# Hot Spot Identification Analysis for Utah Roadways Using Spatial Poisson Linear Mixed Model

Christian Davis

April 18, 2019

**Abstract**

In recent years, the Utah Department of Transportation (UDOT) has not only aimed to diminish, but eradicate fatalities caused by motorized vehicle crashes. As a result, many studies have been conducted to understand the relationship of important road and driving characteristics with the likelihood of a crash, yet few incorporate spatial dependence. In this project, we first strive to more accurately understand the effect of roadway and driving features on the number of crashes. Second, we incorporate a spatial random effect to improve predictions and better quantify the uncertainty associated with the effects of road characteristics. Finally, we use the estimates of the spatial random effects to identify road segments that seem to be more dangerous than would be expected by only accounting for the effects of road and driving characteristics.

## 1 Introduction

In 2006, the Utah Department of Transportation (UDOT) launched the "Zero Fatalities: A Goal We Can All Live With" campaign assertively aiming to eliminate the number of deaths occurring on Utah roads. Since the campaign began, the number of fatal crashes have decreased, on average. However, a staggering number of motorized vehicle crashes resulting in fatalities are still occurring. In 2016, there were a total of 62,471 total crashes which resulted in 281 fatalities across Utah roads. As part of UDOT's effort to increase roadway safety, UDOT collects and maintains several databases. In this way, statistical analysis of roadway and crash data is used as a tool to determine effects (if any) of roadway characteristics (e.g., median width, number of lanes) on the number of crashes occurring along the road. By quantifying such effects, UDOT is able to identify countermeasures that increase overall roadway safety.

In this paper, we seek to facilitate and understand roadway safety in three ways. First, we aim to understand the relationship of important road and driving characteristics with the likelihood of a crash using a fully Bayesian approach. Second, we seek to incorporate spatial dependence to improve predictions for car crashes across Utah roadways and better quantify the uncertainty associated with the effects of road characteristics. Finally, we aim to identify road segments that appear to be more dangerous (so-called "hot spots") than would be expected given the road characteristics.

## 2 UDOT Crash and Roadway Data

The motivating dataset for this project was obtained from DI-9 reports, which are completed by police officers for crashes that result in death, injury, or property damage over $1,500, and is maintained by the UDOT safety division. The data consist of 68 different road and driving

characteristics for approximately 3,000 crashes across over 200 Utah state roadways from 2010-2016. Each roadway is divided into road segments of unequal length where the road and driving characteristics for a particular segment are recorded. Some of these 68 characteristics include, vehicles miles travelled (VMT), speed limit, number of lanes, and total percent trucks along with the total number of crashes for a given segment. Note that VMT is a measure of traffic relative to the length of the segment.

Figure 1 contains a scatter plot matrix for these five variables plotted against one another for all roads. VMT and number of lanes appear to have a moderate positive linear relationship with total crashes while speed limit and total percent trucks appear to have a weak negative relationship with total crashes.

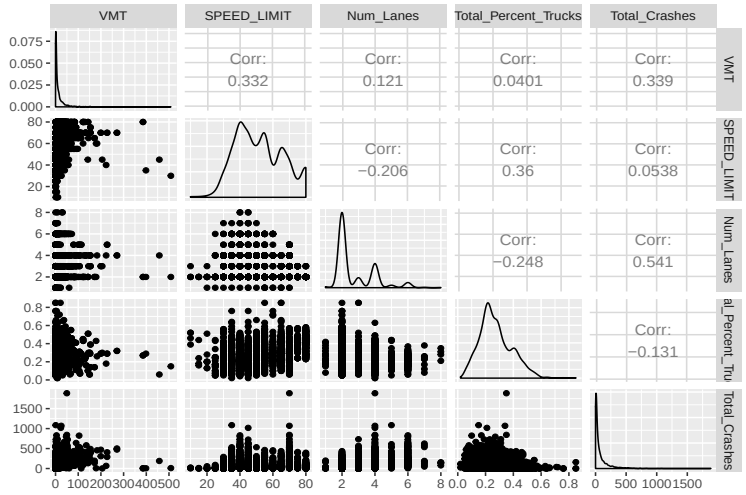


Figure 1: Scatter plot matrix with marginal densities and correlation coefficients

