RESEARCH ARTICLE

CRASH PREDICTION MODELLING OF STATE HIGHWAYS IN ARKANSAS.

Ghulam H. Bham, Ph.D1, Uday R. R. Manepalli2 and Bhanu S. Javvadi3.

1. University of Alaska Fairbanks, AK 99775.
2. Agile Assets Inc. Austin, TX 78749.
3. Verizon, Washington, DC 20006.

Abstract

This paper presents the development of crash prediction models for State highways in Arkansas. Based on over-dispersion, Negative Binomial models were developed, with additional features for dealing with more complex and correlated count data. Six models are presented based on the location and type of the highway, i.e., urban and rural; divided and undivided highways, and junction (intersection and intersection related), and non-junction. Statistical analysis of crash data identified variables associated with crashes. Results indicate that in general Annual Average Daily Traffic (AADT) and number of lanes increased crash frequency.

Introduction:

Traffic crashes are a major burden on society. This burden includes the pain and suffering of the individuals involved and their loved ones, as well as the economic loss to society. According to the National Safety Council, motor vehicle crashes have been the leading cause of unintentional deaths in the United States (NHTSA 2008). With motor vehicles being the main transportation mode in the United States, researchers must seek ways to improve safety on roads. Identifying the root cause of the problem is the first step toward finding a solution. Various factors are responsible for highway crashes. Those recorded by agencies include driver behavior, road conditions, weather, and elements related to highway geometry. The objective of this paper is to identify the crash contributing factors including roadway geometric features based on highway class and type of facility.

Several research efforts have been conducted to assess the relationship between crashes and the factors responsible for them. Past research has been focused on modeling the relationships between total frequency of crashes and the various crash contributing factors responsible. Abdel-Aty et al. (2006a) developed a model that explained the relationships between crash frequency, signalized intersection geometry, and traffic flow characteristics. They fitted general estimating equations (GEE) with a negative binomial function for four correlation structures (independent, exchangeable, autoregressive, and unstructured) using three-year crash data for 208 four-legged signalized intersections. It was observed that intersections with an asphalt surface had fewer crashes than intersections with a concrete surface, and the variable AADT had a high degree of correlation with most other variables, especially the total number of lanes. Another study by Abdel-Aty et al. (2005) explored the hypothesis that different types of collisions are affected by different independent variables. Hierarchical tree-based regression was used to predict the number of crashes reported on both long and short forms for each type of crash. This model split the data into branches on a tree diagram, displaying the average value at each node. Rather than ignoring the entire observation, this model handled missing information by treating a missing independent value as

Corresponding Author:- Ghulam H. Bham.
Address:- Univ. of Alaska Fairbanks, Alaska.
a valid response. This study was based on the rationale that vehicle crashes are common occurrences at signalized intersections. The authors investigated the significant differences among important crash related factors by comparing models based on restricted data sets with those based on complete data sets. For both type of data sets, tree-based regressions were conducted for each type of collision and for the total number of crashes, producing 16 regression models. The results of this research showed that accurate crash frequency prediction demands separate models for each collision type instead of a single model that aggregates crash types.

Abdel-Aty et al. (2006b) developed another model to analyze the spatial effect among signalized intersections along corridors, and identified the variables that significantly influence crash frequency at the intersections. Using GIS, 476 signalized intersections along 41 corridors in three Florida counties were analyzed. They modeled crash frequency for the spatially clustered signalized intersections along a corridor. They identified ten variables as significant and observed that crash frequency increased with an increase in the number of lanes.

Banihashemi and Dimaiuta (2005) developed a crash prediction model based on the interactive highway safety design model (IHSDM). For highways segments, they used an alternative model as a function of AADT. Hadayeghi (2007) developed a series of zonal level collision prediction models most commonly used for urban transportation planning. The data used included collision frequency, road network characteristics, and traffic volume for 481 zones in Toronto. The model was developed to explore the relationship between the number of collisions and traffic intensity on the road network, land use, socio-economic and demographic characteristics. Generalized linear modeling (GLM) procedure was utilized and estimated the model coefficients with the assumption of a negative binomial error distribution.

Abdel-Aty et al. (2006c) estimated a system of models by dividing freeway crash data into multiple mutually exclusive categories, including multiple and single-vehicle crashes, peak and off-peak period crashes, dry and wet pavement crashes, daytime and nighttime crashes, property damage only crashes and injury crashes. The frequency of each crash type was modeled using seemingly unrelated negative binomial regression models. This study found that an increase in AADT causes more vehicle crashes because the increase in traffic volume increases the probability of interaction among vehicles. AADT have no impact on single-vehicle crashes, because such crashes occur due to vehicle malfunction or driver error. The model also evaluated significant factors in both daytime and nighttime crashes. Significant factors in both cases were road curvature, on- and off-ramps, median type, and pavement surface. The frequency of daytime crashes was also affected by the coefficient of variation in speed.

