Accident Prediction Models for Winter Road Safety

Does Temporal Aggregation of Data Matter?

Taimur Usman, Liping Fu, and Luis F. Miranda-Moreno

Most accident prediction models are developed with single-level count data models, such as the traditional negative binomial models with fixed or varying dispersion parameters, assuming independence of data. For many accident data sets in road safety analysis, especially those that are highly disaggregated (hourly data), a hierarchical structure in the data often manifests in some form of correlation. Crash prediction models developed with aggregate data could produce biased results because of the assumption of data independence and inflation of the adequacy of the model's explanation because of the use of aggregate data. The potential effects of data aggregation and correlation on accident prediction models are investigated. The analysis uses an accident database that includes hour-level and storm-level accident counts for individual winter snowstorms at four highway sections in Ontario, Canada. Models of two levels of aggregation, aggregated event-based models and disaggregated hourly based models, were developed. The effect of data aggregation had a significant effect on model results, whereas the difference between conventional regression and multilevel regression was inconsequential.

Road safety is a source of significant concern for transportation officials and researchers. According to a World Health Organization report, about 1.2 million people are killed on roads worldwide each year, and as many as 50 million are injured. Continuation of this trend will make road accidents the third-largest cause of injuries worldwide by 2020 (*1*). Road accidents also result in high social costs. A report by Transport Canada estimates that the annual societal cost due to vehicle collisions exceeds $18 billion in the province of Ontario, Canada, alone (*2*). Significant resources have been allocated to various safety improvement programs involving engineering, education, and reinforcement solutions.

Development of cost-effective safety programs entails two important processes: identification of high-risk locations in the network of interest and development of cost-effective countermeasures. Both processes require accident models that can be used to predict and explain accident occurrences through various explanatory factors related to road geometry, vehicle and driver characteristics, weather, and road conditions.

T. Usman and L. Fu, Department of Civil and Environmental Engineering, University of Waterloo, 200 University Avenue West, Waterloo, Ontario N2L 3G1, Canada. L. F. Miranda-Moreno, Department of Civil Engineering and Applied Mechanics, McGill University, 817 Sherbrooke Street West, Montreal, Quebec H3A 2K6, Canada. Corresponding author: T. Usman, tusman@engmail.uwaterloo.ca.

Transportation Research Record: Journal of the Transportation Research Board, No. 2237, Transportation Research Board of the National Academies, Washington, D.C., 2011, pp. 144–151.
DOI: 10.3141/2237-16

Most accident prediction models belong to the count data regression models, in particular the negative binomial model, which assumes all data or cases are statistically independent. This assumption, however, may be violated when repeated observations over multiple periods (e.g., yearly accident counts) at the same locations (e.g., intersections and road segments) are used as independent cases in model calibration. A common solution for avoiding this problem is to aggregate data from multiple periods for each location into a single observation (e.g., combining monthly data into yearly data or combining multiyear observations into a single-year average). This aggregation treatment addresses the issue of data correlation but will likely result in loss of information and reduction in sample size (*3*).

This paper introduces a multilevel regression approach to capturing the clustered nature of some accident data. The investigation focuses on two specific questions: (*a*) What is the impact of using disaggregate modeling approach? and (*b*) Do multilevel models give substantively different results than a single-level model? A data set compiled for winter road safety is used to examine these questions (*4, 5*). This accident database includes hourly accident counts for individual winter snowstorms on four highway sections in Ontario. This unique data structure allows development and comparison of models of two levels of aggregation: aggregated event-based and disaggregated hourly based models. The hourly observations are used to calibrate and compare single-level and multilevel models.

LITERATURE REVIEW

Road accident modeling has been an area of intensive research in the past few decades. A large number of statistical models have been developed and tested for their suitability to address a variety of complex issues related to accident data. The general consensus is that the negative binomial (NB) distribution is adequate in most cases for modeling road accident counts because of its ability to capture the common nature of overdispersion in accident data (*4–20*).

The NB model structure has been further extended by many researchers to improve its explanatory power and modeling flexibility. For example, a notable extension is the generalized negative binomial (GNB) model, which incorporates a varying dispersion parameter that is a function of a set of covariates. This makes the model capable of controlling for more heterogeneity than does the NB model. It has been shown that use of a varying dispersion parameter could improve model fit (*4, 21–27*).

Another extension is to assume that the error term in the NB model follows a normal distribution instead of a gamma distribution. Models of the resulting structure are known as Poisson lognormal (PLN) models. Such models are good for accident rates with heavier tails

(*25*, *28*). Moreover, multivariate Poisson lognormal models account for both overdispersion and general correlation structure among collision outcomes (*8*, *29*)—for example, correlation among collision categories such as fatal and severity collision counts. This model, in a Bayesian framework, has been extended to account for spatial and temporal correlation among observations. Among previous work dealing with temporal or spatial correlations are that of Lord and Persaud (*30*), Song et al. (*31*), Aguero-Valverde and Jovanis (*32*), and Quddus (*33*). More recently, El-Basyouny and Sayed used multivariate models for accident severity and frequency modeling because of these models' ability to account for overdispersion and correlation, which is present across different levels of severity (*8*). Lenguerrand et al. proposed a hierarchical correlated structure for crashes to model severity with three levels: crash, car, and occupant (*34*). Jones and Jorgensen used multilevel models for accident severity analysis (*35*).

In addition to the spatial, temporal, and among-outcomes correlations, modeling approaches have been proposed to deal with other statistical issues, such as underdispersion, underreporting problems, selection bias (endogeneity), omitted-variables bias, and segmentation (*29*).