In modeling the UDOT crash data, a few complexities need to be considered. The first complexity is overdispersion (i.e., variances are substantially larger than the mean). To see evidence of overdispersion, Figure 2 contains a histogram of the residuals fit using a preliminary Poisson regression model. The residuals are heavily right skewed which suggest that overdispersion will need to be accounted for when modeling the data. The second potential complexity is spatial dependence. Similar spatial (geographical) features, such as elevation, regional climate patterns, terrain, etc., can similarly impact the number of crashes across several adjacent road segments. To see evidence of spatial correlation, we analyzed the spatial effects of Highway 6 found using an exploratory spatial regression model. Highway 6 is a windy mountainous road that connects Utah and Colorado. Figure 3 illustrates the spatial effect of Highway 6 plotted against the centroid (center point) of each road segment. Note that adjacent points tend to stick together suggesting that neighboring segments are more likely to contain similar number of crashes, or spatial correlation. If we do not account for these complexities, our ability to accurately measure uncertainty will be impaired.

# 3 Model

In this section, we specify a fully Bayesian spatial generalized linear mixed model using the Poisson likelihood because it most effectively accounted for overdisperion and spatial dependence. Let $Y_{rs}$ denote the total number of crashes on road $r = 1, \ldots, R$ and segment $S = 1, \ldots, n_r$ where $n_r$ is the

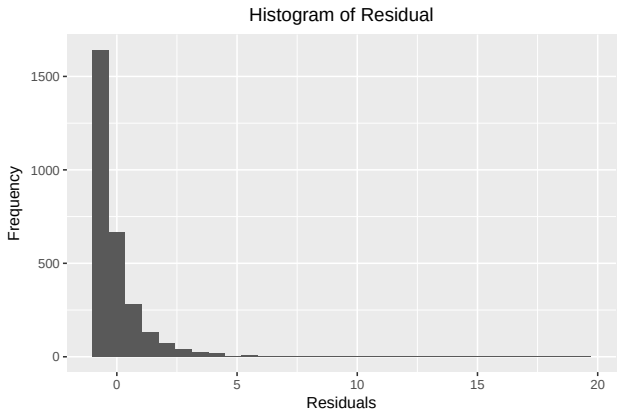


Figure 2: Histogram of the residuals fit using a preliminary Poisson regression model suggesting evidence of overdispersion

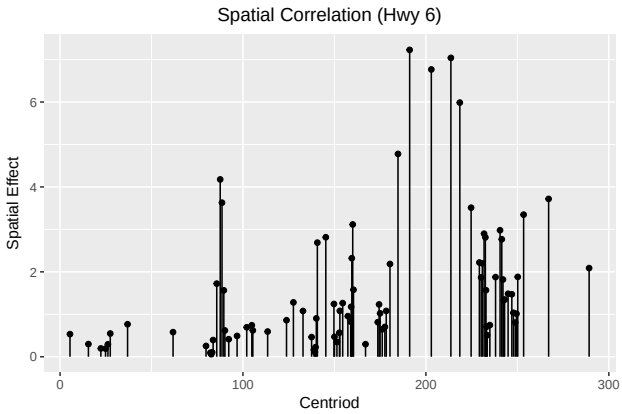


Figure 3: Spatial Effects using an exploratory regression model for Utah State Route 6 suggesting evidence of spatial dependence

number of segments on road $r$. We assume,

$$Y_{rs} \sim \text{Pois}(\mu_{rs}) \tag{1}$$

where $\text{Pois}(\cdot)$ denotes the Poisson distribution, $\mu_{rs}$ is the mean number of crashes on road and segment $(r, s)$.

The spatial Poisson regression model is defined by setting,

$$\log(\mu_{rs}) = \boldsymbol{x}'_{rs}\boldsymbol{\beta} + \boldsymbol{m}'_{rs}\boldsymbol{\theta}_r \tag{2}$$

where $\log(\mu_{rs})$ is the link function for count data and ensures $\mu_{rs} > 0$ as required by the Poisson assumption, $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_P)'$ is a $P+1$-dimensional vector of regression coefficients for the explanatory variables $\boldsymbol{x}_{rs} = (1, x_{rs1}, \ldots, x_{rsP})'$, $\boldsymbol{m}_{rs} = (m_{rs1}, \ldots, m_{rsK})'$ is a vector of spatially defined basis functions with associated coefficients $\boldsymbol{\theta}_r$ designed to capture the spatial correlation between road segments and will be discussed in more detail below. For this analysis, in order to be consistent with previous work to identify high risk regions on Utah roadways, we consider only four explanatory variables in $\boldsymbol{x}_{rs}$: total percent trucks, number of lanes, speed limit, and VMT. VMT was categorized to account for its very right skewed distribution, which can be seen in Figure 1. The fixed effects, $\boldsymbol{\beta}$, provide meaningful interpretations that describe the relationship of important road and driving characteristics with the total number of crashes. For example, let $x_{rs1}$ be the speed limit on segment $s$ of road $r$. Holding all else constant as the speed limit increases by 1 mph, we expect the total number of crashes to be $\exp\{\beta_1\}$ times higher or lower depending on the sign of $\boldsymbol{\beta}_1$. Notably, when $0 < \exp\{\beta_1\} < 1$ (which is equivalent to $\beta_1 < 0$) this would represent a decrease in the expected number of crashes.