A study conducted by Jonsson et al. (2007) developed separate crash prediction models for four major types of crashes: opposite direction, same direction, intersecting direction, and single-vehicle crashes. The research focused on the difference in crash type by examining traffic flow, other variables and crash severity. Data included 999 major intersections and were acquired from HSIS in California. This model assumed that the number of crashes followed a negative binomial distribution. Also, a logarithmic link function was used, with a dispersion parameter estimated using maximum likelihood. The results show that the model for single-vehicle crashes was nearly linear when plotted against traffic flow. The opposite and same direction crash models were very similar with respect to traffic flow and other variables, but very different in terms of crash severity. Intersecting and opposite direction crash models shared some similarities with respect to severity, but the models were very different in terms of traffic flow; therefore, they cannot be combined to form a single model for any two types of crashes.

Deng et al. (2006) analyzed the statistical relationship between head-on crash severity and various potential causal factors such as geometric characteristics of road segments, weather, road surface conditions, and time of occurrence. Data were obtained from the Connecticut DOT. The data included detailed information on crashes on all state maintained highways. Ordered probit modeling was chosen because of the ordinal nature of crash severity. The analysis showed that the frequency of crashes was higher on wet roads and at nighttime. The results also contradicted some expectations. For example, the researchers had assumed that wider lanes would have more severe collisions due to higher speed. However, wide lanes proved to provide a buffer area to drivers to avoid a crash, and the denser access points were prone to more fatal crashes than any other road segments.

These studies suggest that researchers used a range of factors responsible for vehicle crashes. Many have found that geometric conditions on roads have a major effect on crash frequency and severity. AADT is also a significant factor in the analysis, along with various other factors. To avoid over-dispersion of data, most studies assume that the data follows a negative binomial distribution.
Methodology:-
Motor vehicle crashes have been the subject of research for many years. Researchers have developed models in which crash frequency or crash severity are dependent variables, and significant factors that contribute to crashes are independent variables. These models relate dependent variables with independent variables. Statistical analysis of count data can be analyzed by developing explanatory models, prediction models, or both. Explanatory models only attempt to evaluate the significance of covariates on response variable, which, in this case, is crash frequency or crash severity. Predictive models developed, use covariates from the data available. As many variables relate to different set of information about a crash, an exploratory analysis was conducted to find the significant variables that contribute to crashes. The present analysis utilized least-square regression, which assumes that the error term in the model was normally distributed. Next Poisson regression was used for estimating models, based on the assumption that the data were not over-dispersed. Over-dispersion occurs when the data display more variability than predicted by the ratio of variance to mean. Over-dispersed data generally causes all model selection criteria to perform poorly. If over-dispersion is ignored, a model with too many parameters is likely to be selected leading to over interpretation of these parameters. Negative binomial regression was used as it permits dispersion in the data and thus accounts for the violation of this assumption. Model selection criteria included the deviance, scaled deviance, and log-likelihood statistic.

Model Categories:-
Research conducted presented in the preceding section examined specific road segments or intersections to model crash frequency. This paper utilized the data for the State highways in Arkansas. Crash prediction models were developed for both junctions (intersections) and highway segments. This approach takes into account the significant differences in geometric features, traffic control operations, and traffic flow conditions between segments and intersections. Additionally, rural and urban highways, and divided and undivided highways have been modeled. Table 1 presents the different categories of crash prediction models developed.

| Road system | State highways |
|-------------|----------------|
| Vicinity    | Rural, urban   |
| Relation to | Non-junction, intersection, interchange-related |
| Traffic flow| Divided, undivided |

Variables, road system, vicinity, relation to, and traffic flow are listed in the above table. Separate models were developed by classifying the available data according to these variables. Further explanation for categorizing models is provided in the following. In total, eight models were attempted.

1. Rural/Urbam: Crashes occurring on rural roadways were separated from those on urban roadways because roadway characteristics, crash characteristics, causes of crashes are different. Rural and urban roadways differ in geometric features, traffic flow characteristics, speed limits, traffic control, and operations, so it is important to analyze them separately.

2. Divided/Undivided: Crashes occurring on divided highways were separated from those on undivided highways. A divided highway is usually divided by a median. To differentiate any effect as a result of medians, divided highways were examined separately.

3. Junction/Non-Junction/Intersection Related: This is a major categorization as the three types differ from each other in many respects, so combining them would not provide valid results.

Statistical Models:-
Count data can be analyzed using the GENMOD statistical package in Statistical Analysis System (SAS). However, use of an appropriate statistical model is crucial for generating proper inferences. Several model forms have been used in the past to predict traffic crashes. One natural and statistically interpretable measure of occurrence is the crash rate, defined as the number of crashes occurring over a specific period of time divided by AADT and the length of each segment. The challenge in analyzing the crash rates arises when some road segments are more prone to crashes than others i.e., shorter road segments may have a higher number of crashes compared to longer segments with fewer crashes. To overcome this problem, segment length for predicting crash rate was considered.

The GENMOD procedure fits generalized linear models. Previous studies have relied on traditional linear models. Traditional models assume that data is normally distributed, which may not be the case for count data. These models
may not work if the measured mean of a population lies within a range. Traditional models also assume that variance is constant which may not be the case for all situations. Generalized linear models can overcome all of these problems.

