Geographic Information System–Based Integrated Model for Analysis and Prediction of Road Accidents

F. Frank Saccomanno, Liping Fu, and Rajeev K. Roy

The applicability and reliability of accident analysis and prediction models depend on their ability to integrate relevant input from disparate databases in a seamless and automated manner. These inputs include information on road geometry, traffic composition, accident profiles, and spatial referencing. With powerful functionality in spatial referencing, data management, and visualization, geographic information systems (GISs) provide a natural platform for this type of model. An integrated and user-friendly GIS platform for road accident analysis and prediction is described. To demonstrate this platform, it has been applied to safety problems specified at different levels of spatial aggregation, from individual route sections to the overall network. The model was developed by using databases obtained from the Ontario Ministry of Transportation.

Models of road accident prediction require input from a large number of disparate databases, including information on road geometry, traffic volume, accidents, and weather conditions. These databases are collected by different agencies for essentially different purposes. As a result, they tend to lack a common referencing system needed for their integration into accident prediction models. The applicability and reliability of these models depend to a large extent on the ability to integrate these relevant databases in a seamless and automated manner.

With powerful functionality in spatial referencing, data management, and visualization, geographic information systems (GISs) provide a natural platform for the analysis of road accidents (*1*). As a result, many road safety organizations have introduced GIS into their overall road safety management program (*2, 3*). However, existing GIS road safety models are limited to accessing information directly from the raw databases or to drawing inference from overly simplified models regarding the potential for accidents at a given location or route.

This paper describes a GIS-based integrated model of road accident analysis and prediction. This model predicts accidents at different levels of spatial aggregation as specified by the analyst for different problems, and it provides a user-friendly interactive interface with which to develop and evaluate alternative safety countermeasures.

MODEL FRAMEWORK

As illustrated in Figure 1, the main features of the described GIS model are as follows:

- An integrated relational database management system (RDBMS),
- The ability to query to an integrated RDBMS database directly and provide relevant statistics,
- An accident prediction and analysis module applied to different route locations, and
- A GIS platform for visual display of spatial analysis.

This model makes use of GIS ArcView and the Visual Basic program language as the architecture of the GIS road accident model. ArcView was developed by ERIS, Inc.

The GIS platform enables the user to appreciate visually the results of any analysis on the integrated data, whether to predict accidents, analyze their attributes or pattern, or underscore the relationships that give rise to these accidents at a given location at a specific point in time. With the functionality of visualizing the results, the GIS accident prediction module shown in Figure 1 serves as a useful tool for making informed decisions on how to reduce accidents at specific locations, along specific routes, or over the entire road network.

In this study, GIS has been linked to a user-friendly application developed in Visual Basic. The application makes use of a sample of road sections from the Ontario highway network to illustrate the usefulness of the model as a decision-support tool for road accident reduction. Various branches of the Ontario Ministry of Transportation (MTO) collect and manage these databases.

ACCIDENT PREDICTION MODULE

The accident prediction module in the GIS model illustrated in Figure 1 establishes the long-term potential for accidents at a specific location for a given period. The pattern of these accidents can then be investigated to provide insight into how they are caused and what can be done to reduce incidence. Many of the existing GIS road safety models are limited to accessing information directly from the raw databases and to drawing inference from these data to predict the probability of an accident taking place at a given location or route (*4*).

Unfortunately, because of the rare and random nature of accidents, inferring a potential for accidents at a given location solely from the historical accident data will not always yield consistent long-term results. Most locations do not experience many accidents in any given year. Observations tend to be too infrequent and too variable to yield meaningful and reliable long-term analysis. The GIS model presented here uses two statistical accident prediction methods to establish the long-term potential for accidents at a given location. These are the Poisson regression model and the empirical Bayesian (EB) model.

F. F. Saccomanno and L. Fu, Department of Civil Engineering, University of Waterloo, 200 University Avenue West, Waterloo, Ontario N2L 3G1, Canada. R. K. Roy, IBI Group, 5th Floor, 230 Richmond Street West, Toronto, Ontario M5V 1V6, Canada.

