# A Bayesian Semi-parametric Model to Estimate Relationships Between Crash Counts and Roadway Characteristics

Thomas S. Shively
McCombs School of Business, University of Texas at Austin

Kara Kockelman*
Department of Civil, Architectural, and Environmental Engineering
University of Texas at Austin

Paul Damien
McCombs School of Business, University of Texas at Austin

The following paper is a pre-print and the final publication can be found in *Transportation Research Part B* 44 (5):699-715, 2010.

**Abstract:** This paper uses a semi-parametric Poisson-gamma model to estimate the relationships between crash counts and various roadway characteristics, including curvature, traffic levels, speed limit and surface width. A Bayesian nonparametric estimation procedure is employed for the model's link function, substantially reducing the risk of a mis-specified model. It is shown via simulation that little is lost in terms of estimation quality if the nonparametric estimation procedure is used when standard parametric assumptions (e.g., linear functional forms) are satisfied, but there is significant gain if the parametric assumptions are violated. It is also shown that imposing appropriate monotonicity constraints on the relationships provides better function estimates. Results suggest that key factors for explaining crash rate variability across roadways are the amount and density of traffic, presence and degree of a horizontal curve, and road classification. Issues related to count forecasting on individual roadway segments and out-of-sample validation measures also are discussed.

**Keywords**: Forecasting; Poisson-gamma model; Markov chain Monte Carlo, Monotonicity constraints; Regression splines.

*Please send correspondence to*: Kara Kockelman, Department of Civil, Architectural and Environmental Engineering, Mail Code C1767, University of Texas, Austin, Texas, 78712, U.S.A. E-mail: kkockelm@mail.utexas.edu; Telephone: 512-471-0210; Fax: 512-475-8744.

## 1. Introduction

Understanding roadway safety is an important task, but is complicated by the rare nature of crash incidents and the large number of potential causal factors. In addition, there is a lack of subject matter theory regarding the appropriate functional forms to use for relating causal factors to crash rates. As described in Wahba's (1978) classic paper, functional data analysis involves complex curve fitting. The current paper adds to the transportation literature by applying Bayesian nonparametric tools for functional analysis to crash data. Related work includes Mahamassani et al.'s (1988) direct use of cubic regression splines for urban density patterns, Biller's (2000) adaptive computational methods for Bayesian semi-parametric models, and Biller and Fahrmeir's (2001) method to allow for varying coefficients. Fahrmeir and Osuna (2007) used regression splines for a negative binomial model of count data in the context of automobile insurance claims data, while Xie and Zhang (2008) use regression splines in a frequentist context with fixed knot locations to model accident counts at traffic intersections. Neelon and Dunson (2004), Dunson (2005), Schipper et al. (2007) and Shively, Sager and Walker (2009) considered monotonic function estimation in additive models.

This paper uses a Bayesian nonparametric monotone function estimation methodology in the context of a Poisson-gamma model to model and estimate the relationships between the number of crashes on segments of two-lane rural highways and roadway characteristics such as degree of curvature, vertical grade, amount of traffic and speed limit, among others (in all, there are 15 explanatory variables included in the model). The resulting function estimates provide valuable information regarding which characteristics explain the variability in crash counts across roadway segments and therefore which characteristics transportation officials should focus on to improve road safety. The nonlinear relationships can provide a better understanding of the effect roadway characteristics have on crash rates and road safety.

Unfortunately, transportation data sets are often imperfect, with some design variables out of date (e.g., the road was improved but the road file was not updated) and many crashes going unreported. Also, important explanatory variables are often unavailable (e.g., number of snowy

days at a location, roadside clear zone slope and width, and average speeds of travel). Given the issues related to crash data and the inherent noise in the data, it is important to impose a reasonable amount of structure on the analysis without imposing so much that the results are potentially misleading due to incorrect assumptions. A fully parametric analysis lies at one end of the "structure spectrum", and a fully nonparametric analysis at the other end (without any assumptions on either the density function or the functional forms of the covariate relationships). By specifying a Poisson-gamma distribution for the count data and imposing monotonicity constraints on the functional forms of the covariate relationships, we have chosen a compromise. The monotonicity constraints incorporate a significant amount of information into the model but stop short of specifying the functional forms of the relationships.

Nonparametric techniques allow nonlinear relationships to be observed that may not be detected using a parametric analysis. If there is subject matter theory available to specify the appropriate functional forms, then it should be used. However, valuable insights can often be obtained using nonparametric methods that provide limited structure to the analysis without allowing too much flexibility. Nonparametric estimation methods also provide a valuable tool for exploratory data analysis. Such analysis can provide valuable information about relationships that is not available from a parametric approach.

The Bayesian nonparametric methodology used in this paper allows monotonicity constraints to be imposed on the unknown functions when such constraints are appropriate. Strong subject matter arguments can be made that many of the roadway characteristic variables in the data set will be monotonically related to crash counts. We show via simulation that imposing appropriate monotonicity constraints increases the quality of the function estimates and thereby provides more reliable conclusions to be drawn from the analysis. In addition, the monotone nonparametric procedure produces smooth function estimates, without the "wiggles" that often occur with unconstrained nonparametric procedures. The resulting function estimates make more intuitive sense and are easier to interpret and explain to users. Finally, as discussed above,

monotonicity assumptions impose additional structure on the analysis that is particularly useful in the presence of noisy data.

