Accident Prediction Models With and Without Trend: Application of the Generalized Estimating Equations (GEE) Procedure

Dominique Lord
Safety Studies Group
Department of Civil Engineering, University of Toronto
Toronto, Canada M5S 1A4
Tel.: (416) 978-3673
Fax: (416) 978-5054
e-mail: lord@civ.utoronto.ca

Bhagwant N. Persaud
Department of Civil Engineering, Ryerson Polytechnic University
350 Victoria Street,Toronto, Canada M5B 2K3
Tel.: (416) 979-5000, ext. 6464
Fax: (416) 979-5122
e-mail: bpersaud@acs.ryerson.ca

Transportation Research Board
79th Annual Meeting
January 9-15, 2000
Washington, D.C.
Paper No. 00-0496

ABSTRACT

Accident prediction models (APMs) are very useful tools for estimating the expected number of accidents on entities such as intersections and road sections. These estimates are typically used in the identification of sites for possible safety treatment and in the evaluation of such treatments. An APM is, in essence, a mathematical equation that expresses the average accident frequency of a site as a function of traffic flow and other site characteristics. The reliability of an APM estimate is enhanced if the APM is based on data for as many years as possible especially if data for those same years are used in the safety analysis of a site. With many years of data, however, it is necessary to account for the year-to-year variation, or trend, in accident counts because of the influence of factors that change every year. To capture this variation, the count for each year is treated as a separate observation. Unfortunately, the disaggregation of the data in this manner creates a temporal correlation that presents difficulties for traditional model calibration procedures. The objective of this paper is to present an application of a generalized estimating equations (GEE) procedure to develop an APM that incorporates trend in accident data. Data for the application pertains to a sample of 4–legged signalized intersections in Toronto, Canada for the years 1990 to 1995. The GEE model incorporating the time trend is shown to be superior to models that do not accommodate trend and/or the temporal correlation in accident data.

Key words: accident prediction models, temporal correlation, GEE, GLM, signalized intersections

INTRODUCTION

Accident counts on an entity often exhibit trends due to temporal changes in factors such as traffic flow, weather, the economy, and accident reporting practices. Thus, it stands to reason that accident prediction models (APMs) that accommodate these trends should provide better estimates of safety than traditional models in the identification of hazardous entities and in the evaluation of treatments applied to those entities.

In the traffic safety literature, few methods have been proposed on how to estimate the coefficients of APMs with trend. Depending on the outcome sought, these can be grouped into three categories -- marginal models (MM), transition models (TM) and random-effects models (REM). Details of these can be found in Diggle *et al.* (*1*) and Dunlop (*2*).

Most approaches in fact develop marginal models, which appear to be the most appropriate for APMs. For example, Maher and Summersgill (*3*) proposed an iterative solution, based on the method of "constructed variables" presented in McCullagh and Nelder (*4*), to find the proper estimate of the coefficients. A variation of this iterative solution is presented in Mountain *et al.* (*5*) who used an approach proposed by Atkinson (*6*). The difficulty with traditional marginal models is that the extent and type of temporal correlation needs to be known. In fact, Maher & Summersgill suggested that the modelling of year-to-year variation should be avoided whenever possible because of the difficulty in handling the temporal correlation.

Hauer (*7*) put forward a multinomial maximum likelihood function to estimate the coefficients of what is classified as a transition model (TM). However, this approach is cumbersome in that it needs numerous mathematical manipulations which may be out of grasp for the average modeler. In addition, transition models may not always be appropriate for traffic safety modelling.

Shankar *et al.* (*8*) applied a random-effects model (REM) proposed by Guo (*9*) to estimate coefficients for median cross-over accidents using the maximum likelihood method. However, this method assumes that the repeated observations are independent and that all the model coefficients vary from year to year.

In summary, available modelling approaches for incorporating trend in temporally correlated data suffer from one or more of the following limitations:

1. The temporal correlation in the data is ignored (REM and some Marginal Models)

2. The model type may not always be appropriate for APMs (REM and TM)

3. They are too complicated for the average modeller (TM and Marginal Models).

The generalized estimating equations (GEE) procedure proposed by Liang and Zeger (*10*), and Zeger and Liang (*11*) overcomes these difficulties by making it relatively easy for a modeler to develop proper and unbiased marginal models for repeatedly measured data. The procedure can be used even if the extent and the type of correlation is unknown. Several statistical software packages already have a built-in GEE calibration facility.

The objective of this paper is to illustrate the application of the GEE procedure to traffic safety studies when several years of data are available and it is desirable to incorporate trend. The application is for a sample of 4-legged signalized intersections in Toronto Canada, using data for the years 1990-1995. The GEE model with trend is compared to one which does not incorporate trend and to GLMs which do not account for temporal correlation in the accident count data. It is necessary to first provide some background on accident modelling before introducing the GEE concept.