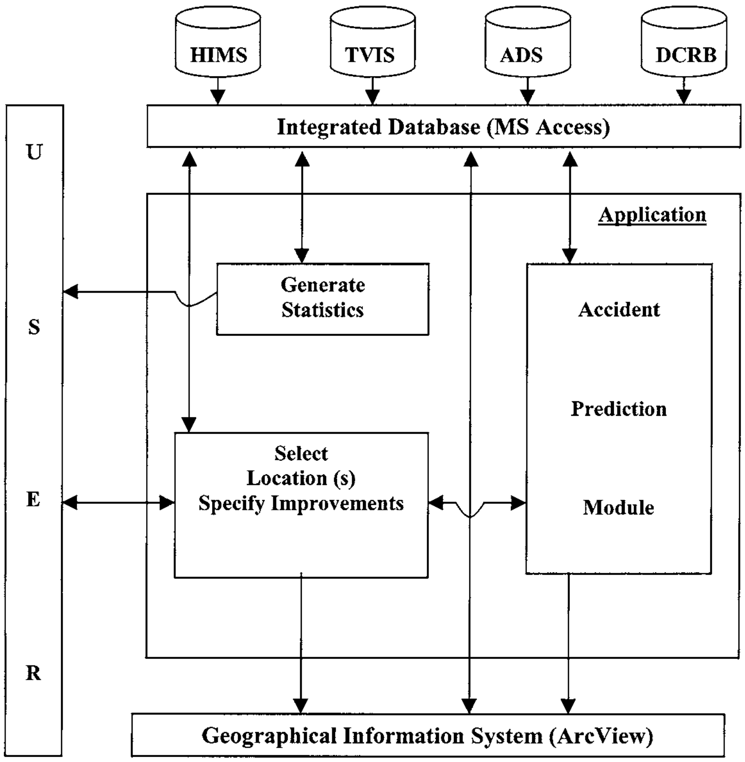

HIMS
TVIS
ADS
DCRB
Integrated Database (MS Access)
U S E R
Application
Generate Statistics
Accident Prediction Module
Select Location (s) Specify Improvements
Geographical Information System (ArcView)

FIGURE 1 GIS-based road accident model.

Both the Poisson and EB models relate accident potentials at specific road locations to various contributing factors.

For the GIS road accident module considered here, a route section–specific Poisson regression model developed by Nassar (*5*) and Nassar et al. (*6*) and based on the same Ontario data was used. This model is of the form

$$E(m)_i = \text{ADTL}^{1.242}\text{LEN}^{0.696} \exp(0.1955\ \text{LN} - 0.1775\ \text{SHW} + 0.2716\ \text{MT2} + 0.5669\ \text{TS} - 0.1208\ \text{PTC} - 0.0918\ \text{Y91}) \quad (1)$$

where

$E(m)_i$ = expected accident frequencies on road section *i*,
ADTL = annual average daily traffic (AADT) per lane in thousands of vehicles on road section *i*,
LEN = length of road section *i* (km),
LN = number of lanes on road section *i*,
SHW = shoulder width of road section *i* (m),
MT2 = median type two of road section *i* (0 = painted, 1 = barrier),
TS = traffic signal on road section *i* (0 = no, 1 = yes),
PTC = pattern type commuter on road section *i* (0 combined, 1 = commuter), and
Y91 = year 1991 (0 = 92, 1 = 91).

Statistical prediction models, such as the Poisson regression models, frequently are plagued by poor specification. Information on factors affecting variations in accident potential at a given location is often incomplete and is insufficient to adequately explain differences in the potential for accidents from year to year. When such models are applied to a given accident database, they often lead to overdispersion error.

The accident model proposed by Nassar and Nassar et al. is useful and applicable when the independent variables with respect to which the proposed analysis is desired are included in the model (*5, 6*). Therefore, to analyze accident involvement based on a random variable not included in the model, a different method must be adopted. The empirical Bayesian method, explained later, was used for this purpose.

Hauer and Persaud suggested that if the expected number of accidents on each road section can be described by a gamma probability distribution, the count of accidents should obey a negative binomial distribution (*7–10*). Dean and Lawless suggested that negative binomial distribution models are most suitable for dealing with count data that display extra-Poisson variation (*11*). In this case, the variation is proportional, rather than equal, to the mean. The GLIM user's guide (*12*) suggests an expression for the variance of the form

$$\text{Var}(X)_i = E(m)_i + [E(m)_i]^2/k \quad (2)$$

where

$\text{Var}(X)_i$ = variance in accident frequencies for road section *i*,
$E(m)_i$ = expected accident frequencies on road section *i* (model estimates), and
k = dispersion parameter.

For sections that behave in a Poisson manner, assume that $\text{Var}(X)_i = E(m)_i$ in Equation 2. The value of the dispersion parameter in this equation is unknown and needs to be established as a preliminary step in accounting for extra-Poisson variation in the accident data.

In fitting the extra-Poisson model, weights are assigned to points (Poisson fitted values) in proportion to the ratio of the Poisson model variance $[E(m)_i]$ over the extra-Poisson variance in Equation 2 such that

$$W_i = \frac{E(m)_i}{E(m)_i + [E(m)_i]^2/k} = \frac{1}{1 + E(m)_i/k} \qquad (3)$$

The weights in Equation 3 are given to each point during fitting so that the variances of the individual points are divided by these weights (*12*). The model can fit as a Poisson model by using a quasi-maximum-likelihood method to estimate the Poisson parameters.

For an individual road section *i* in group *j*, the Bayesian adjusted or estimated number of accident involvement per year ϵ_i is expressed as a combination of $E(m)_i$, the estimated/predicted number of accident involvement based on group *j* to which road section *i* belongs and X_i the observed number of accident involvement for road section *i*, such that

$$\epsilon_i = W_i E(m)_i + (1 - W_i)X_i \qquad (4)$$

The term $E(m)_i$ in Equation 4 represents the estimated/predicted number of accident involvement for road section *i* averaged over all road sections in group *j*. As already stated, $E(m)_i$ is assumed to be gamma distributed and X_i to have negative binomial distribution. The parameter W_i reflects the extent to which the group estimated number of accident involvement and the observed number of accident involvement for a given road section are combined to yield the adjusted expectation of number of accident involvement for this road section *i*. W_i is obtained as explained in Equation 3.