In the traffic safety literature, despite the various issues and proposed modeling approaches, most existing models are single level and calibrated with aggregate data (e.g., by month, year, or multiyear periods). These models assume that noncorrelation exists between disaggregate observations (*36*). This assumption could easily become questionable for many accident data sets that are commonly collected over consecutive periods at the same locations. In these data sets, observations are often clustered in a hierarchical or multilevel fashion with individual observations nested within groups—not necessarily in the form of panel data. In this situation, observations within a group are more likely to have some degree of correlation than are those out of the group (*37*). In addition, some temporal trends can exist. Single-level models ignore the potential within-period variations and the nested effect caused by the repetition of observations belonging to the same locations. This can result in the loss of variability and potentially important explanatory information. For instance, in investigations of the impact of weather (precipitation and temperature) and winter maintenance operations on safety, the variations of weather variables over short periods (hours or days) is likely to be highly influential in generating crashes. Model outcomes can be biased as a result of variations in the data that are not considered. This problem was discussed by Lord and Mannering (*29*) and Washington et al. (*38*). Despite the importance of this issue, very little empirical evidence exists on the data aggregation effect, possibly because of the lack of disaggregate accident and traffic-related data.

The multilevel structure and aggregation problem has been recognized in other studies. Jones and Jorgensen (*35*) and Lenguerrand et al. (*34*) were among the first to recognize the need to consider the hierarchical crash–car–occupant structure of accident data for crash severity modeling. They discussed the potential issues of ignoring the clustering nature of data and the correlation within the clusters, such as erroneous estimates of model coefficients and understated standard errors and confidence intervals for the effects. Their conclusions were similar to those in other disciplines, such as epidemiology, social research, and political science (*3*, *39*–*41*). However, both studies focused only on the data structure for severity.

This research attempts to extend previous studies to evaluate the effect of data aggregation and determine the best approach to represent the multilevel structure of the data. This will increase understanding of the implications of aggregating data, ignoring the correlations and time trends in the disaggregate data. This is done by using a unique hourly data set with totally disaggregated data of accidents, traffic, weather, and winter maintenance operations in several highway sections in Ontario, Canada.

DATA DESCRIPTION

Data used for this study were used previously with some minor modifications (*4*, *5*). Details of study sites, data sources, and their processing are given in the following sections.

Study Sites and Data

Well-instrumented study sites were selected so that detailed data on all major factors of interest are available. Four patrol routes were selected, two on Highway 401 and two on Queen Elizabeth Way (QEW) in the province of Ontario, Canada (*4*, *5*). These are major interurban freeways with multiple lanes in each direction and annual average daily traffic of from 100,000 to more than 400,000. The routes were as follows:

- 401-R1, Highway 400 to Morningside Avenue (28.0 km);
- 401-R2, Trafalgar Road to Highway 400 (31.1 km);
- QEW-R1, Burloak Drive to Erin Mills Parkway (17.4 km); and
- QEW-R2, Erin Mills Parkway to Eastmall (13.1 km).

All relevant data originated from five data sources. The first source was hourly traffic data obtained from loop detectors. The second source was accident data, maintained by the Ontario Provincial Police. These data contained detailed information on each collision, including accident time, accident location, accident type, impact type, severity level, vehicle information, and driver information. The third source of data was road condition and weather information system (RCWIS). This data source contained information about road surface conditions (RSC), maintenance operations, precipitation type and accumulation, visibility, and temperature. RCWIS data are collected by Ontario Ministry of Transportation (MTO) maintenance personnel, who patrol the maintenance routes three or four times during a storm event on average. The fourth source of data was the road weather information system (RWIS). This data source contained information about temperature, precipitation type, visibility, wind speed, road surface conditions, and so forth, recorded by RWIS stations near the selected maintenance routes. All these data were obtained from MTO. The last source of data was obtained from Environment Canada (EC). Data from EC included temperature, precipitation type and intensity, visibility, and wind speed.

Modeling of RSC

MTO reports RSC by using qualitative descriptions, that is, a categorical measure (with seven major categories and 160 subcategories). These categories have intrinsic ordering for severity, which means that a more sensible measure would be an ordinal one. Although binary variables could be used to code ordinal data, this would mean loss of information on the ordering. Therefore, an interval variable was used to map the RSC categories and at the same time to make sure that the new variable would have physical interpretations. Road surface condition index (RSI), a surrogate measure of the commonly used friction level, was therefore introduced to represent various RSC classes described in RCWIS. A friction surrogate was used

because there a number of field studies on the relationship between descriptive road surface conditions and friction provided a basis for determination of boundary friction values for each category. To map the categorical RSC into RSI, the following procedure was used:

1. The major classes of road surface conditions, defined in RCWIS, were first arranged according to their severity in an ascending order as follows:

bare and dry < bare and wet < slushy < partly snow covered < snow covered < snow packed < icy

This order was also followed when sorting individual subcategories in a major class.

2. RSI was defined for each major class of road surface state defined in the previous step as a range of values based on the literature in road surface condition discrimination using friction measurements (*42–45*). For convenience of interpretation, RSI is assumed to be similar to road surface friction values and thus varies from 0.1 (poorest, e.g., ice covered) to 1.0 (best, e.g., bare and dry).

3. Each category in the major classes was assigned a specific RSI value. For this purpose, subcategories in each major category were sorted according to Step 1. Linear interpolation was used to assign RSI values to the subcategories.