We note that the fixed effects $\boldsymbol{\beta}$ are the same across all roadways suggesting that, for example, the effect of increasing the speed limit on segment $s$ of road $r_1$ is the same as the effect of increasing the speed limit on segment $s$ of road $r_2$. This is a fairly strong assumption that we make for a few reasons. First, UDOT wishes to evaluate the effect of road characteristics across the state rather than on individual roads. And, second, the large amount of zeros present in our dataset make it difficult to estimate road-specific effects. Hence, we leave road-varying effects of road characteristics as future work rather than considering it here.

As stated above, $\boldsymbol{m}_{rs}$ is a vector of spatial basis functions (intuitively, spatially defined explanatory variables) designed to account for unobserved explanatory variables (e.g., incline) and

models spatial correlation by imposing spatial structure in the mean function given by Equation (2). Notably, because the associated coefficients $\boldsymbol{\theta}_r$ are road-specific (note the $r$ subscript), the spatial effects are different for every road. This largely follows intuition in that the spatial structure of road $r_1$ could be considerably different than road $r_2$ due to the differences in terrain, weather, etc.

While there are many ways of defining $\boldsymbol{m}_{rs}$, we follow the method proposed by Hughes and Haran (2013) for modeling spatial random effects using the Moran operator. These spatial effects are defined as follows. Let $\boldsymbol{X}$ denote the $n_r \times (P+1)$ design matrix of all observed explanatory variables (road characteristics) with $s^{th}$ row given by $\boldsymbol{x}'_{rs}$. Next define $\boldsymbol{A}_r = \{a_{r,ij}\}$ to be a $n_r \times n_r$ matrix where the $ij^{th}$ element of $\boldsymbol{A}_r$ is given by

$$a_{r,ij} = \begin{cases} 1 & \text{if segment } i \text{ and } j \text{ are adjoining} \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

Next, define $\boldsymbol{P}$ to be the orthogonal projection onto span($\boldsymbol{X}$) such that $\boldsymbol{P} = \boldsymbol{X}_r(\boldsymbol{X}'_r\boldsymbol{X}_r)^{-1}\boldsymbol{X}'_r$ (intuitively, this represents the variability in $\mu_{rs}$ already explained by the explanatory variables in $\boldsymbol{X}_r$ which we don't wish to confound by adding in additional covariates). Finally, let $\boldsymbol{P}^{\perp} = \boldsymbol{I} - \boldsymbol{P}$ represent the orthogonal complement of $\boldsymbol{X}_r$ (intuitively, $\boldsymbol{P}^{\perp}$ represents variability in $\mu_{rs}$ not explained by the explanatory variables in $\boldsymbol{X}_r$).

The Moran operator is defined as $\boldsymbol{P}^{\perp}\boldsymbol{A}_r\boldsymbol{P}^{\perp}$ and represents spatial structure across road $r$ (because $\boldsymbol{A}_r$ contains neighbor information across road $r$) not accounted for by the explanatory variables in $\boldsymbol{X}_r$ (because $\boldsymbol{P}^{\perp}$ is the space orthogonal to $\boldsymbol{X}_r$). Hughes and Haran (2013) suggest defining spatial basis functions $\boldsymbol{M}_r$ as the principal components (eigenvectors) of the Moran operator because these components represent the main sources of spatial structure across road $r$. Specifically, let $\boldsymbol{M}_r$ be a $n_r \times K$ matrix composed of the $K$ eigenvectors of $\boldsymbol{P}^{\perp}\boldsymbol{A}_r\boldsymbol{P}^{\perp}$ associated with all $K$ positive eigenvalues of $\boldsymbol{P}^{\perp}\boldsymbol{A}_r\boldsymbol{P}^{\perp}$. We then define $\boldsymbol{m}_{rs}$ as the $s^{th}$ row of $\boldsymbol{M}_r$.