A natural question that arises for any nonparametric estimation procedure, especially when it is used for models with a large number of functions to estimate and a substantial amount of variability in the dependent variable, is: How good are the function estimates? In particular, does the extra flexibility provided by a nonparametric procedure provide function estimates that are too "noisy"? The answer to this question depends on the complexity of the model, the amount of data, the structure of the explanatory variables and the number of functions to be estimated. It is shown via simulation in section 5 for a Poisson-gamma model with the explanatory variables and number of observations in our data set that the nonparametric monotone estimation procedure does nearly as well as the standard parametric procedure when the parametric assumptions are satisfied, and substantially better when the assumptions are violated.

A second important question is: How well does the estimated model predict future crash counts? In many contexts, questions of this type can be answered using out-of-sample validation procedures. However, we show theoretically and via simulation that no estimation procedure will be able to accurately predict the number of crashes on a specific road segment given the inherent variability in Poisson count data. This is true whether the model is estimated parametrically or nonparametrically, and whether or not the parametric assumptions are satisfied. The lack of forecasting power is shown to hold even in the extreme case when the "true" relationships between crash counts and the explanatory variables are known, i.e., when there is no estimation error in the function estimates. This finding does not invalidate the importance of accurate function estimation because accurate estimates are critically important for developing safer roads. To be more specific, given the rare nature of crashes on any given roadway segment, safety engineers are typically interested in what happens over the long-run, across many road segments with the same or similar characteristics (e.g., same curvature, amount of traffic, and speed limit). The expected number of crashes per segment per year across segments with the same characteristics and across several years (i.e., with random variability averaged out) depends

crucially on the functions relating crash counts to roadway characteristics and therefore on accurate and reliable estimates of the functions.

The empirical results obtained from our model indicate the important factors in explaining the variability in crash rates across road segments are the amount and density of traffic, whether or not there is a curve in the road and if so, how sharp and how long the curve is, and the type of road. The functions associated with eight of the explanatory variables are estimated nonparametrically. The results indicate that four are nonlinear functions and cannot be easily modeled using a parametric functional form. The fifth estimated function is very close to linear, and the variables associated with the other three functions do not appear to be related to crash counts. The remaining seven variables are categorical variables and enter the model linearly. A thorough discussion of the empirical results and their interpretation is given in section 4.

The paper is organized as follows. Section 2 outlines the nonparametric monotone function estimation methodology used in the paper. Section 3 discusses the data used in the analysis. Section 4 gives the empirical results and their interpretation in terms of roadway safety. It also contains a discussion of why it is difficult to make accurate forecasts in the context of Poisson-gamma models. Section 5 provides simulation results to show the significant advantages of using a nonparametric monotone function estimation procedure rather than either a standard parametric procedure or a nonparametric procedure without monotonicity constraints. The MCMC sampling algorithm used to implement the monotone estimation procedure is outlined in the appendix.

## 2. The model and estimation methodology

Crash counts on each roadway segment in a one-year period are assumed to arise from a generalized-additive Poisson-gamma model:

$$Pr(Y_i = y_i \mid x_i, z_i) = exp\{-\phi_i g(x_i, z_i)\} \frac{[\phi_i g(x_i, z_i,)]^{y_i}}{y_i!} \tag{1}$$

where

$$\log[g(x_i, z_i)] = \alpha + f_1(x_{1i}) + \ldots + f_p(x_{pi}) + \gamma_1 z_{1i} + \ldots + \gamma_q z_{qi},$$

and $i$ indexes observations/roadway segments (1, …, $n$). The dependent variable $y_i$ is the number of crashes on the $i^{th}$ road segment, $x_i = (x_{1i}, \ldots, x_{pi})$ are explanatory variables that enter the model nonlinearly and $z_i = (z_{1i}, \ldots, z_{qi})$ are explanatory variables (typically 0/1 indicator variables) that enter the model linearly. The mean function is $\mu(x_i, z_i, \phi_i) = \phi_i g(x_i, z_i)$ where $\phi_i$ are independent $\Gamma(\eta,\eta)$ random variables included to model additional heterogeneity in crash rates across roadway segments. This model is similar to the Poisson-gamma model considered by Dunson (2005) and Fahrmeir and Osuna (2007). The variables included in $x_i$ and $z_i$ include curvature, traffic flow, terrain type, and other variables, as discussed in section 3. Further, $f_1, \ldots, f_p$ are unknown functions to be estimated nonparametrically and $\gamma_1, \ldots, \gamma_q$ are unknown parameter values. For a Bayesian model, prior distributions must be placed on the unknown parameter values. The prior distributions on the $\alpha$ and $\gamma_j$ values are non-informative $N(0, k)$ distributions with $k$ large, while the prior on $\eta$ is a non-informative $\Gamma(0.01, 0.01)$ distribution. The priors for the unknown function spaces are discussed below.

While the negative binomial specification has become the forerunner in crash count modeling, covariates are entered linearly (e.g., Miaou 1994, Abdel-Aty and Radwan 2000, and Kockelman et al. 2006). Extensions to this popular model are largely limited to panel-data settings with random effects (e.g., Chin and Quddus 2003, Ulfarsson and Shankar 2003, and Kweon and Kockelman 2005), zero-inflated specifications (e.g., Gurmu et al. 1999, Kumara and Chin 2003, Lord et al. 2005a), and multivariate applications using Bayesian methods for parameter estimation (Maher 1990, Park and Lord 2007, and Ma et al. 2008). Linearity in parametric expressions for link functions remains the norm acrossmodeling specifications and applications in the transportation discipline.