Nassar et al. investigated the presence of Poisson overdispersion in the Ontario accident data and suggested incorporating an EB adjustment factor (*6*). In this paper, both the Poisson and the EB adjusted estimates were used to reflect the long-term potential for accidents at individual locations or route sections.

By comparing the potential for accidents from either the Poisson regression or EB prediction models with the observed number of accidents at a given location, it can be determined if certain locations should be designated as unsafe; these are referred to as black spot (BS) locations. The approach for identifying BS is discussed at length by Persaud (*9*) and Chong (*13*). A thorough understanding of the causes and consequences of accidents at BS locations should guide decisions on what safety countermeasures should be implemented at these locations.

DATA SOURCES AND THEIR INTEGRATION

As discussed, accident analysis and prediction models require a wide variety of data on road geometry, traffic composition, weather conditions, and accidents. The integration of these disparate databases requires a formal treatment of errors and inconsistencies so that they can be combined in a spatially consistent manner.

Relevant MTO Databases

MTO uses a linear highway reference system (LHRS) to uniquely identify a continuous length of highway with similar geometric and traffic characteristics. Typically, each LHRS section has a length of 0.2 to 18.2 km. An offset distance is used to assign a road accident to a given point location on the LHRS section. The offset is measured from a known section feature point (e.g., a bridge overpass) to point of occurrence of the accident. This known feature point is measured at some distance from the beginning of the LHRS section in which the feature is situated. The LHRS number and the offset distance represent the most disaggregate spatial referencing system for the road network that are presently used in accident reporting.

Four MTO databases are considered relevant in this GIS-based accident model:

- Digital cartographic reference base (DCRB),
- Accident data system (ADS),
- Highway inventory management system (HIMS), and
- Traffic volume inventory system (TVIS).

A sample of information available in these databases is provided in Table 1.

The DRCB is a geocoded database that contains information on road network and other features, such as rail network, lakes, parks, rivers, and streams, in Ontario. Only the road data were taken from this database. This database is in GIS format and can be used for viewing with ArcView.

In Ontario, accident data are collected by police and are compiled yearly by the MTO. The ADS data are stored in the following separate formats:

- Basic accident record: contains information that is unique to each accident, such as date, time, location, number of vehicles and persons involved, number of fatalities, road conditions, and several other details. Each accident is identified by a unique nine-digit number (accident microfilm number) and the LHRS number (referred to as the key point number in ADS).
- Driver and vehicle record: contains information unique to each driver and vehicle involved in the accident, such as plate number, year and make of vehicle, driver license number, driver date of birth,

TABLE 1 Data Available in Different MTO Databases

DATABASE	*VARIABLES*
DCRB	Geo-coded X and Y coordinates of LHRS, Description of starting point, Section length, Route # (for use with GIS software), etc.
ADS **Basic Record**	Accident microfilm number, Accident date and time, Classification of accident, Total driver vehicles/involved persons/Fatalities, Keypoint number (same as LHRS number), Location of accident in reference to a feature point, Road condition, Road type, Environment, etc.
ADS **Driver/Vehicle Record**	Accident microfilm number, Vehicle number, Vehicle make and model, Number of occupants, Driver's license number, Driver's age, Sex of the driver, Result of breath test, Vehicle condition, Approximate speed, Vehicle damage level, etc.
ADS **Involved Person Record**	Accident microfilm number, Involved person number, Vehicle number, Age of the involved person, Sex of the involved person, Injuries, Pedestrian, etc.
HIMS	Road LHRS number, Offset km, Direction of stream, From location, To location, Section length, Highway number, Surface width and type, etc.
TVIS	Existing and projected annual average daily traffic, Summer and winter AADT, Percent commuter traffic, Directional split, etc.

action of driver, damage caused, and condition of the vehicle. Each record is uniquely identified by the driver or vehicle number and the corresponding accident microfilm number.

- Involved person(s) record: contains information that is unique to each occupant of every vehicle involved in an accident. This includes information on injury sustained, seating position in the vehicle, age and condition of the occupant, and use of seat belt. Each record is uniquely identified by the person number and the corresponding accident microfilm number.

Table 1 provides some important variables included in the ADS. Each record included is uniquely identified by the key point number (same as the LHRS number) and the accident microfilm number.

The HIMS database contains information about the geometric features of each LHRS section, including length of each section (subsection), from and to locations, number of lanes, road width, shoulder width, median width, type of shoulder, type of median, and posted speed. A sample of the information available in HIMS is shown in Table 1.

The TVIS database contains information regarding existing traffic volume, projected future traffic volume, summer and winter traffic volumes, directional split, and percent commuter traffic.

Treatment of Errors and Inconsistencies in Integrating Databases

The DCRB database as developed by MTO was geocoded in a spatially usable format for input into GIS. The geocoded DCRB database contains a set of homogeneous highway sections, which are uniquely identified by their LHRS number, and a unique pair of *X* and *Y* coordinates representing the beginning point of the LHRS section. This information combined with section length can be used to identify the section (*X*, *Y*) endpoint coordinates. For this analysis, it was assumed that all sections are linear.

It was assumed that the information available in the DRCB database is correct and all other databases will be edited or modified per the information available in this database.

