties

$$h(t) \geq 0 \qquad \int_0^\infty h(t)\,dt = \infty$$

Finally, in addition to Eq. (4-4), it follows immediately from Eq. (4-3) that

$$f(t) = h(t) \exp\left[-\int_0^t h(x)\,dx\right] \tag{4-5}$$

Because the functions of $f(t)$, $S(t)$ and $h(t)$ are mathematically equivalent specifications, we can undertake our analysis in terms of any one of them. Cox and Oakes (1984) raised a number of reasons why consideration of the hazard function may be a good idea. We prefer the hazard function $h(t)$ to the others since the notion of failure rate is basic and conceptually simple. The function $h(t)$ provides a convenient starting point for undertaking the survival analysis. Presumably, the lifetime of an individual vehicle trip is affected by the concommitant variables. Therefore, in general, we can represent the hazard function as $h(t|X)$, where X is a vector of explanatory variables which are the risk components we mentioned in Section 3. Further components of X may be synthesized to examine interaction effects in a way that is broadly familiar from multiple regression analysis. The hazard function $h(t|X)$ indicates the probability to be involved in an accident at time t for a vehicle with risk components vector X, given that the vehicle trip survives up till t.

B. Types of Hazard Functions and Their Implications

Several types of hazard models for survival analysis have been introduced in the biomedical literature (e.g., Aranda-Ordaz, 1983; Cox and Oakes, 1984). They differ in the way in which the explanatory variables are assumed to influence the underlying hazard. For reasons explained in detail elsewhere [Chang, 1987] we choose the proportional hazards model proposed by Cox as the basis of our formulation.

Specifically the Cox Model is:

$$h(t|X) = h_o(t) * \exp(B*X) \tag{4-6}$$

while $B*X = b_1x_1+b_2x_2+\text{------}+b_px_p$ and the b_p's are unknown regression coefficients. The Cox model possesses the characteristic that the increase of system hazard due to the increase risk of one specific risk component depends on all the other risk components. That is, when the risk component x_i increases Δx_i, the hazard function $h(t|X)$ will increase to $h_0(t)*Exp(B*X)*Exp(b_i * \Delta x_i)$. This characteristic is quite similar to the risk of the driving task in which the risk components work together.

There are several reasons for considering the proportional hazards models (Cox and Oakes, 1984). First, there is a simple easily understood interpretation to the idea that the effect of the risk components vector is to multiply the hazard by a constant factor. Second, censoring time and the occurrence of several types of failure are relatively easily accommodated within this formulation, and in particular the technical problems of statistical inference when $h_0(t)$ is arbitrary have a simple solution. Third, the proportional hazards assumption appears to be reasonable in many situations. Some examples and references to this in the biomedical area are contained in Breslow (1975), and Prentice and Kalbfleisch (1979). In engineering contexts, proportional hazards are considered by Lawless (1976), Mann (1978), and many others.

The effect on accident risk due to the change of one specific risk component depends on other risk components. For example, the accident risk of a mechanical defect (e.g., failure of brake or flat tire) might depend on the vehicle speed and the level of surrounding traffic. The multiplicative risk model can capture this operating characteristic better than an additive risk model.

According to the risk propagation process discussed in Section 3, the effect of risk factors on accident risk can be divided into three sequential stages. In a multiplicative risk model, each stage can be thought of as one multiplicative subfactor. In addition to those three multiplicative subfactors, there are some interactions between risk components across different stages. Those interactions bring additional effects on accident risk and resulting the fourth multiplicative subfactor for hazard function. Therefore, formulating the system hazard function by a multiplicative model, we will have following five elements to be considered:

(1) Nuisance hazard $h_0(t)$

(2) Multiplicative subfactor of driver risk factor

(3) Multiplicative subfactor of vehicle risk factor

(4) Multiplicative subfactor of roadway and environment risk factors and their interaction

(5) Multiplicative subfactor of the interaction between risk components across different risk propagation stages.

Among those five elements, the nuisance hazard $h_0(t)$ can be a time-independent function (i.e., constant) or a time-dependent function of some specific parametric distribution family. The four multiplicative subfactors should be nonnegative and it is natural to suggest the exponential expressions for them.

C. Proposed Model for Accident Occurrence

We consider a population of individual vehicle trips; for each vehicle trip we observe either the time to be involved in an accident (i.e., lifetime) or the time to reach its destination (i.e., censoring time). That is, for the nonaccident vehicle trips we assume that the times to be involved in an accident for those vehicle trips are greater than the times they spent to finish their trips. Hence, an accident trip contributes a factor $f(t|X)$ to our model formulation, but a nonaccident trip contributes a factor $S(t|X)$ to the model. Therefore, the likelihood function for a set of observed data on n vehicle trips can be expressed as follows when the lifetime distribution of an individual trip is considered to be a function of regression vector X_i:

$$\begin{aligned} L &= \prod_{i=1}^{n} \{f(t_i|X_i)\}^{\delta_i} * \{S(t_i|X_i)\}^{1-\delta_i} \\ &= \prod_{i=1}^{n} \{h(t_i|X_i) * Exp[-H(t_i|X_i)]\}^{\delta_i} \\ &\quad * \{Exp\ [-H(t_i|X_i)]\}^{1-\delta_i} \end{aligned} \tag{4-7}$$

where t_i is the lifetime or censoring time for the ith individual and δ_i is the usual indicator variable taking on the value 1 if t_i is lifetime and 0 if t_i is censoring time.