This remainder of this section discusses the nonparametric monotone estimation methodology that will be used to estimate the unknown functions and parameters in (1), and in particular, the prior distribution used on the function space to ensure monotonic function estimates. Section 2.1 provides a brief description of Bayesian nonparametric function estimation

in a regression model with Gaussian errors. Section 2.2 then shows how the methodology can be modified to enforce a monotonicity constraint on an estimated function through the prior distribution on the function space. Given an MCMC sampling algorithm to make the procedure computationally feasible (discussed in the appendix), the monotone function estimation methodology applies to both Gaussian and non-Gaussian models, including the Poisson-gamma model in (1).

## 2.1 Bayesian nonparametric function estimation in Gaussian models

Consider the model

$$y_i = \alpha + f(x_i) + \varepsilon_i \tag{2}$$

where $\varepsilon_i$ are independent $N(0, \sigma^2)$ random variables and $f$ is an unknown function to be estimated nonparametrically. Without loss of generality, we assume $0 < x_1 \leq \ldots \leq x_n \leq 1$. There are numerous nonparametric methods available to estimate $f$, including stochastic splines (Wahba 1978 and Wong and Kohn 1996) and regression splines (Smith and Kohn 1996), among others.

Following Smith and Kohn (1996), we employ a regression spline methodology. More specifically, the quadratic regression spline

$$f_m(x) = \beta_1 x + \beta_2 x^2 + \beta_3 (x - \tilde{x}_1)_+^2 + \ldots + \beta_{m+2} (x - \tilde{x}_m)_+^2 \tag{3}$$

is used to approximate the function $f(x)$ in (2), where $\tilde{x}_1, \ldots, \tilde{x}_m$ are $m$ "knots" placed along the domain of the independent variable $x$ such that $0 < \tilde{x}_1 < \ldots < \tilde{x}_m < 1$ and $(z)_+ = \max(0, z)$. The resulting approximating model is

$$y_i = \alpha + f_m(x_i) + \varepsilon_i. \tag{4}$$

Quadratic regression splines are used rather than cubic splines in order to ensure that the monotonicity constraints (imposed in section 2.2) on the function $f_m(x)$ are tractable and practically feasible.

The smoothness of the function $f_m$ depends on the number and location of the knots. To illustrate this (and to set up the monotonicity constraints developed in section 2.2), consider the function with $m = 1$ knot at $x = \tilde{x}_1$:

$$f_m(x) = \beta_1 x + \beta_2 x^2 + \beta_3 (x - \tilde{x}_1)_+^2 \tag{5}$$

If $\beta_1$, $\beta_2$ and $\beta_3$ are all nonzero (so the knot $\tilde{x}_1$ remains in the model), Figure 1a shows the change in the function at $\tilde{x}_1$. The figure is drawn assuming $\tilde{x}_1 = 0.5$. The corresponding first derivative

$$f_m'(x) = \beta_1 + 2\beta_2 x + 2\beta_3 (x - \tilde{x}_1)_+ \tag{6}$$

is shown in Figure 1b and changes direction abruptly at $x = \tilde{x}_1$ (in other words, the second derivative is discontinuous at $x = \tilde{x}_1$). It is in this sense that the function is "not smooth" at a knot. Figure 1c shows the function $f_m(x)$ with $\beta_3 = 0$. In this case, the knot $\tilde{x}_1$ drops out of the model and the function is a simple quadratic with no change at $\tilde{x}_1$, (i.e., it is "smoother" than the function with $\tilde{x}_1$ included). Intuitively, the more knots there are, the less smooth the function will be. Determining which knots should remain in the model, or equivalently determining which $\beta$'s are nonzero, is central to the analysis. Smith and Kohn (1996) suggested using a Bayesian variable selection technique to determine which $\beta$'s should remain in the model. A variation of their technique is employed in the nonparametric monotone function estimation methodology used in this paper.

**Figure 1 goes here**

To briefly describe Smith and Kohn's variable selection technique, the approximating model in (4) (which is effectively a regression model) is re-written in matrix notation as

$$y = \iota\alpha + X\beta + \varepsilon$$

where $y = (y_1, \ldots, y_n)'$, $\iota = (1, \ldots, 1)'$, the $i^{\text{th}}$ row of $X$ is $(x_i, x_i^2, \ldots, (x_i - \tilde{x}_m)_+^2)$, $\beta = (\beta_1, \beta_2, \ldots, \beta_{m+2})'$, and $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)'$. To construct the prior distributions for the $\beta_j$ values, let $J$ be the $(m +$

2) × 1 vector of indicator variables with the $j^{th}$ element $J_j$ such that $J_j = 0$ if $\beta_j = 0$ and $J_j = 1$ if $\beta_j \neq 0$. Thus, if $J_j = 0$, $\beta_j$ and the corresponding regressor drop out of the model. Further, let $\beta_J$ consist of the elements $\beta_j$ corresponding to those elements of $J$ that equal one, and let $X_J$ consist of the regressor variables in $X$ corresponding to those elements of $J$ that equal one. Using this notation, Smith and Kohn (1996) suggest using the prior $\beta_J \sim N(0,\ c\sigma^2 (X_J' X_J)^{-1})$, where $c$ is typically set to the sample size $n$. In addition, $J_1, \ldots, J_{m+2}$ are independent with $\Pr(J_j = 1) = p_j$, where $p_j$ is set by the user. Finally, non-informative priors are placed on $\alpha$ and $\sigma^2$. The estimate of the function $f_m(x)$, and therefore $f(x)$, is the posterior mean $E(f_m(x) \mid y)$ and can be obtained using an MCMC sampling algorithm. Smith and Kohn show that this function estimation technique provides excellent estimates for a wide range of "true" functions.