SOME BASICS OF ACCIDENT MODELLING

An APM is, in essence, a mathematical equation that expresses the average accident frequency of an entity as a function of traffic flow and other road characteristics. The coefficients of APMs cannot be estimated by the traditional ordinary least squares or weighted least squares regression methods. This is because the assumptions for these methods are violated by the discrete, non-negative nature of accident count data and the reality that the variance in the number of accidents increases as the traffic flow increases. Thus, it is now common to estimate the coefficients of APMs by using maximum likelihood methods to calibrate what are referred to as generalized linear models (GLMs). Extensive descriptions of GLM estimation can be found in Dunlop (*2*), McCullagh and Nelder (*4*), and Myers (*12*).

A GLM usually consists of three components, a *random component*, a *systematic component*, and a *link function* that connects the random and systematic components to produce a linear predictor. The important property of the GLM is the flexibility in specifying the probability distribution for the random component. Thus, GLMs are especially useful in the context of traffic safety, for which the distribution of accident counts in a population often follows the negative binomial distributions (*13, 14*).

There exist many model forms for APMs, but one of the most common ones for intersections is the following:

$$E\{\kappa\} = \alpha F_1^{\beta_1} F_2^{\beta_2}, \qquad (1a)$$

or the GLM linear version

$$\ln(E\{\kappa\}) = \ln(\alpha) + \beta_1 \ln(F_1) + \beta_2 \ln(F_2), \quad (1b)$$

where,

$E\{\kappa\}$ = the expected number of accidents per unit of time;
F_1, F_2 = the entering flows (e.g., vehicles/day, vehicles/hour) on the major and minor roads respectively;
α, β_1, β_2 = coefficients to be estimated.

These coefficients are estimated by the maximum likelihood procedure using a variant of the Newton-Raphson method (*15*). The conventional application of GLMs to estimate APMs without trend is well developed and the reader is referred to the work of Kulmala (*13*), Nicholson and Turner (*14*), Hauer *et al.* (*16*), Miaou (*17*) for more details.

For traffic safety applications, it is desirable to estimate different coefficients for each year for which data are available. For logical reasons, it is usually assumed that the β's, are constant from year to year and it is therefore only necessary to estimate the different α's for each year. In the estimation of these α's, each annual accident count is an observation, which creates a difficulty since these counts are correlated.

To appreciate the difficulty caused by temporal correlation, consider a simple example in which the model defined by Equation (1) is to be developed for longitudinal data for which accidents and traffic flows are available for different time periods (t) at intersections identified from $i = 1$ to I. The model is given by the following equation, with the parameter α in Equation 1 replaced by β_0 to simplify the illustration:

$$E\{\kappa_t\} = \beta_0 F_{1t}^{\beta_1} F_{2t}^{\beta_2}, \quad (2)$$

where,

$E\{\kappa_t\}$ = the expected number of accidents per time period t;
F_{1t}, F_{2t} = the entering flows (e.g., vehicles per day, vehicles per hour) on the major and minor roads respectively for year t;
$\beta_0, \beta_1, \beta_2$ = coefficients to be estimated.

The GLM estimate of β for equation (2) is the solution to the following estimating equation

$$\sum_{i=1}^{I} \boldsymbol{D}_i' \boldsymbol{V}_i^{-1} (\boldsymbol{Y}_i - \boldsymbol{u}_i) = \boldsymbol{0} \quad (3)$$

where,

$$\mu_i = g^{-1}(\boldsymbol{X}_i\beta);$$

$$\boldsymbol{V}_i = \sigma^2\left[(1-\rho)\boldsymbol{I} + \rho\boldsymbol{J}\right] \text{ (covariance matrix);}$$

$$\boldsymbol{D}_i = \frac{\partial \mu_i}{\partial \beta} = \begin{bmatrix} \frac{\partial \mu_{1i}}{\partial \beta_0} & \frac{\partial \mu_{1i}}{\partial \beta_1} & \frac{\partial \mu_{1i}}{\partial \beta_2} \\ \vdots & \ddots & \vdots \\ \frac{\partial \mu_{ni}}{\partial \beta_0} & \frac{\partial \mu_{ni}}{\partial \beta_1} & \frac{\partial \mu_{ni}}{\partial \beta_2} \end{bmatrix}.$$

In equation (3), $\boldsymbol{I}$ is the n_i x n_i identity matrix, $\boldsymbol{J}$ is the n_i x n_i matrix all of whose elements are 1, and ρ is the correlation coefficient between any two measurements at the same intersection i. (Note: the data in this example is assumed to be uniformly correlated, but the reader should be aware that other types of temporal correlation also exist. (See *(1)*)).