RSI values for major road surface classes are as follows:

Road Surface Condition	*RSI Range*
Bare and dry	0.9~1.0
Bare and wet	0.8~0.89
Slushy	0.71~0.79
Partly snow covered	0.5~0.7
Snow covered	0.30~0.49
Snow packed	0.2~0.29
Icy	0.05~0.19

Data Processing

Data from five sources for the winter seasons of 2003 to 2006 were used in this study, all of which have different formats and needed to be preprocessed for merging and integration. Accident data were available as event records and therefore needed to be aggregated into hourly records by totaling the accidents that occurred within each hour of the day. Other attributes associated with accidents were averaged over each hour.

Weather data are from three sources: RCWIS, RWIS, and EC. All data sources were converted to an hourly basis. Precipitation intensity data from EC were available only as a daily total, which is the water equivalent of the total precipitation amount for a day. The data on precipitation type were used to determine the hours with and without precipitation. The total daily precipitation was then uniformly allocated to each hour of the hours with precipitation.

After all the data related to weather and road condition were converted into the hourly format, they were fused into a single data set on the basis of date and time. When multiple data were available for a given field, priority was given to RCWIS and RWIS data over EC data because these data sources are collected near the study sites and therefore are considered to be more representative. Missing RSC data from RCWIS were retrieved from accident data or RWIS data. It was also assumed that the RSC at the hour right after a maintenance treatment was done could be considered as partially snow covered. This data field was then subsequently linearly interpolated for hourly conditions, as discussed in the following section. This produced values of RSCs for all hours over individual storms. If any data were missing for temperature, precipitation, or wind in RCWIS, data from RWIS or EC data were used. This process resulted in a single weather data set based on hour.

In the next step, winter storm events were identified, and a data set called hourly based data (HBD), was formed by extracting hourly events. These hourly events were then aggregated to generate an event-based data set (EBD) through combination of the hourly data of the same events. The events were defined on the basis of not only weather conditions but also of road surface conditions. This approach differs from other event-based studies, in which events are defined on the basis of environmental data alone (*46*). Each event was defined with the following constraints (*4*):

- An event starts when snow or freezing rain is observed.
- An event ends when snow or freezing rain stops and a certain predefined road surface condition is achieved after that time.
- Precipitation must be greater than zero (0 cm/h).
- Air temperature must be less than 5°C.
- The RSI value must not be equal to bare dry conditions.

A total of 883 events were extracted with 483 accidents.

MODEL DEVELOPMENT

The most commonly used approach for modeling accident frequencies is the regression analysis for count data. In particular, the NB model and its extensions have been found to be the most suitable distribution structures for road accident frequency (*19*, *22*, *24*, *25*). In earlier research, it was shown that GNB models have a better fit to the data described in the previous section (*4*). The GNB model is therefore used as a basis for this research.

Following the GNB model framework, let $Y_i \sim \text{Poisson}(\theta_i)$ with $\ln(\theta_i) = \mu_i + \epsilon_i$, where Y_i represents the number of accidents during event or hour i ($i = 1, \ldots, n$), μ_i is the mean accident frequency at event i, and $\exp(\epsilon_i) \sim \text{gamma}(1/\alpha_i, 1/\alpha_i)$, where α_i is the overdispersion parameter. The mean accident frequency (μ_i) is then assumed to be a function of a set of covariates through the log link function commonly used in the road safety literature, that is,

$$\mu_i = (\text{exposure}_i)^{\beta_1} \exp(\beta_0 + \beta_2 x_{i1} + \beta_3 x_{i2} + \cdots + \beta_k x_{ik}) \quad (1)$$

where $(x_{i1}, \ldots, x_{ipk})$ is the jth attribute associated with event and hour i at patrol p, Exposure is as defined in the section on exploratory data analysis, and $(\beta_0, \beta_1, \ldots, \beta_k)$ is a vector of regression parameters. In GNB, the dispersion parameter is assumed to be a function of a set of covariates. With an exponential link function, $\alpha_i = \exp(\gamma_0 + \gamma_1 z_{i1} + \gamma_2 z_{i2} + \cdots + \gamma_k z_{im})$, where $(z_{i1}, \ldots, z_{im})$ is a vector of event and hour-specific factors that may be different from those explaining μ_{ip} and $(\gamma_0, \gamma_1, \ldots, \gamma_m)$ is a vector of parameters.

The second model alternative considered in this research is the PLN model. The PLN differs from the NB model in that instead of gamma distributed error, a lognormal distributed error term is added to the Poisson model to capture the unobserved heterogeneity. This model has the advantage that it can be extended to deal with multilevel data sets. The multilevel model structure is necessary because the disaggregate data set is longitudinal with the hourly records within each storm event forming a set of repeated measures over time. This is different from panel data with the number of periods being constant for each location. The potential within-storm correlation can then be

captured with a multilevel model (*5, 25*). Moreover, the lognormal tails are known to be asymptotically heavier than those of the gamma distribution (*28*). This can be the case when working with a data set in the presence of outliers (*47*).

In a multilevel setting, a PLN model for nested hourly observations at the event level can be represented as

$$Y_{im} \sim \text{Poisson}(\theta_{im}) \qquad \text{with } \ln(\theta_{im}) = \mu_{im} + \gamma_m + \epsilon_{im} \tag{2}$$

where

θ_{im}, μ_{im} = number and mean number of accidents in an hour i belonging to or nested in the storm event m;
γ_m = patrol route-level random effect, following a normal distribution, that is, $\gamma_m \sim N(0, \tau_2)$; and
ϵ_{im} = model error also normally distributed, $\epsilon_{im} \sim N(0, \zeta)$.

ϵ_{im} represents all the unobserved heterogeneities or random variations that are not captured by γ_m, and γ_m represents event-level unobserved factors controlling for the potential within event correlation. In this case, the equation for the mean accident frequency has the following functional form:

$$\mu_{im} = (\text{exposure}_m)^{\beta_1} \exp(\beta_0 + \beta_2 x_{i1m} + \beta_3 x_{i2m} + \cdots + \beta_k x_{ikm}) \tag{3}$$

where m is an index indicating the event level and i the hour index.