We provide an example of the spatial component for Highway 6. Figure 4 displays a map of major highways in Utah. Highlighted in red is Highway 6, which consists of 84 different road segments. The corresponding adjacency matrix $A$ for Highway 6, found in Figure 5, indicates adjoining road segments. We note that adjacent segments, depicted in blue, are found near the diagonals, but that the true diagonal elements of $A$ are always 0, depicted in red. Figures 6 and 7 visually represent the $1^{st}$ and $25^{th}$ eigenvectors of $\boldsymbol{M} = \boldsymbol{P}^{\perp}\boldsymbol{A}\boldsymbol{P}^{\perp}$ for Highway 6. Again, $\boldsymbol{M}_r$ represents the spatial structure across road $r$ after accounting for the explanatory variables in $\boldsymbol{X}_r$. Notably, the spatial structure of $\mathbf{M}_r$ is apparent in both Figures 6 and 7. Figure 6 represents long range spatial dependence since segments 45-84 are associated with positive values of $\vec{m}_{rs}$. In contrast, Figure 7 represents shorter scale dependence because values of $\mathbf{M}$ are similar over shorter distances.

As stated as one of the goals associated with this project, we opt for a fully Bayesian estimation procedure. Hence, we choose the following prior distributions for the unknown parameters $\boldsymbol{\beta}$, $\boldsymbol{\theta}_r$, and $\tau$:

$$\begin{aligned} \boldsymbol{\beta_i} &\overset{iid}{\sim} \text{Laplace}(m_\beta, \tau) \\ \boldsymbol{\theta}_r &\overset{iid}{\sim} \mathcal{N}(\boldsymbol{m}_\theta, \boldsymbol{S}_\theta) \\ \tau &\sim \mathcal{IG}(a, b). \end{aligned} \tag{4}$$

Using the independent Gaussian assumptions, we choose to set $m_\beta = 0$ as a vague, "null hypothesis" prior that the road characteristics in $\boldsymbol{x}_{rs}$ have no effect on the mean number of crashes. The $\tau$ parameter is a scale parameter that is used to shrink the variables that are less important to zero. In other words, $\tau$ allows us to perform a Bayesian Lasso variable selection on the explanatory

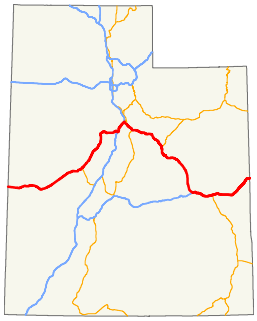

Figure 4: Map of Highway 6

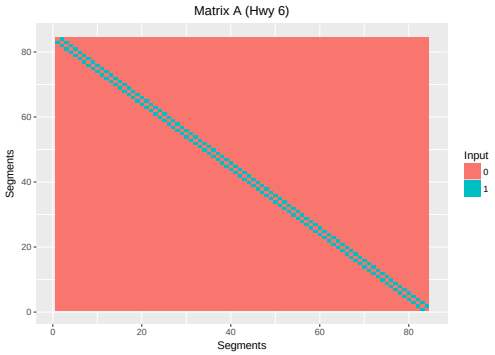


Figure 5: Adjacency matrix $A$

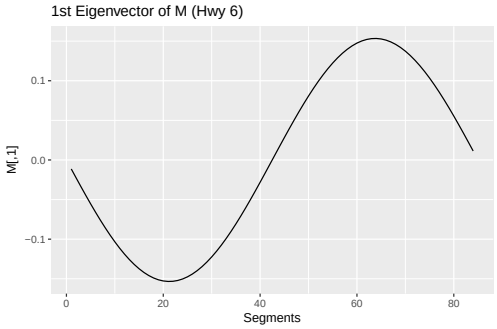


Figure 6: $1^{st}$ Spatial Component of M

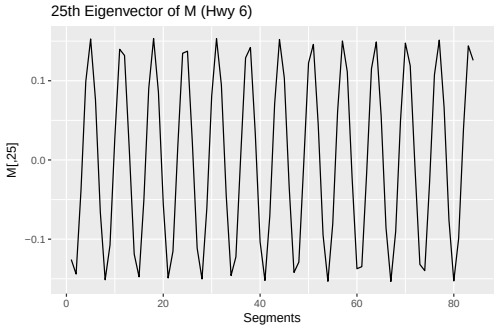


Figure 7: $25^{th}$ Spatial Component of M

variables. Again, following this same strategy, we set $\boldsymbol{m}_\theta = \mathbf{0}$ and $\boldsymbol{S}_\theta = \boldsymbol{I}$. Notably, because $\boldsymbol{\theta}$ is defined on the log-scale, a marginal prior variance of 1 for all coefficients is quite vague. Finally, we place a relatively uninformative Inverse Gamma prior distribution on $\tau$ by setting $a = 2.01$ and $b = 1$.