The variance of the GLM estimate of β then becomes

$$\begin{aligned} Var(\beta) &= \sigma^2\left[\sum_{i=1}^{I} \boldsymbol{D}_i'\boldsymbol{V}_i^{-1}\boldsymbol{D}_i\right]^{-1} \\ &= \sigma^2\left[\sum_{i=1}^{I} \boldsymbol{D}_i'\{(1-\rho)\boldsymbol{I} + \rho\boldsymbol{J}\}^{-1}\boldsymbol{D}_i\right]^{-1}. \end{aligned} \qquad (4)$$

In equation (4), if the observation is positively correlated ($\rho > 0$), which often occurs when the repeated accident counts for the same intersection are used, the variance of $\beta_0, \beta_1, \beta_2$ will be increased by a factor ρ. Thus, the variance will be underestimated if this correlation is ignored. More important, ignoring the temporal correlation may also have an impact on the proper selection of coefficients as some coefficients may be wrongly accepted as significant because of the underestimated variance. For example, one might conclude that the year to year differences in β_0 are significant when in fact they are not.

The coefficients of the GLM incorporating trend in temporally correlated data can still be estimated using the traditional maximum likelihood methods. However, the likelihood function can be very complicated to define and solve. For instance, additional assumptions are routinely needed to specify the likelihood function of non-Gaussian data. And, even if these assumptions are valid, the likelihood often involves numerous nuisance parameters that must be estimated in addition to the explanatory variables. To overcome

this difficulty, an alternative method known as the GEE procedure was proposed by Liang and Zeger (*10*), and Zeger and Liang (*11*).

THE GENERALIZED ESTIMATING EQUATIONS (GEE) PROCEDURE

The GEE procedure is classified as a multinomial analogue of a quasi-likelihood function. The estimate of the coefficients can be found with the same equation as equation (3):

$$\sum_{i=1}^{I} \boldsymbol{D}_i' \boldsymbol{V}_i^{-1} (\boldsymbol{Y}_i - \mu_i) = \boldsymbol{0}, \tag{5}$$

where $\mathbf{D}_i = \partial\mu_i / \partial\beta$, as illustrated in equation (3). The temporal correlation in repeated observations can be described by a n_i x n_i matrix $R(\lambda)$, where λ represents the type of correlation with $\lambda = [\lambda_1, \ldots, \lambda_{n-1}]'$ and $\lambda_i = corr(Y_{it}, Y_{ik})$ *for* $t, k = 1, \ldots, n-1$ $t \neq k$, and n_i is the number of subjects. Therefore, the new covariance matrix now becomes:

$$\boldsymbol{V}_i = \boldsymbol{A}_i^{1/2} \boldsymbol{R}_i(\lambda) \boldsymbol{A}_i^{1/2}, \tag{6}$$

where A_i is an n_i x n_i matrix with $diag[V(\mu_{i1}), \ldots V(\mu_{iT_i})]$. The covariance matrix is given by

$$cov(\hat{\beta}) = \sigma^2 \left[\sum_{i=1}^{I} \boldsymbol{D}_i' \boldsymbol{V}_i^{-1} \boldsymbol{D}_i \right]^{-1}. \tag{7}$$

One can find the solution by simultaneously solving equations (6) and (7) with the iterative reweighted least squares method (*15*). This method is required because the estimates of both β *and* λ need to be found.

In order to solve the GEE correctly, every element of the correlation matrix $\boldsymbol{R}_i$ has to be known. However, in many instances, it is not possible to know the proper correlation type for the repeated measurements. To overcome this drawback, Liang and Zeger (*9*) proposed the use of a "working" matrix $\hat{\boldsymbol{V}}_i$ of the correlation matrix $\boldsymbol{V}_i$ which is based on the correlation matrix $\hat{\boldsymbol{R}}_i$. The estimate of the coefficients is found with the following equation:

$$\sum_{i=1}^{I} \boldsymbol{D}_i' \hat{\boldsymbol{V}}_i^{-1} (\boldsymbol{Y}_i - \mu_i) = \boldsymbol{0}. \tag{8}$$

The covariance matrix of equation (8) is given by

$$cov(\hat{\beta}) = \sigma^2\left[\sum_{i=1}^{I} \boldsymbol{D}_i'\hat{\boldsymbol{V}}_i^{-1}\boldsymbol{D}_i\right]^{-1}\left[\sum_{i=1}^{I} \boldsymbol{D}_i'\hat{\boldsymbol{V}}_i^{-1}\boldsymbol{V}_i\hat{\boldsymbol{V}}_i^{-1}\boldsymbol{D}_i\right]\left[\sum_{i=1}^{I} \boldsymbol{D}_i'\hat{\boldsymbol{V}}_i^{-1}\boldsymbol{D}_i\right]^{-1}. \quad (9)$$

