Application of Latent Class Growth Model to Longitudinal Analysis of Traffic Crashes

Yichuan Peng and Dominique Lord

One of the most important and meaningful tasks in traffic safety is to describe how traffic crash risk changes over time. Over the past 20 years, much work has been done about this task. The recent introduction of latent class models to analyze crash data has created a need to examine how these models could be used for longitudinal data analysis. Latent class models dictate that part of the heterogeneity be attributed by grouping distinct subpopulations into a common data set. Investigation of the commonalities of the subgroups can be useful for targeting specific safety interventions. This paper describes the application of the latent class growth model (LCGM), which is tailored to longitudinal data. Analysis was accomplished with the use of data collected between 1997 and 2007 on rural, two-lane highways in Texas. Trends for all crash severities and injury crashes were examined. It was determined that the crash data could be drawn from three population subgroups for which crash risks were low, medium, and high. The results of this study showed that average shoulder width and speed limit had stronger effects at sites classified as high crash risks, whereas traffic flow had a stronger influence at sites classified as low risks. As expected, higher speed limits increased crash risk, whereas a wider shoulder width reduced risk. In conclusion, the LCGM showed good potential for use in the analysis of longitudinal data, but further research is needed.

Longitudinal data analysis (or panel data analysis) is important in a wide range of areas. In highway safety, the studies that incorporated longitudinal data usually were classified into two groups: those that investigated changes in the number of crashes between two distinct time points, traditionally known as the before–after study (*1–4*), and those that examined time-trend in the data, usually carried out over a long period (*5–12*). The development of time-trend models can be useful to control effects that change over time if the study objective is to establish relationships between crashes and various explanatory factors. Over the last 2 years, a new type of model has been proposed for modeling crash data. This type of model is known either as the latent class model, finite mixture model, or as the Markov switching model (*13*, *14*). This type of model assumes that observations can be generated from distinct subpopulations, which create part of the heterogeneity observed in the data. Thus far, only two studies have examined the application of latent class models to modeling longitudinal data (*15*, *16*). This study adds to the recent work on this topic by examining the application of the latent class growth model (LCGM) to model longitudinal data. This type of model is different from the one proposed by Malyshkina and Mannering (*16*), in which the entire population is divided into several subgroups on the basis of their individual differences to capture such differences by using a series of trajectories over time, without change in subgroup states (discussed further below).

Several statistical techniques or methods can be used to analyze longitudinal data. One method is to study change in raw scores. With this method, the change is computed as the difference between the first time period and the second time period scores, and the resulting raw change values are analyzed as a function of individual or group characteristics (*17*). An alternative approach is to study residual change scores. With this method, the change is computed as the residual between the observed Time 2 score and the expected Time 2 score as predicted by the Time 1 score (*18*), which is similar to what is done in the traditional before–after study proposed by Hauer (*1*). One limitation of this method is that it tends to consider change between only two, discrete time points and thus is more useful for prospective research designs than it is for other kinds (*18*). Researchers sometimes are more interested in modeling developmental trajectories, or patterns of change in an outcome across multiple time points (*19*). Standard growth analyses estimate a single trajectory that averages the individual trajectories of all participants in a given sample. This approach captures individual differences by estimating a random coefficient that represents the variability that surrounds this intercept and slope. By this method, researchers can use categorical or continuous independent variables, which represent potential risk or protective factors, to predict individual differences in the intercept and slope values.

Standard growth models are useful for studying research questions in cases in which all individuals in a given sample are expected to change in the same direction across time, with only the degree of change varying between individuals (*20*). Nagin offers time spent with peers as an example of this monotonic heterogeneity of change (*21*). It is useful to frame a research question in terms of an average trajectory of time spent with peers. Some phenomena may follow a multinomial pattern, however, in which both the dependable variables and the direction of change vary between individuals. Raudenbush used depression as an example by arguing that it was incorrect to assume that all people in a given sample would experience either increasing or decreasing levels of depression (*20*). In a normative sample, many people would never have a high level of depression, whereas some would always have a high level, others would become increasingly depressed, and still others might vacillate between highs and lows. In such instances, a single, averaged growth trajectory could mask important individual differences and lead to the erroneous conclusion that people did not change for a given variable. Such conclusions could be drawn if 50% of the sample increased by the same amount on a particular variable, whereas 50% of the sample decreased by the same amount on that variable. Here, a single growth trajectory would average to zero and thus prompt researchers to assume an

Zachry Department of Civil Engineering, Texas A&M University, 3136 TAMU, College Station, TX 77843-3136. Corresponding author: Y. Peng, pycgcx@tamu.edu.

Transportation Research Record: Journal of the Transportation Research Board, No. 2236, Transportation Research Board of the National Academies, Washington, D.C., 2011, pp. 102–109.
DOI: 10.3141/2236-12

absence of change despite the presence of substantial yet opposing patterns of change for two distinct subgroups in the sample (*22*). Similar phenomena can be seen in traffic crash data. It is not appropriate to assume that the road network is drawn from a single population because of the heterogeneity observed for each road segment (*13*). For this class of problems, alternative modeling strategies are available that consider multinomial heterogeneity in change. One such approach is a group-based statistical technique known as LCGM, which is applied to crash data.