HIMS

In matching HIMS to DCRB, three types of error were investigated:

- Type 1 error. Some LHRS numbers of DCRB were missing in the HIMS database, but the section lengths of the adjoining LHRS sections in HIMS indicate that the two records of DCRB have been merged into one record in HIMS. Therefore, the record in HIMS was split into two records.
- Type 2 error. A few LHRS lengths were unequal in DCRB and HIMS, but the combined length of adjacent LHRS numbers in DCRB had the same length as that of HIMS, and thus the lengths were adjusted accordingly in HIMS.
- Type 3 error. Lengths of a few LHRS sections in HIMS did not exactly match those of DCRB, and the lengths were adjusted per the lengths indicated in DCRB.

Tables 2 through 4 indicate how each error was considered in developing an integrated database for input into GIS for several sample sections in the network.

It should be noted that the preceding adjustments for HIMS errors do not follow a definite pattern, precluding the possibility of correcting these errors through an automated process. The Microsoft Access subform feature was used to expedite this process. The HIMS database was linked to the DCRB database as a child form and fed all the corresponding records for a given LHRS number, which made the editing easy.

TABLE 2 Type 1 Error in HIMS and Adjusted Values

Original Record				
Hwy #	LHRS #	Length in DCRB	Length in HIMS	Offset Distance
1	10014	2.1 km	2.1 km	0 km
1	10017	4.5 km	5.4 km	0 km
1	10020	0.9 km	▨	▨
1	10022	3.4 km	3.4 km	0 km
Record After Adjustment				
1	10014	2.1 km	2.1 km	0 km
1	10017	4.5 km	4.5 km	0 km
1	10020	0.9 km	0.9 km	0 km
1	10022	3.4 km	3.4 km	0 km

TABLE 3 Type 2 Error in HIMS and Adjusted Values

Original Record				
Hwy #	LHRS #	Length in DCRB	Length in HIMS	Offset Distance
402	48115	12.3 km	12.3 km	0 km
402	48120	3.5 km	9.1 km	0 km
402	48123	12.7 km	7.1 km	0 km
402	48127	4.3 km	4.3 km	0 km
Record After Adjustment				
402	48115	12.3 km	12.3 km	0 km
402	48120	3.5 km	3.5 km	0 km
402	48123	12.7 km	12.7 km	0 km
402	48127	4.3 km	4.3 km	0 km

TABLE 4 Type 3 Error in HIMS and Adjusted Values

Original Record				
Hwy #	LHRS #	Length in DCRB	Length in HIMS	Offset Distance
401	47970	6.25 km	6.3 km	0 km
401	47980	6.24 km	6.3 km	0 km
401	47990	7.08 km	7.1 km	0 km
401	47800	7.01 km	7.1 km	0 km
Record After Adjustment				
401	47970	6.25 km	6.25 km	0 km
401	47980	6.24 km	6.24 km	0 km
401	47990	7.08 km	7.08 km	0 km
401	47800	7.01 km	7.01 km	0 km

TVIS

Errors in the traffic volume database were found to be similar to those in HIMS. The only exception was that traffic records for a few LHRS sections were missing.

Because of a lack of other information, traffic volumes on these sections were assumed to be the same as those of the adjacent section. For TVIS, the manual adjustment process was used.

ADS

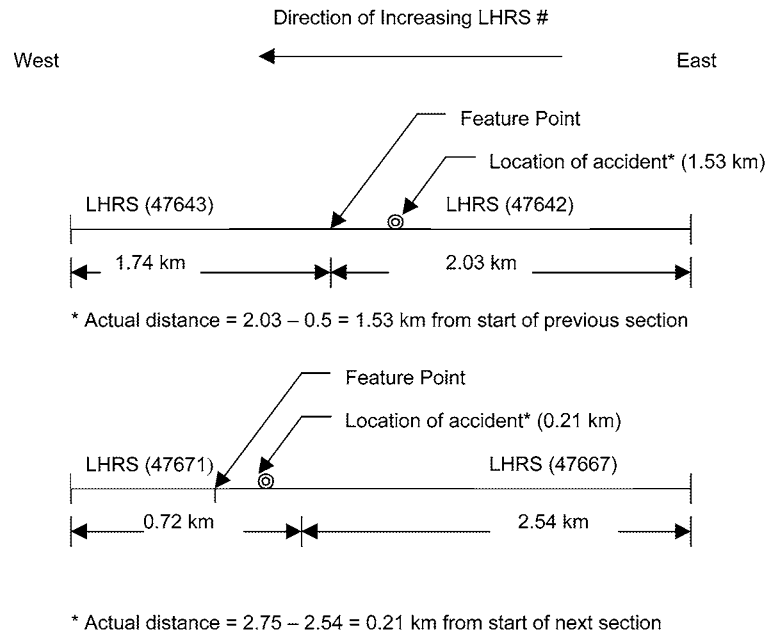
The modification of the accident database was automated by using a Visual Basic program that assigned the accident to the correct LHRS and also calculated the location of the accident from the start of the highway. This was required for plotting on the map with GIS ArcView software.