The proposed methodology above, i.e. in equations (8) and (9), possesses one very useful property in that $\hat{\boldsymbol{\beta}}$ nearly always provides consistent estimates of $\boldsymbol{\beta}$ even if the matrix $\boldsymbol{V}_i$ has been improperly estimated. Thus, the confidence interval for $\boldsymbol{\beta}$ will be correct even when the covariance matrix is incorrectly specified. Therefore, it is not necessary to, a priori, examine the type of temporal correlation (e.g., independent, dependent). Techniques on how to analyze and interpret autocorrelation can be found in books on time series analysis such as the ones by Box and Jenkins (*18*) and Diggle (*19*). One important drawback, however, comes with this property. In order to assume that $\hat{\boldsymbol{\beta}}$ is the proper estimate of $\boldsymbol{\beta}$, it is required that the observation for each subject be known and available. If missing values exist, the estimate of the coefficients may be biased. The extent of the bias is influenced by the type of missing values, e.g. random or informative. Note that in the case of $\hat{\boldsymbol{V}}_i = \boldsymbol{V}_i$, equation (9) becomes the covariance matrix of equation (7).

APPLICATION OF THE GEE

The GEE procedure was applied to develop several APMs that are part of a research project on network safety conducted by the Safety Studies Group at the University of Toronto (*20*). The application presented in this paper pertains to 4-legged signalized intersections in Toronto, Canada.

Data

The accident data base consisted of motor vehicle accidents (fatal, injury, and property damage only) at 868 four-legged signalized intersections in central business district (CBD) and non-CBD areas of Toronto for each of the years 1990 to 1995. The characteristics of the data are presented in Table 1. Exploratory analysis revealed that most of the traffic growth in Toronto happened between 1985 and 1990, and remained fairly constant afterwards; this explains the relative similarity between the various flows shown in Table 1. Traffic counts were not available for all years at any given intersection. The missing counts were estimated according to the methodology developed by Lord (*21*). The application of this procedure was deemed necessary to reduce the possibility of introducing biased estimates with a database that included missing values that may be systematic in the sense that higher volume intersections tend to be counted more often than low volume ones.

TABLE 1. Characteristics of 4-legged signalized intersections sample (1990-1995)

Year	accidents (min-max-total)	major road flow (min-max)	minor road flow (min-max)
1990	0-44-8276	5305-71798	51-41306
1991	0-53-8141	5294-71527	52-41003
1992	0-58-8714	5342-71498	52-41150
1993	0-63-9818	5369-71450	52-41131
1994	0-54-10010	5464-72310	53-42012
1995	0-54-10030	5469-72178	53-42644
overall	0-63-54989	5305-72410	51-42644

Selection of Model Form

The selection of the model form was based on a method proposed by Hauer and Bamfo (*22*), and applied by Lord *et al.* (*23*), known as the **I**ntegral-**D**ifferentiate (ID) method. Basically, the method consists of separating each possible independent variable into a series of bins (e.g., one for each entity) that are placed in increasing order to create an **E**mpirical **I**ntegral **F**unction (EIF). For each entity, the left boundary of the bin is located halfway between the current entity and the previous entity. The right boundary is located halfway between the current entity and the next entity. The bin height is the number of accidents that occurred on that entity. Hence, the value of the EIF at the right boundary of the current bin is the sum of all bin areas from the lowest value up to that boundary. For instance, take three sites that have 5,000, 10,000, and 15,000 vehicles per day and 10, 12, and 15 accidents per year respectively. For the second bin, the left and right boundaries will be equal to 7,500 and 12,500 respectively and the height of the EIF at the right boundary will be equal to 22 (10+12). The goal of the method is to compare the EIF graph created above with pre-established graphs of well-known functions (power, gamma, polynomial, etc.) The graph that has a shape similar to one of the pre-established graphs should indicate the proper relationship between the dependent and the independent variables being investigated. The reader is referred to (*22*) for a better description of this method.

The explanatory variables selected were the entering AADTs (annual average daily traffic) on the major road (F_1) and minor road (F_2) respectively. Thus, e.g., F_1 equals the summation of the entering AADTs on opposite approaches of the major road. The ID method showed that the flow F_1 followed a power relationship and the flow F_2 followed a Gamma relationship. The resulting model form was the following:

$$E\{\kappa\} = \alpha F_1^{\beta_1} F_2^{\beta_2} e^{(\beta_3 F_2)}, \quad (10)$$

where,

$E\{\kappa\}$ = the expected annual number of accidents;
F_1, F_2 = the entering AADTs on the major and minor roads respectively;
$\alpha, \beta_1, \beta_2, \beta_3$ = coefficients to be estimated.

Model Calibration

Five different APM's were calibrated for comparative purposes using the same sample and the GENSTAT software package (*24*). These models are described below.