Generalized linear models allow the mean of a population to depend on a linear predictor through a nonlinear link function, and they allow the response probability distribution to be any member of an exponential family of distributions (Oklahoma State 2009). Generalized linear models include models specific to the data. For example, logistic and probit models are used for binary data, and log-linear models are used for multinomial data. Generalized linear models also include classical linear models with normal errors. Many other useful statistical models can be formulated as generalized linear models by the selection of an appropriate link function and response probability distribution. Generalized linear models are expressed mathematically as:

\[ Y_i = X_i \beta + \epsilon_i \]  

where \( Y_i \) = response variable for the \( i \)th observation and \( X_i \) = column vector of covariates, also called explanatory variables, for the observation and \( \epsilon_i \) is the error term which can be the experimental error, unexplained portion of the model, or anything that can be considered as an error. This error should be as low as possible but models should and cannot be error free.

Equation (1) is a traditional linear model. The vector of unknown coefficients \( \beta \) in the equation is estimated by a least-squares fit to the data \( y \). These vectors are assumed to be independent, normal random variables with zero mean and constant variance. In generalized linear models, the response variable or the variable of interest is usually assumed to follow an exponential distribution. This response can be continuous or discrete. The probability function is expressed as:

\[ f(y) = \exp\{y\theta - b\theta \} \cdot \frac{c(y, \phi)}{a(\phi)} \]  

where \( a, b, c \) are values respective to the distribution, and \( \theta \) is the dispersion parameter.

The dispersion parameter quantifies dispersion or clustering of the response variable, in this case, number of crashes relative to a standard statistical model. The value of the dispersion parameter determines the exponential distribution to be used in the analysis. The exponential distributions can be Poisson, binomial, negative binomial, normal, Inverse Gaussian, gamma, or multinomial. Each distribution has its own probability distribution function. Expected value of \( Y_i \) is related with a linear predictor with an appropriate link function. The response variable has a probability distribution from an exponential family. The variance of the response depends on the mean \( \mu \) through a variance function \( \nu \), and defined as:

\[ \text{var}(Y_i) = \frac{\phi \nu(\mu_i)}{w_i} \]  

where \( \phi \) is the dispersion parameter and \( w_i \) is the weight of each observation. Table 2 presents the probability distributions belonging to the exponential family applied to the data.

**Negative Binomial Regression in Log-Linear Model:**

The dispersion parameter \( k \), used with the negative binomial distribution, is different for other distributions. It applies only to negative binomial distribution, and its value should be fixed or estimated. Detailed discussion on use of negative binomial distribution for count models can be found elsewhere (Lawless (1987), McCullagh and Nelder (1989, Chapter 11), Hilbe (1994) ). The GENMOD procedure estimates \( k \) by using maximum likelihood; alternatively, a constant value can be set. The link function for the three distributions relates the mean \( \mu_i \) of the response in the \( i \)th observation to a linear predictor through a monotonic differentiable link function \( g \). The general form of this link function is \( g(\mu_i) = X_i' \beta \) where, \( x_i \) is a fixed known vector of explanatory variables, and \( \beta \) is a vector of unknown parameters.
Crash Prediction Modeling:

This section estimates crash frequency using statistical techniques based on the crash data. Six models for state highways were developed. For these models, crash frequency was considered as a dependent variable. As discussed above log-linear models are preferable to traditional or lognormal models. Poisson, negative binomial, and zero-inflated Poisson regression models are commonly used models falling under the category of log-linear models. The assumption of a lognormal distribution for the error structure of the response variable, in this case, the crash frequency, is no longer valid for crashes taking place on road segments and intersections. The Poisson model or a negative binomial regression model are a natural choice, as they model the occurrence of rare discrete events well.

The objective in developing a model is to find a relationship between characteristic features of the roadway and crash frequency. Maiou (1994) tested the relationship between highway geometrics and crash frequency using negative binomial regression and suggested that the Poisson regression model should only be used for analyzing count data when the data is not over-dispersed. If the dependent variable, crash frequency, contains more zeroes, then a zero-inflated Poisson regression model or a negative binomial would be appropriate. Several researchers including Kweon and Kockelman (2004), Noland and Qudus (2004) have reached the same conclusion about negative binomial regression models, suggesting that negative binomial regression models are preferred over Poisson for complexities in the data.

Form of the Equation:

Every model was carefully developed. The occurrence of crashes and the characteristic features of these crashes were related using a statistical model. Hauer (1995) was one of the first to draw attention to non-linear modeling of crashes and traffic volume. He suggested the form $N = \alpha V^\beta$, where $N$ is the predicted number of crashes, $V$ is the traffic volume, with $\alpha$ and $\beta$ the parameters to be estimated. The length of a segment was important and considered with AADT to represent traffic volume in the models developed. Segment length (LENGTH) and AADT were considered as exposure variables, and natural logarithm of the AADT was used. On the log-scale, the ratio of crash frequency to AADT and length becomes the difference between log (crash frequency) and log (AADT) + log (AADT) + log (LENGTH). The coefficients of AADT and length were not equal to 1, and determined from the analysis. The model used in the analysis is expressed as:

$$\mu_h = \exp(\beta_0)(ADT)^{\beta_1}(LENGTH)^{\beta_2}\exp(\beta_3X_{h3})\exp(\beta_4X_{h4})\ldots\exp(\beta_qX_{hq})$$

where:

- $\mu_h$ = crash frequency
- $\beta_0, \beta_1, \beta_2, \beta_q$ = estimated parameters
- $X_{hq}$ = variables
- $h$ = represents the number of models
- $q$ = represents the number of variables in the model.