The hazard function $h(t_i|X_i)$ is assumed to be a proportional hazard model:

$$h(t_i|X_i) = h_o(t_i)*Exp[Q(B,X_i)] \tag{4-8}$$

where $Q(B,X_i)$ is the formulation of the risk components vector X_i as a multiplicative factor and B is a vector of parameters to be estimated in the specified model. In this research, only time independent risk components are included in $Q(B,X_i)$; the effect of time dependent risk components are assigned to the nuisance hazard function $h_o(t_i)$. Then, the likelihood function Eq. (4-21) can be formulated as:

$$\begin{aligned} L &= \prod_{i=1}^{n} \{h_o(t_i) * Exp[Q(B,X_i)] * Exp[-H(t_i|X_i)]\}^{\delta_i} \\ &\quad * Exp\ [-H(t_i|X_i)]^{1-\delta_i} \\ &= \prod_{i=1}^{n} \{h_o(t_i) * Exp[Q(B,X_i)]\}^{\delta_i} * \{Exp[-H(t_i|X_i)]\} \end{aligned} \tag{4-9}$$

Usually, for convenience purpose, we take a monotone transformation and make the logarithm of Eq. (4-23) and get the log-likelihood function as

$$LL = \sum_{i=1}^{n} \{\delta_i * [\log(h_o(t_i)) + Q(B,X_i)] - H(t_i|X_i)\} \qquad (4\text{-}10)$$

With the assumed proportional hazards model like Eq. (4-20), the LL(B) is twice differentiable and bounded. We can deduce the existence and uniqueness of the solution of estimated coefficient vector B which maximizes Eq. (4-24), from the literature of survival analysis (Lawless, 1982; Cox and Oakes, 1984).

V. SUMMARY

We have tried to call attention to 2 issues that are important as we consider traffic safety theory and methodology:

1. That there are limited studies that use results from one type of safety methodlogy to enhance other methodologies. A typology of safety methodologies is developed and discussed to illustrate this point.

2. theory and concept should be directly considered before statistical methods are used. A conceptual framework for accident occurrence is developed based upon the principle of the driver as an information processor. The framework underlies the development of a new modeling approach.

3. Survival theory is proposed as an example of a statistical technique that is consistent with the earlier conceptual structure and allows the exploration of a wide range of the factors that contribute to highway operating risk.

It is hoped that other papers support at least some of the ideas discussed in this paper. The authors believe that once the theoretical and conceptual linkages to statistical methods are clarified, more useful empirical assessments will follow.

REFERENCES

Aranda-Ordaz, Francisco J. "An Extension of The Proportional-Hazards Model for Grouped Data". Biometrics 39, pp. 109-117. March, 1983.

Barrett, G.V., Kobayashi, M., and Fox, B.A. "Driving at Requested Speed: Comparison of Projected and Virtual Image Display." Human Factors, 1968, 10, pp. 259-262.

Bears, J., Case, H.W., and Hulbert, S. "Driving Ability as Affected by Age." UCLA-ITTE Report No. 70-18, 1970. (Cited in Barrett, 1971).

Ceder, Avishal and Livneh, Moshe. "Relationship Between Road Accidents and Hourly Traffic Flow --I." Accident Analysis and Prevention, Vol. 14, No. 1, pp. 19-34, 1982.

Chang, Hsin-Li. "A Disaggregate Survival Model of Motor Carrier Highway Accident Occurrence", Ph.D. Dissertation, Department of Civil Engineering, Northwestern University, June, 1987.

Chirchavaia, Thipatal. et. al., "Severity of Large-Truck and Combination-Vehicle Accidents in Over-The-Road Service: A Discrete Multivariate Analysis." A paper presented at the 63rd Annual Transportation Research Board Conference.

Cleveland, Donald E. and Kitamura, Ryuichi. "Macroscopic Modelling of Two-Lane Rural Roadside Accident." Transportation Research Record 681, pp. 53-62, 1978.

Cox, D. R., "Regression Models and Life-Tables." Journal of the Royal Statistical Society, No. 2, pp. 187-220. 1972.

Cox, D.R. and Oakes, D. Analysis of Survival Data, Monographs on Statistics and Applied Probability. Chapman and Hall, New York, 1984.

