MACRO-LEVEL ANNUAL SAFETY PERFORMANCE FUNCTION EVALUATION FOR CITIES, COUNTIES, AND STATE OF CALIFORNIA

Gurdiljot Gill, College of Engineering, California State Polytechnic University - Pomona, 3801 W.Temple Ave, Pomona, 91768, 909-908-7084, gurdiljotg@cpp.edu
Xiaofei Wang, College of Life and Environment, Huangshan University, Huangshan, 245041, Anhui, China, +86-18855913096, wxf@hsu.edu.cn
Wen Cheng, College of Engineering, California State Polytechnic University-Pomona, 3801 W.Temple Ave, Pomona, 91768, 909-869-2957, wcheng@cpp.edu
Jiao Zhou, College of Engineering, California State Polytechnic University - Pomona, 3801 W.Temple Ave, Pomona, 91768, 714-488-7965, jiaozhou@cpp.edu

ABSTRACT

To improve the traffic safety condition, assessment of safety performance of concerned areas is done typically based on crash rate, which is defined as crash count over traffic exposures. However, to gauge the reliability and associated benefits of this approach, it is necessary to test the significance of their impact on certain accident types. The major objective of this study is to use time-series model with auto regression error to examine the annual safety performance of crash rate and victim rate for different area levels in state of California and find the most accurate and reliable factors for crash rate, which will help facilitate the work of traffic safety agencies.

PROJECT DESCRIPTION/OBJECTIVE

Traffic safety has been a major concern in transportation engineering for many years and evaluation of safety performance of concerned sites or areas is an important task for safety management. Crash rate method has been traditionally employed for this purpose [1], which is the crash frequency divided by crash-exposure measures such as million vehicle miles for road sections or million entering vehicles for intersections [5]. This is a useful instrument used by varied governmental agencies due to the fact it is easily interpreted and understood when communicating with the public [2]. Yet, using crash rate for measuring safety performance is based on the assumption that crash exposure correctly represents the risk intensity of a crash. This assumption is always questioned and has not been proved for the safety analysis in most geographical area among US. A study by Hauer et al. [3] raised questions about the qualification for a good crash exposure. It was stated that numerator and denominator of the crash rate should measure the same subject, and relation between crash count and exposure should be positive linear slope. Kim et al. [6] conducted a study in Hawaii which indicated that the crash number may not be only related to population, it might also be influenced by other exposures such as: land uses, employment rate, economic activity, and motor vehicle number. Numerous studies have also been dedicated to exploring the relationship between crash count and crash exposure in different situations. However, most research focused on those conditions that were proposed in a cross-sectional setting, which may not be proper in a time-series setting for large area. Moreover, macrolevel areas such as county of Los Angeles and similar regions in state of California have never used time-series analysis to evaluate crash exposures and safety performance for the crash rate. The crash rate for macrolevel analysis (crash count over population, DVMT etc. factors for cities and counties region) have rarely been use to test the safety performance. Therefore, the major objective of this study is to use time-series model with auto regression error to examine the annual safety performance of crash rate and victim rate for certain areas in state of California. This study attempts empirically to fit a linear positive functional relationship between each of four crash and victim counts (fatality crash, injury crash, fatality victim, injury victim) and each of the four crash exposures (Population, DVMT, Driver number, and Motor vehicle registration number). The study will examine such a relationship over time by using annual time-series crash record from California public roadway report and OTS, and autoregressive error models to determine whether any of the 32 crash and victim rates could be used to represent an annual performance of traffic safety. The authors also expect to use this study to predict the different crash count number with time series model and generate the new crash ranking for different cities and counties, this will be helpful for future research to identify the hotspot for traffic safety since the hot spot identification is becoming essential task in engineering program improvement [4].