Models 1 and 2: These account for time trend by estimating different α's in equation (10) for each year based on the AADTs in that year. For this, equation (10) becomes

$$E\{\kappa_t\} = \alpha_t F_{1t}^{\beta_1} F_{2t}^{\beta_2} e^{(\beta_3 F_{2t})}, \quad (11a)$$

where the subscript *t* indicates the model year. The observations used to calibrate Equation 11 were the accident counts and AADTs for each year for each intersection. For Model 1, the GEE procedure was used; for Model 2, the conventional GLM procedure was used, meaning that the temporal correlation is ignored.

Models 3, 4 and 5: These do not incorporate time trend, i.e., α in equation (10) is the same for each year. For Models 3 and 4, the observations were the accident counts and AADTs for each year for each intersection with the GEE procedure was used for Model 3 and the conventional GLM procedure used for Model 4. For Models 3 and 4, equation (10) becomes

$$E\{\kappa_t\} = \alpha F_{1t}^{\beta_1} F_{2t}^{\beta_2} e^{(\beta_3 F_{2t})}. \quad (11b)$$

The GLM procedure is also used for Model 5, but the difficulty created by the temporal correlation in the annual counts is avoided by using the total accident counts and an average AADT over 6 years. A yearly APM is derived by simply dividing the "6-year" α by 6. Thus, for Model 5, the model is as in equation (10)

$$E\{\kappa\} = \alpha F_1^{\beta_1} F_2^{\beta_2} e^{(\beta_3 F_2)}, \qquad (11c)$$

where, F_1, F_2 are the *average* entering AADTs on the major and minor roads respectively between 1990 and 1995.

Results

The results for the estimates and standard errors of the coefficients are presented in Table 2. These are given for the $\ln(\alpha)$s since GENSTAT actually calibrates a linear version of equation 11. For all of the models, a negative binomial distribution with dispersion parameter γ, was used to specify the error structure. Estimates of γ, which are shown at the bottom of Table 2, were obtained by an iterative maximum likelihood procedure described elsewhere (*20*). The size of γ is in fact a measure of the amount of variation explained by the model since $\hat{\text{Var}}\{\kappa\} = \hat{\kappa}^2/\gamma$.

TABLE 2. Estimates of Model Coefficients (and standard errors)

Models 1 and 2: $E\{\kappa_t\} = \alpha_t F_{1t}^{\beta 1} F_{2t}^{\beta 2} e^{(\beta 3\, F2t)}$,

Models 3 and 4: $E\{\kappa_t\} = \alpha F_{1t}^{\beta 1} F_{2t}^{\beta 2} e^{(\beta 3\, F2t)}$,

Model 5: $E\{\kappa\} = \alpha F_1^{\beta 1} F_2^{\beta 2} e^{(\beta 3\, F2)}$.

Coefficients	GEE with trend **Model 1**	GLM with trend **Model 2**	GEE without trend **Model 3**	GLM without trend	
				Yearly data **Model 4**	6-year data **Model 5**
Year1 [LN(α_1)]	-8.443 (0.599)	-8.443 (0.302)			
Year2 [LN(α_2)]	-8.453 (0.598)	-8.453 (0.302)			
Year3 [LN(α_3)]	-8.392 (0.598)	-8.392 (0.302)			
Year4 [LN(α_4)]	-8.303 (0.597)	-8.303 (0.302)			
Year5 [LN(α_5)]	-8.317 (0.598)	-8.317 (0.302)			
Year6 [LN(α_6)]	-8.321 (0.600)	-8.321 (0.302)			
LN(α)			-8.424 (0.590)	-8.424 (0.303)	-8.048 (0.538)
β_1	0.527 (0.041)	0.527 (0.021)	0.534 (0.041)	0.534 (0.021)	0.534 (0.038)
β_2	0.568 (0.043)	0.568 (0.023)	0.566 (0.043)	0.566 (0.023)	0.518 (0.039)
β_3	8.61E-6 (3.97E-6)	8.61E-6 (2.21E-6)	8.92E-6 (3.97E-6)	8.92E-6 (2.22E-6)	1.34E-5 (4.13E-6)
γ	6.91	6.91	6.91	6.91	6.87

The results in Table 2 lead to three interesting observations. First, as expected, Models 1 and 2 have identical coefficient estimates. This is explained by the fact that no missing values exist in the database and the same data are used to calibrate each model. Likewise, the coefficients for Models 3 and 4 are identical. Second, the table shows that the temporal correlation contributes to about half of the standard errors as evidenced by the fact that the standard errors for the GEE estimates for Models 1 and 3 are approximately twice as large as those for their GLM counterparts (Models 2 and 4). Although the GEE coefficients remain statistically significant at the 5% level, in general, there is no guarantee that this will happen. In theory, some explanatory variables may become insignificant when temporal correlation is considered. Third, if it is not of interest to model the time trend in the data, it is still beneficial to use the GEE procedure as evidenced by the fact that the dispersion parameter γ for Model 3 is slightly higher than that for Model 5, the only GLM that overcomes the problem of temporal correlation. This

result is not unexpected, since the GEE procedure allows for each year's AADT to be different while the GLM procedure in effect uses an AADT averaged over the six years.

