IDENTIFICATION OF CRASH CAUSAL FACTORS: EFFECTS OF SAMPLE DATA SIZE

Uday R. R. Manepalli
Civil, Architectural and Environmental Engineering,
Missouri University of Science & Technology
Rolla, MO, USA, E-mail: umr75@mail.mst.edu

Ghulam H. Bham
Corresponding author
Civil, Architectural and Environmental Engineering,
Missouri University of Science & Technology
Rolla, MO, USA, E-mail: ghbham@mst.edu

Submitted to the 3rd International Conference on Road Safety and Simulation,

September 14-16, 2011, Indianapolis, USA

No. of words in abstract =
No. of words in text = 3716 +2500 (8 tables and 2 Figures) = 6216

ABSTRACT

This paper utilizes data for a county to identify the main crash contributing factors for several counties. For this analysis, the counties of Arkansas are categorized based on crash frequency and crash severity index into five categories. For each category, sample crash data for a county or a group of counties and the remaining data (for several counties) are analyzed and based on the results crash contributing factors are identified. The selection of sample data for each category is based in the order of highest crash severity index (CSI) or highest crash frequency. The crash contributing factors are identified using multinomial logistic regression (MLR). The results indicate that most of the factors identified within each category were also identified for the sample data. Sample size, however, changed for each category. This paper presents the effects of this difference in sample size and the effect of categorization of counties based on crash severity index and crash frequency in identification of crash contributing factors. This study will help better allocate funds by the departments of transportation to identify factors that are positively associated with crash severity. Three years of rural two-lane highway crash data from Arkansas is used in this analysis. Results indicate that division of counties based on crash frequency and identification of crash contributing factors using MLR would ensure better allocation of funds. Rural two-lane undivided highways were selected for analysis as severe crashes are common on these highways.

INTRODUCTION

About 60% of all fatal crashes in Organization for Economic Co-operation and Development (OCED) member countries occur on rural highways (OCED, 1999). In the United States, 40% of all motor vehicle travel on rural highways account for 60% of fatal crashes (NHTSA, 2003a). The ratio of fatalities per 100 million vehicle miles traveled for rural to urban areas is 2.3 to 1 (NHTSA, 2002). In the US, nearly 75% of rural fatal crashes occur on rural two-lane undivided highways (NHTSA, 2003b). Most of the rural highways operate on two-lanes and are undivided (Persaud et. al., 2004). In Arkansas, from 2004 to 2006, 50% of fatal crashes and 49% of major injury crashes occurred on rural two-lane undivided highways. Therefore, it is necessary to identify the causal factors which contribute to these fatal crashes.

Traditional crash prediction models have used negative binomial regression to assess highway safety based on crash counts and crash rates (Shankar et al., 1995; Poch and Mannering, 1996; Abdel-Aty and Radwan, 2000; Harwood et al., 2000; Savolainen and Tarko, 2005). Negative binomial and Poisson models were developed for individual collision types for rural intersections (Kim et al., 2006). Jones and Whitfield (1988) and Liu et al. (1988) presented significant findings using a multivariate approach to crash severity analysis based on logistic regression models. Shankar and Mannering (1996) applied the multinomial logit models for crash severity to study motorcycle safety. Also, Shankar et al. (1995) developed a predictive model of crash severity using a nested logit model of environmental, geometric, weather, vehicular, and human factors.

Pande and Abdel-Aty (2009) presented a new approach to the analysis of severe crash patterns on multilane highways. Segments with severe crashes were compared with segments with no-crashes instead of non-severe crashes to avoid the problem of under-representation of such crashes in the data. Due to the importance of rear-end crashes, Abdel-Aty and Abdelwahab (2004) and Xuedong et al. (2005) performed two different types of analyses, both studies emphasized the greater risk of rear-end crashes than other crash types. The study revealed seven factors that increased the risk of rear-end collisions at junctions (Xuedong et al., 2005). Abdel-Aty and Abdelwahab (2004) found that sight distance, abrupt stopping by the leading vehicles, and close following of heavy-duty vehicles were unsafe for passenger vehicles.

The literature reviewed, indicates that there is significant research carried out in the field of developing models and understanding the factors which influence crashes, and can be applied to specific cases like rural two-lane undivided highways. However, research in identifying crash contributing factors from a sample data representing a large sample size has not been carried out, which is the main objective of this paper. Additionally, this research aims in saving computational time and in better allocation of resources by the departments of transportation (DOT).

The methodology used to achieve the above objective, followed by crash data used is presented in the following sections. The results follow the crash data and the paper ends with conclusions and recommendations for future research.

METHODOLOGY

ArcGIS, a Geographic Information System was used to analyze crash data. The counties were divided into five categories based on crash severity index (CSI) and crash frequency (CF) using Jenk's algorithm (Lewis, 2010), one of six classification schemes in ArcGIS. For each category, county/group of counties (sample data) in the order of highest CSI/highest CF was selected and the crash contributing factors were analyzed using multinomial logistic regression (MLR). The selection of sample data satisfied the condition of minimum sample size of 2000 crashes to achieve statistically reliable estimate for MLR models (Ye and Lord, 2011). The crash contributing factors for the remaining data in each category were identified by MLR. Finally, the percentage of common contributing factors identified was computed with respect to the remaining crash data. Common factors followed a similar trend (increase/decrease) in terms of the odds ratio and those variables that were statistically significant among the sample data were found to be significant for the remaining counties. Though the counties were categorized by crash frequency, the crash contributing factors were analyzed using MLR. This was carried out to check whether the classification of counties based on CSI or CF would affect the results. The choice of using MLR, Jenk's algorithm, and use of CSI are presented in the following in detail.

