CRASH PREDICTION ON RURAL ROADS

Cheng Zhong
Research Assistant, Civil, Construction and Environmental Engineering, School of Engineering,
University of Alabama at Birmingham, Birmingham AL, USA, e-mail: zhongch@uab.edu

Virginia P. Sisiopiku
Associate Professor, Civil, Construction and Environmental Engineering, School of Engineering,
University of Alabama at Birmingham, Birmingham AL, USA, e-mail: vsisiopi@uab.edu

Khaled Ksaibati
Professor, Department of Civil & Architectural Engineering, College of Engineering and
Applied Science, University of Wyoming, Laramie WY, USA, email: khaled@uwyo.edu

Tao Zhong
Transportation Analyst, Transportation Economics & Management Systems, Inc,
Frederick MD, USA, email: transtao@gmail.com

Submitted to the 3rd International Conference on Road Safety and Simulation,
September 14-16, 2011, Indianapolis, USA

ABSTRACT

Historical data confirm that rural roadways carry less than half of America's traffic but account for the majority of the nation's vehicular deaths. According to NHTSA, Wyoming has the highest crash fatality rate in the nation with a reported 2009 road death rate of 24.6 per 100,000 population, more than twice the national average of 11.0. High speed two-lane rural roads are believed to contribute to the fatal crash occurrence in rural states, such as Wyoming. An urgent need exists to systematically examine historical data to better understand contributing factors and develop countermeasures to improve traffic safety in rural settings.

The paper discusses the development of a methodology that utilizes available data from Wyoming (crash records, traffic volume, speed, etc) for crash prediction on rural roads. Prediction models were developed by using regression analysis techniques and data from three counties. Two methods were used in the building process, namely the Negative Binomial Regression (NBR) and the Poisson regression methods. The paper describes the process for selection of candidate roads, data collection and processing, methods employed in model development, and findings and conclusions. Overall, the analysis showed that the NBR method better fitted the over-dispersed crash data available in the study. The proposed model demonstrated that high speed, in conjunction with high volume result in higher crash rates (number of crashes per mile in this study) at high risk locations. The results from the case study can be used to classify rural road segments according to crash risk as well as provide the foundation for similar crash prediction analyses in other states in the future.

Keywords: Rural roads safety, Low volume roads, Crash prediction, Wyoming.

BACKGROUND

Compared to urban roads, rural roads are overall less safe. Historical data confirm that rural roadways carry less than half of America’s traffic but account for over half of the nation’s vehicular deaths (USDOT, 2008). For example, in the year of 2008, 23 percent of U.S. population lived in rural areas whereas rural fatalities account for 56 percent of all traffic fatalities. In 2008, nearly sixty two percent of passenger vehicle occupant fatalities occurred in rural areas (Insurance Institute for Highway Safety, 2008). The fatality rate per 100 million vehicle miles traveled (MVMT) on rural roads was 2.21 compared to urban areas at 0.88 (NHTSA, 2008).

Rural roads face many unique safety challenges that result in higher crash rates. First, roadway design and the presence of roadside hazards (such as utility poles, sharp-edged pavement drops-offs, and trees located close to roadways) create additional safety risks. Second, compared to urban crashes, rural crashes are more likely to be at higher speeds, a contributing factor to higher severity. Third, it often takes longer time for emergency vehicle response to the scene of a rural crash (TRIP, 2005) which affects the survival rate of those injured.

The “Safe, Accountable, Flexible, Efficient Transportation Equity Act: A Legacy for Users” (SAFETEA-LU), contains language indicating that State Departments of Transportation will be required to address safety problems on local and rural roads. The legislation states that it is important for state, county, and city officials to cooperate in producing a comprehensive safety plan to improve safety statewide. This legislation provides an opportunity to implement a more cohesive and comprehensive approach to address rural road safety problems (Evans et al, 2008).

This paper considered Wyoming as a case study. In Wyoming, between 2002 and 2006, the average rural MVMT was 6,654 and the average fatality rate per 100 MVMT was 2.23 (USDOT, 2008). Nearly eighty-six percent of passenger vehicle occupant fatalities occurred in rural areas of Wyoming (NHTSA, 2008), much higher than the national average of sixty-two percent. Due to the high percentage (nearly 70%) of rural population (USDA, 2011) and the extend of the rural roadway network, Wyoming has the highest crash fatality rate in the nation with a reported NHTSA 2009 road death rate of 24.6 per 100,000 population. This is more than twice the national average of 11.0. To address the issue of rural traffic safety, the state of Wyoming initiated a program called Wyoming Rural Road Safety Program (WRRSP) which aimed at helping counties to identify high risk rural locations and develop a strategy to obtain funding for improving safety in the top-ranked locations. One objective of this initiative was to develop a methodology of using available data (crash records, traffic volume, speed, etc) for crash prediction on rural roads.