Variable Selection:

Road log information regarding the routes and sections was considered, and sections were further divided into segments based on begin and end log miles. The region between an end log mile and begin log mile was considered...
a segment. The geometric configuration of a roadway has the following variables defined according to the AHTD Roadway Inventory Manual:

1. Lane width: The actual measurement of a traffic lane. Eight feet is the smallest acceptable lane width on a two-lane route. A lane width of seven feet or less is considered one lane, and the widths of both directions are added to make the lane width. Lane widths usually range from 8 to 12 feet.
2. Surface width: The total actual measurement for through-traffic lanes, including turn lanes/bays, climbing lanes, and passing lanes, but excluding parking lanes, shoulders, and median width.
3. Right shoulder width: Width of the right shoulder.
4. Left shoulder width: Width of the left shoulder.
5. Median width: Width of a median.
6. Access control: Refers to the ability to enter or leave a facility. The value 1 represents full control of access, 2 represents partial control of access, and 3 represents no control of access.
7. Curbs: Indicates the presence of curbs on a facility. A value 0 represents no curbs, 1 represents curbs on left side of the road, 2 represent curbs on right side, and 3 represents curbs on both sides.

Traffic and crash data included variables such as light conditions, road surface conditions, road surface types, roadway alignment, and roadway profiles. Not all variables were considered in the analysis as some variables will make the model redundant with information, and too many variables will not provide a good fit.

Analysis of Parameter Estimates:-
For each parameter in the model, the GENMOD procedure provides the estimated parameter value, the error of the parameter estimate, the confidence intervals, the Wald chi-square statistic and associated p-value for testing the significance of the model parameter. If a column of the model matrix corresponding to a parameter is found to be linearly dependent, or aliased, with columns corresponding to parameters preceding it in the model, the process assigns it zero degrees of freedom and displays a value of zero for both the parameter estimate and its standard error.

Goodness of Fit:-
Statistical inference can be made about the parameters using confidence intervals and hypothesis tests. The GENMOD procedure fits a generalized linear model to the data by maximum likelihood estimation of the parameter vector \( \beta \), which is an iterative fitting process. Fitted generalized linear models can be summarized through statistics such as parameter estimates, their standard errors, and goodness-of-fit statistics. There is, in general, no closed form solution for the maximum likelihood estimates of the parameters. The dispersion parameter \( \phi \) is also estimated by maximum likelihood or, optionally, by the residual deviance or by Pearson’s chi-square divided by the degrees of freedom. Covariance, standard errors, and p-values are computed for the estimated parameters based on the asymptotic normality of maximum likelihood estimators. Log-likelihood functions for the distributions available in the procedure are parameterized in terms of the mean \( \mu_i \), the dispersion parameter, \( \phi \). The term \( y_i \) represents the response for the \( i^{th} \) observation.

An important aspect of generalized linear modeling is the selection of explanatory variables in the model. Many variables can represent causes of an effect. In generalized linear modeling, responsible variables for an effect can be selected by the goodness-of-fit. Changes in goodness-of-fit statistics are often used to evaluate the contribution of subsets of explanatory variables to a particular model. The goodness-of-fit matrix often includes deviance, scaled deviance, log-likelihood, and value/DF.

The value/DF value for the deviance calculates over-dispersion. If it nears 1, it implies that the model is a good fit and the parameters obtained are reliable. The importance of the additional explanatory variable can be measured by the difference in deviances or fitted log likelihoods between successive models.

Data Set Analyzed:-
The data for this study was provided by the Arkansas Highway Transportation Department (AHTD). The data were collected on junctions and non-junctions of State highways from 2005 to 2007. The data included: crash data, person data, and road geometry data. The crash data included surface type (asphalt, concrete), road surface condition (dry, wet), weather (clear, rain, snow, etc.), light conditions (darkness, daylight), etc. Person data reported restraint type, driver’s age, race, gender, alcohol consumption, etc. Road geometry data contained information related to alignment
(straight, curve), profile (grade, level), and intersection (yes, no), etc. The data were merged for analysis using crash number as the primary key to produce a final data set.

Data Preparation:
The data were then analyzed for missing values and extreme or unusual values. Initial exploration of the data revealed a few missing values and inconsistencies. These variables cannot be replaced with any other values that were deleted from the data. A few variables had values that were not defined in the Arkansas Roadway Inventory Manual. Some geometric variables had values of 999, which were invalid and cannot be used in statistical analysis. Certain variables did not provide the necessary information in their original state, were, therefore, recoded and transformed to retrieve useful data. One such variable was AADT. The AADT variable was modified to express it in the same units as the number of crashes in a segment. The crashes were calculated for three years and AADT is the traffic volume.

Data for count models may be censored or truncated. The data were referred as censored if a range of values for the dependent variable is collapsed into a single value. For example, in a survey in which visibility of an object from a distance should be rated, the responses include 0, 1, 2, 3, and 4. If the frequencies of the responses for 0,1,2,3 are very high compared to 4, then the responses for 4 can be converted to 3 and used in the analysis. Now the responses include 0,1,2,3 only. The data can be censored ‘to the right’ or ‘to the left’. For continuous variables, if a variable such as lane width is considered, some of the values in lane width can be missing or can have values like 0, 99, 999, etc. The values were replaced with average values for the remaining values of the lane width.