Multinomial Logistic Regression

Logistic regression models can be used to analyze crash severity with binary response variables (in this case, severe versus non-severe) (Bham et al., 2011). Binary logistic regression models require that the dependent variable have two categories. However, collapsing multiple variables into two categories or removing a category to create a binary logit model can generate misleading conclusions (Abdel-Aty and Abdelwahab, 2009). When crash severity is used, the dependent variable usually has more than two categories. This study, therefore, uses multinomial logistic regression (MLR) to identify factors that contribute to crashes and takes into account multiple levels of crash severity. The odds ratio were also determined, which determines the relative risk of crash severity based on factors that contribute to crashes.

Logistic regression (LR) can be used to predict dependent variables from categorical independent variables, and it can determine the percent of variance in the dependent variable that is explained by the independent variable. LR can also rank the relative importance of independent variables to assess the interaction effects, and explain the impact of covariate control variables. The impact of predictor variables can be explained in terms of the odds ratios. Multinomial logistic regression (MLR) can handle dependent variables with more than two levels. The dependent variable used here, crash severity, has five levels, S1 through S5. Table 1 presents the details of independent variables and their levels.

MLR analysis was performed using the CATMOD procedure with the Statistical Analysis System (SAS) (SAS, 2009) to identify the factors that contribute to crashes and are positively associated with crash severity. The CATMOD procedure has been used in the past for linear modeling, log-linear modeling, logistic regression, and repeated measurement analysis (Bham et al., 2011).

Categorization of Counties

Counties can be categorized based on six classification schemes: equal interval, defined interval, quartile, natural breaks, geometric interval, and standard deviation. In the natural breaks scheme, the classes are based on natural categorizing inherent in the data. The break points are identified by the class breaks that best group similar values and maximize the differences between the classes. The features are divided into classes of the boundaries which correspond to relatively big jumps in the data values. This classification scheme best suits the present study.

In this study, counties were categorized using natural breaks which are based on the Jenks' algorithm (Lewis, 2010). This algorithm is a common method of classifying data in a choropleth map, a type of thematic map that uses shading to represent classes of a feature associated with specific areas (e.g., a population density map). This algorithm generates a series of values that best represent the actual breaks observed in the data as opposed to some arbitrary classificatory scheme; thus, it preserves true clustering of data values. The algorithm creates *k* classes so that the variance within categories is minimized (ArcGIS Online Manual).

Crash Severity Index (CSI)

Crash severity index (CSI) is an index which considers the effect of the severity and compute relative to property damage crashes. A high CSI indicates a large number of fatal crashes or frequency of crashes with various levels of severity. The CSI was computed as:

$$CSI = S1*W1 + S2*W2 + S3*W3 + S4*W4 + S5*W5 \qquad (1)$$

where:

$S1$	= frequency of crashes involving fatalities,
$S2$	= frequency of crashes involving incapacitating injuries,
$S3$	= frequency of crashes involving moderate injury,
$S4$	= frequency of crashes involving complaint of pain,
$S5$	= frequency of crashes involving property-damage-only (PDO), and
WI	= weights assigned to a given crash severity level,

The weights used are based on the comprehensive crash costs per person for each type of crash. The costs are $4,008,900 for a fatality (S1); $216,000 for major injury (S2); $79,000 for a minor injury (S3); $44,900 for complain of pain (S4) and $7,400 for property damage only (S5) (Highway Safety Manual, 2010). The weights represent the ratios of comprehensive crash costs to the cost of PDO crashes (S5). For example, the weight of a fatal crash (S1) is calculated as the cost of such a crash ($4,008,900) divided by the cost of a property-damage-only crash ($7,400) and is thus equal to 542. Other weights were computed similarly and rounded to the nearest zero. They are 29, 11, 6 and 1 for major injury (S2), minor injury(S3); complain of pain (S4) and property damage only (S5), respectively.

DATA

The Arkansas Highway and Transportation Department (AHTD) provided the crash data used in this study. The data were collected from 2004 through 2006. This study is performed for rural two-lane undivided sections which had a crash frequency of 20,359 crashes in three years (2004-2006) which contributes to 18% (20359/112695) of crashes in Arkansas. Among these crashes, 11315 crashes included male drivers and 9044 crashes occurred for female drivers.

Only certain factors were retained for analysis as some factors had missing values. If more than 10% of the values for a factor were missing, that factor was not considered. For, the factors presented in Table 1, no more than 1% had missing values. The county with the highest area is Union with 1038.9 square miles.