This paper is divided into four sections. The second section describes the basic principles associated with the LCGM to analyze the longitudinal trend in traffic crashes. The third section provides a description of the data and the application of the LCGM to investigate the pattern related to longitudinal changes for the total number of crashes and injury crashes that happened on 2,639 rural, two-lane road segments in Texas. The last section of this paper presents the summary results and recommendations for further studies.

METHODOLOGY

This section describes the methodology and the characteristics of LCGM.

Theoretical Basis of LCGM

This study used LCGM to analyze the trend in crash data. The LCGM is a semiparametric technique, which can be used to capture unobserved heterogeneity related to crashes that occurred on road segments and identify distinct subgroups of road segments, which could follow different patterns that change over time.

Unlike standard latent growth modeling techniques, in which individual differences in both the slope and intercept are estimated by using random coefficients, LCGM fixes the slope and the intercept to equality across all subjects or observations within a trajectory. Such an approach is acceptable, given that individual differences are captured by the multiple trajectories included in the model. Given that both the slope and intercept are fixed, the degree of freedom remains available to estimate quadratic trajectories of a variable measured at three time points or cubic trajectories with data available at the fourth time points.

Although LCGM has a wide variety of applications, the rating scale of the instrument used to measure the variable of interest dictates the specific probability distribution used to estimate the parameters. For example, dichotomous data require the use of the binary logit distribution, and frequency data dictate the use of the Poisson distribution. At this point, the Poisson-gamma or other mixed-Poisson models are not available or have not been developed. Because crash data are discrete and nonnegative events, the Poisson distribution was considered a suitable choice for this analysis. Future research will examine other mixed-Poisson mixture models. In the LCGM model, each trajectory is described as a latent variable that represents the predicted score on a given dependent variable of interest for a known trajectory at a specific time and is defined by the following function:

$$Y_{u}^{*} = \beta_{0j} + \beta_{1j} X_{it} + \beta_{2j} X_{it}^{2} + \beta_{3j} X_{it}^{3} + \epsilon_{it} \tag{1}$$

where

X_{it}, X_{it}^2, and X_{it}^3 = independent variable year entered in regular, squared, or cubed term, respectively, for observation *i* and time *t*;
ϵ_{it} = disturbance term assumed to be Poisson distributed with a constant standard deviation; and
β_{0j}, β_{1j}, β_{2j}, and β_{3j} = parameters that define the intercept and slopes (i.e., linear, quadratic, or cubic) of the trajectory for a specific subgroup *j*.

As demonstrated in the above polynomial function, the trajectories are most often modeled by using a linear X_{it}, quadratic X_{it}^2, or cubic trend X_{it}^3, the selection of which depends on the number of time points measured. A linear pattern of change is defined by the X_{it} parameter, and a linear trend may either steadily increase or decrease at varying magnitudes or remain stable. A quadratic pattern of change is defined by the X_{it}^2 parameter and a quadratic trend may increase, decrease, or remain stable up to a certain time point before it changes in either magnitude or direction. A cubic trajectory is defined by the X_{it}^3 parameter, and a cubic trend will have two changes in either the magnitude or direction across different time points. The likelihood function for the LCGM can be defined as follows:

$$L = \prod_{i=1}^{n} p\left(Y_i\right) \tag{2}$$

where

Y_i = trajectory data for subject *i*,
$p(Y_i) = \Sigma_j \pi_j(x_i)\, p^j(y_i)$,
$p^j(y_i)$ = probability of Y_i belonging to group *j*, and
n = number of observations.

For the Poisson model, the probability of observing the data trajectory y_i that is given membership in group *k* is

$$\Pr\left(Y_i = y_i \middle| C_i = k, W_i = w_i\right) = \prod_{y_{ij}>0} \frac{\exp\left(-\lambda_{ijk}\right)\lambda_{ijk}^{y_{ij}}}{y_{ij}!}$$

where

C = risk factors,
W = time-varying covariates,
w = value of time-varying covariates, and

$$\pi_i\left(x_i\right) = \frac{e^{x_i\theta_j}}{\sum_j e^{x_i\theta_j}}$$

where $\pi_i(x_i)$ is probability of belonging to group *j* for covariates (risk) X_i, and θ is the parameter of time-stable covariate effect on each group membership.

SAS Traj Procedure

A SAS procedure, which is performed on the basis of the latent class growth modeling called Traj, was used to conduct the analysis. The Traj procedure provides the option of modeling three different distributions to analyze count, psychometric scale, and dichotomous data (*23*).