# 4 Analysis Results

Table 1 displays the posterior mean, variance, and respective quantiles for the $\boldsymbol{\beta}$ regression coefficients. For example, a 95% credible interval for $\beta_1$, the effect for total percent trucks, is (0.033, 0.646). Since the credible interval does not contain 0, we conclude that the total percent trucks significantly explains the total number of crashes for all roadways. Using this logic, we determine that total percent trucks, number of lanes, speed limit, and VMT are all significantly associated with total number of crashes. Thus, the geometry of the road is associated with the total number of crashes on each road segment. Furthermore, since none of the 95% credible intervals contain zero and are either positive or negative, the intervals can also be used to determine which explanatory variables are positively and negatively associated with number of crashes. VMT between 0.392 and including 0.84 is the only covariate with a negative effect while number of lanes, speed limit, and the remaining VMT categories have a positive effect on the expected number of total crashes.

The transformed regression coefficients provide meaningful interpretations that describe the effect of road and driving characteristics on the expected total number of crashes. For example,

Table 1: Summary statistics for estimated fixed effects

| | | | Quantiles | | | | |
|---|---|---|---|---|---|---|---|
| $\hat{\boldsymbol{\beta}}$ | Mean | Variance | 2.5% | 5% | 50% | 95% | 97.5% |
| Intercept | 2.391 | 0.008 | 2.239 | 2.260 | 2.384 | 2.569 | 2.592 |
| Perc. Trucks | 0.276 | 0.029 | 0.033 | 0.054 | 0.238 | 0.598 | 0.646 |
| Num. Lanes | 0.251 | 0.000 | 0.216 | 0.222 | 0.251 | 0.278 | 0.284 |
| Speed Limit | 0.501 | 0.001 | 0.455 | 0.465 | 0.502 | 0.538 | 0.545 |
| VMT [0.0019,0.392] | 0.012 | 0.000 | 0.010 | 0.010 | 0.012 | 0.014 | 0.015 |
| VMT (0.392,0.84] | -0.048 | 0.000 | -0.090 | -0.083 | -0.048 | -0.011 | -0.002 |
| VMT (0.84,1.52] | 0.204 | 0.000 | 0.164 | 0.172 | 0.204 | 0.239 | 0.245 |
| VMT (1.52,2.73] | 0.154 | 0.000 | 0.113 | 0.118 | 0.153 | 0.193 | 0.200 |
| VMT (2.73,4.38] | 0.144 | 0.000 | 0.106 | 0.111 | 0.145 | 0.177 | 0.184 |
| VMT (4.38,7.43] | 0.226 | 0.000 | 0.190 | 0.193 | 0.226 | 0.257 | 0.263 |
| VMT (7.43,12] | 0.434 | 0.000 | 0.400 | 0.404 | 0.435 | 0.462 | 0.467 |
| VMT (12,18.8] | 0.481 | 0.000 | 0.443 | 0.451 | 0.482 | 0.511 | 0.514 |
| VMT (18.8,34.5] | 0.530 | 0.000 | 0.495 | 0.500 | 0.531 | 0.557 | 0.563 |
| VMT (34.5,508] | 0.777 | 0.000 | 0.740 | 0.745 | 0.777 | 0.807 | 0.810 |

the effect for total percent trucks can be interpreted as there is a 95% probability that as the total percent trucks increases by 1%, the expected total number of crashes will be between $\exp(0.033) = 1.034$ and $\exp(0.646) = 1.908$ times higher.

# 5 Hot Spot Analysis

An integral component to eliminating the number of deaths occurring on Utah roads is to identify road segments that are at higher risk than expected. Pinpointing dangerous roads provides UDOT with the resources to take action, such as changing speed limits or adding shoulders, to minimize the potential for further crashes. In this section, we consider two methods to identify hot spots to assist UDOT in making Utah roads safer.