## 2.2 Monotone nonparametric function estimation

Numerous authors have addressed monotonic function estimation in Gaussian models, including Wright and Wegman (1980), Friedman and Tibshirani (1984), Mammen (1991) and Ramsay (1998). More recently, Neelon and Dunson (2004) and Shively, Sager and Walker (2009) considered monotone function estimation in a Gaussian context from a Bayesian perspective. In addition, several authors have shown how to generalize these methodologies to allow for non-Gaussian specifications, including Dunson (2005), Schipper, Taylor and Lin (2007) and Shively, Walker and Damien (2009).

Shively, Sager and Walker (2009) show how to impose monotonicity constraints on the function $f_m(x)$ in (3), and therefore on the function estimate $E(f_m(x) \mid y)$, through the prior on the $\beta$ values. Their paper considered only the case of Gaussian regression. However, Shively, Walker and Damien (2009) show how the methodology can be generalized to allow for more computationally complex non-Gaussian models such as the Poisson-gamma. To briefly summarize their methodology for monotonicity constraints, we again consider the case where $m = 1$ with $f_m(x)$ and $f_m'(x)$ given in (5) and (6). To impose monotonicity constraints so that the function $f_m(x)$ is non-decreasing requires that $f_m'(x) \geq 0$ for all $x \in (0, 1]$. The key idea is that for

each combination of $J = (J_1, J_2, J_3)$, $f_m'(x)$ is constrained to be non-negative by placing an appropriately constrained prior on $\beta_J$. More specifically, a multivariate normal prior is placed on $\beta_J$ constrained to the region of $\beta$-values that force $f_m'(x) \geq 0$ for all $x \in (0, 1]$.

For example, for $J = (1, 1, 1)$, $f_m'(x) \geq 0$ for all $x \in (0, 1]$ if $f_m'(x) \geq 0$ at $x = 0$, $x = \tilde{x}_1$ and $x = 1$. To see this, note that, if $f_m(x)$ is a quadratic regression spline, then $f_m'(x)$ is a piecewise linear function, as in Figure 1b. This implies that, if $f_m'(x)$ is non-negative at the endpoint of each "piece" of the derivative function, then it is nonzero for all $x \in (0, 1]$. This corresponds to constraining $\beta_{J=(1,1,1)} = (\beta_1, \beta_2, \beta_3)$ to the region defined by

$$\beta_1 \geq 0; \quad \beta_1 + 2\beta_2 \tilde{x}_1 \geq 0; \quad \beta_1 + 2\beta_2 + 2\beta_3(1 - \tilde{x}_1) \geq 0 \tag{7}$$

Therefore, given $J = (1, 1, 1)$, the function $f_m(x)$ is constrained to be non-decreasing by placing a $N(0, c\Omega_{J=(1,1,1)})$ prior on $\beta_{J=(1,1,1)} = (\beta_1, \beta_2, \beta_3)$ constrained to the region defined in (7), where $\Omega_{J=(1,1,1)}$ is an appropriately defined covariance matrix. The specific structure of this matrix is discussed in more detail in the appendix.

Similarly, for $J = (1, 1, 0)$, $f_m'(x) \geq 0$ for all $x \in (0, 1]$ if $f_m'(x) \geq 0$ at $x = 0$ and $x = 1$, as in Figure 1d. This corresponds to constraining $\beta_{J=(1,1,0)} = (\beta_1, \beta_2)$ to the region defined by

$$\beta_1 \geq 0 \quad \text{and} \quad \beta_1 + 2\beta_2 \geq 0 \tag{8}$$

Note that since $J_3 = 0$, the prior assigns $\beta_3 = 0$ with probability one. Therefore, given $J = (1, 1, 0)$, the function $f_m(x)$ is constrained to be non-decreasing by placing a $N(0, c\Omega_{J=(1,1,0)})$ prior on $\beta_{J=(1,1,0)} = (\beta_1, \beta_2)$ constrained to the region defined in (8). Other combinations of $J = (J_1, J_2, J_3)$ can be handled similarly. Also, the methodology generalizes to allow for any number of knots.

The technique of setting the priors for $\beta_J$ for each $J$ effectively puts a prior on the function space for $f_m(x)$ that places non-zero probability only on non-decreasing functions. Given this prior on the function space, the posterior mean $E(f_m(x) \mid y)$, and therefore the estimate of $f(x)$ will be a non-decreasing function. Shively, Sager and Walker (2009) provide a technique for handling the changing constraints as variables drop in and out of a Gaussian regression model

that can be implemented via an MCMC sampling algorithm. And Shively, Walker and Damien (2009) develop a new MCMC sampling algorithm so the constrained spline methodology can be applied in the context of non-Gaussian models. They also show the substantial gain that is achieved by incorporating appropriate monotonicity constraints in non-Gaussian models, and show that the methodology outperforms existing nonparametric monotone function estimation methodologies for models where such methods have been developed. The MCMC sampling algorithm used to implement the methodology in the context of this paper's Poisson-gamma model is given the appendix.