LITERATURE REVIEW

Crash prediction models offer an estimate of expected accident frequency as a function of traffic flow characteristics and roadway geometries. Regression equations that relate crash experience to traffic and other geometric conditions are widely used in modern highway safety analysis (NCHRP, 2001). Extensive research had been performed to examine the relationship between vehicle crashes and traffic flow features (e.g. traffic volume, speed) or geometric designs (e.g.

lane width, shoulder width). In previous safety studies, linear regression, Poisson regression and Negative Binomial Regression (NBR) were three techniques used to develop regression models (Wang, 2008).

Previous safety studies such as Miaou et al (1993) used multiple linear regression techniques to study the relationships between vehicle accident and geometric features. Japanese researchers (Okamoto, 1989) tried to use multiple linear regression to analyze accident rates related to geometric design elements. They found that linear regression was not suitable to model vehicle accidents. The underlying assumption of linear regression is that events follow a normal distribution. Therefore, the linear model may predict a negative value. However, in real life, traffic crash data are always discrete and regarded as a random variable that takes non-negative integer values. These characteristics imply that crash data may follow more closely the Poisson distribution, instead.

Miaou and Lum tried using the Poisson regression to model truck accident data (Miaou et al, 1992). From their analysis they found that truck accidents were strongly related to traffic volume and the roadway geometric factors, such as vertical grade and horizontal curvature. Poisson regression was used to analyze traffic count data. This technique can be used to model the number of occurrences (or the rate) of an event of interest, as a function of some independent variables. In Poisson regression, it is assumed that the dependent variable Y that corresponds to the number of occurrence of an event (number of crashes per mile in this study), has a Poisson distribution given the independent variables $X_1, X_2, \ldots.., X_i$. The general form of the Poisson regression is as following:

$$f(Y) = \frac{\mu^{Y} exp\,(-\mu)}{Y!} \qquad (1)$$

Where: $f(Y)$ is the probability that the outcome is Y, and

In exponential form, equation 1 can be rewritten as:

$$\mu_i = exp\,(\beta_0 + \sum_{j=1}^{n} X_i\beta_j) \qquad (2)$$

Where: μ_i is the expected crash per mile on road *i*

$X_1, X_2 \ldots.. X_i$ are the values of the roadway variables (traffic volume, speed, etc) on road *i*

$\beta_{1,} \ldots. \beta_j$ are the coefficients to be estimated by modeling.

The expected crash rate is the number of crashes adjusted for intensity and it is assumed to be an exponential value applied to a suitable combination of roadway variables. Thus, the model falls under the heading of a Generalized Linear Model (GLM). The exponential function guarantees that the mean (the number of expected crashes) is non-negative. The most widely accepted way to estimate the parameters β_j is to use a Maximum Likelihood Estimation (MLE) procedure. The likelihood function can be written as:

$$L(\bar{\beta}) = \prod_{i=1}^{n} f_i\,(Y_i) = \prod \frac{[\mu(X_i,\beta)]^{Y_i}\, exp\,[-\mu(X_i,\beta)]}{Y_i!} \qquad (3)$$

Where: $\mu(X_i, \beta)$ is the function which relates μ_i to X_i.

Miaou and Lum (Miaou et al, 1993) also pointed out the limitations of using the Poisson Regression approach. The Poisson distribution's fundamental assumption is that the variance should be equal to its mean. However, real crash data rarely comply with this assumption. In most cases, the variance is larger than its mean. This phenomenon causes what is called over-dispersion. The consequence of the over-dispersion is that the variances of the estimated parameters tend to be underestimated. In other words, the estimated β from MLE under the Poisson regression model is still close to the true parameter, but the significance levels of the estimated parameters may be overstated.

In dealing with the over-dispersion in crash data, NBR, an alternative to Poisson regression, has been used in accident modeling. In 1995, Shankar (Shankar, 1995) tried to use the NBR to overcome the over-dispersion problem. He used both Poisson regression and NBR to model the effects of road geometry and environmental factors on the number of crashes. He found that NBR modeled the crash data better than Poisson regression when the crash data were over-dispersed. Caliendo (Caliendo et al, 2007) used both Poisson regression and NBR to examine the relationship between geometric features and accident frequency on multilane roadways in Italy. They found that Poisson regression was inappropriate to model the random variation of the number of crashes if there was clear evidence that over-dispersion was present.