The ADS database contained two types of errors that were considered in integrating the databases for GIS input. In one error, there was an unmatched key point number when the record was compared to the DCRB. Because there was no way to ascertain the correct LHRS number, these records were deleted from the accident database. Occurrence of this kind of record varied from 1 to 2 percent, and it is assumed that the deletion will not have a significant effect on the result. In the other error, an accident was attributed to a wrong key point number (LHRS), because the distances mentioned for the location of the accident from the start of the key point were based on the mileage of the feature point and distance and the direction of the accident from the feature point. Therefore, it was necessary to establish the direction in which the LHRS numbers increased so that the exact location of the accident on the highway and the correct LHRS number to which the accident is attributed could be determined. This was necessary for locating the accident in ArcView. Figure 2 demonstrates how the accidents were allocated to the correct LHRS number.

ILLUSTRATIVE EXAMPLES

To illustrate the GIS model, five types of transportation query were postulated:

MICROFILM #	KEYPOINT #	HWY #	SECTION LENGTH (km)	OFFSET KM OF FEATURE POINT (km)	DIRECTION FROM FEATURE POINT	DISTANCE FROM FEATURE POINT (km)	DISTANCE FROM START OF LHRS (km)
Original Records							
520179	47643	401	1.74	0	East	0.5	-0.5
811952	47667	401	2.54	2.8	East	0.05	2.75
Adjusted Records							
520179	47642	401	2.03	0	West	1.53	1.53
811952	47671	401	0.72	0	West	0.21	0.21



* Actual distance = 2.03 – 0.5 = 1.53 km from start of previous section

* Actual distance = 2.75 – 2.54 = 0.21 km from start of next section

FIGURE 2 Allocation of accident to correct LHRS number.

- Simple spatial or general query of the integrated database to retrieve attributes of roadway network and accident pattern,
- Generating accident statistics for the selected location(s),
- Predicting accident potential for the selected location by using sources for the estimates (Poisson and EB models or observed),
- Designating safety BS and assessing the effect of safety countermeasures, and
- Visualizing the results of analysis spatially.

The following sample highway section was selected for analysis:

- Highway number: 401;
- From location: Highway 404 and Don Valley Parkway Interchange;
- To location: North York, Leslie Street and IC-373;
- LHRS number: 47635;
- Number of lanes: 12;
- Section length: 2.01 km; and
- AADT (1992):276,500.

Figures 3 and 4 illustrate a few trends in accident experience, which can be generated for the preceding selected route sections along Highway 401. Figure 3 illustrates the yearly variation in the observed number of accidents from 1990 to 1993 for the highway section considered in this application. In general, between 600 and 700 accidents can be expected per year along this stretch of Highway 401. The total number of accidents per year does not vary appreciably from year to year for the period 1990–1993.

Figure 4 illustrates the month-to-month variation in observed number of accidents for the period 1990–1993. If one assumes that travel exposure (vehicle-kilometers per month) does not vary much from month to month in a given year, one may note that the number of accidents tends to be higher during November and December. This could reflect the onset of winter driving conditions in southern Ontario, which is expected to increase the potential for accidents.

After generating and analyzing accident statistics for the selected locations, the expected number of accidents at the selected locations for the period of interest can be estimated. Figure 5 provides an indication of the observed and expected number of accidents along the selected section of Highway 401 based on the two prediction models, Poisson regression and EB. As expected, the number of accidents predicted by the EB model is closer to the observed number than are the values predicted by the Poisson model. Although this is true for the entire selected section, it does not necessarily hold for individual sections.

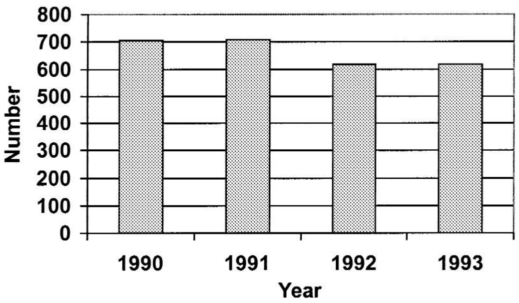


FIGURE 3 Yearly variation in observed number of accidents.

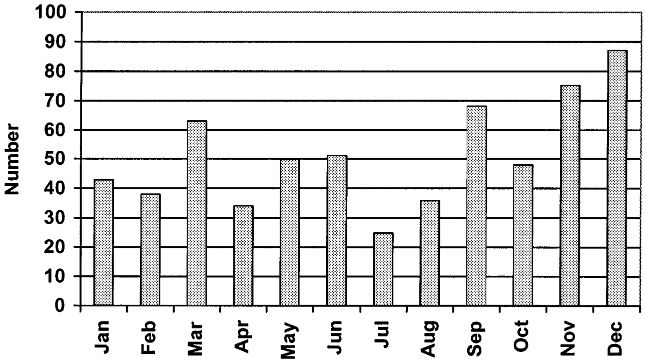


FIGURE 4 Monthly variation in observed number of accidents, 1992.

Again for the purpose of illustration, a number of BS sections have been designated along the selected highway (401). The output from this analysis is presented in scaled map form. A BS section is defined as any section where the observed number of accidents exceeds the predicted number by at least one standard deviation from either the Poisson or the EB model estimate. Figure 6 shows the Poisson model–designated BS sections along the test highway, and Figure 7 shows the BS sections from the EB model.

By comparing Figures 6 and 7, it can be seen that the EB model yielded fewer BS than did the Poisson regression model. The EB model BS sections are largely included in the Poisson BS model BS sample.