The random term in Equation 3 accounts only for the random effect on the intercept. A more complex extension would consider the random effects in the slopes, that is, the slopes could be assumed to vary by events. This variation is left for future investigation.

To measure the intraclass correlation (correlation among observations within the same storm event), the intraclass correlation coefficient (ICC), denoted by ρ, is used. This coefficient is computed on the basis of the variance components of model defined previously and ranges from 0 to 1. If all hourly accident count observations are independent of one another, $\rho = 0$. $\rho \neq 0$ implies that the observations are not independent, for example, ICC > 0 implies that the accident occurrence in the same storm is influenced by similar unobserved storm factors. In this case, for the two-level model defined in Equation 2, intracorrelation between accident count observations coming from the same storm (denoted ρ_p) is obtained as (*40*)

$$\rho_m = \text{cor}(Y_{im}, Y_{i'm}) = \frac{\tau_2}{\zeta + \tau_2} \tag{4}$$

In this paper, only four models are considered:

1. Event-based GNB model using EBD,
2. Hourly multilevel GNB using HBD,
3. Hourly multilevel PLN using HBD, and
4. Hourly single-level PLN using HBD.

The first two models are used to investigate the effects of data aggregation, and the third and fourth are used to examine the implication of ignoring data correlation. STATA version 9 is used to calibrate all models.

Exploratory Data Analysis

Box plots of individual data fields and correlation among variables were used to check data sets for any outliers. A number of two-way interactions were considered for some of the variables. These interaction terms were identified on the basis of some possible physical interpretation.

A correlation analysis was carefully conducted for each individual and combined data set. As suspected, it was found that precipitation type and maintenance operations were consistently correlated with RSI (with a correlation coefficient greater than .60) and were therefore excluded from further analysis. Descriptive statistics are presented in Table 1 for the variables found to be significant.

An exploratory analysis of disaggregate data indicated a possible trend in the observed collisions over individual storms. To statistically test this observation, four models with different trend forms were tested, although only three are listed: a model without a trend, a model with a linear trend component, and a model with a dummy variable indicating if the hour is the first or second hour of the storm (after other variations of time indicators were considered), which was found to be significant.

In both data sets, a dummy variable termed as a site-specific factor was included in the analysis to capture the possible effect on road safety of other route-specific factors, such as location, driver population, and road geometry.

Within-groups correlation is assessed with an ICC, which is calculated as the ratio of within-group's variance to total variance as defined in Equation 4. Again, when ICC is close to 0, single-level and multilevel models will have no difference in results. In this case, ICC was calculated for the disaggregate data to confirm the presence of correlation within events, which turns out to be 8% for this data set.

The following factors and variables were used in the analysis:

- Total number of accidents for the event or hour,
- Average wind speed (km/h) for the event or hour,
- Average air temperature (°C) for the event or hour,
- Average visibility (km) for the event or hour,
- Average RSI for the event or hour,
- Exposure—product of total traffic volume (sum of the hourly traffic volumes of an event) during the event and segment length, converted into millions of vehicle kilometers (for HBD analysis, this was the product of segment length and hourly traffic, converted into millions of vehicle kilometers),
- Precipitation intensity (cm/h),
- Hourly traffic volume,
- Site-specific variable (401-R1 = 1, 401-R2 = 2, QEW-R1 = 3, and QEW-R2 = 4), and
- Storm stage—0 if first or second hour, 1 otherwise (for HBD analysis only).

MODEL CALIBRATION AND RESULTS

The compiled data sets in STATA were used to calibrate the models in this study. A stepwise elimination process was followed to identify the significant factors. Table 2 presents the calibration results, and the major findings are summarized in the following.

Effects of Data Aggregation and Correlation

The event-based GNB model (using aggregated data) was compared with the hourly based GNB to assess the effect of data aggregation. Because the two models used data of different aggregation levels, some commonly used quality-of-fit statistics such as the Akaike information criterion (AIC) are not applicable (*48*). (The AIC statistic

TABLE 1 Descriptive Statistics

Variable	No. of Observations	Mean	SD	Min.	Max.
Hourly Data					
Accidents	6,551	0.07	0.31	0.00	4.00
Visibility (km)	6,551	11.08	8.22	0.00	24.10
Wind speed (km/hr)	6,551	16.30	10.25	0.00	59.00
Temperature (°C)	6,551	–4.36	5.06	–23.90	8.00
Precipitation (cm/h)	6,551	1.89	2.35	0.00	18.00
Road surface index	6,551	0.74	0.24	0.05	1.00
Hourly traffic	6,551	16,921	13,763	653	88,696
Lane exposure	6,551	–1.36	0.88	–4.69	0.63
Aggregate Data					
Accidents	883	0.55	1.42	0.00	19.00
Visibility (km)	883	12.86	6.56	0.80	24.10
Wind speed (km/h)	883	16.38	9.46	0.00	50.00
Temperature (°C)	883	–3.29	4.35	–20.73	4.58
Hourly precipitation	883	1.63	1.64	0.04	12.40
Total precipitation	883	14.07	24.42	0.12	246.00
Road surface index	883	0.79	0.19	0.13	0.99
Hourly traffic	883	17,782	13,196	816	77,330
Lane exposure	883	0.38	1.13	–3.51	3.46

NOTE: SD = standard deviation; min. = minimum; max. = maximum.