Having shown that the GEE is useful for both APMs with and without trend, it remains remains to address the issue of whether or not the modelling of trend is useful. According to equation (11a), α_t is in fact an index of the relative year to year variation in accidents. From the values of $\ln(\alpha_t)$ in Table 2, one can calculate the following values of α_t from Model 1 (Table 3).

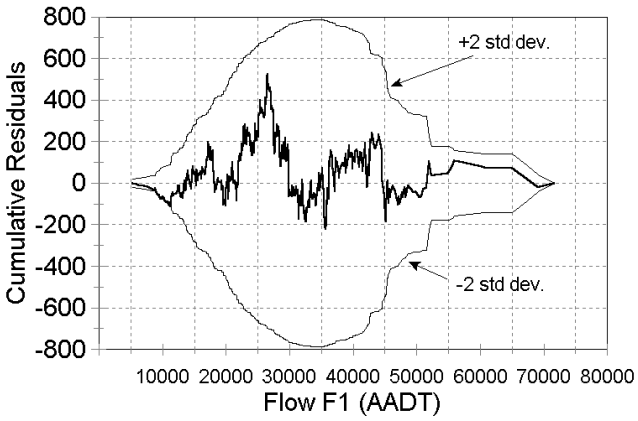
TABLE 3. Estimates of the coefficient α_t for Model 1

YEAR	1	2	3	4	5	6
α_t	0.000215	0.000213	0.000227	0.000248	0.000244	0.000243

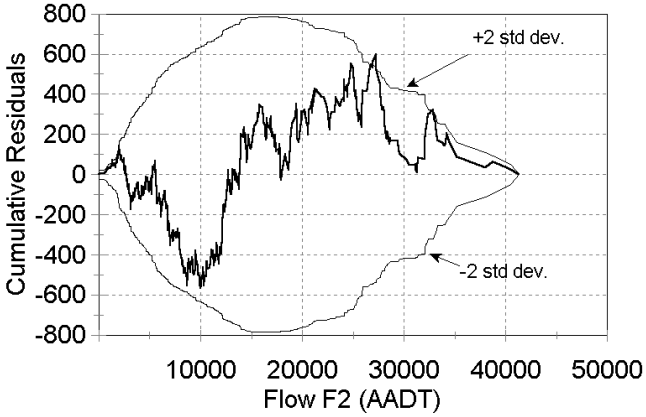
Based on the standard errors of $\ln(\alpha_t)$ in Table 2, it would tempting for a modeller to conclude that the year to year differences are not statistically significant and therefore one should select Model 3 with a common α for each year. This decision would in fact be supported by the result that the values of γ are identical. However, note that the values of α_t for years 4 to 6 are approximately 15% higher than those for years 1 and 2. Thus, a model with a common α for each year would overestimate accidents in years 1 and 2 and underestimate accidents in years 4, 5, 6. This would create some difficulty in longitudinal studies. To see this, imagine that intersections were treated in Year 3 and, for a proper before and after study (See Hauer, *25*), Model 3 is used in the estimation of the number of accidents that would have occurred in Years 4, 5, 6 without the treatment. Since this value is underestimated, the treatment effectiveness obtained by comparing it to the actual number of accidents in years 4, 5 and 6 would be underestimated. This difficulty would be avoided by using Model 1 since this model captures the increased accident experience in the "after" period that would have materialized without the treatment. Another benefit of using the time trend is that it allows the jurisdiction to identify and investigate potentially dangerous trends such as the 15% increase in accidents noted above. Thus, on balance, it seems beneficial to incorporate trend in developing APMs since trends that are insignificant in the statistical sense, such as the one in our application, still require consideration.

Model Validation

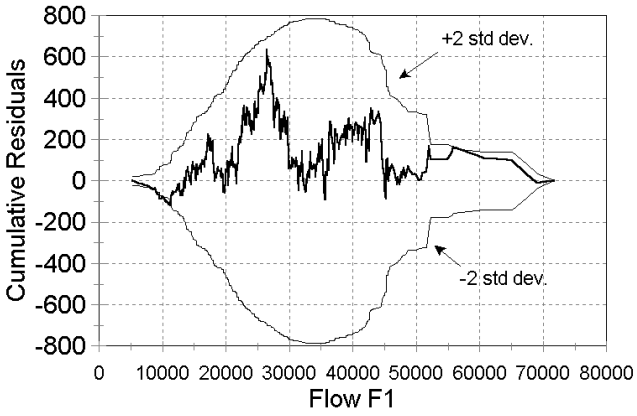
The quality of fit was investigated with the **Cu**mulative **Re**siduals (CURE) method (*22*, *23*). This method consists of plotting the cumulative residuals for each independent variable. The goal is to graphically observe how well the function fits the data set. The CURE method has the advantage of not being dependent on the number of observations as are many other traditional statistical procedures (e.g., R^2, etc.). The cumulative residuals were produced for Models 1 and 3. The cumulative residuals are presented for the flows F_1 and F_2 respectively in Figure 1.