DATA COLLECTION

The data that used in this research is based on four major data sources: Statewide Integrated Traffic Records System (SWITRS); Highway Performance Monitoring System (HPMS); California Department of Finance, and California Office of Traffic Safety. The data from SWITRS focus on the collision and victim number of each city and county in state of California. The collision data include 7 different categories of collision: total fatal and injury, alcohol involved collision (fatal & injury), pedestrian involved collision, bicycle involved collision (fatal & injury), motorcycle involved collision, and property damage collision. For the victim data, 6 categories' data were provided: vehicle driver, vehicle passenger, bicyclist, motorcyclist, and total victims count; each category includes the number of people killed and injured. The data were collected in a range of 10 years from 2001 to 2010. This will help the research to incorporate more reliable data and accurate calculation; we will be able to see the change of collision numbers in different regions from every year. California Highway Performance Monitoring System (HPMS) provided the vehicle travel data (Daily Vehicle Miles Travel) for 57 counties and 485 cities of State of California in their public roadway data report. One of the major exposure that is needed for model set up is the population count of each county, city, and the entire state of California. We collected this data from California Department of Finance. The other related data were collected from the Federal Census website, which includes more different factors: area, household number, yearly maintains mileage, and lane miles.

METHODOLOGY

Time-series regression models seem to be a reasonable choice for this research because of the related data. The dependant variable for this study is a nonnegative integer, Poisson distribution or negative binomial model seems more appropriate. However, based on the previous research [7], since the data aggregation level in this analysis is high (e.g., aggregation for county and city over 1 year), and the results are very high average value of the data, a linear regression model is actually applicable. When the data are recorded over time, a typical linear regression is not likely to be appropriate, because the potential serial correlation often found in data [7]. Therefore a typical linear model with Auto Regression errors correcting for an AR type of serial correlation is employed.

Autoregression Error Model

A linear relationship between a crash or victim count and an exposure over time is fitted by means of a linear regression model. When an error term of the model is serially correlated, autoregression error can be included in the model to correct standard errors of the structural coefficients of the model. This type of regression model is AR error model; it will show as follows:

$$y_t = x_t' \beta + v_t \tag{1}$$

$$v_t = \epsilon_t - \emptyset_1 * v_{t-1} \ldots - \emptyset_m * v_{t-m} + \epsilon_t \tag{2}$$

Where: t = time index; y = dependent variable at time t (e.g., number of fatality/injury/victims in year t); x_t = independent variables; $v_t \ldots v_{t-m}$= m+1 serially correlated errors; ϵ_t= normal and independent error term; β = set of structural coefficients; $\emptyset_1 \ldots \emptyset_m$= m AR coefficients. The formula indicates that the model will have a time series feature, while the typical linear regression model has a constant error, the AR model will describe error term as a function. If v_t is presented as the error term for the current year, it will have a correlation of $\emptyset_1$ to the last year**.** These sets of time series AR model that specifically takes account of properties of count data will produce more accurate and reliable result to measure safety performance than normal crash rate.

Model Development

The Durbin-Watson test (AR-1) and the Breusch-Godfrey Lagrange multiplier test (AR-2) are used for this study to detect serial correlations in residual of an ordinary least-squares regression model. If a correlation is found, the order of AR error terms that should enter into the model needs to be determined. To suggest an appropriate AR order, a stepwise Autoregression method, which is analogous to a stepwise regression method for a typical cross sectional regression model was used in this study. By applying the model into project R, the test will present with the parameter that shows linear relationship between yt, the dependent variable (fatal collision number, fatality, injury collision number, and injury victims) and independent variables includes population, DVMT, Driver number, and Motor vehicle registration number.

RESULTS

A total of 64 models for multiple exposures and 480 models for single exposures functions were created and estimated. The coefficient estimates presented here were statistically significant at the 0.05 significance level. The AR model includes four counties (Los Angeles, Riverside, Santa Barbara, Sierra) and four cities (San Diego, Berkeley, Walnut, Riverside), each city and county includes seven types of models with three independent variables: Population, Licensed Driver, Motor Vehicle Registration (for city model, there are only two variables: Population in millions and DVMT in 10 millions), and four dependent variables: number of fatal collision, number of injury collision, number of fatalities and number of injured victim. To evaluate the effectiveness of the model, the study focus on the T-value and P-value for the model, with smaller P-value and larger T-value, the model will be more reliable to predict the safety performance. Also, since AIC value presents the goodness of the data, we only choose the models with smallest AIC values for evaluation. For brevity purpose, only Los Angeles county results are discussed. We chose three exposures: population, number of licensed driver and number of motor vehicle registration for generating the safety performance function with the 4 different crash type.

