CRASH PREDICTION BASED ON TRAFFIC ANALYSIS ZONE-RELATED INFORMATION

Mohammad Elhocheimi, College of Engineering, California State Polytechnic University-Pomona, 3801 W.Temple Ave, Pomona, 91768, 714-334-0005, melhocheimi@cpp.edu
Wen Cheng, College of Engineering, California State Polytechnic University - Pomona, 3801 W.Temple Ave, Pomona, 91768, 909-869-2957, wcheng@cpp.edu

ABSTRACT

The main focus of this study is to introduce Crash Estimation to the regional planning stage. This study was conducted based on various Traffic Analysis Zones (TAZ) within three different counties in Southern California Area which include Los Angeles, Orange, and San Bernardino. The data used for this study was obtained from three main sources: Southern California Association of Governments (SCAG), California Statewide Integrated Traffic Records System (SWITRS), and Census Bureau. Negative Binomial Regression modeling and Multicollinearity were utilized to find the theoretical crash estimation relations between crash counts and other TAZ-related attributes such as socioeconomic, demographic, land use, transportation network, and exposure.

Project Description/Objectives

Although Transportation Equity Act for the 21st century (TEA-21) required agencies to develop studies on crash reduction, yet minimal studies had actually focused on the regional planning stage for crash estimation [5]. Crash forecasting has grown to play a major role in the transportation planning field. Allocating funds to the right TAZ and efficiently locating hot spots is always a challenge for OTS that is constantly looking for easier and more efficient ways. An approach to subdivide the study area into equally 0.1 square miles was used by Kim, where he found that using an equally subdivided TAZ had an advantage over other dividing methods [2]. Another significant model that utilized TAZ was done by Nedran and Shashi who were able to predict crashes based on trip generation [3] .The significance of crash forecasting in the planning stage has increased recently, but still has a lot of research areas -in such field. Thus, our duty as civil engineers is to find the best, most efficient, and most affordable way to reach zero casualties on our streets. The Negative Binomial Regression will be utilized in our paper as well as the Multicollinearity function. Excel, GIS and R program are used in this research paper as well. The R program was used for modeling (refer to Binomial Regression Model Section). The area of study was determined to be Orange County, Los Angeles, and San Bernardino in California. These counties were chosen because together they tend to have the leading number of car crashes in Southern California [4]. For this paper the three types of crashes were considered: Fatal, Injury, and Total Crash. Los Angeles is more condensed than the other two counties, a separate analysis was done on Los Angeles County to determine if the same independent variables were significant for both cases. One of the main objectives of this paper is to provide the user and law makers with a tool, in this case, a theoretical equation that has a solid proof of how socioeconomic factors, traffic exposure, land use, and employment affects crash rate. This is very significant since the general public and lawmakers would realize how certain factors that might be considered minor can actually play a role in forecasting crashes. The most important objective of this paper is to provide this study and the theoretical equations that could be used to assist in reducing car crashes every day and potentially saving lives.

Study Design

In order to successfully find the theoretical equations, the research study was conducted through the following five stages:

- **Stage one**

This was the initial stage when the data was divided into main groups: Dependent and Independent. The independent variable data was requested from SCAG and census, and the dependent variable data was requested from SWITRS.

- **Stage two**

During this stage the raw data was sorted in Excel, a check for data error was performed, and then the independent variables were ranked according to their importance or relevance to the scope of the research paper.

- **Stage three**

Data was loaded to R and the initial negative binomial model was determined. However, this model could not be counted because it was too specific and many independent variables showed insignificant while studies showed they are incredibly significant.

- **Stage four**

Collinearity analysis was performed and the highest VIF independent value was omitted. Then another negative binomial regression was performed; this iterative process continued until VIF for each variable was less than five.

- **Final Stage**

Final negative binomial regression model and the equation was ready to be derived. In this process two separate approaches were used:1) any independent variable with a confidence of 95% or higher was used for LA equations, 2) only independent variables with confidence of 99% and 99.99 % were used for the combined counties. Once the confidence level was determined the theoretical equations could be written.

Data Collection

The data was obtained from three different sources: Southern California Association of Governments (SCAG). California Statewide Integrated Traffic Records (SWITRS), and United States Census Bureau. For Independent variables SCAG and Census websites were utilized, while the Dependent variables were requested from the SWITRS website and were classified into 3 different types: Total Crash, Fatal, and Injury. The following tables summarize the Independent and Dependent variables and classify the variables into each group such as socio economic factors, employment, diversity and mixed use of land, access to transit and bus, land development, biking and walking characteristics, and traffic exposure. The following tables summarizes the acronyms used in R as well.