When the LCGM is applied, the number of distinct trajectories to be extracted from the data has to be first determined and then the model selected with the number of trajectories that best fits the data. It is preferable to have prior knowledge about the number and the shape of trajectories whenever theory and literature exist in the area of study, similar to the Poisson mixture model proposed by Park and Lord (*13*) and Park et al. (*4*). There are many reasons to expect the existence of different subpopulations, because the crash data are generally collected from various geographic, environmental, and geometric design contexts over some fixed time periods.

Another important objective of this research was to establish whether the crash risk was related to the measured covariates that included risk factors and time-varying covariates. Some previous applications of the semiparametric approach in social science categorized subjects by latent trait from observable behavior (*24*). The group assignments were then fitted to the covariates with standard linear models. This classify–analyze procedure did not, however, account for the uncertainty in group assignment and could lead to a certain level of bias. This study illustrated the inclusion directly into the model of risk factors (e.g., speed limit, surface width, and average shoulder width) and of the time-varying, covariate traffic volume of the segments. In so doing, the assignment of uncertainty could be accounted for automatically and the relationship between these measured covariates and traffic crashes could be identified.

ANALYSIS AND RESULTS

This section describes the characteristics of the data and the modeling results for all crashes and for injury crashes.

Description of Data

The data analyzed in this paper were provided by the Texas Department of Transportation and the Texas Department of Public Health. The road network consisted of 2,639 rural, two-lane segments. The data were collected for traffic flow, segment length, surface width, average shoulder width, and speed limit. Only traffic flow changed during the entire time period (hence it was classified as a time-varying covariate), but all the other variables remained constant. Crash data were collected for the following severity levels: K = Fatal, A = Injury Type A or incapacitating injury, B = Injury Type B or nonincapacitating injury, C = Injury Type C or possible injury, and O = property damage only (PDO). As discussed above, crashes were analyzed for all crashes (KABCO) and for injury crashes (KABC). Because crash data were not available for 2002, that year was excluded from the analysis. The definition of a reportable collision changed in 2003 and increased the number of PDO crashes. Its effects on the models are described below. Table 1 provides a summary of the data.

TABLE 1 Summary Statistics of Variables Used in This Study

Variable Name	Min.	Max.	Mean	Standard Deviation
1997 crashes	0	9	1.625	1.501
1998 crashes	0	12	1.760	1.448
1999 crashes	0	15	1.851	1.685
2000 crashes	0	12	1.780	1.450
2001 crashes	0	10	1.795	1.523
2003 crashes	0	22	2.119	2.091
2004 crashes	0	18	2.084	1.900
2005 crashes	0	21	2.110	2.070
2006 crashes	0	14	2.101	1.817
2007 crashes	0	20	2.228	2.034
1997 average daily traffic	10	13,000	1,409.526	1,922.179
1998 average daily traffic	10	13,000	1,409.526	1,922.179
1999 average daily traffic	10	13,300	1,438.719	1,954.421
2000 average daily traffic	20	15,300	1,490.364	2,069.204
2001 average daily traffic	10	18,400	1,507.2	2,114.893
2003 average daily traffic	10	22,000	1,584.244	2,239.84
2004 average daily traffic	10	19,400	1,604.975	2,223.261
2005 average daily traffic	10	21,000	1,668.799	2,329.92
2006 average daily traffic	10	22,000	1,636.999	2,308.724
2007 average daily traffic	10	29,000	1,670.811	2,354.088
Speed limit (mph)	15	70	57.849	8.720
Surface width (ft)	14	48	24.659	8.504
Right shoulder width (ft)	0	10	0.504	1.505

NOTE: min. = minimum; max. = maximum.

Longitudinal Analysis of All Traffic Crashes from 1997 to 2007

As has been mentioned, the number of distinct trajectories had to be first determined. In the analysis, the data were divided into two, three, and four subpopulations and then the Bayesian information criterion (BIC) values were compared to obtain the best number of subgroups [see Jones et al. for a description of how the BIC is used to compare models and determine the number of subgroups (*23*)]. The BIC values for the models are shown in Table 2. The results showed that the ΔBIC value became negative when the number of subgroups increased from three to four. It was therefore determined that the data could be subdivided into three groups (or classes).

After the number was determined for the subgroups (hereafter referred to as groups), it was necessary to determine the trajectory form of each group. As a general rule, a quadratic trajectory model should first be tested. The output of the tested quadratic trajectory is shown in Tables 3 and 4. Each component of Group 1 was not significant, and the linear components of both Groups 2 and 3 were not significant at the 5% level (because the *p*-value of these parameters was much higher than .05). Thus it was determined that the linear trajectory provided the best option. The results of the three-group linear trajectory modeling process are shown in Tables 5 and 6.

Figure 1 illustrates the trend for KABCO traffic crash risk for 1997 to 2007. This figure shows that all road segments in this data set could be divided into three groups. The number of crashes per mile

TABLE 2 Tabulated BIC Values

Number of Subgroups	BIC	ΔBIC
1	−25,063.63	—
2	−22,900.44	2,163.19
3	−22,146.63	753.81
4	−22,256.33	−109.7

NOTE: — = not applicable.

