Accident Prediction Models for Freeways

BHAGWANT PERSAUD AND LESZEK DZBIK

The modeling of freeway accidents continues to be of interest because of the frequency and severity of these accidents and the congestion associated with them. Some difficulties with conventional modeling techniques are identified. A distinctive approach is presented, whereby generalized linear modeling is used with both macroscopic and microscopic data to develop regression model estimates of a freeway section's accident potential and an empirical Bayesian procedure is used for refining these estimates.

Freeway accidents are a source of concern not only because of their frequency and severity but also because of the resulting traffic congestion. It is, therefore, not surprising that attention continues to be focused on the modeling of these accidents to identify associated factors and to enable analysts to predict their frequency. Recent papers (*1–3*) are evidence of this ongoing interest. This paper is based on recent research (*4,5*) that applied an accident modeling approach that is somewhat distinct from those used by others.

FREEWAY ACCIDENT MODELING ISSUES

In this section, a number of difficulties with previous models are reviewed. The approach adopted for this paper is then introduced.

The first difficulty with existing models is that they tend to be macroscopic in nature since they relate accident occurrence to average daily traffic (ADT) rather than to the specific flow at the time of accidents. The difficulty with the macroscopic approach is that a freeway with intense flow during rush periods would clearly have a different accident potential than a freeway with the same ADT but with flow evenly spread out during the day, but an ADT-based model would indicate that the two freeways have identical accident potentials.

Second, some modelers assume, a priori, that accidents are proportional to traffic volume and go on to use accident rate (accidents per unit of traffic) as the dependent variable. There is much research to suggest that this assumption is not only incorrect but can also lead to paradoxical conclusions (*6*). Similarly, though accidents should increase with traffic intensity, the model form should not, a priori, assume that accidents are a linear function of traffic volume (*7*).

Third, conventional regression modeling assumes that the dependent variable has a normal error structure. For accident counts, which are discrete and nonnegative, this is clearly not the case; in fact, a negative binomial error structure has been shown to be more appropriate (*8*). Most regression packages in use cannot accommodate such a structure.

Finally, it is impossible for regression models to account for all of the factors that affect accident occurrence. This difficulty can lead to paradoxical conclusions when, as is often done, such models are used to imply cause and effect. Also, when these models are used for accident prediction, the estimates tend to be unreliable if the unexplained variation is relatively large.

The need to overcome these difficulties was fundamental to the modeling approach adopted in the work described in this paper. To this end, use was made of a generalized linear modeling package that allows the flexibility of a nonlinear accident-traffic relationship and a user-specified error structure for the dependent variable and of a complementary empirical Bayesian procedure for improving the accuracy of regression model accident predictions. The approach was applied to both microscopic data (hourly accidents and hourly traffic) and macroscopic data (yearly accident data and average daily traffic).

THEORETICAL ASPECTS OF REGRESSION MODELING

Generalized linear modeling using the GLIM computer package (*9*) was used to obtain a regression model for estimating P, the accident potential per kilometer per unit of time, given a freeway section's physical characteristics, the volume (T) per unit of time, and a set of variables that describe operating conditions during the time period. The model form used was

$$E(P) = aT^b \quad (1)$$

where a and b are model parameters estimated by GLIM. Models were so constructed that the parameters a and b could depend on the values of the factorial variables. This model form ensures that predicted accidents would be zero if there is no traffic, but does not, a priori, assume a linear relationship between accidents and traffic volume. Scatter plots of raw data confirmed that this model form is reasonable for both macroscopic and microscopic models.

The accident count on a section was used as an estimate of the dependent variable. GLIM allows the specification of a negative binomial error structure for a dependent variable, which, as noted earlier, is more appropriate for accident counts than the traditional normal distribution. Although the error structure pertains to the accident counts, a log link function could be specified to allow GLIM to estimate models of the form

$$\ln[E(P)] = \ln(a) + b\ln(T) \quad (2)$$

B. Persaud, Department of Civil Engineering, Ryerson Polytechnical Institute, 350 Victoria Street, Toronto, Ontario M5B 2K3, Canada. L. Dzbik, Department of Civil Engineering and Engineering Mechanics, McMaster University, Hamilton, Ontario L8S 4L7, Canada.