Table 1.0 Description of Independent and Dependent Variables

Variable	Description	Minimum	Maximum	Mean	Std. dev*
Socioeconomic Characteristics					
POP_DEN	Population density	0.00	282.23	42.75	37.99
EMPLOY_DEN	Employment density	0.00	6109.01	21.40	114.48
HSH_DEN	Household density	0.00	156.61	14.13	14.37
Employment Characteristics					
RET_DEN	Retail job density	0.00	105.37	2.20	4.52
RETSERV_DEN	Retail and services job density	0.00	237.24	7.01	14.21

Diversity and Mixed Use of Land					
JOBMIX	Highest jobs	0.00	10.10	2.08	1.28
EMPLOY_HSD	Employment/household	0.00	88608	120.11	1688.04
EMPLOY_POP	Employment/population	0.00	88608	80.81	1604.32
Access to Transit Characteristics					
INT_DEN	Intersection density	0.00	4.48	0.56	0.42
EXBUS_DEN	Stop density for Bus and BRT	0.00	16.02	0.05	0.38
TOTBUS_DEN	Total bus stop density	0.00	41.61	0.57	1.27
Land Development Characteristics					
HSDMLTP_PCT	Percentage of household living in multiple units	0.00	4.62	0.58	0.54
HQTA_PCT	%of TAZ area in high quality transit area	0.00	1.05	0.43	0.44
Biking and Walking Data					
BKLNIND_DEN	Bike lane density indicator	0.00	116.74	12.15	10.84
BLK_LENG	Estimated block length	0.00	44.42	1.17	0.91
WLK_ACC	Walk accessibility	0.00	376.84	12.42	5.12
MTR_IMP	All roads officially as motorways	0.00	80.04	0.50	0.20
MJRD_LI	Major roads less important than motorways	0.00	23.40	0.34	1.11
MJRD_OTR	Other major roads	0.00	69.28	0.08	1.47
SECRD_IMP	Secondary Roads	0.00	42.56	0.44	1.91
RDCONN_IMP	Local Connecting Roads	0.00	42.47	2.04	2.35
LCLRD_HIMP	Important local roads	0.00	70.48	1.26	2.04
LCLRD	Local Roads	0.00	318.08	8.41	16.07
LCRD_MIMP	Minor importance Local Roads	0.00	1014.37	7.96	36.07
TOT_RDSEG	Total miles of all road segments	0.00	1170.65	21.06	50.04
BKLN_1	Class I bikeway	0.00	12.98	0.21	0.73
BKLN_2	Class II bikeway	0.00	17.60	0.57	1.27
BKLN_3	Class II bikeway(Bike Route)	0.00	22.29	0.24	0.88
TOTMI_BKLN	Total miles of bikeways	0.00	22.29	1.02	1.94
Traffic Exposure					
VMTLM1	Total estimated VMT of light-medium duty vehicles	0.00	1927650.00	102384	135728.7
VMTLM2	AM peak estimated VMT for light-medium duty vehicles	0.00	419452.00	21554	28666.81
VMTLM3	PM peak estimated for light-medium duty vehicles	0.00	651190.00	34745	43780.07
VMTLH4	Total estimated VMT of light-heavy duty vehicles	0.00	50270.00	1233	2680.00
VMTLH5	AM peak estimated VMT for light- heavy duty vehicles	0.00	9439.00	232.50	492.92
VMTLH6	PM peak estimated VMT for light-heavy duty vehicles	0.00	10260.00	253.70	534.99

VMTMH7	Total estimated VMT of medium-heavy duty vehicles	0.00	35979.00	920.80	1930.15
VMTMH8	AM peak estimated for medium - heavy duty vehicles	0.00	6464.00	166.30	339.95
VMTMH9	PM peak estimated VMT for medium -heavy duty vehicles	0.00	5562.00	145.20	294.75
VMTH10	Total estimated VMT of heavy duty vehicles	0.00	271076.00	4444.0	14364.46
VMTH11	AM peak estimated VMT for heavy duty vehicles	0.00	37618.00	625.40	1988.71
VMTH12	PM peak estimated VMT for heavy duty vehicles	0.00	45321.00	770.30	2418.32
Dependent Variables					
TOT_CRASH	Number of Total crashes	0.00	209	20.35	17.20
FATAL	Number of Fatal type crashes	0.00	7	0.27	0.59
INJURY	Number of Injury type crashes	0.00	208	20.08	17.03