TABLE 3 Poisson Model Maximum Likelihood Estimates for KABCO Crashes for Model with Three Quadratic Trajectories

Group	Parameter	Estimate	Standard Error	*t* for H_0: Parameter = 0	Prob >|*t*|
1	Intercept	–3.001	140.309	–0.021	0.983
	Linear	–0.023	0.139	–0.165	0.869
	Quadratic	0.00001	0.00003	0.339	0.735
2	Intercept	–0.707	43.687	–0.016	0.987
	Linear	–0.068	0.044	–1.567	0.117
	Quadratic	0.00003	0.00001	3.203	0.001
3	Intercept	0.546	34.676	0.016	0.987
	Linear	–0.047	0.035	–1.357	0.175
	Quadratic	0.00002	0.00001	2.794	0.005

NOTE: t = t-statistic; H_0 = assumption that parameter equals zero.

TABLE 4 Group Membership for KABCO Crashes for Model with Three Quadratic Trajectories

Group	Road Segments %	Standard Error	*t* for H_0: Parameter = 0	Prob >|*t*|
1	59.297	1.477	40.151	0.000
2	34.893	1.380	25.293	0.000
3	5.811	0.544	10.678	0.000

for the first group, which consisted of 59.3% road segments, was almost equal to zero for the entire time period. [Note that the long-term mean was never equal to zero; see Lord and Mannering (*14*).] This group could be classified as low-crash-risk roads. The number of crashes per mile for the second group, which consisted of 34.9% road segments, was not high. This group could be classified as sites with a medium crash risk. Little increase in crash risk occurred at these sites from 2003 to 2007, compared with 1997 to 2001. The third group, which consisted of 5.8% of the road segments, could be considered high risk, because the number of crashes per mile was significantly higher than for the other two groups.

Figure 1 shows that, although there was a significant overall increase in the number of crashes per mile for high-risk locations, the variations between 1997 and 2001 and between 2003 and 2007 were relatively small (compared with the increase). The variation was still higher, however, than it was for the other two groups for the same time periods. The increase observed in 2003 can be explained by the change in the definition of reportable PDO collisions, as discussed earlier. From 1997 to 2001, a reportable PDO crash included only crashes in which at least one vehicle had to be towed away from the site of a collision. After 2003, a reportable PDO crash included all vehicles that sustained at least $1,000 in damages. Given the change in definition of a reportable collision, the PDO collisions were removed, and the analysis focused on injury crashes (KABC). The results are presented in the next section.

Longitudinal Analysis of KABC Traffic Crashes from 1997 to 2007

The same procedure as the one described in the above section was used to determine the form of model. The best number of subgroups was still equal to three. Similar to what the KABCO data showed, the results deemed the linear trajectory model as the best model (Tables 7 and 8). The graphical representation of the results is shown in Figure 2.

Figure 2 shows that all road segments in this data set still could be divided into three groups, which could be defined as low-, medium-, and high-crash-risk sites. The first group (low number of crashes per mile) contained 75.0% of the road segments, while the second group (medium number of crashes per mile) included 21.5%. The remaining sites accounted for 3.5% of all road segments. Compared with the KABCO data, the data in this case identified more roads as relatively safe. The trend showed that the number of injury crashes had been falling steadily, a trend that has been observed elsewhere in the United States (*25*). The decrease in injury crashes was more significant at sites that experienced a high number of crashes per mile.

The reduction observed and illustrated in Figure 2 could be explained by various factors. For instance, given the price of gas, which has been increasing since about 2003, the average traveling speed may have been lower from 2003 to 2007. Higher gas prices could also have resulted in a reduction in the total miles driven (not captured in this model). The Texas Department of Transportation also implemented some safety measures, such as rumble strips and raised pavement markers, on several rural, two-lane highways. This information was not available at the time this study was done. The decrease

TABLE 5 Poisson Model Maximum Likelihood Estimates for KABCO Crashes for Model with Three Linear Trajectories

Group	Parameter	Estimate	Standard Error	*t* for H_0: Parameter = 0	Prob >|*t*|
1	Intercept	–50.393	21.269	–2.369	0.018
	Linear	–0.024	0.011	2.293	0.022
2	Intercept	–138.729	10.956	–12.663	0.000
	Linear	0.069	0.005	12.683	0.000
3	Intercept	–95.172	11.003	–8.650	0.000
	Linear	0.048	0.005	8.789	0.000

TABLE 6 Group Membership for KABCO Crashes for Model with Three Linear Trajectories

Group	Road Segments %	Standard Error	t for H_0: Parameter = 0	Prob > |t|
1	59.297	1.477	40.158	0.000
2	34.893	1.379	25.295	0.000
3	5.811	0.544	10.679	0.000

in annual injury crashes may also be attributed in part to improvements in vehicle design, driving behavior, or both. Finally, on the basis of the results, the sites in Group 3 could be further investigated to determine the factors that influenced the reduction in injury crashes.