## 5.1 Method 1: Posterior Predictive Percentile

The first approach we implement is similar to previous hot spot identification analysis done on Utah roads, namely Schultz et al. (2013). Using the posterior predictive distribution, we generate a distribution of simulated data, or total number of crashes, for each road segment. Figure 8 is a graphical representation of the posterior predictive distribution for segment 2 on road 269. The red dotted line represents the actual number of crashes on the road segment. We rank each road segment by calculating the area to the left of the observed number of crashes, or a quantile for the observed value. A quantile close to zero means that there were considerably fewer crashes than expected, while a quantile close to one indicates that there were considerably more crashes than expected. Thus, hot spots will have quantiles near one.

Table 2 displays the top 10 highest risk road segments according to our first criteria. Note that most of the high risk roads are in the $100^{th}$ percentile, meaning that these road segments have observed significantly more crashes than expected. For instance, the $1^{st}$ ranked road segment, between mile markers 25 and 27 on highway 6, observed 8 total crashes. However, the total number

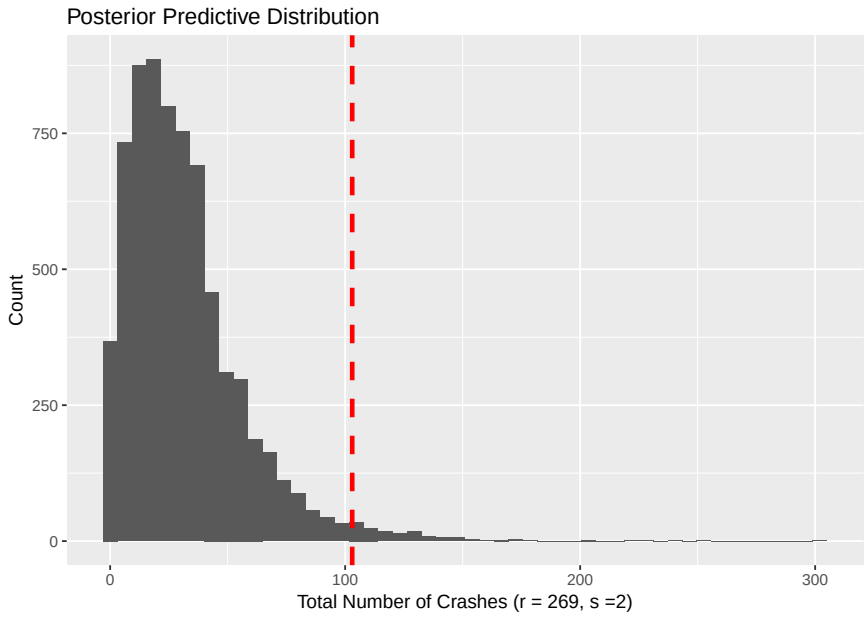


Figure 8: Posterior predictive distribution for road 269 segment 2

of crashes expected is significantly less at about 0.970.

It is critical to mention that this methodology identifies road segments that are at higher risk of a crash, statistically. However, it does not identify road segments that are the most practically important. For example, the $3^{rd}$ ranked road segment found on segment 57 of highway 6 observed 436 total crashes but only expected roughly 298 crashes. Because this road segment observed considerably more crashes than the $1^{st}$ most dangerous road segment, it may be more of a priority.

### 5.2 Method 2: Pseudo $Z$-score

The following method uses a similar methodology to identify hot spots, but with the intent to detect road segments that are more practically significant. Hot spots are determined by standardizing each observation using an approach analogous to the $Z$-score. We call it the pseudo $Z$-score. The pseudo

Table 2: Highest risk roads found using the posterior predictive percentile method

| Rank | Road | Segment | Beg_MP | End_MP | Prob | Expected | Observed |
|---|---|---|---|---|---|---|---|
| 1 | 6 | 5 | 25.248 | 27.100 | 1 | 0.970 | 8 |
| 2 | 6 | 48 | 173.446 | 173.984 | 1 | 141.744 | 193 |
| 3 | 6 | 57 | 194.774 | 211.158 | 1 | 298.944 | 436 |
| 4 | 6 | 60 | 220.927 | 228.510 | 1 | 105.801 | 172 |
| 5 | 8 | 1 | 0.000 | 1.180 | 1 | 164.913 | 254 |
| 6 | 9 | 5 | 4.248 | 7.309 | 1 | 62.034 | 121 |
| 7 | 9 | 8 | 8.998 | 9.776 | 1 | 75.828 | 142 |
| 8 | 9 | 11 | 10.883 | 12.458 | 1 | 30.939 | 67 |
| 9 | 9 | 28 | 44.850 | 57.075 | 1 | 56.599 | 95 |
| 10 | 10 | 21 | 41.233 | 46.319 | 1 | 33.253 | 71 |