Table 1. List of variables

Terms	**Definition**	**Description of levels**
ATM	Atmospheric Conditions	Clear, Rain
LGT	Light Conditions	Dark, Daylight
RSUR	Roadway Surface	Dry, Wet
RALI	Roadway Alignment	Curve, Straight
RPRO	Roadway Profile	Grade, Level
TOC	Collision Types	Angle, Head-On, Rear End, Sideswipe Same Direction (SSSD), Single Vehicle Crashes (SVC), Sideswipe Opposite Direction (SSOD)
WK	Day of the week	Weekdays (M-F), Weekends (Sat, Sun)
DUI	Driving Under Influence	Yes, No

RESULTS

In this section, first the data used in both cases (by categorization based on CSI and CF) are presented, followed by presentation of analysis using MLR for CSI and CF categorization. This section ends with the discussion on the choice of categorization which leads to a small sample size and effective identification of crash contributing factors.

Figures 1 and 55 represent the categorization of counties by CSI and CF for rural two-lane undivided highways in Arkansas. Table 2 lists the number of counties by each category. It also provides information on crash severity (and also crash frequency) for each category and indicates the county/group of counties in the order of the highest CSI in each category. Table 3 presents categorization with crash frequency. It should be noted that for statistically reliable results using MLR, the minimum data used was 2000 crashes (Ye and Lord, 2011). Hence, in Table 2 for first, fourth and fifth categories, total data for the category were used in the analysis i.e., sample data cannot be used. This is similar to the first and fifth categories in Table 3.

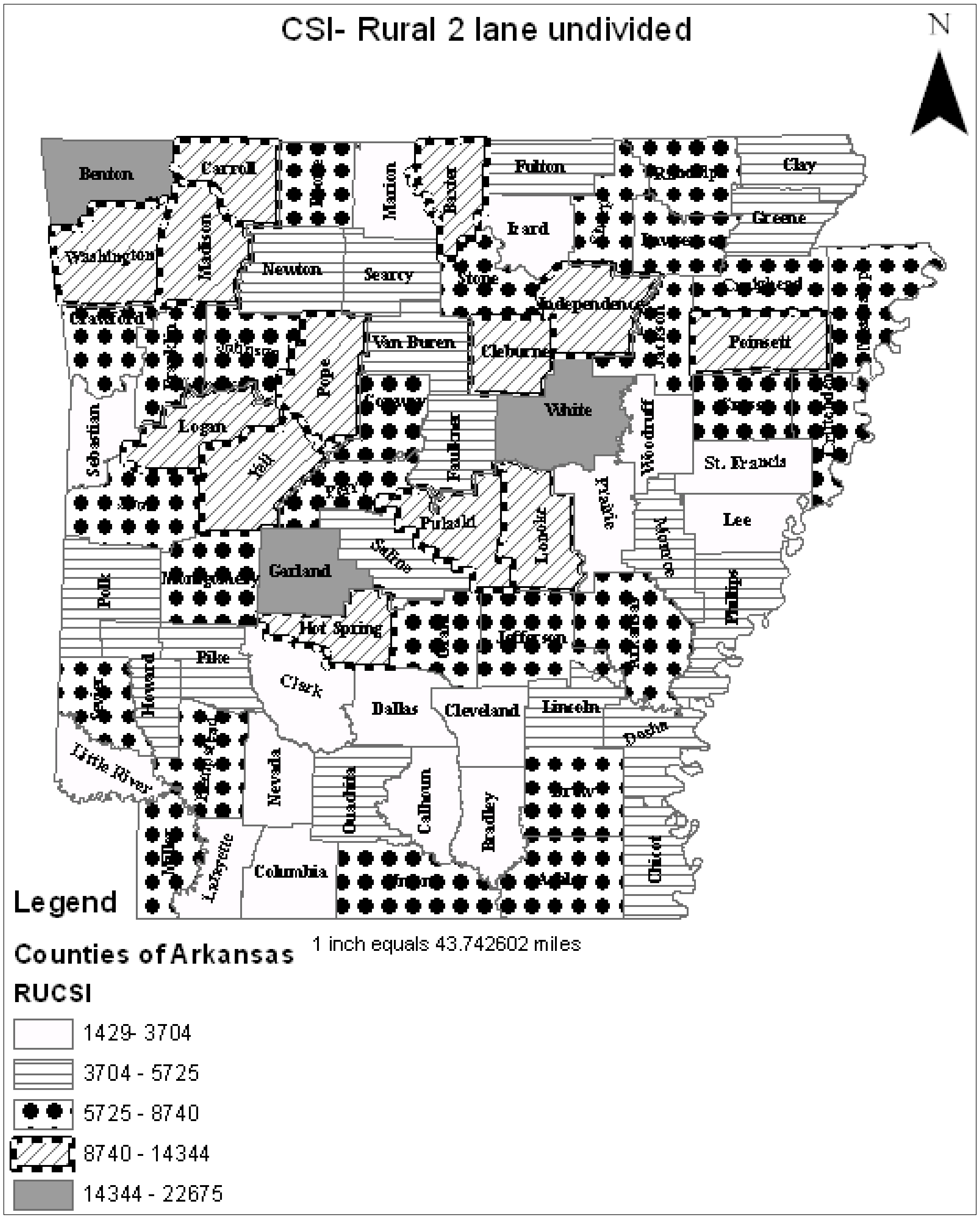

Figure 1. Categorization of counties by CSI (crash severity index) for rural two-lane undivided highways in Arkansas