is defined as $-2LL + 2p$, where LL is the log likelihood of a fitted model and p is the number of parameters, which is included to penalize models with higher number of parameters: a model with smaller AIC value represents a better overall fit.) Nevertheless, the following observations could still be made from Table 2 on the basis of sample size and intuition:

- Because of data aggregation, some variables that are expected to have a statistically significant effect on accident frequency could become statistically insignificant. For example, precipitation intensity and storm stage were both found to be significant in the hourly GNB model with intuitively reasonable effect directions but were proved to be insignificant in the event-based GNB model. This was likely because of information lost to data aggregation.
- There are noticeable differences in the model coefficients for those variables that are significant in both models, and for most variables the absolute values of the coefficients (size of effect) decreased from the aggregated model to the disaggregated model. This pattern of change in the model coefficients is a sign of the confounding effect of some variables caused by data aggregation. For example, the coefficient associated with RSI changed from −1.938 in the hourly model to –2.724 in the event model, a 41% increase in effect size. This increase could be caused by the confounding effects of hourly precipitation and storm stage.

The effects of data correlation can be observed in the single-level and multilevel PLN models calibrated with the hourly data. As shown in Table 2, the models were quite similar in variables and the associated coefficients, possibly because correlation within events is not very strong (*49*). However, the AIC statistics did show that the multilevel PLN fitted the data slightly better than did the single-level model (Table 2).

Factors Affecting Winter Road Safety

In general, results from the models are consistent, as shown in Table 2. Most results obtained in the research for winter road safety and associated factors are consistent with those reported in the literature, with a few exceptions. The following specific observations are made from the modeling outcomes:

- The most interesting result is that the RSI was found to be a statistically significant factor influencing road safety across all sites. The negative sign associated to the factor suggests that higher accident frequencies are associated with poor road surface conditions. This result makes intuitive sense and has confirmed the findings of many studies (*42*, *50*), mostly those in Nordic countries. However, this research is the first to show the empirical relationship between safety and road surface conditions at a disaggregate level, making it feasible to quantify the safety benefit of alternative maintenance goals and methods.
- Visibility was found to have a statistically significant effect on accident frequency during a snowstorm. The negative model coefficient also makes intuitive sense, suggesting that reduced visibility was associated with more accidents. This result is different from those in a previous statistical study (*51*), which used data from 37 sites and found that visibility was significant only at two sites. That study considered collisions occurring at different roadways related to a single weather station. This approach may have masked the effect of visibility because of confounding of missing factors and large aggregation levels in both space (coastal areas versus inter cities) and time (seasonal variation).
- As expected, exposure, defined as millions of vehicle kilometers traveled (product of the total traffic volume over a storm event and route length for aggregate data and product of the traffic volume per

TABLE 2 Summary Results of Model Calibration

	Aggregated Versus Disaggregated				Multilevel Versus Single-Level			
	GNB Event-Based Model		GNB Hourly Model		PLN Multilevel Model		PLN Single-Level Model	
Variable	Coeff.	Sig.	Coeff.	Sig.	Coeff.	Sig.	Coeff.	Sig.
Intercept	1.350	.002	−1.774	.000	−2.195	.000	−2.420	.000
Wind speed (km)								
Temperature (°C)					−0.021	.055	−0.015	.120
Visibility (km)	−0.033	.003	−0.026	.000	−0.029	.000	−0.028	.000
Hourly precipitation			0.054	.018	0.049	.038	0.059	.008
Road surface index	−2.724	.000	−1.938	.000	−1.875	.000	−1.854	.000
Hourly traffic	−0.00005	.000						
Lane exposure	0.686	.000	0.141	.019	0.186	.003	0.149	.013
S2E1			−0.546	.000	−0.499	.000	−0.560	.000
S2E2			0.000		0.000		0.000	
401-R1	1.286	.000	1.993	.000	2.066	.000	2.002	.000
401-R2	0.644	.037	1.346	.000	1.549	.000	1.356	.000
QEW-R1	−0.147	.676	−0.114	.686	−0.097	.735	−0.097	.708
QEW-R2	0.000		0.000		0.000		0.000	
Number of observations	883		6,551		6,551		6,551	
LL (constant only)	−802.919		−1,689.8					
LL (model)	−671.405		−1,581.2		−1,583.5		−1,585.3	
AIC	1,370.81		3,190.42		3,189.05		3,192.56	
Overdispersion Model								
Intercept	3.503	.000	1.427	.069				
Road surface index	−2.932	.013	1.304	.067				
Lane exposure	−0.580	.002						
401-R1	−2.114	.002	−2.261	.001				
401-R2	−1.282	.049	−1.552	.031				
QEW-R1	0.466	.514	−0.144	.874				
QEW-R2	0.000		0.000					

NOTE: Coeff. = coefficient; sig. = statistical significance; S2E1 and S2E2 are storm stage or trend factors, with S2E2 representing the first or second storm hour and S1E1 others; LL = log likelihood.

hour and route length for disaggregate data), was found to be significant, suggesting that an increase in traffic volume, storm duration, or route length would lead to an increase in the number of accidents. Inclusion of this term ensures that traffic exposure is accounted for during estimation of the safety benefits of specific policy alternatives. The coefficient associated with the exposure term has a value of less than 1, suggesting that the moderating effect of exposure is nonlinear with a decreasing rate. This result is consistent with those from road safety literature (*24, 30, 52–58*).