Another way to designate BS is to establish sections where fatal accidents were observed. Figure 8 illustrates sections of the 400-level highways in southern Ontario where fatal accidents were observed in 1992. For the Highway 401 test section, it can be seen that many sections with fatal accidents fall within sections that have been classified as BS by either the Poisson regression or the EB model.

CONCLUSIONS

This paper described a GIS-based model for road accident prediction and analysis. This model provides a seamless platform for integrating and validating disparate data sources. Unlike many existing transportation GIS models, the model presented here predicts accidents by using the state-of-the-art methods applied at a different level of spatial aggregation as specified by the analyst.

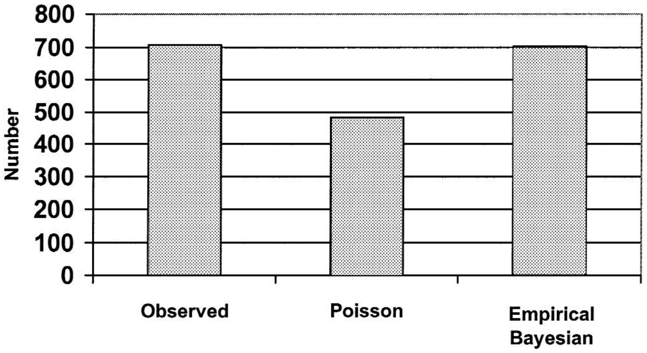


FIGURE 5 Model estimated/predicted number of accidents, 1992.

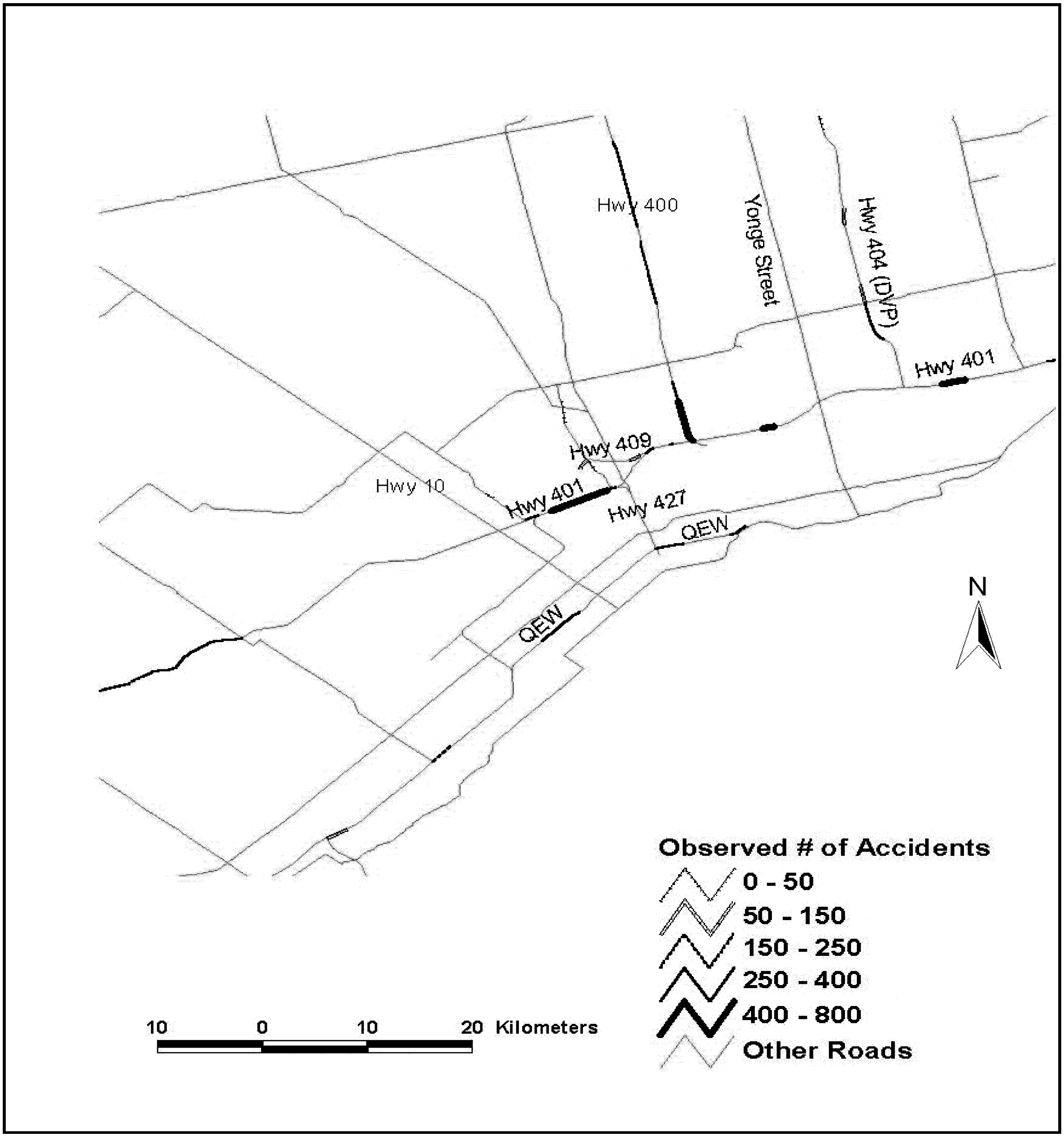


FIGURE 6 Black spot sections as predicted by Poisson model.