NBR generalizes the Poisson regression by permitting the variance to be over-dispersed. In the NBR model, the variance equals to the mean plus a quadratic term in the mean whose coefficient is called the over-dispersion parameter α (Equation 4).

$$Var\ [Y_i] = E\ [Y_i][1 + \alpha E[Y_i] = E\ [Y_i] + \alpha E[Y_i]^2 \tag{4}$$

Where: a= over-dispersion parameter.

The selection between the two models, i.e., Poisson regression or NBR, depends on the value of α. When this parameter is equal or close to zero, a Poisson model is appropriate. When it is larger than zero, it represents the variance above and beyond the mean. The over-dispersion phenomenon is commonly due to the variation of the highway variables present in the model, such as accident-related factors pertaining to drivers, vehicles, and location not encompassed by the highway variables (Miaou et al, 1993). For the NBR model, the expected accident frequency for a section i is written as:

$$\mu_i = exp\left(\beta_0 + \sum_{j=1}^{n} X_i \beta_j\right) \tag{5}$$

Where: μ_i = $EY_i|X_i\ for\ Y_i|X_i$ distributed as a negative random binominal variable.

One of the forms of NBR distribution can be written as:

$$f(Y) = \frac{\Gamma\left(\frac{1}{\alpha}+Y_i\right)}{\Gamma(Y_i+1)\Gamma\left(\frac{1}{\alpha}\right)}\left(\frac{\frac{1}{\alpha}}{\frac{1}{\alpha}+\mu_i}\right)\left(\frac{\mu_i}{\frac{1}{\alpha}+\mu_i}\right)^{Y_i} \tag{6}$$

Where: Γ is a gamma function.

From the literature review, it can be found that Poisson regression or NBR are suitable candidate options to model crash data. Therefore, this study focused on applying these two methods in developing crash prediction models using data from rural roads in Wyoming.

CASE STUDY: CRASH PREDICTION ON RURAL ROADS IN WYOMING

Candidate Roads Selection and Crash Data

In order to develop a crash prediction model for low volume rural roads in Wyoming, thirty six rural roads were considered for inclusion in the evaluation from three Wyoming counties, namely Laramie, Carbon, and Johnson. All study roads included in developing the prediction model were identified by the WRRSP as high risk roads. The reported crash records over the 1995 to 2005 time period were obtained from the Wyoming Department of Transportation (WYDOT). This dataset contains all types of crashes that occurred on all roadway classifications. Since, this project focused on rural roads, only the crashes that occurred on rural county roads were included in the analysis. The crash records from WYDOT contain various attributes for every crash, including accident route number and name, accident mile point, accident year, number of vehicles involved in the accident, number of injuries and fatalities in the accident, accident severity, light condition, weather conditions and road surface types. In this study, the key attribute retrieved from the crash records for modeling was the total number of all severity levels of crashes that occurred during the ten year period. Table 1 summarizes the crashes on all the roads included in this experiment.

Traffic Counts and Speeds

One of the objectives of this safety study was to determine the correlation between traffic volume and speed and the number of crashes. Therefore, traffic volume and the 85th percentile speed data were considered as key factors in developing the crash prediction model. Unfortunately, Wyoming local government did not collect traffic data on these roads on a routine basis. Therefore, traffic data on all the candidate roads were collected by the research team. The traffic counter locations were determined mainly based on the risk locations identified from the crash analysis. Another consideration was the existence of major intersections which may result in changing traffic volumes. As an example, if a rural road stretches a very long distance and intersects with higher functional class of roads, it is very likely that the intersection areas will have high traffic volume. Two or more automatic traffic counters were installed at these locations. When developing the prediction model, the traffic data collected from the highest traffic volume spots were used.

Automatic traffic counters were used to collect traffic data for this study including traffic volume, speed and vehicle classification data. The specific type involved in the study is “TRAX RD”, which is manufactured by JAMAR Technology Inc. Properly installed traffic counters can collect TRAX RD employs two road tubes to record the traffic data. The tubes connected with TRAX RD were placed perpendicular to the flow of the traffic and set to 8 feet apart. When vehicles crossed over the road tubes, air impulses were generated to trigger the two air-impulse switches inside the traffic counter.