## 3. Data

The data used in this paper were collected in Washington State in 2002 and stored through the Highway Safety Information System. To keep the data size manageable, we examine traffic crashes on two-lane rural roadways in the Puget Sound region, as assembled by Ma et al. (2008). A total of 7710 rural two-way highway segments in this region are used in the analysis, with an average segment length of 0.0665 miles for a total of 513 centerline miles. The sample contains 913 police-logged crashes, including crashes that resulted in property damage only. Table 1 shows all variables used in the analysis, along with their summary statistics.

Table 1's first variable, *Number of crashes*, is our model's dependent variable. The next eight variables are continuous explanatory variables that enter the model via unknown functional forms (i.e., the $x$-variables in equation (1)), while the remaining seven variables are categorical variables that enter the model linearly (the $z$-variables in (1)).

**Table 1: Summary statistics**

| Variable name | Mean | Std. Dev. | Min | Max |
|---|---|---|---|---|
| $y_i$: Number of crashes | 0.118 | 0.426 | 0 | 5 |
| $x_{1i}$: Vehicle miles traveled in 2002 | 8,8671 | 143,134 | 21 | 267,965 |
| $x_{2i}$: Average annualized daily traffic (# of vehicles) | 3,752.7 | 2,727.3 | 254 | 28,624 |
| $x_{3i}$: Horizontal curve length (feet) [1] | 666.8 | 575.6 | 20 | 4715 |
| $x_{4i}$: Degree of horizontal curvature (degrees/100 feet) [1] | 6.268 | 7.433 | 0.17 | 100.52 |
| $x_{5i}$: Vertical curve length (feet) [2] | 484.1 | 373.1 | 20 | 3,200 |
| $x_{6i}$: Vertical grade (percentage) [2] | 1.992 | 2.004 | 0.01 | 16.13 |
| $x_{7i}$: Average shoulder width on each side (feet) | 2.087 | 1.297 | 0 | 16.5 |
| $x_{8i}$: Posted speed limit (miles/hour) | 49.620 | 8.152 | 25 | 60 |
| $z_{1i}$: Surface width (feet) | 23.979 | 4.387 | 16 | 70 |
| $z_{2i}$: Indicator for horizontal curve: 1=yes; 0=no | 0.372 | 0.483 | 0 | 1 |
| $z_{3i}$: Indicator for vertical curve: 1=yes; 0=no | 0.627 | 0.484 | 0 | 1 |
| $z_{4i}$: Indicator for minor arterial: 1=yes; 0=no | 0.285 | 0.451 | 0 | 1 |
| $z_{5i}$: Indicator for collector: 1=yes; 0=no | 0.246 | 0.430 | 0 | 1 |
| $z_{6i}$: Indicator for rolling terrain: 1=yes; 0=no | 0.596 | 0.491 | 0 | 1 |
| $z_{7i}$: Indicator for mountainous terrain: 1=yes; 0=no | 0.039 | 0.195 | 0 | 1 |

[1] Summary statistics for horizontal curve length and degree of curvature are only for the 2868 road segments that include a horizontal curve.

[2] Summary statistics for vertical curve length and vertical grade are only for the 4834 road segments that include a vertical curve (i.e., only segments that are on a hill).

The variables average shoulder width, posted speed limit and surface width are fairly self-explanatory. Surface width is treated as a linear *z*-variable because over 90% of the values are 22, 23 or 24 feet, rendering it a nearly categorical variable.

Average annualized daily traffic (AADT) is defined as the average number of vehicles per day on the road segment in 2002. Since all highway segments in the data set are two lane facilities, it serves as a measure of traffic intensity here. And since segment speeds are likely in the 50 and 60 mph range (but variable and not reported/known, as is typical in such data sets), AADT is nearly proportional to – and thus serves as a measure of – traffic density. Moreover, since capacity values across these segments are nearly constant, AADT also serves as a volume-to-capacity variable (as discussed by Lord et al. [2005b], in their study of crash rate model

specifications). Vehicle miles traveled (VMT) in 2002 is simply AADT times the length of the roadway segment times 365 days.

The indicator variable for horizontal curve ($z_2$) is zero if the road segment is straight and one if the segment has a curve. Horizontal curve length is the distance along a segment's centerline from the start of horizontal curvature (either rightward or leftward) to the end of such curvature. In other words, it is the distance in centerline stationing between the point of curvature to the point of tangency. Degree of horizontal curvature is the number of degrees of curvature per 100 feet of curve. Essentially, higher degree-of-curve turns are "tighter" (of lower radius) and result in greater centrifugal forces acting on vehicles, if speeds are not reduced.

Similarly, the indicator for vertical curve ($z_3$) is zero if the road segment is flat and one if it is on a hill. Vertical curve length is the stationing distance from the start of a section that departs one gradeline and leads to another – via curvature in the vertical sense (e.g., a hilltop or valley bottom). Vertical grade is the rise or fall in elevation for every 100 feet of horizontal distance. For example, a road segment that rises or falls two feet for every 100 feet of horizontal distance has a two percent grade and is denoted as 2.00 in the data set. Since traffic travels in both directions on the two-lane road segments in our sample, all grades are shown in the positive sense.

For the remaining indicator variables, a minor arterial road is a relatively high-speed highway, but not as important as or at the design standards and flow volumes of an interstate highway or freeway. Collector roads are the lowest class of highway, providing more access (via driveways and local street connections) than arterials and interstate highways.