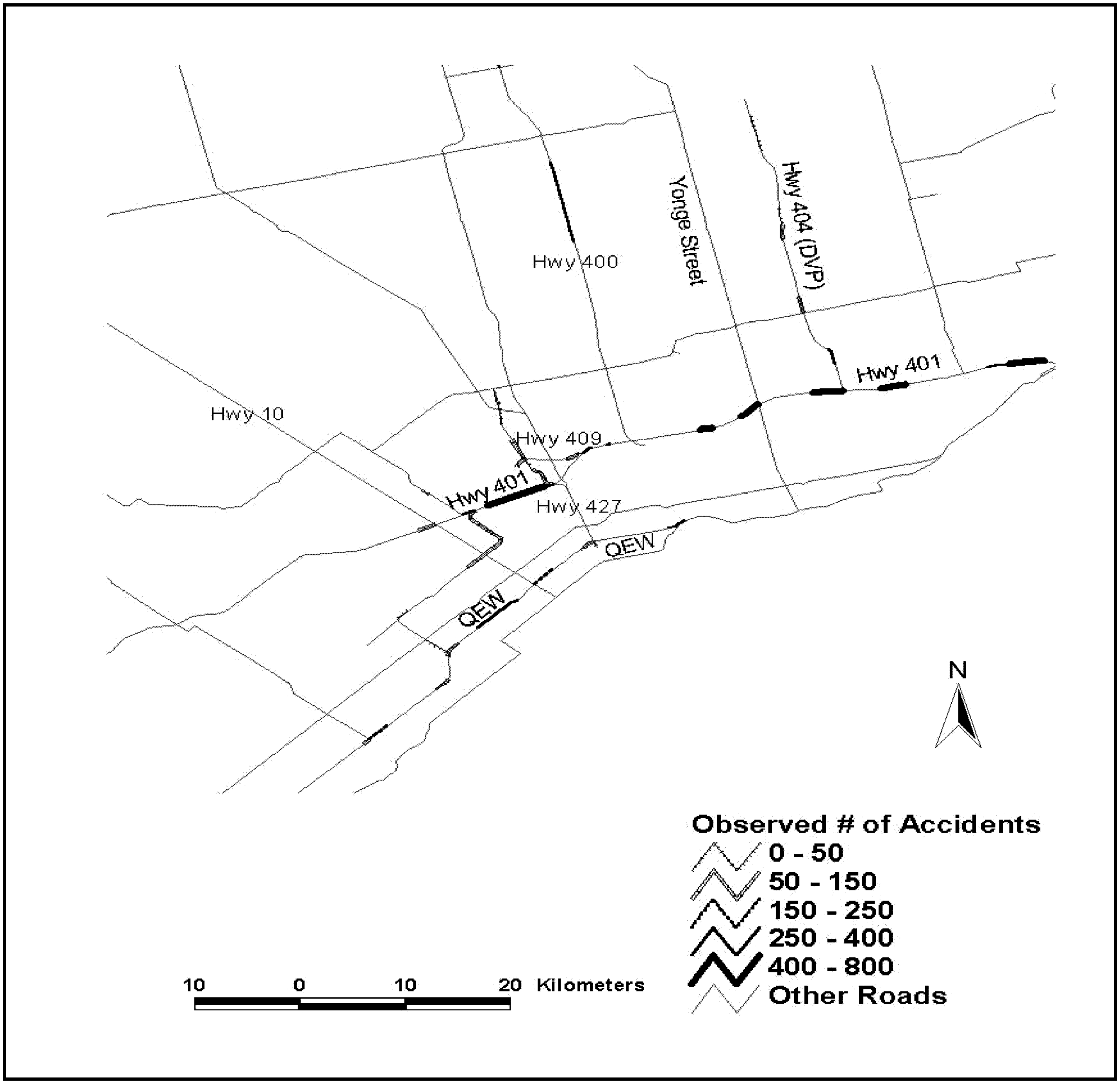


FIGURE 7 Black spots as predicted by EB model.