* Standard Deviation

Negative Binomial Regression Model

In this paper, the Negative Binomial regression model, also known as Log-Linear regression model, is used to forecast the number of crashes. The Negative Binomial model is widely used in the traffic accident modeling since these studies usually have an overdispersed data set. Additionally, the Negative Binomial regression provides a statistical significance of each independent variable. For this paper, six different negative binomial models were performed which include the following three types: Fatal, Injury and Total crash. Three models were performed on Los Angeles county data alone, and another three models were performed for Los Angeles county, Orange county and San Bernardino county data combined. However, if we were only to do one iteration, we will end up with what we started with, 42 independent variables. It is important to find which variables are correlated with each other. Guevara explain that certain variables share common characteristics and suggests caution when explaining the results [1].The correlation procedure is done in R using Collinearity analysis. The generalized linear model outcome could be rewritten as:

$$y_i = e^{f_x} \quad (1)$$

Where $f_{(x)}$ = the linear regression model which includes various independent variables
i=ith observation
y_i= crash number for various types of crashes

Collinearity Analysis

Variance Inflation factor (VIF) is one of the ways in measuring collinearity; it provides an estimation of how much the variance will increase if the entities are correlated. According to Vo, VIF would be equal to one if there is no correlation among them [6]. Using R, a collinearity analysis is performed and a VIF value is found for each variable. For the purpose of our study, the VIF of less than five was considered acceptable. This iteration process will continue until all VIF values are less than five for each independent variable, Fig 1.0 will further demonstrate the process.

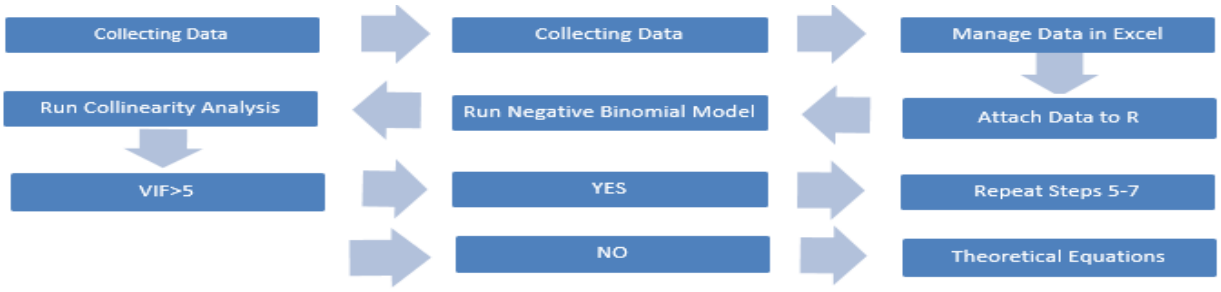

Fig1.0 Model Flow Chart

Research Results

The following theoretical equations summarizes the most significant variables that contributed to the total crashes, Fatal Type crashes, and Injury Type crashes in LA and in the three counties combined

Total Traffic Crashes Theoretical Equation (2)

$$f(x) = 1.988 + 0.001479(POP\_DEN) + 0.1335(RETSERV\_DEN) + 0.2326(HSDMLTP_{PCT}) + 0.3763(HQTA_{PCT}) + 0.008343(BKLNIND_{DEN}) - 0.05384(BLK_{LENG}) + 0.04978(SECRD_{IMP}) + 0.06185(RDCONN_{IMP}) + 0.03452(LCLRD_{HIMP}) - 0.003252(LCRD_{MIMP}) + 0.03558(BKLN_3) + 0.00001367(VMTLM2)$$

Total Traffic Crashes-(*County of Los Angeles*-) (3)

$$f(x) = 1.956 - 0.053710(JOBMIX) - 0.00001924(EMPLOY_{HSD}) + 0.2049(HSDMLTP_{PCT}) + 0.3686(HQTA_{PCT}) + 0.007778(BKLNIND_{DEN}) - 0.05691(BLK_{LENG}) + 0.09348(MTR_{IMP}) + 0.07908(MJRD_{LI}) + 0.1039(SECRD_{IMP}) + 0.08089(RDCONN_{IMP}) + 0.03424(LCLRD_{HIMP}) - 0.006796(LCRD_{MIMP}) + 0.1070(BKLN_2) + 0.1060(BKLN_3) - 0.07276(TOTMI_{BKLN}) + 0.000009551(VMTLM3) - 0.0001554(VMTH12)$$

Fatal Traffic Crashes Theoretical Equation (4)

$$f(x) = -1.977 + 0.9913(JOBMIX) + 0.1326(TOTBUS_{DEN}) + 0.0167(BKLNIND_{DEN}) + 0.06783(SECRD_{IMP}) + 0.008069(LCRD) - 0.004081(LCRD_{MIMP}) + 0.000008954(VMTLM2)$$

Fatal Traffic Crashes-(*County of Los Angeles*-) (5)

$$f(x) = -2.114 - 0.3829(INT_{DEN}) - 0.7433(EXBUS_{DEN}) + 0.1426(TOTBUS_{DEN}) + 0.02698(BKLNIND_{DEN}) + 0.07705(MTR_{IMP}) + 0.08503(MJRD_{LI}) + 0.1039(SECRD_{IMP}) - 0.01103(LCRD) + 0.0004487(VMTLM3)$$