With a negative binomial error specification, it can be shown that the variance of the regression estimates can be estimated from

$$\text{Var}(P) = \text{E}(P)^2/k \quad (3)$$

where the parameter k was estimated using a maximum likelihood procedure that assumes that each squared residual of the regression model is an estimate of Var(P) and that each count comes from a negative binomial distribution with mean E(P) and variance given by Equation 3. This equation indicates that, in comparing two models with the same dependent variable, the one with the larger value of k would give more accurate predictions.

Because ordinary least squares regression was not used, goodness-of-fit of a model could not be assessed in the conventional way, using the coefficient of determination. Instead, goodness-of-fit was assessed by using a generalized Pearson chi-squared statistic (*8,9*) to estimate the amount of variation explained by the systematic component of a model.

MACROSCOPIC MODELS

Data

The data originated in computer files obtained from the Ontario Ministry of Transportation and consists of accident, inventory, and traffic data for approximately 500 freeway sections in Ontario. Some characteristics of the macroscopic data set are summarized in Table 1.

Model Calibration and Results

For each section, the accident count for the each of the years 1988 and 1989 (in effect, the log of this value) was used as an estimate of the dependent variable. To account for varying section lengths, the term log (section length) was specified as an "offset" that GLIM subtracts from each point estimate of ln[E(P)]. Thus, in effect, models were estimated for prediction of the number of accidents per kilometer per year.

Tables 2 and 3 give the estimated regression model coefficients for total accidents and severe (injury and fatal) accidents.

MICROSCOPIC MODELS

Data

The data pertain to a 25-km segment of Highway 401 in Toronto, Canada, part of which has a Freeway Traffic Management System (FTMS). The sections, which range in length from 0.7 to 3 km, are separated by interchanges, and all have express and collector roadways typically with three lanes each per direction.

For the microscopic modeling, it was necessary that conditions pertaining to each data record used in the regression analysis be fairly homogeneous. Thus, it was decided to disaggregate each day into 24 periods of 1 hr each and to derive data for each hour, for express and collector lanes separately, and for day and night. For the accident data this task was straightforward. For the traffic data, it was necessary to derive hourly and seasonal variation factors and collector/express lane distribution factors and apply these factors to the average daily traffic. To maintain a reasonable level of homogeneity, only data pertaining to weekdays were used for the models presented in this paper.

After preliminary data analysis that indicated different accident patterns for congested and uncongested periods, it was decided to build the regression models using, for each section, only data for off-peak hours for which that section tended to be uncongested. To make this determination, we used 5 days of traffic data for sections in the FTMS and used a procedure described elsewhere (*10*). Congested and uncongested hours for a section were identified as hours for which the applicable condition existed on all 5 days. It was assumed that any errors in this process would have a negligible effect since the amount of incorrectly classified data was likely to be relatively small.

TABLE 1 Data Summary for Macroscopic Models

	4-lane	> 4 lanes
Total km	1594	397
1988-89 total accidents	13725	24464
1988-89 severe accidents	4999	7519
ADT (Weighted by length)	19621	87896

TABLE 2 Macroscopic Model for TOTAL Accidents per Kilometer per Year

Model Parameter	Estimated Parameter Adjustment	Standard Error*	Adjusted Parameter Estimate
ln(a) for ADT/1000:			
4 lanes	0 (Base)	0.087	-1.920
> 4 lanes	0.271	0.062	-1.649
***b* for:**			
all lanes	0 (Base)	0.028	1.135
k = 3.52; Variation Explained = 98%; Observations = 1012			

* Applies to coefficient estimate for the base case; otherwise, applies to the adjustments.

TABLE 3 Macroscopic Model for SEVERE Accidents per Kilometer per Year

Model Parameter	Estimated Parameter Adjustment	Standard Error	Adjusted Parameter Estimate
ln(a) for ADT/1000:			
4 lanes	0 (Base)	0.126	-2.776
> 4 lanes	-0.417	0.254	-3.193
***b* for:**			
4 lanes	0 (Base)	0.040	1.082
> 4 lanes	0.124	0.068	1.206
k = 4.55; Variation Explained = 93%; Observations = 1012			

The regression data set contained, for each section and each uncongested hour, the hourly average traffic volume, the number of applicable hours in the 2-year period 1988–1989, a tally for each accident type of interest for those hours, and a code to indicate the light condition (day/night). For some hours (e.g., 6:00 to 7:00 p.m.), it was necessary to have separate sets of data for day and night conditions. A summary of information in the regression data set is given in Table 4.