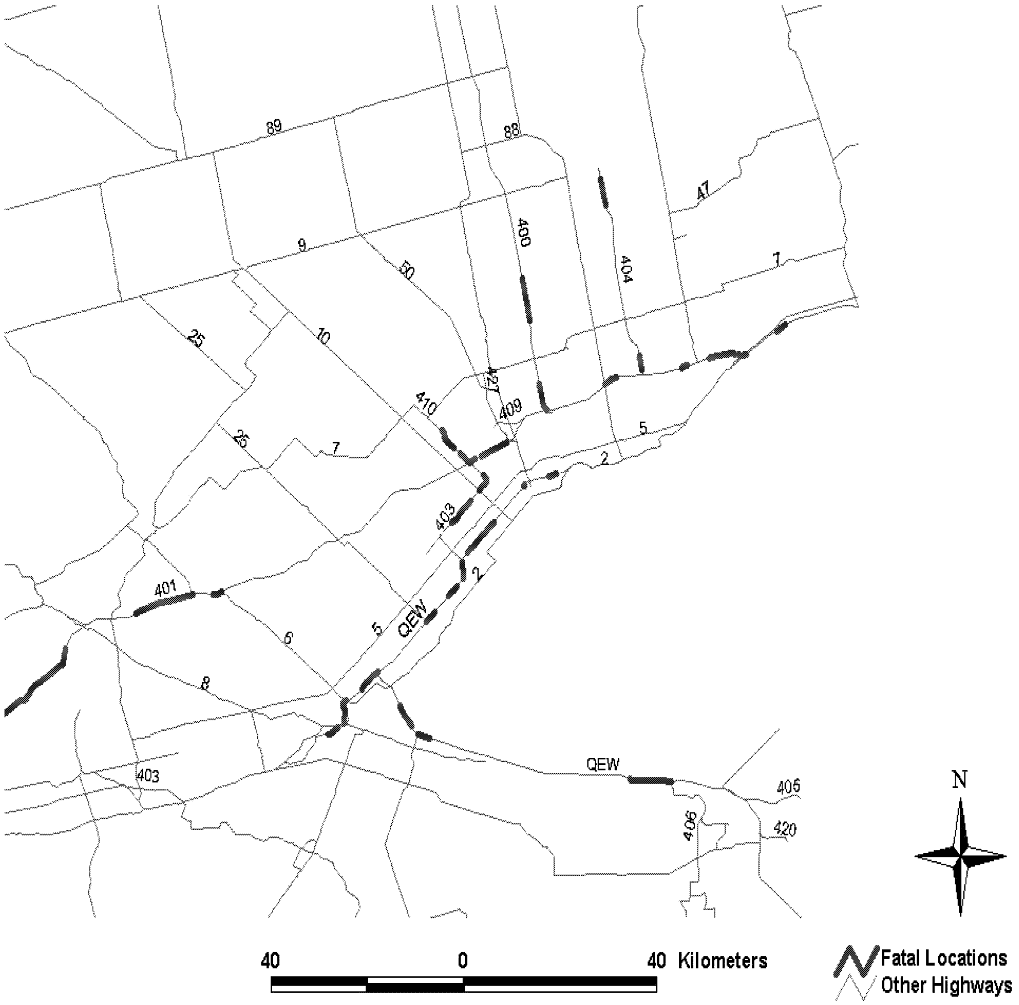


FIGURE 8 Fatal spots on 400-level highways in 1992.

The model can be used to evaluate the effectiveness of alternative safety countermeasures designed to reduce accidents at unsafe locations or routes.

A comparison was included of two methods for predicting accidents and designating BS route sections. The EB method was found to yield fewer BS locations. Accidents at BS locations can be analyzed further for their causes and consequences. This should help analysts make decisions on safety countermeasures to implement at individual locations.

REFERENCES

1. Harkey, D. L. Evaluation of Truck Crashes Using a GIS-Based Crash Referencing and Analysis System. In *Transportation Research Record: Journal of the Transportation Research Board, No. 1686,* TRB, National Research Council, Washington, D.C., 1999, pp. 13–21.
2. Feng, C., H. Wei, and J. Lee. World Wide Web GIS Strategies for Transportation Applications. Presented at the 78th Annual Meeting of the Transportation Research Board, 1999.
3. Jobes, B., and V. Papayannoulis. Integrated Traffic Simulation/GIS Platform. Presented at the 77th Annual Meeting of the Transportation Research Board, 1998.
4. Hakkart, A. S., and A. Peled. A PC-Oriented GIS Application for Road Safety Analysis and Management. *Traffic Engineering Control,* 1993.
5. Nassar, S. A. *Integrated Road Accident Risk Model (ARM).* Ph.D. thesis. University of Waterloo, Waterloo, Ontario, Canada, 1996.
6. Nassar, S., F. Saccomanno, and J. Shortreed. Disaggregate Analysis of Road Accident Severities. *International Journal for Impact Engineering,* Vol. 15, No. 6, 1994.
7. Hauer, E. Empirical Bayes Approach to the Estimation of Unsafety: The Multivariate Regression Method. *Accident Analysis and Prevention,* Vol. 24, No. 5, 1992, pp. 457–477.
8. Hauer, E. On the Estimation of the Expected Number of Accidents. *Accident Analysis and Prevention,* Vol. 18, No. 1, 1986, pp. 1–12.
9. Persaud, B. N. *Black Spot Identification and Treatment Evaluation.* Working Paper, Department of Civil Engineering, Ryerson Polytechnic Institute, 1990.
10. Miaou, S. P. The Relationship Between Truck Accidents and Geometric Design of Road Sections: Poisson vs. Negative Binomial Regressions. *Accident Analysis and Prevention,* Vol. 26, No. 4, 1994, pp. 471–482.
11. Dean, C., and J. F. Lawless. Test for Detecting Overdispersion in Poisson Regression Models. *Journal of the American Statistical Association,* Vol. 84, 1986, pp. 467–472.
12. Baker, R. J., and J. A. Nelder. *GLIM System Release 3.77 Manual,* 2nd ed. Numerical Algorithm Group, Oxford, 1987.
13. Chong, K.-C. *An Integrated GIS Interface Technology for Road Accident Risk Modeling.* M.S. thesis. University of Waterloo, Waterloo, Ontario, Canada, 1996.

Publication of this paper sponsored by Committee on Spatial Data and Information Systems.