Analysis of Risk Factors

It is important to know whether and to what degree some geometric features are associated with elevated levels of crash risk once it is determined into which subgroups the road sections are classified. The three variables investigated were the speed limit, average shoulder width, and the total paved surface width. Traffic volume was analyzed separately, because it was a time-varying variable, and the results are presented in the next section.

Table 9 presents the summary results for the three risk factors. This table shows that the coefficient for the speed limit was positive, which meant that the risk would increase at higher speed limits and the likelihood of belonging to the higher-crash-risk group also increased. Use of an LCGM therefore may help identify high-risk segments and implement strategies to reduce operational speeds, including the speed limit. The same figure shows, however, that the coefficients for surface width and average shoulder width were both negative, which meant that the likelihood of belonging to the higher-crash-risk group decreased as the surface width and shoulder width increased. The widening of surface and shoulder widths is another effective method to lower crash risk, which has been documented elsewhere (*26*).

Because the average shoulder width was closely linked to the paved surface width, it was possible that the variables could be correlated. Only speed limit and average shoulder width were analyzed, therefore, and the results are presented in Table 10. The effects of the shoulder width were more important when surface width was removed. It could be reasonably assumed that shoulder width and surface width might be correlated to a certain degree (*27, 28*).

Figure 3 illustrates the marginal relationships of the risk factors, speed limit and shoulder width, to the likelihood of belonging to the highest-crash-risk category versus the lowest-crash-risk category. Similar results can be seen in these three figures. As speed limit increased from 30 to 70 mph, so did the likelihood of safety-related problems with Group 3. As the shoulder width increased from zero to 10 ft, however, the likelihood of belonging to the high-opposition group decreased.

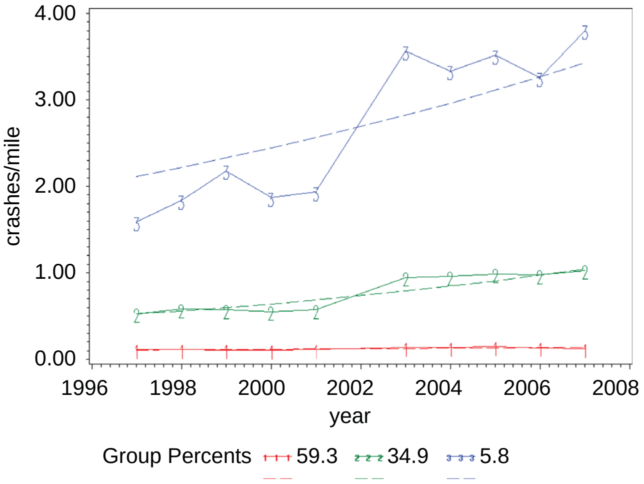

FIGURE 1 Three linear trajectories for KABCO crashes, 1997–2007.

Analysis of Time-Varying Covariates

The trajectories described above defined the developmental course for crash risk over different years. The trajectories, however, were not deterministic functions of time. External events during this period may have deflected a trajectory. Traffic volume has been recognized as one of the most important factors to affect crash risk. Because the traffic volume changed from 1997 to 2007, it was necessary to analyze the effect of traffic volume as a time-varying covariate. This study extended the basic three-group model by including this time-varying covariate to evaluate whether and how traffic volume was associated with the number of crashes.

Table 11 presents the results of the analysis for the time-varying covariate traffic volume. This table shows that the coefficient of traffic volume in each group was positive, which meant that the possibility of a crash increased as traffic volume increased, and the likelihood of belonging to the higher-crash-risk group increased. In addition, the *p*-value of the test for the parameter of the traffic volume was much lower than .05, which meant that the effect of the traffic volume was significant. The parameter of the traffic volume for Group 1 was .00010, whereas for Group 3 it was only .00003. This meant that, as the traffic volume increased, the effect on the number of crashes per

TABLE 7 Poisson Model Maximum Likelihood Estimates for KABC Crashes for Model with Three Linear Trajectories

Group	Parameter	Estimate	Standard Error	t for H_0: Parameter = 0	Prob > |t|
1	Intercept	85.721	23.868	3.591	0.0003
	Linear	–0.044	0.012	–3.717	0.0002
2	Intercept	80.297	13.728	5.849	0.000
	Linear	–0.040	0.007	–5.890	0.000
3	Intercept	34.123	15.245	2.238	0.000
	Linear	–0.016	0.008	–2.184	0.025

TABLE 8 Group Membership for KABC Crashes for Model with Three Linear Trajectories

Group	Road Segments %	Standard Error	*t* for H_0: Parameter = 0	Prob > |*t*|
1	74.990	1.288	58.230	0.000
2	21.479	1.157	18.565	0.000
3	3.531	0.456	7.736	0.000

mile was more important on road segments with low crash risk than on segments classified as high risk.