- The model also suggests that when other factors (RSC, visibility, exposure) are controlled for, Highway 401 (401-R1 and 401-R2) is more susceptible to crashes than is QEW (QEW-R1 and QEW-R2), whereas the difference in risk between the two maintenance routes on the same highway is quite small. The discrepancy between Highway 401 and QEW needs further investigation; however, a possible explanation is that Highway 401 has more interchanges per kilometer than QEW, which is known to be an important factor influencing freeway safety in general.
- In addition to exposure, hourly traffic was found to be significant in the aggregate data analysis, suggesting that traffic variation within events is an important factor in accidents.
- Air temperature was found to be significant only in the disaggregate data analysis using PLN. This result confirms some previous finds (*59*). The negative sign suggests that the mean number of accidents will increase as temperature decreases.
- Precipitation was consistently found to be significant in all the models when disaggregate data were used, and this confirms some previous results (*60*). The positive sign suggests that the mean number of accidents will increase with an increase in precipitation.
- The analysis confirmed the significance of a trend component in road collisions for the duration of individual storms. Specifically, it was found that the first 2 h of an event had a statistically higher collision rate than the remaining hours of the event, which is consistent with findings reported in the literature.

CONCLUSIONS AND FUTURE WORK

This paper provided empirical evidence about the effects of data aggregation and correlation on disaggregated accident prediction models. The analysis was conducted with a winter hourly accident data set that had a hierarchical event-hour structure. The data set

included hourly accident counts of all snowstorms that occurred during the three winter seasons from 2003 to 2006 at four instrumented freeway sections in Ontario, Canada.

1. The study showed that temporal aggregation of accident data matters. Data aggregation that ignores data correlation could result in loss of information and models of distorted risk factors and effect size. Some important factors could turn out to be insignificant, whereas they should be significant and would have been found significant if the data were not aggregated. Also, effects of these insignificant variables could be distributed to the significant variables, distorting their parameter estimates.
2. The effect of data correlation for the specific data set used in this study was found to be small with inconsequential differences in significant factors and coefficients. A possible reason for this indifference is that the event-level correlation in this data set is weak. Thus, the conventional single-level models may be used for data with weak or no within events correlation. Use of single-level models for multilevel or hierarchical data with a large number of observations can also prove to be time-efficient for analysis because multilevel models are normally data intensive and are computationally expensive, requiring much time for analysis (*61*). In case of high correlation, however, multilevel models should be considered.

This analysis was carried out with data from only one type of highway (urban freeway). Future efforts will concentrate on examining the validity of these findings across a wider spectrum of road section locations.

ACKNOWLEDGMENTS

This research was supported by Ontario Ministry of Transportation (MTO) in part through the Highway Infrastructure and Innovation Funding Program. The authors acknowledge the assistance of Max Perchanock, Steve Birmingham, Zoe Lam, and David Tsui of MTO and Brian Mills of Environment Canada.