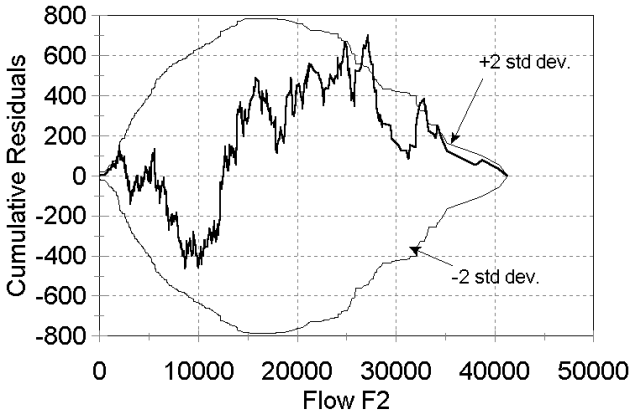
A) Model 1



B) Model 1



C) Model 3



D) Model 3

FIGURE 1. Cumulative residuals with ±2σ for Models 1 and 3: a) flow F_1 for Model 1; b) flow F_2 for Model 1; c) flow F_1 for Model 3; d) and flow F_2 for model 3

Figure 1 shows that Model 1 fits the data with better accuracy than Model 3. Indeed, the cumulative residuals reveal that the curves in Figures 1a and 1b oscillate closer to the value of 0 than the curves in Figures 1c and 1d. In fact, there is a range in Figure 1c for traffic flows above 25,000 veh/day where the curve slightly exceeds the two standard deviation boundaries for the flow F_2. The flows do not exceed the boundary by much but it is still worse than what is shown in Figure 1b for the same range of flows. Finally, the

CURE method supports the earlier conclusion that Model 1 is the best model -- that it is better to incorporate the time trend.

SUMMARY AND CONCLUSIONS

The purpose of this paper was to describe the application of the GEE procedure to develop accident prediction models which reflect the variation in accident occurrence from year to year. Approaches used in other research are either very cumbersome or do not account for the temporal correlation in annual accident counts. The GEE procedure overcomes these difficulties in developing adequate and unbiased APMs. The procedure also has the advantage that many statistical software packages already have a built-in GEE function.

The mathematics of the GEE procedure are summarized in the paper. It is seen that, while it is important (but not necessary) to have a complete data set, it is not fundamental to know *a priori* the proper temporal correlation. This is because the GEE procedure uses a "working" correlation matrix that enables the procedure to converge to the exact solution even when the correlation matrix is incorrect.

The GEE procedure was applied to develop APMs for a sample of 4-legged signalized intersections in Toronto, Canada for the years 1990-1995. Two GEE models were developed -- one which accommodated annual trend by estimating different models for every year and another which assumed identical coefficients for each year. These models were compared to equivalent models which do not account for the temporal correlation in annual accident counts. The results demonstrate that not accounting for temporal correlation does not affect the coefficient estimates but considerably underestimates the variances of these estimates. This means that explanatory variables may be incorrectly attributed as significant if the temporal correlation is not considered. The results also show that APMs which incorporate time trend usually perform better than APMs which do not.

The main disadvantage of using the GEE procedure to estimate APMs with trend lies in the requirement for each site to have observations for all variables in each year. Unfortunately, traffic flows and other characteristics are not always available for every year at each site. This reality often leads to the removal of these sites from the sample population -- an action that reduces the quality of fit of the APM. The missing values can be estimated but the efficiency of the GEE can be drastically reduced if these errors of omission are systematic, as is likely the case for missing traffic flows. It would therefore be of great value to conduct further research to refine the GEE procedure to accommodate missing values, especially for cases where these constitute a systematic bias.

REFERENCES

1. Diggle, P.J., K.-Y. Liang, and S.L. Zeger. *Analysis of Longitudinal Data*. Clarendon Press, Oxford, U.K., 1994.

2. Dunlop, D.D. Regression For Longitudinal Data: A Bridge from Least Squares Regressions. *The American Statistician*, Vol. 48, No. 4, 1994, pp. 299-303.

3. Maher, M., and I. Summersgill. A Comprehensive Methodology for the Fitting of Predictive Accident Models. *Accident Analysis & Prevention*, Vol. 28, No. 3, 1996, pp. 281-296.