SUMMARY AND FURTHER RESEARCH

This study showed that the LCGM can be a useful technique for analyzing variables that change over time. It is a good alternative to raw and residual change scores as well as to standard growth approaches whenever multiple and somewhat contradictory patterns of change can be observed in the data. This paper documents the following results:

- The rural, two-lane road segments could be divided into three groups: sites with low crash risk (but the long-term mean was not equal to zero), sites with medium crash risk, and high-risk locations.
- It was shown that 59.3% of all road segments were identified as low-crash-risk sites and 34.9% were identified as medium-crash-risk sites, while the remaining 5.8% were categorized as high-risk locations for KABCO crashes. There was a significant increase in the number of crashes per mile for medium- and high-risk locations from 1997 to 2007. The main reason for the increase appeared to be the change in definition of a reportable collision.
- For injury crashes, about 75.0% of all road segments were identified as low-crash-risk sites, and 21.5% were identified as medium-crash-risk sites, while the remaining 3.5% were categorized as high-risk locations. As opposed to KABCO, there was a decrease in all three subgroups, and the decrease was more important for sites classified as high-crash-risk sites.

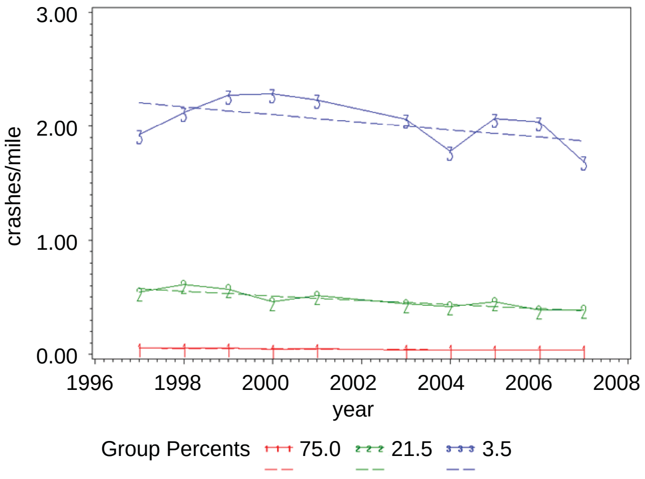

FIGURE 2 Three linear trajectories for KABC traffic crash risk, 1997–2007.

- The risk factor analysis showed that crash risk increased at higher speed limits, and this relationship was more important for the higher-crash-risk group. The likelihood of belonging to the higher-crash-risk group decreased, however, as the surface width and average shoulder width increased.
- The analysis showed that the time-varying covariate traffic volume was highly linked to the number of crashes per mile. In addition, the effects were more significant for road segments classified as low-crash-risk sites.

This paper provides initial insights into the effects of some geometric features and traffic volumes as a function of the groups. Although the research presented was somewhat theoretical, the methodology and results have some practical applications. For instance, more accurate policy decisions could be made by analyzing the trend in

TABLE 9 Risk Factors Related to Speed Limit, Surface Width, and Shoulder Width

Group	Parameter	Estimate	Standard Error	*t* for H_0: Parameter = 0	Prob > |*t*|
Poisson Model Maximum Likelihood Estimates					
1	Intercept	−56.299	21.017	−2.679	0.007
	Linear	0.027	0.011	2.602	0.001
2	Intercept	−138.759	10.963	−12.657	0.000
	Linear	0.069	0.005	12.677	0.000
3	Intercept	−95.728	11.044	−8.667	0.000
	Linear	0.048	0.005	8.806	0.000
Group Membership					
1	Constant	0.000	—	—	—
2	Constant	0.802	0.411	1.953	0.051
	Shoulder width	−0.107	0.031	−3.397	0.0007
	Speed limit	0.017	0.006	3.045	0.0023
	Surface width	−0.015	0.006	−2.386	0.0171
3	Constant	−0.553	0.703	−0.786	0.562
	Shoulder width	−0.156	0.051	−3.051	0.023
	Speed limit	0.029	0.010	2.892	0.0038
	Surface width	−0.0069	0.010	−0.667	0.505

NOTE: — = not applicable.

TABLE 10 Risk Factors Related to Speed Limit and Shoulder Width

Group	Parameter	Estimate	Standard Error	t for H_0: Parameter = 0	Prob > |t|
Poisson Model Maximum Likelihood Estimates					
1	Intercept	–55.176	21.099	–2.615	0.009
	Linear	0.026	0.010	2.538	0.011
2	Intercept	–138.828	10.973	–12.651	0.000
	Linear	0.069	0.005	12.671	0.000
3	Intercept	–95.802	11.022	–8.692	0.000
	Linear	0.048	0.005	8.831	0.000
Group Membership					
1	Constant	0.000	—	—	—
2	Constant	0.218	0.332	0.656	0.512
	Shoulder width	–0.111	0.031	–3.536	0.0007
	Speed limit	0.014	0.006	2.500	0.0023
3	Constant	–0.829	0.567	–1.463	0.143
	Shoulder width	–0.159	0.051	–3.109	0.002
	Speed limit	0.028	0.010	2.806	0.005

NOTE: — = not applicable.

the number of crashes over time in distinct groups. Once these groups were identified, it would be possible to investigate each one individually, and this could help to identify unknown factors that influenced the number and severity of crashes for each group.