Table 1 Summary of Crash Data
Source: Crash Data of 1995-2005 from WYDOT

County	Road Number	Road Length (miles)	Property Damage Only (PDO)	Injury	Fatal	Total Crashes	Crashes per Mile
Carbon	385	16.25	1	6	0	7	0.431
Carbon	291	57.43	25	14	3	42	0.731
Carbon	603	3.67	3	0	0	3	0.817
Carbon	702	7.32	7	0	0	7	0.956
Carbon	353	6.6	2	1	0	3	0.455
Carbon	550	1.48	1	0	0	1	0.676
Carbon	203	7.62	5	1	0	6	0.787
Carbon	660	14.52	5	4	0	9	0.620
Carbon	500	23.94	10	5	1	16	0.668
Carbon	561	8.13	5	3	0	8	0.984
Carbon	504	16.05	4	11	0	15	0.935
Carbon	324	5.17	6	2	0	8	1.547
Carbon	401	34.53	25	12	2	39	1.129
Carbon	710	3.09	4	0	0	4	1.294
Carbon	701	19.13	4	4	0	8	0.418
Carbon	700	17.2	3	5	0	8	0.465
Laramie	210	10.8	11	19	0	30	2.778
Laramie	109	9.48	13	12	1	26	2.743
Laramie	136	8.23	5	6	0	11	1.337
Laramie	143-2	28.38	10	6	2	18	0.634
Laramie	212-1	4.11	4	5	0	9	2.190
Laramie	102-1	7.32	7	8	0	15	2.049
Laramie	120-1	22.73	14	8	1	23	1.012
Laramie	124	10.84	9	8	0	17	1.568
Laramie	215	18.47	17	24	1	42	2.274
Laramie	209	7.33	10	6	0	16	2.183
Laramie	203-1	36.8	14	16	0	30	0.815
Laramie	164-1	12.26	4	5	0	9	0.734
Laramie	162-2	10.95	15	13	1	29	2.648
Laramie	A149-1	0.69	4	0	0	4	5.797
Johnson	212	1.6	2	1	0	3	1.875
Johnson	14	8.49	4	2	0	6	0.707
Johnson	91H	12.2	19	6	0	25	2.049
Johnson	3	32.7	8	1	0	9	0.275
Johnson	132	12.94	7	0	0	7	0.541
Johnson	40	8.32	5	3	0	8	0.962
Johnson	85	5.9	4	1	0	5	0.847
Johnson	256	1.69	4	4	0	8	4.734

Various tube layouts could be selected to record different traffic flow patterns. In this safety study, the selected tube layout is shown in Figure 1. In this layout, the traffic data were recorded separately in each direction.

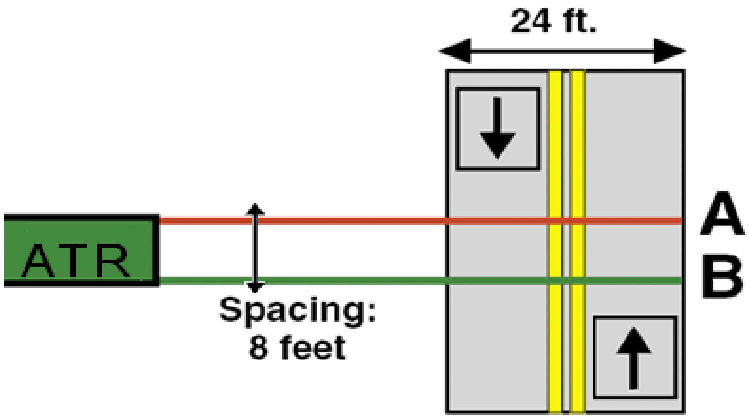

Figure 1 Tube Layout for Collecting Traffic Data
Source: JAMAR Technology, Trax RD Manual

TRAX RD is solar powered and its battery can last more than one week. In this study, traffic counters were installed for approximately one week to collect the weekday and weekend traffic data at each data collection site. The simple axle vehicle classification scheme was used to classify vehicles. Any type of vehicle that has more than or equal to three axles was categorized as a truck. Table 2 shows an example of the traffic data collected on each section.

Table 2 Traffic Data on County Road 324

	Volume		Vehicle Classification				85th Percentile Speed, MPH	
	Direction 1	Direction 2	Direction 1		Direction 2		Direction 1	Direction 2
	Cars &Trucks	Cars &Trucks	Cars	Trucks	Cars	Trucks	Cars &Trucks	Cars &Trucks
Wed 7/11/2007	90	91	89	1	91	0	61	60
Thu 7/12/2007	83	82	78	5	80	2	63	61
Fri 7/13/2007	98	96	97	1	94	2	62	62
Sat 7/14/2007	168	172	166	2	170	2	57	59
Sun 7/15/2007	99	96	99	0	96	0	59	61
Mon 7/16/2007	70	67	67	3	65	2	59	58
Tue 7/17/2007	75	75	74	1	75	0	60	59
Average	98	97	96	2	96	1	60	60
	Directional Distribution (%)		Percent of Vehicles (%)					
	47	53	98	2	99	1		

The collected traffic data indicated that truck volumes account for only a small percentage. Therefore, it is not necessary to consider truck volumes separately and thus combined average daily traffic (ADTs) were used in this study. The traffic counters recorded traffic volume separately for each direction. Traffic volume used in this study was the sum of both directions of daily average over the traffic counter duration period (approximately one week). The daily 85th percentile speed was obtained from TRAX RD software after processing the data collected by the traffic counter. Similar to the traffic volume, the 85th percentile speed used for this study was the average of the daily 85th percentile speed of the traffic counter duration period.