Since the data were censored, the distribution can be truncated, if no values were observed in a certain range. For example, if observations with values of 4 were discarded because the frequency for that response was low, the remaining observations would come from a truncated distribution. Explanatory variables or response variables are important factors in forecasting crash frequency. Variables used in a model should be significant to predict crash frequency. Continuous variables were also used in crash frequency modeling. The main variables were lane width, surface width, median width, length of the segment, and AADT. Certain variables were eliminated, like the variables used only for identification, the variables with a lot of missing data, and the variables with unacceptable values. Cleaned data were used in statistical analysis.

Results:-
State highways are a system of highways throughout a state, which primarily serve arterial, or through traffic. With the exception of urban areas, the highway system includes both primary and secondary roads. Highways, which are not included in the primary system of State highways, serve as collectors and feeder routes. These roadways connect local service roads to the state highway system. Studies have shown that Arkansas State Highways are more prone to crashes especially in the vicinity of a major city, such as the cities of Fayetteville and Little Rock. Table 3 presents that crash frequency was higher near urban areas. With 23,624 crashes, urban highways captured more than 67 percent of all the crashes on State highways during 2005-07. Two categories of intersections related crashes were found, i.e. ‘intersection related’ crashes and ‘intersections’, because of the low sample size of ‘intersection related’ crashes, they were merged with ‘intersections’. There are 239 State highways in Arkansas, categorized into 8 different models as indicated in Table 3. Table 3 shows that the crash frequency is highest for urban highways, undivided and non-junctions. Such models are, therefore, very important.

Table 3:- Categories of Model Development for State Highways

| TYPE  | MEDIAN | JUNCTION TYPE               | MODEL | CRASH FREQUENCY | CRASH PERCENTAGE |
|-------|--------|----------------------------|-------|-----------------|------------------|
| Urban | Undivided | Intersection and Intersection Related | S1    | 13928           | 39.86%           |
|       |         | Non-junction               | S2    | 5806            | 16.61%           |
|       | Divided | Intersection and Intersection Related | S3    | 2782            | 7.96%            |
|       |         | Non-junction               | S4    | 1108            | 3.17%            |
| Rural | Undivided | Intersection and Intersection Related | S5    | 2843            | 8.13%            |
|       |         | Non-junction               | S6    | 7665            | 21.93%           |
Individual Model Analysis:-

Table 3 shows the models developed for eight different categories. Rural divided models had poor goodness-of-fit due to small sample size and are not included in this paper. This paper presents models that provided good fit. Tables 4 and 6 show the significant variables with the parameter estimates for the different models. Upper bound (UB) and lower bound (LB) values represent 95 percent confidence intervals of the estimates. Selection of significant variables was based on the p-value. Variables with p-value less than 0.005 were considered significant with a level of 95 percent. Table 5 presents the goodness-of-fit criteria and the corresponding values for fitness of the models.

Urban Highways:-

First model (S1): Urban, Undivided, Intersection and Intersection Related:

This model represents all the undivided intersections and intersection-related segments in the urban vicinity. From Table 3, it can be observed that this highway type has the highest number of crashes, 13,928. This is almost 59 percent of the entire urban crash frequency and just under 40 percent of all the crashes on State highways in Arkansas, which is very high. Several variables were used to model crash frequency. Table 4 represents all the significant variables obtained from the analysis.

Table 4 indicates high traffic volume causes an increase in crash frequency. Negative estimate for partial and full control of access indicates that higher access control, lowers the crash frequency. Increase in the number of lanes indicates an increase in crash frequency. This could be attributed to excess lane changing, which may effect flow conditions. Lane width, surface width and road width are geometric variables. It is also observed that all of these variables have negative estimates. This indicates that, as the width of the lane, the surface of the road increases, the frequency of crashes decreases. These variables are inter-related and, ultimately, indicate the impact of the width of a roadway section. The higher the width, the larger the driving space, the fewer crashes will occur. Table 5 provides different goodness-of-fit criteria. Since the deviance and the scaled deviance were very near to 1, it can be observed that the Model S1 is a good fit.

Second Model (S2): Urban, Undivided, Non-Junctions:

Urban undivided non-junction segments also have high crash frequency, accounting for just under 17 percent of the total frequency. Table 4 represents all the significant variables obtained from this analysis. It can be observed that geometric variables are significant for Model S2. As expected and similar to Model S1, AADT has a positive estimate. This indicates high crash frequency in high volume segments. Once again, partial or full access control shows a significant negative estimate, indicating that more the control, the fewer the crashes. Similar to Model S1, an increase in the number of lanes and a decrease in lane width also contributes positively to the crash frequency. Further, it can be observed that presence of a right shoulder decreased crash frequency, but the presence of a left shoulder increased crash frequency. This could be attributed to lower vehicle speed on the right lanes compared to in the left lanes. As expected, variable length has a significant positive estimate. This indicates that, as expected, longer highway sections have higher crash frequency. It can be observed from Table 5 that the deviance and the scaled deviance are close to 1. Therefore, Model S2 is a good fit.