Fatal collision model

Licensed Driver, Motor Vehicle Registration number and Population amount were found statistically related to the Fatality. The final fatality model with Licensed Driver number and population count had an AR Order 2, the equation shows in the following as equation (3):

$$Fatalityt = 1.4 + 1.41*(Population)^2 - 1.43*(License\ Driver)^2 - 3.59*(Motor\ Vehicle)^2 \quad (3)$$

Population Factor: T-Value = 2.264; P-Value = 0.0642; Licensed Driver: T-Value = -2.365; P-Value = 0.0559; Motor Vehicle: T-Value =-2.290; P-Value = 0.0619; AIC:-363.928

Where T-value and P value measures the fit of the transformed model and the fit of next value predict by the model. The signs for Licensed Drivers and Motor Vehicles were negative, which is intuitive according to the definition of exposure: higher exposure results in more occurrences of crashes and victims. The negative sign implies that the number of fatalities collision in county of Los Angeles decrease as the number for Licensed Driver and Motor Vehicle increases. Thus, Licensed Driver and Motor Vehicle did not satisfy the linear condition (i.e., a straight line with a positive slope). Residual of the models was normally distributed, and their value were constant over time. If we run the model separately with only population factor, we will have the result shown in equation (4) below:

$$Fatalityt = 1.4 + 6.92*(Population\ in\ millions)^2 + v_t \quad (4)$$

Therefore, we conclude that only one of the three exposures (i.e., population) met the linearity condition and appropriate as a fatality exposure. This situation implies that a fatality rate could use the number of population as an annual safety performance measure for county of LA for the purpose of comparisons in traffic facilities over 10 year long term (2001 ~10). However, statistical uncertainty should be taken into account when the fatal crash rate per population is used as a safety performance measure for LA County.

Injury Crash Model

As shown in the following R code function population, licensed driver, and motor vehicle registration number amount were found statistically related to the injury collision count, the relation shows as linear function. The final injury model with Licensed Driver number and population count had an AR Order 1 shown in equation (5):

$$Injury = 3 + 1.21*(Population)+0.011*(License\ Driver) - 2.21*(Motor\ Vehicle) \quad (5)$$

Population Factor: T-Value = 1.069; P-Value = 0.326; Licensed Driver: T-Value = 1.106; P-Value = 0.311; Motor Vehicle: T-Value =-1.849; P-Value = 0.114; AIC:-59.32

The negative sign for Motor Vehicle implies that the number of injury collision in county of Los Angeles decrease as the number for Motor Vehicle increases. Thus, Motor Vehicle did not satisfy the linear condition (i.e., a straight line with a positive slope). Furthermore, the T-Value and P-Value shown in the function did not meet with the condition; T-Value is much smaller than absolute value of 2.5 and P-Value is much larger than absolute value of 0.05. Therefore, none of the three exposures met the linearity condition and appropriate as a fatality exposure.

Fatality Victim Model

Licensed Driver and Population amount were found statistically related linearly to the fatality victim, the relation shows as significant linear function. The final fatality model with Licensed Driver number and population count had an AR Order 2 in equation (6):

$$Victimf = -20+0.803*(Population) +0.046*(License\ Driver) –2.16*(Motor\ Vehicle) \qquad (6)$$

Population Factor: T-Value = 2.795; P-Value = 0.0313; Licensed Driver: T-Value = 26.333; P-Value = 0.0001; Motor Vehicle: T-Value =-7.493; P-Value = 0.0003; AIC: -59.24 . Two of the three exposures (i.e., population and Licensed Driver) met the linearity condition and appropriate as a fatality exposure. This situation implies that a fatality victim rate could use the number of population and number of licensed driver as annual safety performance measures for county of LA.

Injury Victim Model

Licensed Driver and Population amount were found statistically related linearly to the injury victims, the relation shows as significant linear function. The final injury victim model with Licensed Driver number and population count had an AR Order 2 in equation 7:

$$Victiminjury = 117+20.42*(Population)+0.756*(License\ Driver) –18.19*(Motor\ Vehicle) \qquad (7)$$

Population Factor: T-Value = 4.665; P-Value = 0.0034; Licensed Driver: T-Value = 5.911; P-Value = 0.001; Motor Vehicle: T-Value =-7.493; P-Value = 0.00032; AIC: -37.12 . Motor Vehicle did not satisfy the linear condition (i.e., a straight line with a positive slope). Two of the three exposures (i.e., population and Licensed Driver) met the linearity condition and appropriate as a fatality exposure. The table below exhibits the overall results for all four counties.