Model Calibration and Results

For each section, the accident count for the 2-year period 1988–1989 for each uncongested hour (in effect, the log of this value) was used as an estimate of the dependent variable. To account for varying section lengths and number of hours of data, the term ln(section length * number of hours) was specified as an "offset" that GLIM subtracts from each point estimate of ln[E(P)]. Thus, in effect, models were estimated for prediction of the number of accidents per kilometer per hour.

Tables 5 and 6 give the estimated regression model coefficients for total accidents and severe (injury and fatal) accidents. In the calibration process it was found that there was no significant difference between day and night accident frequencies.

As indicated, the estimated coefficients are for a 1-km section for 1 hr. Thus, the regression estimate of total accident potential for a 2-km collector section during an hour with a volume of, say, 8,000 vehicles, is given by $E(P) = 2*e^{-6.276}*8^{0.717} = 0.01671$ accidents/hour and $Var(P) = 0.01671^2/2.59 = 0.000108$.

DISCUSSION OF REGRESSION MODEL RESULTS

Figure 1 shows plots of regression predictions per kilometer per year for the macroscopic models. These plots indicate

TABLE 4 Data Summary for Microscopic Models

	Hourly Volume		Day	Night	Collector Accidents				Express Accidents			
					Severe		Total		Severe		Total	
Hr	Collector	Express	Hrs	Hrs	Day	Night	Day	Night	Day	Night	Day	Night
00	1542	2167	-	521	-	17	-	46	-	14	-	31
01	802	1239	-	521	-	6	-	29	-	9	-	20
02	465	848	-	521	-	6	-	23	-	5	-	19
03	351	646	-	521	-	1	-	11	-	0	-	11
04	446	815	-	521	-	2	-	6	-	3	-	15
05	1382	2128	-	325	-	1	-	3	-	4	-	11
06	6642	9196	44	127	2	2	9	21	2	5	6	16
07	10477	12825	264	-	32	-	101	-	24	-	78	-
08	10316	12015	521	-	73	-	218	-	46	-	177	-
09	7946	9860	521	-	34	-	127	-	31	-	108	-
10	6708	9042	521	-	32	-	111	-	24	-	83	-
11	7012	9023	521	-	33	-	108	-	30	-	84	-
12	6944	8858	521	-	27	-	84	-	20	-	71	-
13	7357	9330	521	-	28	-	89	-	52	-	110	-
14	8153	10062	521	-	45	-	119	-	27	-	106	-
15	9826	11543	521	-	54	-	185	-	60	-	180	-
16	10538	11950	391	-	78	-	231	-	68	-	212	-
17	10047	11042	307	43	65	6	192	28	58	8	161	29
18	8686	10334	220	171	25	24	84	81	23	16	64	65
19	6425	8103	198	260	13	21	34	66	9	20	20	75
20	4683	6126	-	151	-	12	-	32	-	14	-	56
21	4463	5545	-	521	-	26	-	68	-	21	-	74
22	3793	4821	-	521	-	21	-	74	-	27	-	104
23	3057	3845	-	521	-	25	-	55	-	35	-	89
TOTALS					541	170	1692	543	474	181	1460	615

TABLE 5 Microscopic (Off-Peak) Model for TOTAL Accidents per Hour per Kilometer

Model Parameter	Estimated Parameter Adjustment	Standard Error*	Adjusted Parameter Estimate
ln(a) for Vol/Hr/1000:			
Collector	0 (Base)	0.081	-6.276
Express	-0.258	0.077	-6.534
***b* for:**			
Collector/Express	0 (Base)	0.045	0.717
k = 2.59; Variation Explained = 87%; Observations = 684			

* Applies to coefficient estimate for the base case; otherwise, applies to the adjustments.