Models S1 and S2 represent undivided highway sections in urban areas with higher traffic volumes. As a result these highway sections have higher crash frequency, with a total of 19,734 crashes representing 56.47 percent of the total number of crashes on State Highways between 2005 and 2007. This indicates that undivided highways are more prone to crashes, and require more attention while driving. This is due to the fact that undivided highways are separated by pavement markings between the lanes and lack any space between traffic in the opposite direction. An imbalance in traffic will not only affect the lanes with traffic in the same direction, but also in the opposite direction. Unlike the divided highway segments, this increases the possibility of crashes in the opposite lane. Divided highway segments have barriers or medians to keep safe distance between the opposite lanes. This keeps the traffic disturbance confined to one direction, without interrupting the flow in the opposite direction.
Third Model (S3): Urban, Divided, Intersection and Intersection related:–

Urban divided segments have low crash frequency compared to that of undivided segments. This is attributed to the available space between the two roadways. Table 4 represents the significant variables obtained from the analysis. All geometric variables are significant in Model S3 and the significant variables are very similar to that of Models S1 and S2. The presence of a right shoulder decreases crash frequency, but presence of a left shoulder increases the same. An increase in lane width or road width decreases crash frequency, which indicates that fewer crashes took place where the width of lane or road segment is high. Once again, as expected, variables length and AADT are significant with a positive estimate. This indicated high crash frequency for longer highway segments and segments with high traffic volume. Table 5 shows the goodness-of-fit criteria for Model S3. The deviance and scaled deviance values are close to 1. Hence, Model S3 is a good fit.

Fourth Model (S4): Urban, Divided, Non-Junctions:–

Similar to Model S3, urban divided non-junction segments also have low crash frequency. This indicates that driving on divided segments are comparatively less prone to crashes when compared to undivided segments. The entire urban divided network contributes to just over 11 percent of total crashes. Table 4 represents all the significant variables obtained from this analysis.

Similar to Model S3, Model S4 shows all geometric variables as significant in Table 4. As explained earlier, highway segments with high traffic volume have high crash frequency. This is the reason for a significant positive AADT estimate. Similar to Models S1, S2 and S3, an increase in the number of lanes increases the frequency of crashes. Highway segments with narrower lane width or surface width have higher crash frequency, and longer highway segments also have a higher number of crashes. Table 5 indicates that the deviance and the scaled deviance are close to 1. Therefore, Model S4 is a good fit.

By comparing Models S1, S2, S3 and S4, it can be observed that 83.5 percent of the total crashes in urban areas occurred on undivided segments of State highways. The entire urban network consists of 23,624 crashes, which is 67.6 percent of the total crash frequency. This high frequency indicates that safety in urban areas is of high concern, with undivided highway segments as a higher priority due to their high crash frequency. Moreover, crash frequency in urban areas is higher due to high traffic volume on the highway segments. For all models, access indicated a negative estimate, which indicated reduction in crash frequency as a result of availability of access.

Table 4: Urban Highways: Significant Variables and Parameter Estimates

| Parameter | Estimate | Standard Error | LB     | UB     | Wald Chi-Square | p-value |
|-----------|----------|----------------|--------|--------|----------------|---------|
| **First model (S1): Urban, Undivided, Intersection and Intersection Related** |
| OTHERS    | 0.2783   | 0.1302         | 0.0231 | 0.5334 | 4.57           | 0.0326  |
| LITdark   | 0.1258   | 0.0636         | 0.0011 | 0.2505 | 3.91           | 0.048   |
| LnAADT    | 0.2685   | 0.0121         | 0.2448 | 0.2923 | 491.96         | <.0001  |
| ACCES_Y   | -0.3676  | 0.0681         | -0.501 | -0.2341| 29.15          | <.0001  |
| NOLAN     | 0.1203   | 0.0131         | 0.0946 | 0.146  | 84.09          | <.0001  |
| LNWID     | -0.0029  | 0.0008         | -0.0044| -0.0013| 13.42          | 0.0002  |
| SURWD     | -0.0021  | 0.0005         | -0.003 | -0.0012| 19.6           | <.0001  |
| RDWID     | -0.0026  | 0.0006         | -0.0038| -0.0014| 17.39          | <.0001  |
| **Second Model (S2): Urban, Undivided, Non-Junctions** |
| LnAADT    | 0.204    | 0.017          | 0.1706 | 0.2374 | 143.55         | <.0001  |
| ACCES_Y   | -0.1175  | 0.03           | -0.1762| -0.0587| 15.35          | <.0001  |
| NOLAN     | 0.2271   | 0.0259         | 0.1762 | 0.278  | 76.6           | <.0001  |
| LNWID     | -0.0345  | 0.0021         | -0.0385| -0.0304| 280.03         | <.0001  |
| RSHOS     | -0.0078  | 0.0017         | -0.0111| -0.0046| 22.44          | <.0001  |
| LSHOS     | 0.0124   | 0.0019         | 0.0087 | 0.0162 | 41.96          | <.0001  |
| Lnlength  | 0.1654   | 0.0122         | 0.1415 | 0.1894 | 183.11         | <.0001  |
| **Third Model (S3): Urban, Divided, Intersection and Intersection related** |
| ACCES_Y   | -0.3062  | 0.0426         | -0.3897| -0.2227| 51.62          | <.0001  |
| NOLAN     | 0.1018   | 0.0409         | 0.0217 | 0.1819 | 6.21           | 0.0127  |