In Table 3, surface type indicates on which type of road surface the traffic counter was installed. It was defined as a categorical variable. As seen from Table 3, “0” indicates that the traffic counter was installed on gravel or dirt surface, while “1” indicates an asphalt surface.

Table 3 Summary of Traffic Data

County	Road Number	Road Length (miles)	Surface Type	Volume (ADT)	Speed (mph)
Carbon	385	16.25	0	37	49.5
Carbon	291	57.43	0	35	47.5
Carbon	603	3.67	0	200	50.5
Carbon	702	7.32	0	48	38
Carbon	353	6.6	0	99	29.5
Carbon	550	1.48	0	247	47
Carbon	203	7.62	0	161	35.5
Carbon	660	14.52	0	112	48
Carbon	500	23.94	0	293	44.5
Carbon	561	8.13	0	192	33.5
Carbon	504	16.05	1	218	62.5
Carbon	324	5.17	1	195	60
Carbon	401	34.53	1	324	66.5
Carbon	710	3.09	1	112	47
Carbon	701	19.13	0	722	51.5
Carbon	700	17.2	1	164	49
Laramie	210	10.8	0	173	42
Laramie	109	9.48	0	357	46
Laramie	136	8.23	0	238	46.2
Laramie	143-2	28.38	0	308	51.5
Laramie	212-1	4.11	0	46	55.5
Laramie	102-1	7.32	0	138	52
Laramie	120-1	22.73	0	256	42.8
Laramie	124	10.84	1	747	51.1
Laramie	215	18.47	1	395	56.5
Laramie	209	7.33	1	898	52.2
Laramie	203-1	36.8	1	156	68.5
Laramie	164-1	12.26	1	200	61.3
Laramie	162-2	10.95	1	160	68
Laramie	A149-1	0.69	1	373	68.5
Johnson	212	1.6	1	583	36.5
Johnson	14	8.49	0	174	44.5
Johnson	91H	12.2	1	1468	51.3
Johnson	3	32.7	1	125	39.4
Johnson	132	12.94	1	253	52.9
Johnson	40	8.32	0	229	33
Johnson	85	5.9	0	350	31.3
Johnson	256	1.69	1	510	42.7

Difficulties of Installing Traffic Counters on Gravel and Dirt Roads

A significant portion of the rural roads in this study were gravel or dirt roads. This added to the difficulty of installing traffic counters. The major problem was fixing the road tubes on the road surface. There are no traffic counters specifically designed to collect traffic data on gravel or dirt roads and experience shows that road tubes work well on paved roads but not so on gravel or dirt roads. The rubber tubes need special treatment before installation. Otherwise, it is very likely that the tubes could be pierced by sharp gravel. If the tubes leak, they cannot generate accurate air impulses to the counter.

One method of protecting the tubes is enclosing the rubber tube inside a cover such as a fire hose. However, this causes another problem of being able to fix the tubes on the ground. Without any cover, the tubes can be easily fixed by metal clamps on asphalt. But a tube inside a fire hose is difficult to be fixed. Sometimes, the tubes are displaced from their original installed position. In order to calculate the speeds of the vehicles, the traffic counter needs the precise time stamp (generated by the air impulse) with an accurate distance of the two tubes. Tubes' displacement changes the distance between the two tubes. As a result, the traffic counter will not get the accurate vehicle classification and speed data. For this reason, the speed data from some roads may be unavailable or inaccurate. However, from the collected traffic data, it was found that at most locations, the daily traffic volumes and speeds were consistent and the variation could be neglected. Moreover, the inaccurate data due to the displacement of the tubes were deleted. At these locations, two or three days data were used to calculated ADT and 85th percentile speed.

Data Analysis and Prediction Model Developing

Traffic data from the three counties were combined in one dataset for developing the crash prediction model. The dataset consisted of a total of 38 records. Table 3 summarizes the traffic and surface type data. It was clear from the traffic data collected in this study that the measured 85th percentile speeds were significantly higher than the posted speed limits (up to 15 MPH).