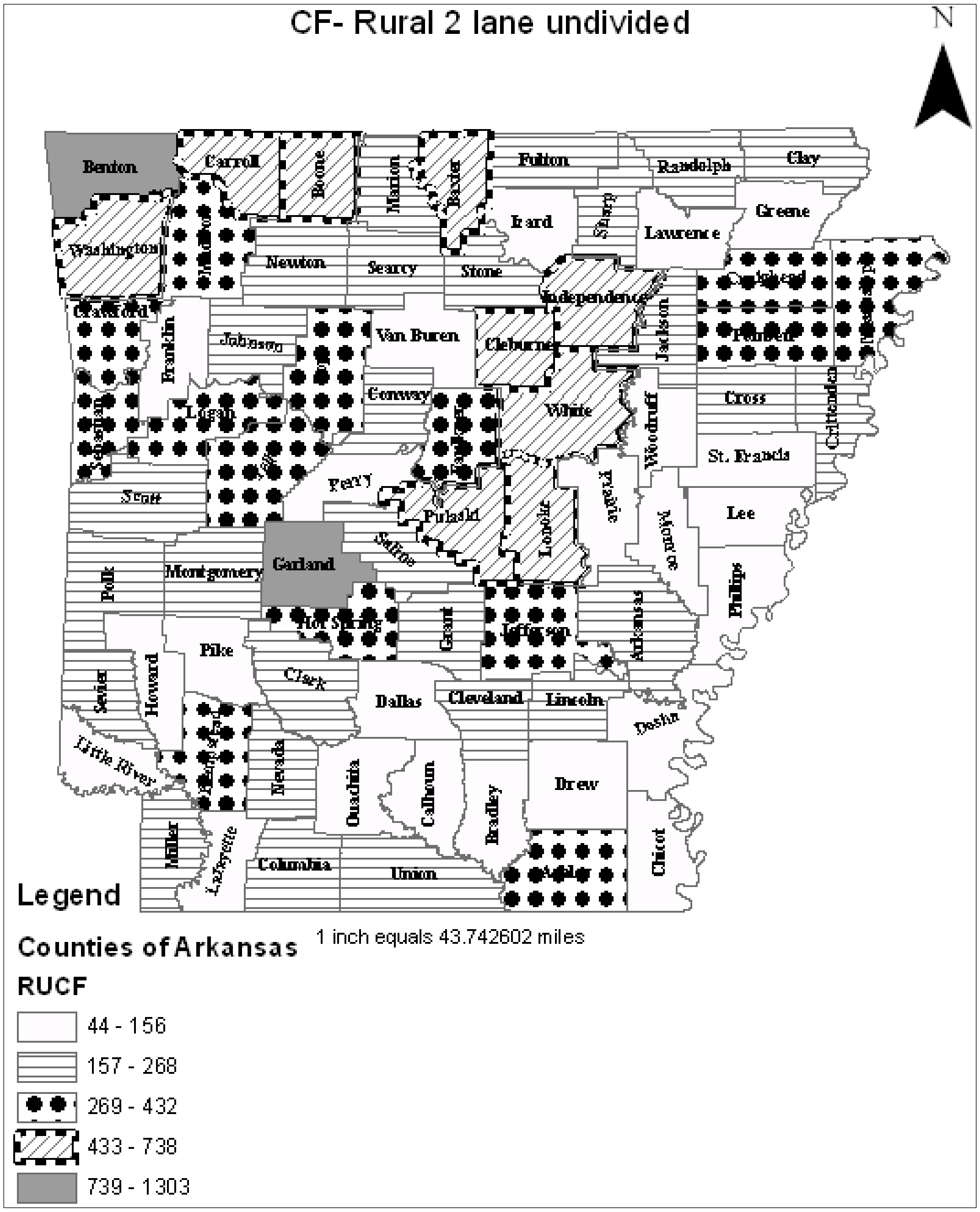

Figure 2. Categorization of counties by CF (crash frequency) for rural two-lane undivided highways in Arkansas

Table 2. Results by category, highest CSI in each category, and percentage of crash data

Category	Number of Counties	County/ Counties with highest CSI@	CSI for county in col. ‘C’ (CF)	Total CSI for counties in col. ‘B’ (CF)	Percentage* by CSI (CF)	Range of CSI
(A)	(B)	(C)	(D)	(E)	(F)	(G)
First	3	Benton, Garland, White	62708 (2911)	62708 (2911)	-	14345-22675
Second	13	Madison, Lonoke, Carroll, Baxter	54039 (2088)	151267 (5874)	36 (36)	8740-14345
Third	26	Craighead, Crawford, Union, Mississippi, Boone, Jefferson	43464 (2184)	178969 (6369)	27 (34)	5725-8740
Fourth	17	All 17 counties	81393 (2966)	81393 (2966)	-	3704-5725
Fifth	16	All 16 counties	43464 (2239)	43464 (2239)	-	1429-3704
Total	75	-	285068 (12388)	517801 (20359)	55 (61)	-

@should satisfy the condition of min sample size of 2000 crash frequency

* represented by county in col. C, in terms of CSI

Table 3. Results by category, highest CF in each category, and percentage of crash data