Crandall, F., Duggar, B., and Fox, B. "A Study of Driver Behavior During a Simulated Driving Task." Bio-Dynamics, Contract No. PH103-64-79, 1966. (Cited in Barrett, 1971).

Foldvary, L.A. "Road Accident Involvement per Miles Travelled." Accident Analysis and Prevention. Vol. II, pp 75-99. 1979.

Hakkinen, Sucil. "Traffic Accident and Professional Driver Characteristics: A Follow-up Study." Accident Analysis and Prevention, Vol. II, No. 1, pp. 7-18, 1979.

Hall, J. W. and Dickinson, L.V. "Truck Speeds and Accidents on Interstate Highways." Transportation Research Record 486, pp 19-32, 1974.

Hamerslag, R., Roos, J.P. and Kwakernaak, M. "Analysis of Accidents in Traffic Situations By Means of Multiproportional Weighted Poisson Model." Transportation Research Record 847, pp 29-36. 1982.

Haver, E. "Traffic Conflicts and Exposure", Accident Analysis and Prevention, Vol. 14, No. 5, pp. 359-364, 1982.

Heimstra, N.W. "The Effects of Stress Fatigue on Performance in A Simulated Driving Situation." Ergonomics, 1970, 13, pp. 209-218.

Helander, M. "Vehicle Control and Driving Experience: A Psycholphysiological Approach." Proceedings of The 6th Congress of The International Ergonomics Association. College Park, Md., July, 1976.

Hulbert, S. and Wojcik, C. Driving Task Simulation. In T. W. Ferbes (Ed.), Human Factors in Highway Traffic Safety Research. New York: Wiley, 1972.

Institute of Traffic Engineering, Transportation and Traffic Engineering Handbook, Third Edition, 1976.

Ivey, D.L., et. al., "Predicting Wet Weather Accidents." Accident Analysis and Prevention, Vol. 13, pp. 83-99, 1981.

Jovanis, Paul P. and Chang, Hsin-Li. "Modelling the Relationship of Accidents to Miles Traveled." Paper presented at the 65th TRB Annual Meeting, Washington, D.C., January 1986. Forthcoming paper in Transportation Research Record.

Jovanis, Paul P. and Delleur, James. "Exposure-Based Analysis of Motor Vehicle Accident." Transportation Research Record 910, 1984.

Koornstra, Matthijs J. "Multivariate Analysis of Categorical Data with Applications to Road Safety Research." Accident Analysis and Prevention, Vol. 1, pp. 217-221, 1969.

Lavette, Robert A "Development and Application of Railroad-Highway Accident Prediction Equation." Transportation Research Record 628 pp. 12-19, 1977.

Lawless, Jerald F. Statistical Models and Methods for Lifetime Data. John Wiley & Sons, New York, 1982.

Meyers, Warren S. "Comparison of Truck and Passenger-Car Accident Rates on Limited-Access Facilities". Transportation Research Record 808, pp. 48-60, 1981.

Montgomery, D.C. and Peck, E.A. Introductory to Linear Regression Analysis, New York, John Wiley & Sons Inc., 1982.

Oppe, S. "The Use of Multiplicative Models for Analysis of Road Safety Data." Accident Analysis and Prevention, Vol. II, pp. 101-115, 1979.

Perkins, S.R. GMR Traffic Conflicts Techniques -- Procedures Manual. General Motors Research Publication 895, 1969.

Platt, F.N. The Highway Systems Safety Car. Detroit, Michigan, Ford Motor Co., 1970.

Prentice, R.L. and Kalbfleisch, J.D., "Hazard Rate Models with Covariates", Biometrics, 35, pp. 25-39, March, 1979.

Ruijgrok, C.J. and Van Essen, P.G. "The Development and Application of Disaggregate Poisson Model for Trip Generation." Paper Submitted for Presentation at 59th Annual Meeting of TRB, January, 1980.

Salvatore, S., "Effect of Removing Acceleration Cues on Sensing Vehicular Velocity." Perceptual and Motor Skills, 1969, 28, pp. 615-622.

Shinar, David. Psychology on The Road: The Human Factor in Traffic Safety. John Wiley and Sons, New York, 1978.

Shinar, D., Rockwell, T.H., and Malecki, J., "Rural Curves: Designed for The Birds? Or The Effect of Changes in Driver Perception on Rural Curve Negotiation." Paper presented at the 8th Summer Meeting of the Transportation Research Board, Ann Arbor, Michigan, August, 1975.

Snyder, James C. "Environmental Determinants of Traffic Accidents: An Alternate Model." Transportation Research Record 486, pp. 11-18, 1974.

Treat, J.R., Tumbas, N.S., McDonald, S.T., Shinar, D., Hume, R.D. Mayer, R.E., Stansifer, R.L., and Castellan, N.J. Tri-level Study of The Causes of Traffic Accidents. Report No. DOT-HS-034-3-535-77 (ATC), Indiana University, March 1977.