Injured Traffic Crashes Theoretical Equation (6)

$$f(x) = 1.983 + 0.001536(POP_{DEN}) + 0.2302(HSDMLTP_{PCT}) + 0.3803(HQTA_{PCT}) + 0.008511(BKLNIND_{DEN}) - 0.05561(BLK_{LENG}) + 0.04802(SECRD_{IMP}) + 0.06258(RDCONN_{IMP}) + 0.03457(LCLRD_{HIMP}) - 0.003212(LCRD_{MIMP}) + 0.03767(BKLN_3) + 0.00001376(VMTLM2)$$

Injured Traffic Crashes-(*County of Los Angeles*-) (7)

$$f(x) = 1.901 - 0.001300(POP_{DEN}) - 0.00001924813(EMPLOY_{HSD}) + 0.2184(HSDMLTP_{PCT}) + 0.3806(HQTA_{PCT}) + 0.006422(BKLNIND_{DEN}) - 0.06003(BLK_{LENG}) + 0.09710(MTR_{IMP}) + 0.08169(MJRD_{LI}) + 0.1037(SECRD_{IMP}) + 0.08125(RDCONN_{IMP}) + 0.03143(LCLRD_{HIMP}) - 0.007646(LCRD_{MIMP}) - 0.07605(BKLN\_1) + 0.000009764(VMTLM3) - 0.0001647(VMTH12)$$

For brevity purpose, we can take Equation 2 as an example to further understand the relationship between various contributing factors and crash number. In Equation 2 it is known that *POP_DEN,RETSERV_DEN,HSDMLTP_PCT,HQTA_PCT,BKLNIND_DEN,SECRD_IMP,RDCONN_IMP,LCLRD_HIMP,BKLN_3,* and *VMTLM2* have a positive relationship with the total traffic crashes in three counties data set, while *BKL_LENG*, and *LCRD_MIMP* have a negative relationship. Specifically, for one unit increase in *POP_DEN* the total traffic crashes are expected to increase by exp (0.001479) =1.0015 unit increase in total crashes.

Conclusion and Recommendations

For the purpose of this study only independent variables with a confidence of 99.99 % and 99% were used for the general theoretical models. In addition to the 99.99 and 99% confidence, 95% confidence was used for County of Los Angeles theoretical equations. From our results we found that the LA theoretical equations and the combined theoretical equations share a lot of similarities. We found that both the general theoretical equation and LA equation for total crash model have the following independent variables in common: Percentage of TAZ area in high quality transit area, percentage of household living in multiple units, bike lane density indicator, estimated block length, secondary roads, local connecting roads, important local roads, minor importance local roads for fatal type crashes we found that LA equation. Then we say that the general equations share the following independent variables: total bus stop density, bike lane density indicator, secondary roads, and local roads. It was also noticed that for fatal type crashes buses and transit played a major role in both equations. Also, bike lanes density factor was another important variable. For injury type crashes we found that the LA equation and the general equation share the following independent variables: population density, percentage of household living in multiple units, percentage of TAZ in high quality transit area, bike lane density indicator, estimated block length, secondary roads, local connecting roads, important local roads, minor importance local roads. An important result we found is that bike lane density indicator, local connecting roads, minor and major roads, population density and percentage of TAZ in transit area, and estimated block length are proven to have a significant role in the three types of crashes. Now we have a better understanding of the relation between crash types and other factors, we can recommend allocation of funds to be directed towards TAZ that have problems in the above independent variables.

REFERENCES

[1] Guevara, F. L. De, Washington, S. P., & Oh, J. Forecasting Crashes at the Planning Level: Simultaneous Negative Binomial Crash Model Applied in Tucson, Arizona. *Transportation Research Record: Journal of the Transportation Research Board*, 191-199. 2004.

[2] Kim, K., Brunner, I., & Yamashita, E. Influence of Land Use, Population, Employment, and Economic Activity on Accidents. *Transportation Research Record*, *1953*(1), 56–64:10.3141/2006.

[3] Naderan, A., & Shahi, J. (2010). Crash Generation Models. *Transportation Research Record: Journal of the Transportation Research Board*,101–106.2010

[4] Statewide Integrated Traffic Records (SWITRS) http://iswitrs.chp.ca.gov/Reports/jsp/userLogin.jsp

[5] TEA-21 - A Summary - An Overview. https://www.fhwa.dot.gov/tea21/sumover.htm

[6] Vo, Tom.M. *Analyzing and Forecasting Traffic Crashes In Southern California Region (Master Thesis)*. California Polytechnic University Pomona-Department of Civil Engineering 2016