Now that this paper has documented the feasibility of applying LCGM to the analysis of longitudinal data, much more work needs to be done on the topic. Many factors (e.g., the presence of a guardrail, utility poles, side slope, horizontal curve density) should be included in future analyses. Other research should be done to determine the common characteristics associated with each group. Perhaps safety treatments could be targeted for each. Finally, because overdispersion has been observed in crash data, an LCGM might be developed that

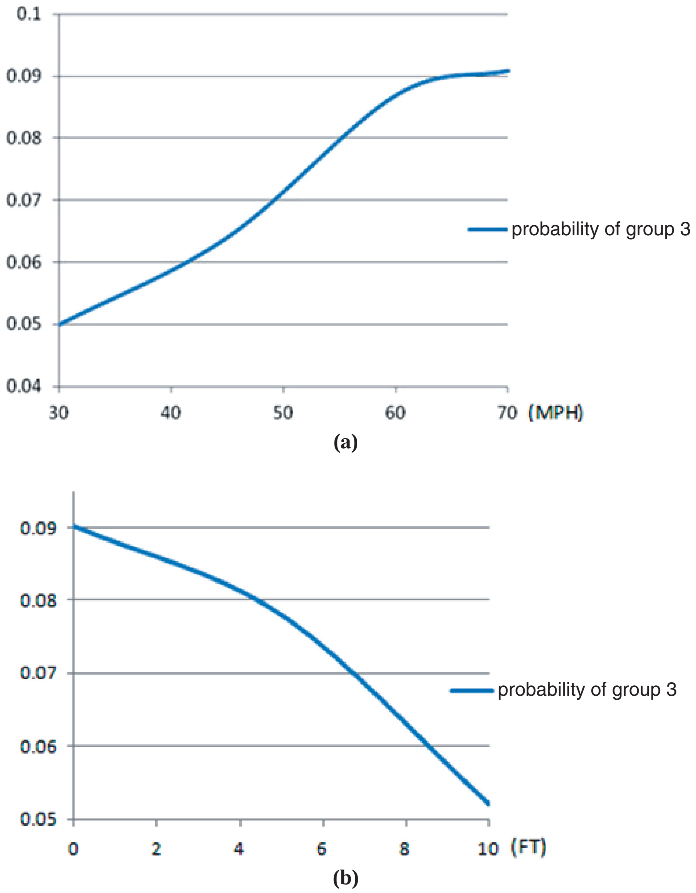

FIGURE 3 Probability of belonging to Group 3 on basis of (*a*) speed limit and (*b*) average shoulder width.

TABLE 11 Results of Analysis for Time-Varying Traffic Volume

	Poisson Model Maximum Likelihood Estimates				
Group	Parameter	Estimate	Standard Error	t for H_0: Parameter = 0	Prob >|t|
1	Intercept	30.494	23.244	1.312	0.189
	Linear	–0.016	0.012	–1.388	0.165
	ADT current year	0.00010	0.00001	10.012	0.000
2	Intercept	–157.062	12.021	–13.066	0.000
	Linear	0.078	0.006	13.077	0.000
	ADT current year	0.00005	0.00001	6.309	0.000
3	Intercept	–103.562	10.797	–9.591	0.000
	Linear	0.052	0.005	9.725	0.000
	ADT current year	0.00003	0.000	6.591	0.000

could accommodate Poisson-gamma models. At this point, these models cannot be used for LCGM.

ACKNOWLEDGMENTS

The authors thank Bobby L. Jones of Carnegie Mellon University for offering suggestions on this analysis and for providing the tools to analyze the data. The authors also acknowledge Debra Vermillion and the Texas DOT for providing the data.