Outlier Identification

Outliers are extreme observations in the dataset. They may stem from errors in data collection or miscalculation. The negative binominal regression method uses the maximum likelihood method to estimate the predictor variables' coefficients. As a result, outliers may lead to serious distortions in the estimated regression function (Kutner, 2003). During the model development process, two outliers were identified. One outlier was the County Road 701 in Carbon County, and the other was County Road A149 in Laramie County. County Road 701 has a relatively high traffic volume but a very low crash rate. It is very likely that new developments around this road have occurred in recent years, which resulted in increasing traffic flow. However, the recent high traffic volume has not yet been translated into high crash rates. County Road A149 is a unique section. It is very short, less than one mile. The crash records indicate that only four PDO crashes occurred on this road in the ten-year analysis period. This extremely short length was behind another section with abnormally high crash rate. Due to the reasons explained above, these two observations were discarded from the dataset, which resulted in 36 records remaining in the final dataset for modeling.

Crash Prediction Model Development

As stated in the literature, previous safety studies had used geometric factors such as, lane width, shoulder width, horizontal and vertical distance as the predictor variables in the prediction model. However, such information was not available for this safety study. More importantly, the developed crash prediction model needed to be simple and practical enough to be used by the local governments. From the roadway classification survey, traffic volume and traffic speed collection were common in studies conducted by counties. Therefore, traffic volume, traffic speed, road surface type, and an interaction variable (the product of traffic volume and speed) were used as the predictor variables in modeling. Crash rate (number of crashes per mile) was the response variable in the model. In this study, the statistical analysis software, SAS (proc genmod), was used for modeling.

As stated before, one interest of this study was to evaluate the combined and individual effects of traffic volume and speed on crash rates of rural roads. Therefore, various combinations of the predictor variables were tested in modeling. The basic process was as follows:

1. Put one predictor variable alone in the model and use SAS to run this model.
2. Add the surface type into the model while keeping the predictor variable and rerun the model to see if there is any interaction between the predictor variable and surface type.

Similar steps were performed on traffic volume and traffic speed. Finally, traffic volume and speed were analyzed in the model simultaneously.

When using different combinations of the predictor variables to develop a crash prediction model, Poisson regression and NBR were evaluated separately. Table 4 and Table 5 summarize these results. The estimated coefficients of the predictor variables are summarized in the estimate column. The p-values of the predictor variables reflect the goodness of fit. Simply speaking, the p-value indicates a predictor variable's probability of being associated with the response as strongly as is seen in the observed data set. In other words, small p-values indicate that a predictor variable should probably be included in the model. The usual convention for p-value is to be smaller than 0.05 (95% significance level) to keep a predictor variable in the model.

Goodness of Fit

The standard Poisson regression and NBR are both forms of GLM (Dobson et al, 2008). In the generalized linear model, one of the goodness of fit criteria, namely deviance,

Table 4 Using Poisson Regression to Fit the Crash Data

Model Number	Predictor Variables	Estimate	P-Value	Goodness of Fit		
				Deviance	Degree of Freedom (DF)	Deviance/DF
1	Volume*Speed	15.8596	<.0001	157.0424	34	4.6189
2	Volume*Speed Surface	16.5071 -0.0519	<.0001 0.5981	156.7640	33	4.7504
3	Speed	0.0117	0.0061	184.4524	34	5.4251
4	Speed Surface	0.0105 0.0407	0.0528 0.7150	184.3195	33	5.5854
5	Volume	0.0001	<.0001	158.5255	34	4.6625
6	Volume Surface	0.0008 0.0018	<.0001 0.9853	158.5251	33	4.8038
7	Volume Speed	0.0008 0.0105	<.0001 0.0164	152.8154	33	4.6308

Table 5 Using Negative Binominal Regression to Fit the Crash Data

Model Number	Predictor Variables	Estimate	P-Value	Goodness of Fit			
				Deviance	Degree of Freedom (DF)	Deviance/ DF	Log Likelihood
1	Volume*Speed	16.0736	0.0267	36.3341	34	1.0686	975.8060
2	Volume*Speed	30.2164	0.3093	36.3908	32	1.1372	975.9298
	Surface	0.1381	0.7064				
	Volume*Speed* Surface	-15.2914	0.6200				
3	Speed	0.0122	0.2522	36.7000	34	1.0794	973.7859
4	Speed	0.0196	0.3413	35.2631	32	1.1020	974.3200
	Surface	1.2329	0.4108				
	Speed* Surface	-0.0218	0.4579				
5	Volume	0.0008	0.0267	36.1447	34	1.0631	975.8185
6	Volume	0.0011	0.4164	36.1312	32	1.1291	975.8663
	Surface	0.1123	0.7572				
	Volume*Surface	-0.0003	0.8162				
7	Volume	0.0008	0.0286	36.0422	33	1.0922	976.4679
	Speed	0.0111	0.2540				

has an approximate chi-square distribution with n-p degrees of freedom, where n is the number of the observations and p is the number of predictor variables (including the intercept). The expected value of a chi-square random variable is equal to the degrees of freedom (df). If the model fits the data well, the ratio of the deviance to df should be close to one. If this ratio is significantly larger than one, it may indicate that the model fails to account for the data's variability.