4. McCullagh, P., and J.A.Nelder. *Generalized Linear Models: Second Edition*. Chapman and Hall, Ltd., London, U.K., 1989.

5. Mountain, L., M.J. Maher, and B. Fawaz. The Influence of Trend on estimates of Accidents at Junctions, *Accident Analysis & Prevention*, Vol. 30, No. 5, 1998, pp. 641-49.

6. Atkinson, A.C. *Plots, Transformations and Regressions: an Introduction to Graphical Methods of Diagnostic Regression Analysis*. Clarendon Press, Oxford, U.K., 1985.

7. Hauer, E. *The Effect of Resurfacing on the Safety of Two-Lane Rural Roads in New York State*, Report to the New York State Transportation Department, Department of Civil Engineering, University of Toronto, Toronto, Canada, 1993. (see also Hauer, E., D. Terry, and M.S. Griffith. The effect of Resurfacing on the Safety of Two-Lane Rural Roads in New York State. In *Transportation Research Record 1467*, TRB, National Research Council, Washington, D.C., 1995, pp. 30-37.)

8. Shankar, V.N., R.B Albin, J.C. Milton, and F.L. Mannering. Evaluating Median Cross-Over Likelihoods With Clustered Accident Counts: An Empirical Inquiry Using the Random Effects Negative Binomial Model. In *Transportation Research Record 1635*, TRB, National Research Council, Washington, D.C., 1998, pp. 44-48.

9. Guo, G. Negative Multinomial Regression Models for Clustered Event Counts. *Sociological Methodology*, Vol. 26, 1996, pp. 113-132.

10. Liang, K.-Y., and S.L. Zeger. Longitudinal Data Analysis Using Generalized Linear Models. *Biometrika*, Vol. 73, 1986, pp. 13-22.

11. Zeger, S.L., and K.-Y. Liang Longitudinal Data Analysis for Discrete and Continuous Outcomes. *Biometrics*, Vol. 42, 1986, pp. 121-130.

12. Myers, R. *Classical and Modern Regression with Applications*. Duxbury Press, Belmont, U.S., 1990.

13. Kulmala, R. *Safety at Rural Three- and Four-Arm Junctions: Development and Applications of Accident Prediction Models*. VTT Publications 233, Technical Research Centre of Finland, Finland, 1995.

14. Nicholson, A., and S. Turner. Estimating Accidents in a Road Network. In *Proceedings of Roads 96 Conference*, Part 5, New Zealand, 1996, pp. 53-66.

15. Green, P. Iterative Reweighted Least Squares for Maximum Likelihood Estimation and Some Robust and Resistant Alternative (with discussion). *Journal of the Royal Statistical Society*, *Series B*, Vol. 46, 1984, pp. 149-162.

16. Hauer, E., J.C.N Ng, and J. Lovell. Estimation of Safety at Signalized Intersections. In *Transportation Research Record 1185*, TRB, National Research Council, Washington, D.C., 1989, pp. 48-61.

17. Miaou, S.-P. *Measuring the Goodness-of-Fit of Accident Prediction Models*. Federal Highway Administration, Publication No. FHWA-RD-96-040, Virginia, U.S., 1996.

18. Box, G.P., and G.M. Jenkins. *Time Series Analysis: Forecasting and Control* (Revised Edition). Holden-Day, San Francisco, U.S., 1970.

19. Diggle, P.J. *Time Series: a Biostatistical Introduction*. Oxford University Press, Oxford, U.K., 1990.

20. Lord, D. *Safety Issues in Urban Transportation Networks*. Ph.D. Thesis, Department of Civil Engineering, University of Toronto, 2000.

21. Lord, D. Procedure to Estimate Missing Year-to-Year Traffic Counts at Intersections, In *proceedings of the 2000 CSCE Annual Meeting*, London, Canada, 2000.

22. Hauer, E., and J. Bamfo. Two Tools for Finding What Function Links the Dependent Variable to the Explanatory Variables. *Proceedings of the ICTCT 1997 Conference,* Lund, Sweden*., 1997.*

23. Lord, D., E. Hauer, and J. Bamfo. Application de deux nouvelles méthodes pour examiner la relation entre les accidents et les variables explicatives. *Routes et Transports*, Vol. 28, No. 3, Association québécoise du transport et des routes, pp. 11-20, 1999. (jn French).

24. Payne, R.W., P.W. Lane, *et al. Genstat 5: Release 3*. Clarendon Press, Oxford, U.K., 1993.

25. Hauer, E. *Observational Before-After Studies in Road Safety: Estimating the Effect of Highway and Traffic Engineering Measures on Road Safety*. Pergamon, Elsevier Science Ltd, Oxford, U.K., 1997.