Category	Number of Counties	County/ Counties with highest CF@	CF for county in col. 'C' (CSI)	Total CF for counties in col. 'B' (CSI)	Percentage* by CF (CSI)	Range of CF
(A)	(B)	(C)	(D)	(E)	(F)	(G)
First	2	Garland, Benton	2173 (44557)	2173 (44557)	-	739-1303
Second	9	White, Lonoke, Washington, Pulaski	2550 (54331)	5009 (109987)	51 (49)	433-738
Third	14	Craighead, Faulkner, Poinsett, Madison, Mississippi	2052 (48176)	4893 (121049)	42 (40)	269-432
Fourth	27	Newton, Sevier, Union, Saline, Fulton, Polk, Montgomery, Randolph, Marion	2219 (52460)	5728 (147245)	39 (36)	157-268
Fifth	23	All 23 counties	2556 (94963)	2556 (94963)	-	44-156
Total	75	-	11550 (294487)	20359 (517801)	57 (57)	-

@should satisfy the condition of min. sample size of 2000 crash frequency

* represented by county in col. C, in terms of CF

MLR identified factors that contributed to crashes and were associated their severity. Table 4 presents the results for Madison, Lonoke, Carroll, and Baxter Counties (second category) based on CSI categorization (the counties in the order of highest CSI in the second category), and Table 5 shows the collective results for the remaining counties in the second category. The factors common to Madison, Lonoke, Carroll, and Baxter Counties and the other counties in the second category that follow similar trend (i.e., an increase or decrease in the odds ratio) are shaded. Details of the analysis of the odds ratio are presented below, along with a few examples for each case.

Table 4. Factors that Positively Associate with Severity of Crashes in Madison, Lonoke, Carroll, and Baxter Counties (Second Category) based on CSI categorization

Parameter		Estimate	Standard Error	Chi-Square	Pr>ChiSq	Odds Ratio
Fatal vs Property Damage Only						
Intercept		-6.3148	0.3408	343.28	<.0001	
RPRO	Grade vs Level	0.3312	0.147	5.08	0.0242	1.39
TOC	Angle vs SVC	4.5851	0.3422	179.57	<.0001	98.01
TOC	Head-on vs SVC	8.5489	0.5229	267.33	<.0001	5161.07
DUI	No vs Yes	-0.7621	0.1721	19.62	<.0001	0.47
Major Injuries vs Property Damage Only						
Intercept		-0.5855	0.2045	8.2	0.0042	
ATM	Clear vs Rain	0.5084	0.2007	6.41	0.0113	1.66
RPRO	Grade vs Level	0.2746	0.0679	16.36	<.0001	1.32
TOC	Head-on vs SVC	3.2525	0.6264	26.96	<.0001	25.85
TOC	Rear-end vs SVC	-1.1246	0.2348	22.93	<.0001	0.32
TOC	SSOD vs SVC	-1.16	0.3458	11.25	0.0008	0.31
TOC	SSSD vs SVC	-1.2144	0.4721	6.62	0.0101	0.30
DUI	No vs Yes	-0.5394	0.0979	30.36	<.0001	0.58
Minor Injuries vs Property Damage Only						
Intercept		-0.9564	0.2592	13.61	0.0002	
TOC	Head-on vs SVC	1.8439	0.746	6.11	0.0135	6.32
TOC	SSSD vs SVC	-2.2314	0.8637	6.67	0.0098	0.11
DUI	No vs Yes	-0.3613	0.1007	12.87	0.0003	0.70
Complain of Pain vs Property Damage Only						
Intercept		-0.5965	0.2042	8.53	0.0035	
TOC	Head-on vs SVC	1.4633	0.7026	4.34	0.0373	4.32
TOC	SSOD vs SVC	-0.9415	0.3105	9.19	0.0024	0.39

*shading indicates common factors in Tables 4 and 5 (includes similar increase/decrease)

Table 5. Factors that Positively Associate with Severity of Crashes (Second Category#) based on CSI categorization

Parameter		Estimate	Standard Error	Chi-Square	Pr>ChiSq	Odds Ratio
Fatal vs Property Damage Only						
Intercept		-2.0491	0.3041	45.41	<.0001	
RALI	Curve vs Straight	0.2813	0.1159	5.89	0.0152	1.32
TOC	Head-on vs SVC	3.5016	0.4305	66.15	<.0001	33.17
TOC	Rear-end vs SVC	-1.5568	0.5373	8.4	0.0038	0.21
DUI	No vs Yes	-1.0635	0.1219	76.12	<.0001	0.35
Major Injuries vs Property Damage Only						
Intercept		-1.0122	0.1781	32.3	<.0001	
RALI	Curve vs Straight	0.2972	0.0704	17.81	<.0001	1.35
RPRO	Grade vs Level	0.2693	0.0656	16.83	<.0001	1.31
TOC	Head-on vs SVC	2.432	0.3643	44.57	<.0001	11.38
TOC	Rear-end vs SVC	-0.5177	0.2066	6.28	0.0122	0.60
TOC	SSSD vs SVC	-1.2929	0.5071	6.5	0.0108	0.27
DUI	No vs Yes	-0.4058	0.0941	18.59	<.0001	0.67
Minor Injuries vs Property Damage Only						
Intercept		-0.2835	0.1298	4.77	0.0289	
RSUR	Dry vs Wet	0.2694	0.1123	5.76	0.0164	1.31
TOC	Head-on vs SVC	1.5407	0.3477	19.64	<.0001	4.67
TOC	Rear-end vs SVC	-0.6377	0.1387	21.14	<.0001	0.53
TOC	SSSD vs SVC	-0.7167	0.2532	8.01	0.0046	0.49
DUI	No vs Yes	-0.3438	0.0735	21.86	<.0001	0.71
Complain of Pain vs Property Damage Only						
ATM	Clear vs Rain	-0.2625	0.1233	4.53	0.0333	0.77
LGT	Dark vs Daylight	0.1981	0.0531	13.92	0.0002	1.22
RALI	Curve vs Straight	0.1866	0.0535	12.16	0.0005	1.21
TOC	Head-on vs SVC	0.9342	0.3711	6.34	0.0118	2.55
TOC	SSOD vs SVC	-0.6769	0.2081	10.58	0.0011	0.51