411
2.4 undivided segments contribute

Table 6:

It should be noted that not many rural sections have curbs, however, the variable was significant and therefore kept 
significant variables obtained for undivided intersections. Presence of 
30 percent, leaving divided segments, with 
percent of total crashes, it can be stated that driving in rural segments is safer. Rur 
Rural segments have comparatively low crash frequency when compared to that of urban highways. With 3 
Fifth Model 
Rural Highways: 

Table 5:- Goodness-of-Fit for All Models

Table 6:- Rural Highways: Significant Variables and Parameter Estimates

Rural Highways:

Fifth Model (S5): Rural, Undivided, Intersection and Intersection Related:

Rural segments have comparatively low crash frequency when compared to that of urban highways. With 32.4 percent of total crashes, it can be stated that driving in rural segments is safer. Rural undivided segments contribute 30 percent, leaving divided segments, with 2.4 percent of crashes, as the safest segments in the entire State highway network. Their crash frequency was so low that it was insufficient to fit any model. Table 6 represents all the significant variables obtained for undivided intersections. Presence of curbs indicated a positive impact on crashes. It should be noted that not many rural sections have curbs, however, the variable was significant and therefore kept in the model.

| Parameter   | Estimate | Standard Error | LB    | UB    | Wald Chi-Square | p-value |
|-------------|----------|----------------|-------|-------|----------------|---------|
| LnAADT      | 0.3665   | 0.022          | 0.3234| 0.4095| 278.53         | <.0001  |
| SURWD       | -0.0066  | 0.0018         | -0.0011| -0.003| 13.28          | 0.0003  |
| RDWID       | -0.0263  | 0.0022         | -0.0307| -0.022| 141.58         | <.0001  |
| RSHOW       | 0.0578   | 0.0176         | 0.0234| 0.0923| 10.83          | 0.001   |
| LSHOW       | -0.0661  | 0.0195         | -0.1044| -0.0278| 11.46          | 0.0007  |
| CURBS_Y     | 0.3491   | 0.0707         | 0.2105| 0.4877| 24.36          | <.0001  |
| ACCESS_N    | -0.3881  | 0.0975         | -0.5791| -0.1971| 15.86          | <.0001  |

Sixth Model (S6): Rural, Undivided, Non-Junctions

| Parameter   | Estimate | Standard Error | LB    | UB    | Wald Chi-Square | p-value |
|-------------|----------|----------------|-------|-------|----------------|---------|
| LnAADT      | 0.2806   | 0.0155         | 0.2502| 0.311 | 327.63         | <.0001  |
| Lnlength    | 0.259    | 0.01           | 0.2395| 0.2786| 671.95         | <.0001  |
From Table 6, it can be observed that there are similarities between urban and rural models. Length and AADT, as expected, have significant positive estimates. This indicates that long highway sections and sections with high volume have a high crash frequency. Similar to the urban models, an increase in surface width or road width decreases crash frequency. The presence of a right shoulder increases crash frequency, whereas a left shoulder offers a negative estimate decrease in crash frequency in rural highway sections. This indicates the dissimilarity between urban and rural highway sections. Since the deviance and scaled deviance in Table 5 are close to 1, Model S5 provides a good fit.

Sixth Model (S6): Rural, Undivided, Non-Junctions:
Undivided non-junction segments in rural highways had the highest crash frequency in the entire rural highway network, consisting of 67.7 percent of all the rural crashes. Table 6 represents all the significant variables obtained for undivided non-junctions.

It can be observed from Table 6 that Model S6 has similarities with Model S5. AADT and length offer significant positive estimates, indicating that crash frequency is higher for longer sections and sections with higher volume. This has been consistent in all the models. Also, decrease in crash frequency with an increase in either road width or surface width indicated the consistency of these variables. As expected, a straight roadway profile has a significant negative estimate, which indicates that fewer crashes took place on tangent sections of highways. It can also be observed from Table 6 that fewer crashes occur on level roadways. Similar to Model S5, presence of right shoulders increases crash frequency, but left shoulders decrease crash frequency. Also, no access to the facility indicated a negative estimate, indicating reduction in the frequency of crashes. From Table 5, deviance and scaled deviance are close to 1, hence, Model S6 is a good fit.

Discussion:
From the analysis of State highways, it can be observed that crash frequency was highest on undivided segments. Moreover, the frequency was higher on high volume segments. Level and straight highway segments were comparatively safe. The goodness-of-fit values for all models developed show that they are a good fit. Lane width is shortest compared to surface and road widths. Surface width includes the total width, excluding median and shoulder widths. Road width includes everything related to the road. Shoulder width selection could reflect the traffic volume. It should be noted that lane, surface and road width have the same signs on urban highways. Further, for Models S2 and S3 on urban highways, the left shoulder has positive estimates whereas the right shoulder showed negative estimates. The signs for shoulders were opposite for rural highways, Models S5 and S6.