REFERENCES

1. World Health Organization. Geneva, 2004. http://www.who.int/world-health-day/2004/infomaterials/world_report/en. Accessed April 16, 2007.
2. *Analysis and Estimation of the Social Cost of Motor Vehicle Collisions in Ontario.* Transport Canada, Ontario, Canada, 2007.
3. Hutchings, C., S. Knight, and J. C. Reading. The Use of Generalized Estimating Equations in the Analysis of Motor Vehicle Crash Data. *Accident Analysis and Prevention,* Vol. 35, No. 1, 2003, pp. 3–8.
4. Usman, T., L. Fu, and L. F. Miranda-Moreno. Quantifying Safety Benefit of Winter Road Maintenance: Accident Frequency Modeling. *Accident Analysis and Prevention,* Vol. 42, No. 6, 2010, pp. 1878–1887.
5. Usman, T., L. Fu, L. F. Miranda-Moreno, and M. Perchanock. A Multilevel Disaggregate Model for Quantifying the Safety Effect of Winter Road Maintenance. Presented at 13th Winter Road Congress Meeting, Quebec City, Canada, 2010.
6. Anastasopoulos, P. C., and F. L. Mannering. A Note on Modeling Vehicle Accident Frequencies with Random-Parameters Count Models. *Accident Analysis and Prevention,* Vol. 41, No. 1, 2009, pp. 153–159.
7. Carson, J., and F. Mannering. The Effect of Ice Warning Signs on Accident Frequencies and Severities. *Accident Analysis and Prevention,* Vol. 1, No. 33, 2001, pp. 99–109.
8. El-Basyouny, K., and T. Sayed. Collision Prediction Models Using Multivariate Poisson-Lognormal Regression. *Accident Analysis and Prevention,* Vol. 41, No. 4, 2009, pp. 820–828.
9. El-Basyouny, K., and T. Sayed. Accident Prediction Models with Random Corridor Parameters. *Accident Analysis and Prevention,* Vol. 41, No. 5, 2009, pp. 1118–1123.
10. Fridstrøm, L., J. Ifver, S. Ingebrigtsen, R. Kulmala, and L. K. Thomsen. Measuring the Contribution of Randomness, Exposure, Weather, and Daylight to the Variation in Road Accident Counts. *Accident Analysis and Prevention,* Vol. 27, No. 1, 1995, pp. 1–20.
11. Geedipally, S. R., and D. Lord. Investigating the Effect of Modeling Single-Vehicle and Multi-Vehicle Crashes Separately on Confidence Intervals of Poisson-Gamma Models. *Accident Analysis and Prevention,* Vol. 42, No. 4, 2010, pp. 1273–1282.
12. Johansson, P. Speed Limitation and Motorway Casualties: A Time Series Count Data Regression Approach. *Accident Analysis and Prevention,* Vol. 28, No. 1, 1996, pp. 73–87.
13. Kim, D., and S. Washington. The Significance of Endogeneity Problems in Crash Models: An Examination of Left-Turn Lanes in Intersection Crash Models. *Accident Analysis and Prevention,* Vol. 38, No. 6, 2006, pp. 1094–1100.
14. Lambert, D. Zero-Inflated Poisson Regression, with an Application to Defects in Manufacturing. *Technometrics,* Vol. 34, No. 1, 1992, pp. 1–14.
15. Lord, D., S. P. Washington, and J. N. Ivan. Poisson, Poisson-Gamma and Zero-Inflated Regression Models of Motor Vehicle Crashes: Balancing Statistical Fit and Theory. *Accident Analysis and Prevention,* Vol. 37, No. 1, 2004, pp. 35–46.
16. Maher, M. J. A Bivariate Negative Binomial Model to Explain Traffic Accident Migration. *Accident Analysis and Prevention,* Vol. 22, No. 5, 1990, pp. 487–498.
17. Maher M. J., and I. Summersgill. A Comprehensive Methodology for the Fitting Predictive Accident Models. *Accident Analysis and Prevention,* Vol. 28, No. 3, 1996, pp. 281–296.
18. Milton, J., and F. Mannering. The Relationship Among Highway Geometrics, Traffic-Related Elements and Motor Vehicle Accident Frequencies. *Transportation,* Vol. 25, 1998, pp. 395–413.
19. Shankar, V., F. Mannering, and W. Barfield. Effect of Roadway Geometrics and Environmental Factors on Rural Freeway Accident Frequencies. *Accident Analysis and Prevention,* Vol. 27, No. 3, 1995, pp. 371–389.
20. Miranda-Moreno, L. F. *Statistical Models and Methods for Identifying Hazardous Locations for Safety Improvements.* PhD dissertation. University of Waterloo, Ontario, Canada, 2006.
21. Hauer, E. Overdispersion in Modelling Accidents on Road Sections and in Empirical Bayes Estimation. *Accident Analysis and Prevention,* Vol. 33, No. 6, 2001, pp. 799–808.
22. Miaou, S.-P., and D. Lord. Modeling Traffic Crash-Flow Relationships for Intersections: Dispersion Parameter, Functional Form, and Bayes Versus Empirical Bayes Methods. In *Transportation Research Record: Journal of the Transportation Research Board, No. 1840,* Transportation Research Board of the National Academies, Washington, D.C., 2003, pp. 31–40.
23. Miranda-Moreno, L. F., L. Fu, F. F. Saccomanno, and A. Labbe. Alternative Risk Models for Ranking Locations for Safety Improvement. In *Transportation Research Record: Journal of the Transportation Research Board, No. 1908,* Transportation Research Board of the National Academies, Washington, D.C., 2005, pp. 1–8.
24. El-Basyouny, K., and T. A. Sayed. Comparison of Two Negative Binomial Regression Techniques in Developing Accident Prediction Models. In *Transportation Research Record: Journal of the Transportation Research Board, No. 1950,* Transportation Research Board of the National Academies, Washington, D.C., 2006, pp. 9–16.
25. Miranda-Moreno, L. F., and L. Fu. A Comparative Study of Alternative Model Structures and Criteria for Ranking Locations for Safety Improvements. *Networks and Spatial Economics,* Vol. 6, No. 2, 2006, pp. 97–110.
26. Mitra, S., and S. Washington. On the Nature of Over-Dispersion in Motor Vehicle Crash Prediction Models. *Accident Analysis and Prevention,* Vol. 39, No. 3, 2007, pp. 459–468.
27. Cafiso, S., G. Di Silvestro, B. Persaud, and M. A. Begum. Revisiting Variability of Dispersion Parameter of Safety Performance for Two-Lane Rural Roads. In *Transportation Research Record: Journal of the Transportation Research Board, No. 2148,* Transportation Research Board of the National Academies, Washington, D.C., 2010, pp. 38–46.
28. Kim, H., D. Sun, and R. K. Tsutakawa. Lognormal vs. Gamma: Extra Variations. *Biometrical Journal,* Vol. 44, No. 3, 2002, pp. 305–323.
29. Lord, D., and F. Mannering. The Statistical Analysis of Crash-Frequency Data: A Review and Assessment of Methodological Alternatives. *Transportation Research Part A,* Vol. 44, No. 5, 2010, pp. 291–305.