*shading indicates common factors in Tables 4 and 5 (includes similar increase/decrease); # excludes Madison, Lonoke, Carroll, and Baxter Counties

In this paper, the dependent variable for MLR is the crash severity level i.e., property damage only. Results from Table 4 indicate that during grade roadway profile compared to level grade, the fatal crashes were more likely to occur than property damage only crashes, and the odds ratio increased by a factor of 1.39 if other variables remained constant. Similarly, the relative risk of fatal crash increased by a factor of 98 compared to property damage crashes when type of collision was angled compared to single vehicle crashes (SVC) if other variables remained constant. From Table 4, for all cases i.e., from fatal to complain of pain compared to property damage crashes, the relative risk of a fatal crash increased when type of collision was head-on compared to SVC. For DUI, the odds of fatal crash to property damage crash decreased by a factor of 0.47. This indicates that DUI leads to more severe crashes.

Table 5 indicates that curved roadway alignment compared to straight, major injury crashes were more likely to occur than property damage only crashes, and the odds ratio increased by a factor of 1.35. Similarly, during dark conditions compared to daylight, the odds of complain of pain crashes to property damage crashes increased by a factor of 1.22. For all cases i.e., from fatal to complain of pain compared to property damage crashes the relative risk increased when the type of collision was head-on compared to SVC. To reduce the repetitive nature of these results, the results of second category based on CSI categorization are presented only.

Table 6 summarizes these results for second and third categories based on CSI categorization and each crash severity level. It presents the contributing factors common to both the county/group of counties with the highest CSI and the remaining counties within a single category, and shows similar trends in the odds ratio for both.

Table 7 summarizes these results for second, third and fourth categories based on CF categorization and each crash severity level. It presents the contributing factors common to both the county/group of counties with the highest CSI and the remaining counties within a single category, and shows similar trends in the odds ratio for both.

When all the categories were considered, in both categorizations (CSI and CF), it was observed that the occurrence of crashes lead to more severe crashes than non-severe crashes on rural two-lane undivided highways. This can also be observed from results in Tables 6 and 7 for selected categories. This indicates that on rural two-lane undivided highways a crash has a higher probability of being severe given that a crash has occurred. Also, in both categorizations (CSI and CF) the factor head-on collision type compared to SVC, and DUI was positively associated with higher severity of a crash. In both categorizations (CSI and CF), the contributing factors for complain of pain versus property damage only crashes were fewer in number compared to other severities. For counties categorized by CF and CSI, the rear-end collision and the sideswipe opposite direction crashes with respect to SVC increased the relative risk of severe injury crashes compared to property damage crashes given a crash has occurred.

Table 6. Summary results showing factors that contribute to crashes and are positively associated with severity for selected county/counties and category based on CSI categorization*

S. No (1)	**Description (2)**	**Fatal (3)**	**Major Injury (4)**	**Minor Injury (5)**	**Complain of Pain (6)**
I.	**Category 2**				
1	No. of contributing for factors Benton, Garland and White counties	4	7	3	2
2	No. of contributing factors Category 2 (excluding Benton, Garland, and White counties)	4	6	5	5
3	No. of contributing factors common to I.1 and I.2	2	4	3	2
4	Percentage of all crashes resulting from factors common to I.1 and I.2	50	67	60	40
5	Cumulative percentage of commonly identified factors	**55%** **((2+4+3+2)/((4+6+5+5))**			
II	**Category 3**				
1	No. of contributing for factors Craighead, Crawford, Union, Mississippi, Boone and Jefferson counties	3	7	4	1
2	No. of contributing factors Category 3 (excluding Craighead, Crawford, Union, Mississippi, Boone and Jefferson counties)	5	9	6	1
3	No. of contributing factors common to II.1 and II.2	3	6	3	0
4	Percentage of all crashes resulting from factors common to II.1 and II.2	60	67	50	0
5	Cumulative percentage of commonly identified factors	**57**			
	Consolidated Total percentage of common identified factors	**56**			

*Note: The first, fourth and fifth category results are not presented as there is no sample data for those categories

%computed as (summation of the I.3)/ (summation of the I.2)*100

Table 7. Summary results showing factors that contribute to crashes and are positively associated with severity for selected county/counties and category based on CF categorization*