Table 1. Model Effectiveness for Counties

County	**Numerator**	**Denominator**		
		Population	**License Driver**	**Motor Vehicle**
Los Angeles	Fatal Crash	Positive	Negative	Negative
	Injury Crash	Not Significant	Not Significant	Not Significant
	Fatal Victim	Positive	Positive	Negative
	Injury Victim	Positive	Positive	Negative
Riverside	Fatal Crash	Negative	Negative	Positive
	Injury Crash	Negative	Negative	Positive
	Fatal Victim	Negative	Positive	Not Significant
	Injury Victim	Positive	Negative	Positive
Santa Barbara	Fatal Crash	Negative	Negative	Not Significant
	Injury Crash	Not Significant	Not Significant	Not Significant
	Fatal Victim	Positive	Negative	Negative
	Injury Victim	Positive	Positive	Not Significant
Sierra	Fatal Crash	Not Significant	Not Significant	Not Significant
	Injury Crash	Positive	Negative	Not Significant

	Fatal Victim	Not Significant	Not Significant	Not Significant
	Injury Victim	Positive	Negative	Not Significant

CONCLUSIONS AND RECOMMENDATIONS

The AR model indicates that for the most selected counties, population and licensed driver appear to be appropriate as the annual safety performance measure for the long-term comparison (2001 to 2010). While the motor vehicle registration was not appropriate for the three counties regardless of injury severity or subject, it interestingly appears to be appropriate for the crash and victim rate in the County of Santa Barbara. This suggests that any collision or victim rate based on motor vehicle registration should not be used to evaluate annual safety performance for the other three counties for long term except Santa Barbara County. Analyses of the data for the past decade revealed that the total crash rate per DVMT seems the most appropriate for measuring California safety performance for the past decade (1990s). The crash rate per population appeared to be the second most appropriate measure, and all other rates were not appropriate for a short term safety performance measure. In summary, for California county level annual safety performance, the crash and victim rate per driver appears appropriate for long-term comparison (2001 to 2010); while the crash rate per population appears appropriate for long-term comparison in city level annual safety performance analysis. However, statistical uncertainty should be taken into account when these rates are used to measure annual safety performance. When this uncertainty is taken into account, certain county and city (i.e., Riverside, Berkeley) seems to have experienced considerable increases in the fatal crash and victim rate per driver in 2004 and 2005. Generally, the predictions of the safety performance function generated by AR model satisfy the trend for the fatal and injury crash rate per exposures that shows in the graphic for the chosen cities and counties. However, due to the statistical uncertainty, the condition and function may vary for different cities; future investigation is suggested for other city and counties with different size type. Municipal, county, and city in state of California should examine the appropriateness of the safety performance measures that they have been using for their governing areas. It is hoped that this study can help researchers, law authority, and highway landscape policy maker, etc. to take better measure toward preventing accidents.

ACKNOWLEDGEMENTS

The research work was partially supported by the grant of Key Project of Science of Education Department in Anhui (gxfxZD2016234).

REFERENCES

[1]. Hauer, E. (1995). On Exposure and Accident Rate. *Transportation Research Board, 36,* 134-138.
[2]. Hauer, E. (2004). On the Estimation of the Expected Number of Accidents. *Accident Analysis and Prevention, 18,* 1–12.
[3]. Hauer, E., Harwood, W. D., Council, M. F. & Griffin, S. M. *(*2001). Estimating Safety by the Empirical Bayes Method: A Tutorial. *Transportation Research Board*, *1784*, 126-130. doi: *http://www.ctre.iastate.edu/educweb/docs/Bayes_tutor_hauer*
[4]. Huang, H. C. (2009). Empirical Evaluations of Alternative Approaches in identifying Crash Hotspots: Naive Ranking, Empirical Bayes and Full Bayes. *National Research Council, 9,* 32- 41
[5]. Lovegrove, R. G. & Sayed T. (2007). Macrolevel collision prediction models to enhance traditional reactive road safety improvement programs. *Transportation Research Board Business*

Office, 33 (5), 65-73.
[6]. Kim, K., Brunner, M. & Yamashita, Y. E. (2006). The Influence of Land Use, Population, Employment, and Economic Activity on Accidents. *Transportation Research Board Business Office, 1953,* 56-64
[7]. Kweon, Y. (2006). Traffic Fatality and Injury Forecasts and Reduction Goals in Virginia. *Virginia Transportation Research Council, 6,* 20-23