$Z$-score is calculated for the Poisson distribution by subtracting the mean from the observed values and dividing by the standard deviation, given by the following equation,

$$z\text{-score} = \frac{y_{rs} - \mu_{rs}}{\sqrt{\mu_{rs}}} \tag{5}$$

where $\mu_{rs} = e^{\boldsymbol{x}'_{rs}\boldsymbol{\beta}+\boldsymbol{m}'_{rs}\boldsymbol{\theta_r}}$. We compute the pseudo $z$-score for each posterior draw and take the average to obtain an overall $z$-score for each road. Intuitively, the pseudo $z$-score represents the expected number of standard deviations the observed number of crashes, $y_{rs}$, is from the mean. Negative pseudo $z$-scores mean that fewer crashes are observed than expected, while positive pseudo $z$-scores mean that more crashes are observed than expected. Thus, very high positive $z$-scores flag high risk roads. As noted, each observation is standardized so that each road segment is weighted equally.

Table 3: Highest risk roads found using the pseudo $z$-score method

| Rank | Road | Segment | Beg_MP | End_MP | $Z$-score | Expected | Observed |
|---|---|---|---|---|---|---|---|
| 1 | 15 | 236 | 324.447 | 328.640 | 21.872 | 334.374 | 734 |
| 2 | 15 | 212 | 295.616 | 297.920 | 20.157 | 1189.921 | 1885 |
| 3 | 15 | 223 | 307.956 | 309.333 | 17.404 | 149.497 | 362 |
| 4 | 108 | 4 | 1.568 | 3.002 | 17.239 | 92.014 | 257 |
| 5 | 80 | 33 | 128.607 | 131.869 | 15.679 | 116.794 | 286 |
| 6 | 248 | 1 | 0.000 | 1.071 | 15.253 | 113.758 | 276 |
| 7 | 15 | 81 | 298.916 | 300.300 | 15.210 | 312.351 | 581 |
| 8 | 15 | 217 | 303.414 | 304.691 | 15.086 | 490.203 | 824 |
| 9 | 15 | 201 | 276.462 | 278.528 | 14.946 | 281.450 | 532 |
| 10 | 77 | 11 | 8.868 | 9.069 | 14.909 | 47.127 | 149 |

Table 3 displays the top 10 hot spots using the pseudo $z$-score approach. The top ranked road segment, found on highway 15 between mile markers 324 and 328, has a pseudo $z$-score of 21.872. This means that the observed number of total crashes, 734, is approximately 22 standard deviations away form the expected number of crashes, on average.

### 5.3 Method 3: Spatial Effect

The third method for identifying hot spots takes advantage of the unique construction of our model to account for spatial correlation. Again, $\boldsymbol{m}_{rs}$, represents spatial basis functions designed to account for the variability in the mean total number of crashes not explained by the explanatory variables in $\boldsymbol{x}_{rs}$. The coefficients, $\boldsymbol{\theta}_{rs}$, are the associated spatial effects for these spatial explanatory variables. We rank the road segments that may appear safe by road geometry, but are expected to have considerably more crashes associated with factors not related to road geometry using these spatial components. We define a spatial score using $\boldsymbol{m}'_{rs}\boldsymbol{\theta_r}$. For each road segment, we computed the spatial score for all posterior draws using $\boldsymbol{m}'_{rs}\boldsymbol{\theta_r}$ and obtain an overall score by taking the average. Positive spatial scores are associated with more crashes than would be expected by only considering the the road characteristics, while negative spatial scores are associated with fewer crashes than would be expected by only considering the road characteristics. Hence, high spatial scores flag roads that may appear to be at greater risk of a crash than the road and driving characteristics would imply.