S. No (1)	Description (2)	Fatal (3)	Major Injury (4)	Minor Injury (5)	Complain of Pain (6)
I	**Category 2**				
1	No. of contributing for factors White, Lonoke, Washington, and Pulaski counties	4	6	5	2
2	No. of contributing factors Category 2 (excluding White, Lonoke, Washington, and Pulaski counties)	4	7	4	3
3	No. of contributing factors common to I.1 and I.2	2	6	3	1
4	Percentage of all crashes resulting from factors common to I.1 and I.2	50	86	75	33
5	Cumulative percentage of commonly identified factors	**67**			
II	**Category 3**				
1	No. of contributing for factors Craighead, Faulkner, Poinsett, Madison and Mississippi counties	5	6	5	2
2	No. of contributing factors Category 3 (excluding Craighead, Faulkner, Poinsett, Madison and Mississippi counties)	4	4	4	1
3	No. of contributing factors common to II.1 and II.2	2	4	4	0
4	Percentage of all crashes resulting from factors common to II.1 and II.2	50	100	100	0
5	Cumulative percentage of commonly identified factors	**77**			
III	**Category 4**				
1	No. of contributing factors for Newton, Sevier, Union, Saline, Fulton, Polk, Montgomery, Randolph and Marion counties	3	6	4	4
2	No. of contributing factors Category 4 (excluding Newton, Sevier, Union, Saline, Fulton, Polk, Montgomery, Randolph and Marion counties)	4	7	5	0
3	No. of contributing factors common to III.1 and III.2	2	4	4	0
4	Percentage of all crashes resulting from factors common to III.1 and III.2	50	57	80	0
5	Cumulative percentage of commonly identified factors	**63**			
	Consolidated Total percentage of common identified factors	**68**			

*Note: The first and fifth category results are not presented as there is no sample data for those categories

When both the categorizations (CSI and CF) were compared, Tables 2 and 3, the sample data were 55% (61%) when categorized by CSI (CF); the sample data were 57% (57%) when categorized by CF (CSI).

Table 8. Summary Results

Categorization	Total Percentage of Sample Date used	Total Percentage of Contributing factors identified
CSI	55	56
CF	57	68

Based on the summary results from Table 8, higher percentage of factors can be identified when the categorization is based on crash frequency. However, it should be noted, though the counties were categorized by CF, the various levels of crash severity were considered for analysis. This indicates that 57% of the data were sufficient to determine 68% of the crash contributing factors. The analysis of 18 counties was sufficient to identify 68% of crash contributing factors for 50 counties, when the counties were categorized by CF. The major findings of this research are summarized below:

- On rural two-lane undivided highways, a crash will lead to a higher severity crash.
- Head-on collisions and DUI were associated with higher severity crashes on rural two-lane undivided highways. Also, 50% of the fatal crashes in Arkansas (2004-2006) occurred on these highways. Further, in Arkansas (2004-2006) 53% of the total head-on collisions which lead to fatal crashes occurred on rural two-lane undivided highways.
- Sample data can be used to identify more than 50% of the crash contributing factors when the counties are divided by CF or CSI. However, the sample data should satisfy the condition of minimum sample size of 2000 crashes to determine statistically reliable estimates when MLR is used. Use of sample data rather than total data can save computational time in identifying crash contributing factors.
- Crash contributing factors can be identified using other statistical models likes Poisson, negative binomial, etc. for each crash severity level. When MLR is used, the results can be interpreted with respect to property damage crashes. This indicates that researchers can identify factors that are positively associated with the severity of a crash in addition to crash frequency. Hence, for allocation of funds, the factors that are positively associated with the severity of a crash can be given higher priority rather than factors which positively impact crash frequency.
- It is advantageous to categorize the counties by crash frequency and later identify crash contributing factors based on crash severity. In this way, both crash frequency and crash severity are considered. However, this result is limited to the data set used in the current study. More data is required to generalize this statement.
- DOTs can allocate funds and concentrate on counties with higher crash frequency, rather than allocate equal importance to all counties. For specific details, each highway should be analyzed separately; however, the entire length of each highway need not be analyzed. For instance, S-7 (state highway) passes through Garland County and runs through a major part of the state of Arkansas (308 miles). Hence, priority should be given to those sections of S-7 in Garland County rather than its entire length.

CONCLUSIONS AND RECOMMENDATIONS

The main objective of this research was to identify crash contributing factors from a sample data set by dividing the state into different categories based on crash severity index (CSI) and crash frequency (CF). Percentage of factors identified by the sample data were compared to the remaining data using multinomial logistic regression (MLR). Results indicated that categorization by CF lead to 9% higher identification of crash contributing factors than categorization based on CSI.

The methodology used in this paper saves computational time i.e., data for 18 counties were sufficient to identify nearly 68% of the crash contributing factors for 50 counties. Factors contributing to crashes were identified using MLR. The odds ratio was used to identify the factors positively associated with crash severity. The use of MLR is recommended as the factors that are positively associated with crash severity can be identified. The allocation of funds by DOT can be prioritized based on the results of MLR i.e., factors can be identified that are positively associated with the severity of a crash rather than crash frequency.

Results indicated that head-on collisions led to more severe crashes when compared to other collision types on rural two-lane undivided highways. Also, DUI lead to more severe crashes. Several other variables like employment rate, education levels, etc. were not considered for this paper and are scope for future research. Three years of crash data (2004-2006) on rural two-lane undivided highways were used in this paper that comprised of 20,359 crashes. The proposed methodology should be tested with other data sets to evaluate its effectiveness. The use of different crash severity weights and its effects on categorization of counties and identification of crash contributing factors requires evaluation and is the scope of future research.

ACKNOWLEDGEMENTS

The authors are grateful to the Arkansas Highway and Transportation Department for providing data for this research and the anonymous reviewer for his/her comments.