30. Lord, D., and B. N. Persaud. Accident Prediction Models With and Without Trend: Application of the Generalized Estimating Equations Procedure. In *Transportation Research Record: Journal of the Transportation Research Board, No. 1717*, TRB, of the National Research Council, Washington, D.C., 2000, pp. 102–108.
31. Song, J. J., M. Ghosh, S. Miaou, and B. Mallick. Bayesian Multivariate Spatial Models for Roadway Traffic Crash Mapping. *Journal of Multivariate Analysis*, Vol. 97, No. 1, 2006, pp. 246–273.
32. Aguero-Valverde, J., and P. P. Jovanis. Analysis of Road Crash Frequency with Spatial Models. In *Transportation Research Record: Journal of the Transportation Research Board, No. 2061*, Transportation Research Board of the National Academies, Washington, D.C., 2008, pp. 55–63.
33. Quddus, M. A. Modelling Area-Wide Count Outcomes with Spatial Correlation and Heterogeneity: An Analysis of London Crash Data. *Accident Analysis and Prevention*, Vol. 40, No. 4, 2008, pp. 1486–1497.
34. Lenguerrand, E., J. L. Martin, and B. Laumon. Modelling the Hierarchical Structure of Road Crash Data—Application to Severity Analysis. *Accident Analysis and Prevention*, Vol. 38, No. 1, 2006, pp. 43–53.
35. Jones, A. P., and S. H. Jorgensen. The Use of Multilevel Models for the Prediction of Road Accident Outcomes. *Accident Analysis and Prevention*, Vol. 35, No. 1, 2003, pp. 59–69.
36. McCullagh, P., and J. A. Nelder. *Generalized Linear Models*, 2nd ed. Chapman and Hall, London, 1989.
37. West, B. T., K. B. Welch, and A. T. Galecki. *Linear Mixed Models: A Practical Guide Using Statistical Software*. CRC Press, Boca Raton, Fla., 2007.
38. Washington, S. P., M. G. Karlaftis, and F. L. Mannering. *Statistical and Econometric Methods for Transportation Data Analysis*, 2nd ed. Chapman Hall and CRC, Boca Raton, Fla., 2010.
39. Heck, R. H., and S. L. Thomas. *An Introduction to Multilevel Modeling Techniques*. Psychology Press, London, 2000.
40. Newsom, J. T., and M. Nishishiba. *Nonconvergence and Sample Bias in Hierarchical Linear Modeling of Dyadic Data*. 2004. http://www.upa.pdx.edu/IOA/newsom/mlrdyad4.doc. Accessed March 29, 2010.
41. Gelman, A., and J. Hill. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, New York, 2007.
42. Wallman, C. G., P. Wretling, and G. Oberg. *Effects of Winter Road Maintenance*. Swedish National Road and Transport Research Institute, Linköping, Sweden, 1997.
43. Wallman, C. G., and H. Astrom. *Friction Measurement Methods and the Correlation Between Road Friction and Traffic Safety: A Literature Review*. Swedish National Road and Transport Research Institute, Linköping, Sweden, 2001.
44. *Winter Maintenance Performance Measurement Using Friction Testing*. Transportation Association of Canada, Ottawa, Ontario, 2008.
45. Feng, F., L. Fu, and M. S. Perchanok. Comparison of Alternative Models for Road Surface Condition Classification. Presented at 89th Annual Meeting of the Transportation Research Board, Washington, D.C., 2010.
46. Knapp, K. K., D. L. Smithson, and A. J. Khattak. The Mobility and Safety Impacts of Winter Storm Events in a Freeway Environment. Presented at Mid-Continent Transportation Symposium, Iowa State University, Ames, 2000.
47. Winkelmann, R. *Econometric Analysis of Count Data*. Springer, Berlin, 2003.
48. Akaike, H. A New Look at the Statistical Model of Identification. *IEEE Transactions on Automatic Control*, Vol. 19, No. 6, 1974, pp. 716–723.
49. Goldstein, H. Multilevel Mixed Linear Model Analysis Using Iterative Generalized Least Squares. *Biometrika*, Vol. 73, No. 1, 1986, pp. 43–56.
50. Norrman, J., M. Eriksson, and S. Lindqvist. Relationships Between Road Slipperiness, Traffic Accident Risk and Winter Road Maintenance Activity. *Climate Research*, Vol. 15, No. 3, 2000, pp. 185–193.
51. Brijis, T., C. Offermans, E. Hermans, and T. Stiers. Impact of Weather Conditions on Road Safety Investigated on Hourly Basis. Presented at 85th Annual Meeting of the Transportation Research Board, Washington, D.C., 2006.
52. Andrew, V., and J. Bared. Accident Models for Two-Lane Rural Segments and Intersections. In *Transportation Research Record 1635*, TRB, National Research Council, Washington, D.C., 1998, pp. 18–29.
53. *NCHRP Synthesis 295: Statistical Methods in Highway Safety Analysis*. TRB, National Research Council, Washington, D.C., 2001.
54. Roozenburg, A., and S. Turner. Accident Prediction Models at Traffic Signals. Presented at Annual Technical Conferences of the Institution of Professional Engineers, Auckland, New Zealand, 2005.
55. Bin Mustakim, F., B. D. Daniel, and K. Bin Ambak. Accident Investigation, Blackspot Treatment and Accident Prediction Model at Federal Route FT50 Batu Pahat-Ayer Hitam. *Engineering e-Transaction*, Vol. 1, No 2, 2006, pp. 19–32.
56. Sayed, T., and G. R. Lovegrove. Macrolevel Collision Prediction Models to Enhance Traditional Reactive Road Safety Improvement Programs. In *Transportation Research Record: Journal of the Transportation Research Board, No. 2019*, Transportation Research Board of the National Academies, Washington, D.C., 2007, pp. 65–73.
57. Jonsson, T., J. N. Ivan, and C. Zhang. Crash Prediction Models for Intersections on Rural Multilane Highways: Differences by Collision Type. In *Transportation Research Record: Journal of the Transportation Research Board, No. 2019*, Transportation Research Board of the National Academies, Washington, D.C., 2007, pp. 91–98.
58. Lord, D., S. D. Guikema, and S. Geedipally. Application of the Conway-Maxwell-Poisson Generalized Linear Model for Analyzing Motor Vehicle Crashes. *Accident Analysis and Prevention*, Vol. 40, No. 3, 2008, pp. 1123–1134.
59. Fu, L., M. S. Perchanok, L. F. M. Moreno, and Q. A. Shah. Effects of Winter Weather and Maintenance Treatments on Highway Safety. Presented at 85th Annual Meeting of the Transportation Research Board, Washington, D.C., 2006.
60. Andrey, J., B. Mills, and J. Vandermolen. *Weather Information and Road Safety*. Institute for Catastrophic Loss Reduction, Toronto, Ontario, Canada, 2001.
61. Steenbergen, M. R., and B. S. Jones. Modeling Multilevel Data Structures. *American Journal of Political Science*, Vol. 46, No. 1, 2002, pp. 218–237.

The Safety Data, Analysis, and Evaluation Committee peer-reviewed this paper.