Table 4: Highest risk roads found using the spatial effect method

| Rank | Road | Segment | Beg_MP | End_MP | Spatial Effect | Expected | Observed |
|---|---|---|---|---|---|---|---|
| 1 | 91 | 20 | 27.148 | 28.425 | 3.642 | 805.842 | 1086 |
| 2 | 172 | 5 | 3.993 | 5.985 | 2.969 | 839.033 | 1018 |
| 3 | 89 | 114 | 336.030 | 337.878 | 2.958 | 675.690 | 831 |
| 4 | 173 | 4 | 3.189 | 4.738 | 2.949 | 575.361 | 727 |
| 5 | 266 | 1 | 0.000 | 0.777 | 2.900 | 660.017 | 671 |
| 6 | 209 | 15 | 11.694 | 12.179 | 2.848 | 624.146 | 637 |
| 7 | 48 | 8 | 3.865 | 12.981 | 2.807 | 459.073 | 729 |
| 8 | 15 | 212 | 295.616 | 297.920 | 2.737 | 1189.921 | 1885 |
| 9 | 89 | 115 | 337.878 | 338.543 | 2.656 | 412.004 | 437 |
| 10 | 15 | 211 | 293.634 | 295.616 | 2.622 | 827.041 | 1070 |

Table 4 displays the top 10 hot spots given their spatial score. The road segment that is expected to have the highest number of crashes not associated with road geometry is highway 91 between mile markers 27 and 28. The spatial effect corresponding to this road segment is 3.642. We see that this road observed 1086 total crashes but was only expected to have approximately 805 crashes.

Note that it's plausible to identify road segments that observe fewer crashes than would be expected using the spatial effect method. However, since road segments with fewer observed crashes than expected are considered "safe," we removed all such road segments to reduce confusion.

### 5.4 Method 4: Road Effect

The last approach we use seeks to classify hot spots based upon the road characteristics. The $\boldsymbol{\beta}$ regression coefficients describe the effects of road and driving characteristics (total percent trucks, number of lanes, speed limit, etc.) on the total number of crashes. We can rank road segments that are expected to be more prone to crashes related to road geometry using these regression coefficients. We define a road effect score by taking the average of $\boldsymbol{x}'_{rs}\boldsymbol{\beta}$ for each posterior draw across all road segments. High road effect scores indicate road segments that are expected to be at higher risk due to poor road geometry while low road effect scores are associated with road segments that are at lower risk due to good geometry.

Table 5 displays the highest risk road segments corresponding with the largest road effect scores. The top ranked road segment due to poor road construction is highway 154 between mile marker 10 and 11. This particular highway segment has a road effect score of 4.460. The expected number of crashes is roughly 222, which is less than the observed number of total crashes of 247. Again, we only consider road segments with more observed crashes than expected.

## 6 Conclusion

Although there has been great advancements in safety and engineering of vehicles and roadways in recent years, crash related fatalities in Utah continue to occur at an unacceptable rate. With the goal to eradicate fatalities and improve roadway safety, we quantified the relationship of important road and driving characteristics with the likelihood of a crash while accounting for spatial dependence. We found that as the VMT between 0.392 and 0.84 increases, the number of total crashes decrease. We also found that as the number of lanes, speed limit, and the remaining VMT

Table 5: Highest risk road segments found using the road effect method

| Rank | Road | Segment | Beg_MP | End_MP | Road Effect | Expected | Observed |
|---|---|---|---|---|---|---|---|
| 1 | 154 | 10 | 10.384 | 11.390 | 4.460 | 222.100 | 247 |
| 2 | 154 | 8 | 8.274 | 9.600 | 4.438 | 231.924 | 270 |
| 3 | 15 | 118 | 346.719 | 349.354 | 4.421 | 83.446 | 124 |
| 4 | 15 | 251 | 346.719 | 349.354 | 4.421 | 84.027 | 118 |
| 5 | 15 | 120 | 351.853 | 357.554 | 4.415 | 356.497 | 415 |
| 6 | 15 | 253 | 351.853 | 357.554 | 4.415 | 350.424 | 392 |
| 7 | 154 | 17 | 17.936 | 18.946 | 4.390 | 189.177 | 279 |
| 8 | 15 | 247 | 343.067 | 343.852 | 4.364 | 100.135 | 104 |
| 9 | 15 | 212 | 295.616 | 297.920 | 4.345 | 1189.921 | 1885 |
| 10 | 80 | 26 | 123.231 | 124.125 | 4.342 | 192.737 | 273 |

categories increase, the number of total crashes also increases. Finally, we targeted high risk road segments using four methods to help UDOT recognize where to best concentrate their efforts.

A shortcoming in this analysis is poor convergence which most likely poorly affected our hot spot results. In the future, we suggest considering more variables in the variable selection procedure and we suggest removing the road segmentation process to consider spatial dependence based on longitude and latitude.