TABLE 6 Microscopic (Off-Peak) Model for SEVERE Accidents per Hour per Kilometer

Model Parameter	Estimated Parameter Adjustment	Standard Error	Adjusted Parameter Estimate
ln(a) for Vol/Hr/1000:			
Collector	0 (Base)	0.116	-7.608
Express	-0.276	0.098	-7.883
***b* for:**			
Collector/Express	0 (Base)	0.063	0.777
k = 2.30; Variation Explained = 87%; Observations = 684			

that, for the same total traffic volume, four-lane freeways have a lower accident risk than those with more lanes. This result is possibly explained by the tendency for freeways with more than four lanes to be found in urban areas that are generally associated with rush hour congestion and an accompanying greater collision risk.

Figure 2 shows plots of microscopic model regression predictions per kilometer per hour for the two accident types and for express and collector roadways. It is evident that, for a given traffic volume level, collector roadways have a higher accident potential than the express roadways. It is important to note that, for these regression lines, the slope is decreasing as hourly volume increases, perhaps capturing the influence of decreasing speed. This is in contrast to the macroscopic plots in Figure 1, which all show increasing slopes. It is possible that the macroscopic plots are reflecting the increasing probability of risky maneuvers, such as passing and lane changing, with higher ADT levels.

The issue of how accident risk is related to the quality of traffic operation was examined separately. Recall that the microscopic regression models were based on data for hours without congestion. For congested hours, the average hourly traffic volume was calculated along with the average number of accidents per hour per kilometer. Separate values were calculated for the collector and express systems, and in the case of total accidents it was possible to calculate separate

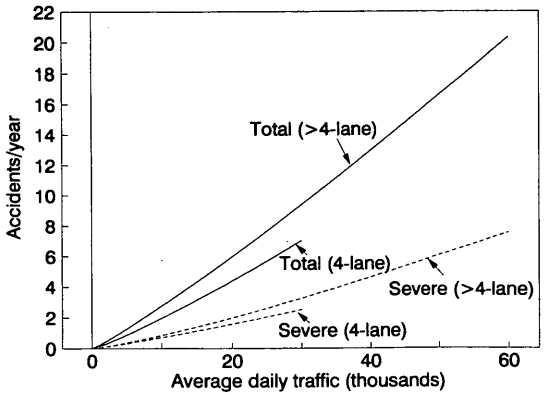

FIGURE 1 Macroscopic regression model predictions.

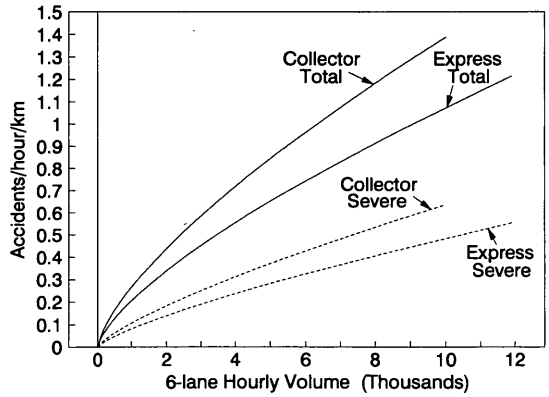

FIGURE 2 Microscopic regression model predictions.

values for morning and evening congested periods. This supplementary work was exploratory in nature, and the results are to be interpreted with caution since they are based on untested assumptions, in particular about the time and location of congestion and about the propriety of using an average traffic volume for congested periods. The results of the analysis of the effect of traffic operation are shown in Figures 3 and 4. Subject to the cautions mentioned, the following conclusions are indicated:

- Congestion is associated with a higher risk of accidents than high-volume uncongested operation.
- The afternoon congested period has a higher accident risk than the morning rush period, but the difference is only significant for the express system.
- Collector system congestion is associated with a higher accident risk than express system congestion.

ACCIDENT PREDICTION—REFINEMENT OF REGRESSION MODEL ESTIMATES

It is now accepted among safety analysts that the underlying long-term accident potential, rather than the commonly used short-term count, is more proper for identifying unsafe sections and for evaluating safety effectiveness of improvements. Regression model predictions have been used as an estimate of this value, but the difficulty with this is that, in general, two road sections that are similar in all of the independent variables used in a regression model will still be different in true accident potential even though they will have the same model predictions. This, in turn, is because it is not possible to account in the regression model for all the factors that cause differences in accident potential (e.g., weather, geometrics).