## 4. Empirical results and their interpretation

Given the variety of factors at play, infrequency of crashes, and high noise-to-signal ratio in Poisson, negative binomial and related specifications, proper modeling of crash data is challenging. However, thoughtful nonparametric designs that allow for adequate behavioral flexibility can open a variety of new doors for covariate effects and uncover previously

unobserved relationships. Further, incorporating appropriate monotonicity constraints into the model increases the quality of the estimated relationships and leads to more reliable and conclusive results. As discussed below, intuition, subject matter knowledge and the empirical results support monotonicity assumptions on many of the relationships between roadway characteristics and crash counts.

Section 4.1 discusses the subject matter reasons for imposing monotonicity constraints on the relationships and gives the empirical results of our analysis. Section 4.2 interprets the results while Section 4.3 gives out-of-sample validation results and discusses the difficulty in obtaining accurate forecasts for a Poisson-gamma model.

## 4.1 Empirical results

The natural logarithms of the continuous *x*-variables are used as the explanatory variables in the model rather than the untransformed variables (i.e., log(*x*) values are used rather than *x*). Histograms (not shown here) for many of the continuous variables exhibit strong right skew. Using a logarithmic transformation puts all the variables on a similar scale and provides numerical stability in the nonparametric estimation procedure. The transformation does not affect the final inference since the function estimates adjust appropriately to the natural log transformation and can be easily back-transformed to the original *x*-scale.

*log*(*VMT*) was originally included in the model as a variable with an unknown function to be estimated nonparametrically. However, the resulting function estimate was nearly linear. So, for ease of interpretation, as discussed in section 4.2, *log*(*VMT*) is included in the model as a linear *z*-variable. The final model we estimate is given in (1) with

$$\begin{aligned}\log[g(x_i, z_i)] = {} & \alpha + \gamma_1 \log(VMT_i) + \gamma_2 \log(SurfaceWidth_i) \\ & + \gamma_3 MinorArterial_i + \gamma_4 Collector_i + \gamma_5 RollingTerrain_i + \gamma_6 Mountainous_i \\ & + HorizCurve_i \times [\alpha_H + f_1(\log(HorizontalCurveLength_i)) + f_2(\log(HorizontalCurveDegree_i))] \\ & + VerticalCurve_i \times [\alpha_V + f_3(\log(VerticalCurveLength_i)) + f_4(\log(VerticalCurveGrade_i))]\end{aligned}$$

$$+ f_5(\log(AADT_i)) + f_6(\log(SpeedLimit_i)) + f_7(\log(AverageShoulderWidth_i)). \quad (9)$$

The indicator variable $HorizCurve_i$ (as defined in section 3) effectively removes the variables $HorizontalCurveLength_i$ and $HorizontalCurveDegree_i$ from the model when the $i^{th}$ road segment does not contain a horizontal curve. The indicator variable $VerticalCurve_i$ has a similar effect for road segments without vertical curves (i.e., without hills or valleys).

The fully linear version of the model in (9), where the functions $f_1$ through $f_7$ are all assumed to be linear, was estimated initially. The linear coefficients were then used (along with subject matter reasoning) to set the monotonicity constraints in the nonparametric version of the model. The linear coefficients are given in column 2 of Table 2.

**Table 2: Estimated coefficients**

| Variable | Estimated coefficients | |
|---|---|---|
| | Fully linear model | Nonparametric model |
| Intercept | **-13.14** (1.41) | **-15.30** (1.97) |
| log(*VMT*) | **0.69** (0.04) | **0.69** (0.04) |
| log(*SurfaceWidth*) | -0.16 (0.10) | -0.16 (0.35) |
| *MinorArterial* | -0.05 (0.10) | -0.03 (0.10) |
| *Collector* | 0.15 (0.11) | 0.20 (0.11) |
| *RollingTerrain* | -0.08 (0.09) | -0.07 (0.09) |
| *Mountainous* | -0.03 (0.34) | 0.03 (0.35) |
| *HorizCurve* ($\alpha_H$ coefficient) | -1.35 (0.83) | -1.65 (1.65) |
| *VerticalCurve* ($\alpha_V$ coefficient) | **0.93** (0.43) | 0.16 (0.44) |
| log(*HorizontalCurveLength*) | 0.12 (0.12) | --- |
| log(*HorizontalCurveDegree*) | **0.37** (0.10) | --- |
| log(*VerticalCurveLength*) | **-0.18** (0.07) | --- |
| log(*VerticalCurveGrade*) | -0.00 (0.07) | --- |
| log(*AADT*) | **0.45** (0.07) | --- |
| log(*SpeedLimit*) | 0.03 (0.21) | --- |
| log(*AverageShoulderWidth*) | -0.08 (0.10) | --- |
| $\eta$ | 0.67 (0.10) | 0.71 (0.10) |

Standard errors are in parentheses. Bolded coefficients lie more than two standard errors from zero. The dashed lines (---) indicate the relationship for this variable is estimated nonparametrically, as shown in Figure 2.

The coefficients in column 2 associated with *HorizontalCurveLength*, *HorizontalCurveDegree*, *AADT* and *SpeedLimit* are positive, so the functions $f_1$, $f_2$, $f_5$ and $f_6$ were constrained to be monotonically increasing functions. The coefficients associated with *VerticalCurveLength*, *VerticalCurveGrade* and *AverageShoulderWidth* are negative, so the functions $f_3$, $f_4$ and $f_7$ were constrained to be monotonically decreasing.