Based on the examination of the Poisson regression results summarized in Table 4, it was found that the crash data is over-dispersed (i.e., the ratio of the deviance/df is significantly larger than 1). Thus using Poisson regression, the independent variables may seem significant in the model (with p-value smaller than 0.05), however, the results may be misleading due to the over-dispersion. Standard errors of the estimated coefficients are incorrectly estimated, implying an invalid chi-square test (UCLA, 2007). In contrast, Table 5 shows that the NBR fits the data reasonably well (i.e., the ratio of the deviance/df is very close to 1). Therefore, in this study, NBR was selected as the best option for modeling.

Interpretations of the Results

It is clear from Table 5 that when the interaction variable (the product of volume and speed) is analyzed in the model alone, it was found to be significant. However, if the interaction variable and the surface type were both in the model, none of them were significant. As an example, in Model 2, "Volume*Speed", "Surface", and "Volume*Speed*Surface" were all in the model, but according to their p-values, none of them were significant in the model. This suggests that there was no interaction between the interaction variable and the surface type. Similar observations can be made for the traffic volume and speed variables.

From another aspect, the speed variable alone in the model was found to be statistically insignificant. However, when it was combined with traffic volume as the interaction variable and added in the model, it became significant. This implies that on the analyzed rural roads in Wyoming, traffic speed has a significant effect on road safety but its effect is masked unless it is combined with higher traffic volume.

From Table 5, it can be found that the Model 1 and Model 5 have very close Deviance/df and log likelihood values. NBR is one of GLMs. A common comparator of GLM that accounts for model complexity is the Akaike Information Criterion (AIC). Simply stated, smaller AIC value of a model generally means this model is better than the other. It is expressed as:

$$AIC = -2*Log\ likelihood + 2*k \qquad (7)$$

Where: k is the number of parameters in the model.

For example, from Table 5, the AIC value for Model 1 that includes the "Volume*Speed" predictor is -2*975.8060+ 2*2 =-1947.612. The AIC value for Model 5 that includes the "Volume" predictor is -2*975.8185+2*2= -1947.637. From the AIC value, there is no clear superiority when comparing Models 1 and 5. Therefore, both Model 1 and Model 5 are proposed based on the NB regression analysis. The total number of crashes will occur in ten years are:

$$Total\ crash= exp\ (-0.0340+16.0736*\ Volume*Speed\ /1{,}000{,}000)*\ Road\ Length \qquad (8)$$

$$Total\ crash= exp\ (-0.0428+0.0008*\ Volume)*\ Road\ Length \qquad (9)$$

Where: *exp* is the exponential function

Road length is the length of the analyzed road, and

Constants -0.0340 and -0.0428 are estimates of constant β_0 in Equation (5)

Another concern of the model's goodness fit is the Proportionate Reduction in Variation (PRV) and it is usually evaluated by the value R^2. It measures the proportionate reduction of total variation in response variable associated with the use of the set of predictor variables (Kutner, 2003). In ordinarily least square (OLS) regression, R^2 takes the value between 0 and 1. Larger R^2 indicates that the model can explain more observed variability. In GLM, no such equivalent R^2 exists. In the GLM, the coefficients of the predictor variables are estimated from the maximum likelihood procedure (UCLA, 2007). Therefore, unlike the OLS regression, the coefficients are not calculated to minimize variance. However, to evaluate the goodness of fit of the GLM, several pseudo-R^2 were proposed. Although all pseudo-R^2 measures are imperfect, they still help describe PRV in a general way. One pseudo-R^2 proposed by Cox & Snell (COX et al, 1989) is expressed as following:

$$R^2 = 1 - exp\ [-\tfrac{2}{n}\{l(\hat{\beta}) - l(0)\}] \qquad (10)$$

Where: $l(\hat{\beta})$ is the log likelihood of the fitted model,

$l(0)$ is the log likelihood of the null model, and

n is the sample size

For Model 1, the log likelihood of the null model is 973.1323. The pseudo-R^2 of the fitted model is $1\text{-}exp[-\frac{2}{36}\{975.8060\text{-}973.1323\}]$= 0.138. This means that the model can explain the 13.8% of the observed variability. Using the same equation, the pseudo-R^2 of Model 5 is 0.1386. The relatively low pseudo-R^2 may result from two respects, namely number of prediction variables and sample size. Introducing other prediction variables such as geometric features (road width, shoulder width) to the model may be helpful in improving the predictability of the model. It should be kept in mind that the objective of this safety project was to help counties in Wyoming to identify high risk locations. Therefore, the developed model was not meant for predicting the precise number of crashes but rather be used to evaluate if a road is potentially high risk. Meanwhile, a simplified model will be easier to be used by counties. Relatively small sample size may also have effects on pseudo-R^2 value. If more comprehensive and complete data could be obtained from a future study, the predictability of the model is expected to improve.