REFERENCES

Abdel-Aty, M., and Abdelwahab, H. (2004). Modeling rear-end collisions including the role of drivers' visibility and light truck vehicle using a nested logit structure. Accident Analysis and Prevention, Vol. 36, No. 3, pp. 447-456.

Abdel-Aty, M., and Abdelwahab, H. (2009). Analysis and prediction of traffic fatalities resulting from angular collisions including the effect of vehicles' configuration and compatibility. Accident Analysis and Prevention, Vol. 36, No. 3, pp. 457-469.

Abdel-Aty, M., and Radwan, A, E. (2000). Modeling traffic accident occurrence and involvement. Accident Analysis and Prevention, Vol. 32, pp. 633-42.

ArcGIS 9.2 Desktop Help Online Manual. http://webhelp.esri.com/arcgisdesktop/9.2/index.cfm?TopicName=Standard_classification_schemes Accessed, June 23, 2010.

Bham, G. H., Javvadi, B. S., Manepalli, U. R. R. A Multinomial Logistic Regression Model for Single- and Multi-Vehicle Collisions. Journal paper in review.

Harwood, D.W., Council, F. M., Hauer, E., Hughes, W, E., and Vogt, A. (2000). Prediction of the expected safety performance of rural two-lane highways. Technical Report, FHWA-RD-99-207.

Highway Safety Manual, First Edition, 2010. Vol. 1 Section 4-66, Table 4A-1.

Jones, I. S., and Whitfield, R. A. (1988). Predicting injury risk with "new car assessment program crashworthiness ratings. Accident Analysis and Prevention, Vol. 20, No. 6, pp. 411-419.

Kim, D., Washington, S., and Oh, J. (2006). Modeling crash types: New insights into the effects of covariates on crashes at rural intersections. Journal of Transportation Engineering, Vol. 32, No. 4, pp. 282-292.

Krygier, J. B. (2010). http://go.owu.edu/~jbkrygie/krygier_html/geog_353/geog_353_lo/geog_353_lo07.html Accessed, June 23, 2010. J. B. Krygier: Geography lecture notes.

Lewis, D. J. (2010). http://danieljlewis.org/2010/06/07/jenks-natural-breaks-algorithm-in-python/ Accessed, June 25, 2010. University College London, Department of Geography.

Lui, K. J., McGee, D., Rhodes, P., and Pollock, D. (1988). An application of a conditional logistic regression to study the effects of safety belts, principal impact points, and car weights on riders' fatalities. Journal of Safety Research, Vol. 19, No. 4, pp. 197-203.

NHTSA. (20002). Traffic Safety Facts, 2001. Report no. HS-809-484. U. S. Department of Transportation, Washington, D. C.

NHTSA. (2003a). Traffic Safety Facts, 2001: Rural/urban Comparison. Report no. HS-809-524. U.S. Department of Transportation, Washington, D. C.

NHTSA. (2003b). Fatality Analysis Reporting System. U.S. Department of Transportation, Washington, D. C.

OCED. (1999). Safety Strategies for Rural Roads. Organization for Economic Co-operation and Development. Paris.

Pande, A., and Abdel-Aty, M. (2009). A Novel approach for analyzing severe crash patterns on multilane highways. Accident Analysis and Prevention, Vol. 41, No. 5, pp. 985-994.

Persaud, B., Retting, A. R., and Lyon, A. C. (2004). Crash reduction following installation of centerline rumble strips on rural two-lane roads. Accident Analysis and Prevention, Vol. 36, pp. 1073-1079.

Poch, M., and Mannering, F, L. (1996). Negative binomial analysis of intersection accident frequencies. Journal of Transportation Engineering, Vol. 122, pp.105-113.

SAS. http://support.sas.com/onlinedoc/912/docMainpage.jsp SAS online documentation manual 9.1.2. Accessed, December, 2009.

Savolainen, P, T., and Tarko, A, P. (2005). Safety Impacts at Intersections on Curved Segments. Transportation Research Record, 1908, Transportation Research Board of the National Academies, Washington, D.C., 130-140.

Shankar, V. N., Mannering, F. L., and Barfield, W. (1995). Effect of roadway geometrics and environmental conditions on rural accident frequencies. Accident Analysis and Prevention, Vol. 27, pp. 371-390.

Shankar, V. N., Mannering, F. L., and Barfield, W. (1995). Effect of roadway geometrics and environmental conditions on rural accident frequencies. Accident Analysis and Prevention, Vol. 27, No. 3, pp. 371-390.

Shankar, V., and Mannering, F. L. (1996). An exploratory multinomial logit analysis of single-vehicle motor cycle accident severity. Journal of Safety Research, Vol. 27, No. 3, pp. 183-194.

Xuedong, Y., Radwan, E., and Abdel-Aty, M. (2005). Characteristics of rear-end accidents at signalized intersections using multiple logistic regression model. Accident Analysis and Prevention, Vol. 37, No. 6, pp. 983-995.

Ye, F., and Lord, D. (2011). Comparing Three Commonly Used Crash Severity Models on Sample Size Requirements: Multinomial Logit, Ordered Probit and Mixed Logit Models. Proceedings of the 90th Annual Meeting of the Transportation Research Board, Washington, D. C.