Conclusions:
This paper describes the development of crash prediction models for State highways in Arkansas. Crashes were divided into subgroups according to the location of the roads (urban or rural), and the division of the roads (divided or undivided). Roadways were also classified as junctions and non-junctions. In total, six models were developed. Negative binomial regression models were preferred over Poisson regression models.

Negative binomial regression models are extensions of the Poisson models which include properties to address complex data that is over-dispersed and correlated. To avoid complications during analysis, the data were cleaned, and only appropriate data were used. Any invalid or unacceptable values in the data were omitted. Variables were selected depending upon the quality of the data provided, the purpose of the variables, and the significance of the variables in calculating crash frequency. Correlated variables were not used in the analysis, i.e., if two variables provide the same information about the data, then one of the variables was used. This reduced redundancy in the model and provided a good fit.
Variables identified as significant varied among the models. Length of segment and AADT proved significant for almost all models, mainly because the number of crashes were often directly proportional to these two variables. Geometric variables like lane width, surface width, and road width were found to be significant in a number of models. Right shoulder width and left shoulder width turned out to be significant for a few models. As expected, segment-related crashes and intersection-related crashes showed different significant variables. All models were developed using the parameter estimates of these variables. It would be supportive for the models developed to be calibrated and validated with other data sets to determine their strength and usefulness in predicting future crashes.

Acknowledgments:
This study was funded by the Mack-Blackwell Rural Transportation Center (MBTC) and the Arkansas Highway and Transportation Department (AHTD). AHTD also provided the datasets used in the study.

References:
1. Abdel-Aty, M., Xuesong, W. (2006b) “Crash Estimation at Signalized Intersections along Corridors: Analyzing Spatial Effects and Identifying Significant Factors”, Transportation Research Record 1953, Transportation Research Board, 98-111.
2. Abdel-Aty, M., Keller, J., Brady, P. A. (2005) “Analysis of Types of crashes at Signalized Intersections by Using Complete Crash Data and Tree-Based Regression”, Transportation Research Record 1908, Transportation Research Board, 37-45.
3. Abdel-Aty, M., Pemmanaboina, R., Hsia, L. (2006c) “Assessing Crash Occurrence on Urban Freeway Applying a System of Interrelated Equations.” Transportation Research Record 1953, Transportation Research Board, 1-9.
4. Abdel-Aty, M., Xuesong, W., Brady, P. A. (2006a) “Crash Estimation at Signalized Intersection: Significant Factors and Temporal Effects.” Transportation Research Record. 1953, Transportation Research Board, 10-20.
5. Banihashemi, M., Dimaista, M. (2005) “Maximizing Safety Improvement Benefits in Crash Prediction Models with Crash Modification Factors.” Transportation Research Record. 1908, Transportation Research Board, 9-18.
6. Deng, Z., Ivan, J. N., Garder, P. (2006) “Analysis of Factors Affecting the Severity of Head On Crashes; Two-lane Rural Highway in Connecticut.” Transportation Research Record, 1953, Transportation Research Board, 137-146.
7. Hadayeghi, A., Shalaby, A. S., Persaud, B. N. (2007) “Safety Prediction Models: Proactive Tool for Safety Evaluation in Urban Transportation Planning Applications.” Transportation Research Record. 2019, Transportation Research Board, 225-236.
8. Hauer, E. (1995) “On exposure and Crash Rate”, Traffic Engineering and Control, 36, (3), 134-138.
9. Hilbe, J. (1994) “Log Negative Binomial Regression Using the GENMOD Procedure.” Proceedings of the Nineteenth Annual SAS User’s Group International Conference, Cary, NC, SAS Institute Inc.
10. Jonsson, T., Ivan, J. N., Zhang, C. (2007) “Crash Prediction Models for Intersections on Rural Multilane Highway: Differences by Collision Type”. Transportation Research Record. 2019, Transportation Research Board, 91-98.
11. Kweon, Y. J., Kockelman, K. M. (2004) “Spatially disaggregate panel models of crash and inquiry counts: the effect of speed limits and design.” 83rd Annual Meeting Transportation Research Board, Washington, D.C.
12. Lawless, J. F. (1987) “Negative Binomial and Mixed Poisson Regression.” Canadian Journal of Statistics, 15, (3), 209-225.
13. McCullagh, P. and Nelder, J. A. (1989) Generalized Linear Models, Second Edition.
14. Miaou, S. P. (1994) “The Relationship between truck crashes and geometric design of road sections: Poisson versus negative binomial regressions.” 73rd Annual Meeting Transportation Research Board Washington, D.C.
15. National Highway Transportation Safety Association (NHTSA, 2008), http://list.nsc.org/defensivedriving/images/uploads/NHTSA_Traffic_Safety_Facts_-_August_2008.pdf, Accessed April 20, 2015.
16. Noland, R. B., Quddus, M. A. (2004) “A spatially disaggregate analysis of road causalities in England.” Accident Analysis and Prevention, 36, (6), 973-984.
17. Oklahoma State, http://www.okstate.edu/sas/v8/saspdf/stat/chap29.pdf. Accessed, April 20, 2015.
18. Pedan, A. (2001) “Analysis of Count Data Using SAS System.” Statistics, Data Analysis, and Data Mining. Vasca Inc., Tewksbury, MA.