This regression model in this study was developed based on the crash and traffic data from the roads, selected by the WRRSP. These roads have the highest crash rates (number of crashes per mile) among the county rural roads in the three counties included in the pilot study. The developed model was successful in providing counties with a useful and practical tool to determine if a specific road has a higher than normal crash rate. As an example, if a road in a county has actual 7 crashes in a ten-year period and the model predicts 15 crashes based on the prevailing traffic condition, then this road should not be considered as a high risk road. However, if a road has 20 actual crashes and the model predicts only 15 crashes, then this road should be considered as a high risk road.

CONCLUSIONS AND RECOMMENDATIONS

The NBR and the Poisson regression methods were both examined in the study. The NBR was found to be superior to the Poisson regression in fitting the overdispersed Wyoming crash data. The p-value of the surface type in the model was not found significant when interaction with other traffic variables took place. The type of road surface type (gravel vs. paved) showed statistically similar crash rates in the dataset analyzed in this study. According to the regression model findings, high speed by itself did not significantly correlate with high crash rates. However, high traffic volume in conjunction with high speed resulted in higher crash rates. It should be noted that the prediction model is recommended to be used to determine if a specific rural road should be considered as high risk.

The dataset used for developing the prediction model contained only 36 effective observations. The absence of adequate traffic data on Wyoming rural roads made it difficult to increase the sample size. The relatively small size of the dataset may have reduced the predictability of the model. It is recommended that local government and state DOTs should focus on collecting traffic data on rural roads in a more systematic way. The availability of such data should help in confirming and refining the prediction model developed in this study in the future.

REFERENCES

Caliendo, Ciro and Guida, Maurizio (2007). “A Crash-Prediction Model for Multilane Roads”. *Accident Analysis and Prevention,* Vol. 39.

Cox, D.R. and E. J. Snell (1989). “Analysis of Binary Data (2nd edition)”. London: Chapman & Hall.

Dobson, AJ and Pavneh, AG (2008). "An Introduction to the Generalized Linear Models, 3rd Edition". Chapman & Hall/CRC.

Evans, B. and Ksaibati, K. (2008). "Carbon County Rural Road Safety Evaluation Program". WYT^2 Center Report.

Insurance Institute for Highway safety (2008). "Fatality Facts 2008" http://www.iihs.org/research/fatality_facts_2008/statebystate.html

Kutner, Michael H (2003). "Applied Liner Regression Models, 4th Edition". McGraw-Hill Irwin.

Miaou and Wright (1992). "Relationships between Truck Accidents and Highway Geometric Design: A Poisson Regression Approach".

Miaou, Shaw-Pin and Harry, Lum (1993). "Modeling Vehicle Accidents and Highway Geometric Design Relationships" *Accident Analysis and Prevention* Vol. 25.

National cooperative Highway Research Program (NCHRP) (2001). "Statistical Methods in Highway Safety Analysis". Synthesis 295.

National Highway Traffic Safety Administration (NHTSA) (2008). "Traffic Safety Facts 2008 Data: Rural/Urban Comparison"
http://www-nrd.nhtsa.dot.gov/Pubs/811164.pdf

Okamoto, H. and Koshi, M. (1989). "A Method to Cope with Random Errors of Observed Accidents Rates in Regression Analysis". *Accident Analysis Prevention*, Vol. 21.

Shankar (1995). "Effect of Roadway Geometrics and Environmental Factors on Rural Freeway Accident Frequencies". *Accident Prevention and Analysis*, Vol. 27.

The Road Information Program (TRIP) (2005). "Safety, Mobility and Economic Challenges in America's Heartland".

UCLA: Academic Technology Services, Statistical Consulting Group (2007). http://www.ats.ucla.edu/stat/mult_pkg/faq/general/Psuedo_RSquareds.htm

U.S. Department of Transportation (DOT) (2008). "The U.S. Department of Transportation Rural Safety Initiative". http://www.dot.gov/affairs/ruralsafety/ruralsafetyinitiativeplan.htm

U.S. Department of Agriculture (USDA) (2011). "Economic Research Service: State Fact Sheets: Wyoming". http://www.ers.usda.gov/StateFacts/wy.HTM

Wang, Yinhai (2008). "Cost Effective Safety Improvement for Two-Lane Rural Roads". TNW2008-04.