Strong subject matter reasons also justify the monotonicity constraints on *HorizontalCurveLength*, *HorizontalCurveDegree*, *AADT* and *SpeedLimit*. For example, a road segment's degree of curvature is inversely proportional to the horizontal curve's radius, and thus directly proportional to the centrifugal force experienced by the vehicle and its occupants. Further, roadway banking and side friction can be exceeded by natural forces occurring on high-degree curves, particularly on slick roadways (e.g., on rainy days) at the start and end of curves where banking is generally not fully developed. These factors should lead to an increase in crash rates and imply a monotonically increasing relationship between *HorizontalCurveDegree* and the expected number of crashes. In addition, the longer the curve is, the greater the opportunity for a driver to lose control and be involved in a crash. This should imply a monotonically increasing relationship between *HorizontalCurveLength* and the expected number of crashes.

Since all road segments in the data set are rural two-lane highways, and speed and flow values are not given by time of day or day of year (and certainly are not constant), *AADT* serves as a measure of traffic intensity or congestion. The reason is that, if vehicle miles travelled are held constant, higher *AADT* values imply more traffic on the road segment and a tighter spacing between the vehicles. Shorter spacings should increase crash counts, suggesting a monotonically increasing relationship between *AADT* and the expected number of crashes.

Finally, higher posted speed limits are expected to be associated with higher crash rates because drivers traveling at higher speeds have less time to react to avoid danger. This suggests a monotonically increasing relationship between *SpeedLimit* and the expected number of crashes.

The coefficients associated with *VerticalCurveGrade* and *AverageShoulderWidth* are negative, but both lie less than one standard error from zero. This indicates that neither variable

is statistically significant in the linear model, in terms of explaining variability in crash counts across roadway sections. This is confirmed by the flat function estimates that are obtained from the nonparametric model. The coefficient associated with *VerticalCurveLength* is negative and more than two standard errors from zero. It is not clear from a subject matter perspective what the direction of the relationship should be between *VerticalCurveLength* and crash counts, because vertical curves include both uphill and downhill road segments. However, for the same change in grade longer curves will allow for longer sight distances and ostensibly safer driving conditions. Given the strong evidence in results from the linear model, the function $f_3$ for *VerticalCurveLength* was constrained to be monotone decreasing.

The estimates of the $\gamma$-coefficients in (9), when $f_1$ through $f_7$ are estimated nonparametrically, are given in the third column of Table 2, along with their standard errors (in parentheses). Figures 2 and 3 plot the estimated functions $f_1$ through $f_7$ for the continuous explanatory variables. In both Figures 2 and 3 all variables (other than the specific $x_j$ being plotted against) are set to their median values. To interpret these figures, note that $g(x, z)$ is the mean function if $\phi = 1$. The solid curves in Figures 2a-g represent estimates of the function $\log[g(x, z)]$ plotted against $\log(x_j)$ for each of the seven $x_j$-variables. Similarly, the solid curves in Figures 3a-g represent estimates of $g(x, z)$ plotted against $x_j$ (not $\log(x_j)$). The two sets of figures represent similar information, although on different scales (in Figure 2 on a log-log scale and in Figure 3 on the original scale of the data), and are useful for different purposes, as discussed below. Finally, the dashed lines in Figures 2 and 3 are 50% confidence bands and provide a measure of uncertainty regarding the estimated function. One of the general advantages of Bayesian methods is the ready availability of confidence bands that provide meaningful measures of uncertainty. The coefficient and function estimates were obtained by combining the results from four independent runs of the MCMC sampling algorithm, each with warm-up and sampling periods of 5,000 and 20,000 iterations.

It is useful to note that Figure 2's shapes of the estimated functions for $\log[g(x, z)]$ do not depend on the median values of the non-$x_j$-variables used to create the plots. These values affect

only the vertical placement of the estimated functions. (This is not true for the estimated mean functions in Figure 3 due to the exponentiation required to obtain $g(x, z)$ rather than $\log[g(x, z)]$.) This means the inferences drawn from Figure 2 regarding shape and nonlinearity do not depend on the values of the non-$x_j$-variables used to represent typical values. The invariance property in Figure 2 to the choice of "typical" values is important when determining whether the $f_j$ functions can be modeled using linear functions, as discussed below.

**Figures 2 and 3 go here**

A comparison of column 2 and column 3 results, alongside Figures 2a-g, illustrates the similarity regarding which variables contribute explanatory power in the linear and nonparametric models. In particular, the variables log(*VMT*), log(*AADT*), log(*HorizontalCurveDegree*) and log(*VerticalCurveLength*) contribute substantially to explaining crash count variability in both model specifications. The only substantial difference between the two models is in the coefficient associated with the indicator variable for the presence of a *VerticalCurve*. In the fully linear model, the associated coefficient is more than two standard errors above zero, while in the nonparametric model it is effectively zero.

For the variables *HorizontalCurveLength*, *HorizontalCurveDegree*, *AADT*, *VerticalCurveGrade* and *AverageShoulderWidth*, the estimated functions are plotted for *x*-values through their 98$^{th}$ percentile values. The highly skewed nature of these *x*-variables and the sparcity of very large values mean that the function estimates for large *x*-values are unreliable and add little to our understanding of the core (and most common) relationships. Compounding the estimation problem for large *x*-values is the "endpoint" effect that typically occurs with nonparametric function estimation. For an *x*-value in the "middle" of the data, there are data on either side to provide information about the function estimate (because the estimation methodology provides smooth function estimates which borrow strength from nearby points). However, for large values there are no data on the right to help estimate the function value with the result that function estimates for large values are often poor. In our data set the endpoint

effect for small values of $x$ is tempered for many of the $x$-variables because there are so many small values. In most data sets, however, there is typically an endpoint problem at both ends of the $x$-range.