To mitigate this problem, use can be made of an empirical Bayesian technique that combines the regression prediction with the observed short-term accident count for a section of interest. This method has previously been used in estimating the long-term accident potential of rail-highway grade crossings (*11*), Toronto intersections (*12*), and Ontario drivers (*13*) and road sections (*4*).

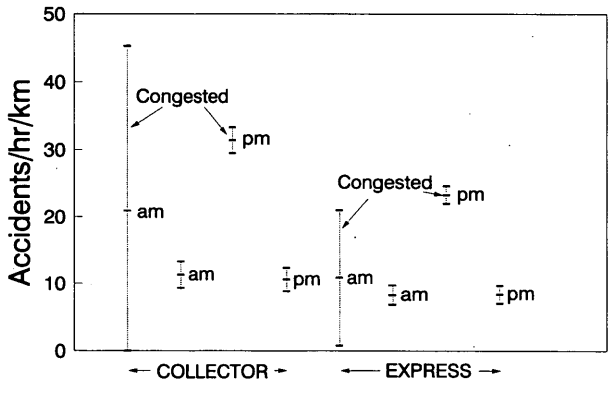

FIGURE 3 Point estimates (with 95 percent confidence intervals) of accident potential during high-volume operation—TOTAL ACCIDENTS.

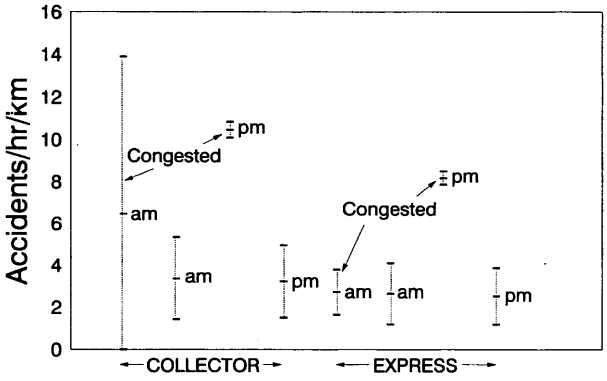

FIGURE 4 Point estimates (with 95 percent confidence intervals) of accident potential during high-volume operation—SEVERE ACCIDENTS.

Theory

Using the empirical Bayesian procedure, E(P) from Equation 1 can be refined for an individual road section using the accident count, x, in n units of time (years in the case of macroscopic models and hours for the microscopic case) on that section to give E($m|P, x, n$), a revised estimate of accident potential. It can be shown (*11*) that, under reasonable assumptions, the revised estimate of accident potential (per unit of time) is

$$\mathrm{E}(m|P, x, n) = q[w\mathrm{E}(P) + (1 - w)x] \quad (4)$$

where

E(P) = regression estimate for one unit of time,
w = [1/(1 + Var(P)/E(P))] = [1/(1 + E(P)/k)], and
q = [(1 + E(P)/k)/(1 + nE(P)/k)].

It can also be shown that the variation in ($m|P, x, n$) can be estimated by

$$\mathrm{Var}(m|P, x, n) = \{\mathrm{E}(m|P, x, n)/[n + k/\mathrm{E}(P)]\} \quad (5)$$

Equation 4 indicates that the estimated accident potential of a section is a combination of what is observed, x, and of E(P)—what is predicted on the basis of its characteristics (traffic volume, etc.).

Illustration

Suppose the collector section in the earlier example recorded six accidents in 80 hr with an average hourly traffic volume of 8,000. Thus, in Equation 4, x = 6 and n = 80. Recall from the earlier example that E(P) = 0.01671, Var(P) = 0.000108, and k = 2.59. Thus, for Equations 4 and 5, q = 0.6638 and w = 0.9936.

These values give the Bayesian estimate of accident potential as E($m|P$, 6, 80) = 0.03651 accidents per hour and Var($m|P$,

TABLE 7 Macroscopic Model Validation Results—Mean Squared Difference Between Predicted and Observed Accidents per Kilometer per Year

Estimation method	Total accidents	Severe accidents
Accident count	32.2	5.94
Regression model	31.9	5.84
Empirical Bayesian	23.0	3.78

TABLE 8 Microscopic Model Validation Results—Mean Squared Difference Between Predicted and Observed Accidents per Hour-Kilometer

Estimation method	Total accidents ($*10^{-6}$)	Severe accidents ($*10^{-6}$)
Accident count	0.928	3.573
Regression model	0.771	2.682
Empirical Bayesian	0.743	2.469

6, 80) = 0.03651/(80 + 2.59/0.01671) = 0.000155. Note that the accident potential estimate is between the regression estimate (0.01671) and the observed accidents per hour (6/80 = 0.075).