REFERENCES

1. Hauer, E. *Observational Before-After Studies in Road Safety.* Pergamon Press, Elsevier Science Ltd., Oxford, England, 1997.
2. Persaud, B. N., R. A. Retting, P. E. Garder, and D. Lord. Safety Effect of Roundabout Conversions in the United States: Empirical Bayes Observational Before-After Study. In *Transportation Research Record: Journal of the Transportation Research Board, No. 1751,* TRB, National Research Council, Washington, D.C., 2001, pp. 1–8.
3. Carriquiry, A. L., and M. Pawlovich. *From Empirical Bayes to Full Bayes: Methods for Analyzing Traffic Safety Data.* White paper. Iowa State University, Ames, 2004.
4. Park, E.-S., J. Park, and T. J. Lomax. A Fully Bayesian Multivariate Approach to Before-After Safety Evaluation. *Accident Analysis and Prevention,* Vol. 42, No. 4, 2010, pp. 1118–1127.
5. Maher, M. J., and I. Summersgill. A Comprehensive Methodology for the Fitting of Predictive Accident Models. *Accident Analysis and Prevention,* Vol. 28, No. 3, 1996, pp. 281–296.
6. Mountain, L., M. J. Maher, and B. Fawaz. The Influence of Trend on Estimates of Accidents at Junctions. *Accident Analysis and Prevention,* Vol. 30, No. 5, 1998, pp. 641–649.
7. 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, National Research Council, Washington, D.C., 2000, pp. 102–108.
8. Wang, X., and M. Abdel-Aty. Temporal and Spatial Analyses of Rear-End Crashes at Signalized Intersections. *Accident Analysis and Prevention,* Vol. 38, No. 6, 2006, pp. 1137–1150.
9. Giuffrè, O., A. Granà, T. Giuffrè, and R. Marino. Improving Reliability of Road Safety Estimates Based on High Correlated Accident Counts. In *Transportation Research Record: Journal of the Transportation Research Board, No. 2019,* Transportation Research Board of the National Academies, Washington, D.C., 2007, pp. 197–204.
10. Lord, D., and M. Mahlawat. Examining Application of Aggregated and Disaggregated Poisson–Gamma Models Subjected to Low Sample Mean Bias. In *Transportation Research Record: Journal of the Transportation Research Board, No. 2136,* Transportation Research Board of the National Academies, Washington, D.C., 2009, pp. 1–10.
11. Holder, W. Mandated Server Training and Reduced Alcohol Involved Traffic Crashes: A Time Series Analysis of the Oregon Experience. *Accident Analysis and Prevention,* Vol. 26, No. 1, 1994, pp. 89–97.
12. Rehm, J., and G. Gmel. Aggregate Time-Series Regression in the Field of Alcohol. *Addiction,* Vol. 96, No. 7, 2001, pp. 945–954.
13. Park, B.-J., and D. Lord. Application of Finite Mixture Models for Vehicle Crash Data Analysis. *Accident Analysis and Prevention,* Vol. 41, No. 4, 2009, pp. 683–691.
14. 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.
15. Malyshkina, N. V., F. L. Mannering, and A. P. Tarko. Markov Switching Negative Binomial Models: An Application to Vehicle Accident Frequencies. *Accident Analysis and Prevention,* Vol. 41, No. 2, 2009, pp. 217–226.
16. Malyshkina, N., and F. L. Mannering. Zero-State Markov Switching Count-Data Models: An Empirical Assessment. *Accident Analysis and Prevention,* Vol. 42, No. 1, 2010, pp. 122–130.
17. McGinnis, R. G., M. J. Davis, and E. A. Hathaway. Longitudinal Analysis of Fatal Run-Off-Road Crashes, 1975 to 1997. In *Transportation Research Record: Journal of the Transportation Research Board, No. 1746,* TRB, National Research Council, Washington, D.C., 2001, pp. 47–58.
18. Curran, P. J., and B. O. Muthén. The Application of Latent Curve Analysis to Testing Developmental Theories in Intervention Research. *American Journal of Community Psychology,* Vol. 27, 1999, pp. 567–595.
19. Nagin, D. S. Group-Based Modeling of Development. *Psychology Compass,* Vol. 2, 2005, pp. 302–317.
20. Raudenbush, S. W. Comparing Personal Trajectories and Drawing Causal Inferences from Longitudinal Data. *Annual Review of Psychology,* Vol. 52, 2001, pp. 501–525.
21. Nagin, D. S. Overview of a Semi-Parametric, Group-Based Approach for Analyzing Trajectories of Development. *Proc., Statistics Canada Symposium 2002: Modelling Survey Data for Social and Economic Research.* Ottawa, Ontario, Canada, 2002.
22. Roberts, B. W., K. E. Walton, and W. Viechtbauer. Patterns of Mean Level Change in Personality Traits Across the Life Course: Analysis of Longitudinal Studies. *Psychological Bulletin,* Vol. 132, 2006, pp. 1–25.
23. Jones, B. L., S. Daniel, and K. R. Nagin. A SAS Procedure Based on Mixture Models for Estimating Developmental Trajectories. *Sociological Methods and Research,* Vol. 29, No. 3, 2001, pp. 374–393.
24. Nagin, D. S. Analyzing Developmental Trajectories: A Semi-Parametric, Group-Based Approach. *Psychological Methods,* Vol. 4, 1999, pp. 139–157.
25. NHTSA, U.S. Department of Transportation. *2008 Traffic Safety Annual Assessment: Highlights.* www-nrd.nhtsa.dot.gov/pubs/811172.pdf. Accessed July 28, 2010.
26. *Highway Safety Manual,* 1st ed. American Association of State Highway and Transportation Officials, Washington, D.C., 2010.
27. Gross, F., and P. P. Jovanis. Estimation of the Safety Effectiveness of Lane and Shoulder Width: Case-Control Approach. *Journal of Transportation Engineering,* Vol. 133, No. 6, 2007, pp. 362–369.
28. Li, X., D. Lord, and Y. Zhang. Development of Accident Modification Factors for Rural Frontage Road Segments in Texas Using Results from Generalized Additive Models. *Journal of Transportation Engineering.* Vol. 137, No. 1, 2010, pp. 74–83.

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