Validation

The validation of the overall approach involved a comparison of the prediction accuracy resulting from the use of 1988 accident counts or regression predictions as estimates of the 1987 counts, as opposed to using the empirical Bayesian procedure based on 1988 data.

For the macroscopic case, this required the calculation of the mean squared difference between estimated and observed 1987 total and severe accidents for each of approximately 1500 km of freeways for each of the three estimation methods. It is assumed that the better estimate is the one with the smallest mean squared difference. The results of the validation exercise for the macroscopic models are given in Table 7.

For the microscopic case, validation required the calculation of the squared difference between estimated and observed 1987 counts per squared hour-kilometer for each cell (see Table 4) for each section, and, in essence, averaging this value over all sections and cells. The results of the validation exercise for the microscopic models are given in Table 8.

The results in both cases show that the empirical Bayesian method appears to be best followed, as expected, by the regression model prediction method.

ACKNOWLEDGMENTS

The research for this paper was supported under an operating grant from the Natural Sciences and Engineering Council of Canada. The assistance of the Ministry of Transportation of Ontario in providing the data and in funding related research is greatly appreciated.

REFERENCES

1. Y. Huang, R. Cayford, and A. D. May. Accident Prediction Models for Freeway Segments. Presented at 71st Annual Meeting of the Transportation Research Board, Washington, D.C., 1992.
2. Kraus et al. Epidemiological Aspects of Fatal and Severe Injury Urban Freeway Crashes. *Accident Analysis and Prevention* (forthcoming).
3. A. Ceder and M. Livneh. Relationships Between Road Accidents and Hourly Traffic Flow—I. Analysis and Interpretation. *Accident Analysis and Prevention*, Vol. 14, No. 1, 1982, pp. 19–34.
4. B. N. Persaud. *Accident Potential Models for Ontario Road Sections*. Ministry of Transportation of Ontario, March 1991.
5. L. Dzbik. *Accident Prediction Models for Freeways*. Master's thesis. McMaster University, Hamilton, Ontario, Canada, 1992.
6. D. Mahalel. A Note on Accident Risk. In *Transportation Research Record 1068*, TRB, National Research Council, Washington, D.C., 1985.
7. S. P. Satterthwaite. *A Survey of Research into the Relationships Between Traffic Accidents and Traffic Volumes*. Transport and Road Research Laboratory Supplementary Report 692, United Kingdom, 1981.
8. E. Hauer, J. Lovell, and B. N. Persaud. *New Directions for Learning About Safety Effectiveness*. Report FHWA/RD-86-015. Federal Highway Administration and National Highway Traffic Safety Administration, Jan. 1986.
9. R. J. Baker and J. A. Nelder. *The GLIM System—Release 3*. Rothamsted Experimental Station, Harpenden, United Kingdom, 1978.
10. L. Aultman-Hall and F. L. Hall. *Demonstration of a Method Using Only Detector Data To Evaluate the Effectiveness of Highway 401 FTMS*. Department of Civil Engineering, McMaster University, Nov. 1991.
11. E. Hauer and B. N. Persaud. How To Estimate the Safety of Rail-Highway Grade Crossings and the Safety Effect of Warning Devices. In *Transportation Research Record 1114*, TRB, National Research Council, Washington, D.C., 1986.
12. E. Hauer, J. C. Ng, and J. Lovell. Estimation of Safety at Signalized Intersections. In *Transportation Research Record 1185*, TRB, National Research Council, Washington, D.C., 1988.
13. A. Smiley, B. N. Persaud, E. Hauer, and D. Duncan. Accidents, Convictions and Demerit Points—An Ontario Driver Records Study. In *Transportation Research Record 1238*, TRB, National Research Council, Washington, D.C., 1989, pp. 53–64.

Publication of this paper sponsored by Committee on Traffic Records and Accident Analysis.