We now consider which functions $f_j(\log(x_j))$ in (9) can be reasonably assumed to be linear in $\log(x_j)$ and which should be estimated nonparametrically. The closer an estimated function in Figure 2 is to a straight line, the stronger the evidence that it is a linear function. More specifically, if a straight line "fits inside" the 50% confidence bands this provides evidence that the true relationship is linear. Conversely, if a straight line does not fit inside the confidence bands there is evidence the relationship is nonlinear and should be estimated nonparametrically. As an example, for the variable *HorizontalCurveDegree* in Figure 2b, a straight line does not fit into the confidence bands so this provides evidence that the true relationship is nonlinear. Further, the estimated function for *HorizontalCurveDegree* does not have an obvious nonlinear parametric functional form such as quadratic or logarithmic. This illustrates the importance of using nonparametric estimation procedures to uncover nonlinear relationships that may not be apparent from a parametric analysis. From Figure 2, there is strong evidence that the functions associated with *AADT* and *HorizontalCurveDegree* are nonlinear and moderately strong evidence that the function associated with *VerticalCurveLength* is nonlinear. Also, neither of the estimated functions for *AADT* or *HorizontalCurveDegree* has an obvious nonlinear parametric functional form.

We use 50% confidence bands rather than a higher percentage (such as 90% bands) because, from a statistical point of view, it is a more serious error to assume a function is linear when it is not than the converse. If the function is linear, the nonparametric monotone estimation procedure will still provide a good estimate, assuming sample size is adequate. This is confirmed by section 5's simulation results. Conversely, if a function is actually nonlinear, the parametric procedure assuming linearity will result in a mis-specified model and give poor function estimates relative to the nonparametric monotone procedure. This is also confirmed by section 5's simulation results. This implies that given sufficient data an analyst is often better off using the

nonparametric monotone procedure unless there is strong evidence the parametric assumptions of the model are satisfied.

We can also use the confidence bands to provide evidence regarding whether or not a specific $x$-variable is related to crash counts. If a flat (horizontal) line falls inside the 50% confidence bands for a specific $x_j$-variable then it is reasonable to conclude $x_j$ is not related to crash rates. It is fairly clear from observing Figures 2 and 3 that the variables *HorizontalCurveDegree*, *AADT* and *VerticalCurveLength* are related to crash counts while *VerticalCurveGrade*, *SpeedLimit* and *AverageShoulderWidth* are not. There is some evidence, though not overwhelming, that *HorizontalCurveLength* is related to crash counts. As with determining whether a relationship is linear, we use 50% confidence bands to be conservative and therefore reduce the probability of incorrectly concluding there is no relationship when there is one. Incorrectly removing a variable when it is related to crash counts creates a mis-specified model with potentially serious consequences. For example, in a standard Gaussian regression model, incorrectly omitting variables can result in estimated coefficients with the wrong sign (see Maddala, 1977, pp. 155-156).

The three variables that do not appear to be related to crash counts (*VerticalCurveGrade*, *SpeedLimit* and *AverageShoulderWidth*) were removed from the model, and the remaining functions were re-estimated. The function estimates for the modified model are similar to the original model with all variables included. They are not shown here, to conserve space.

We also note the lack of "wiggle" in the estimated functions. One of the advantages of imposing appropriate monotonicity constraints in the estimation procedure is the smooth functions that result. Unconstrained nonparametric procedures often produce function estimates that are "noisy," particularly if the noise-to-signal ratio in the data is high and the sample size is not sufficiently large. Unsmooth and noisy function estimates can be difficult to interpret and explain to end users.

## 4.2 Interpretation of the empirical results

This section provides an interpretation of the empirical results obtained from the model in (9) when the unknown functions and coefficients are estimated using the nonparametric monotone estimation procedure. We begin with a discussion of the estimates of the $\gamma$-coefficients and then discuss the estimated functions. First, in theory, crash counts should be directly proportional to *VMT* which means the coefficient $\gamma_1$ should be one. However, in practice this is often not the case, including in our model where the estimate is 0.69. The most likely reason $\gamma_1$ is less than one is that longer segments tend to be more homogenous in design, offering fewer surprises to drivers and therefore resulting in lower crash counts. This means that if *AADT* remains the same but segment length (and thus *VMT*) doubles, the crash counts may fall, thanks to more consistency in design features.

Second, the coefficient associated with the indicator variable for *Collector* is positive and 1.8 standard errors from zero. The positive coefficient indicates a higher number of crashes on collector roads than on arterials (both primary and secondary), which implies two-lane rural collector roads in the Seattle area are less safe than arterials. This makes good sense if speed limits and other attributes are held constant, because arterials are typically built to a higher safety standard and collectors provide more access to local land use, resulting in a higher share of vehicles entering (and leaving) the facility via driveways and cross streets, which increases the likelihood of vehicle interactions. The remaining $\gamma$-coefficients in column 3 are less than one standard error from zero. These results are consistent with the coefficients given in column 2 for the fully linear model, with the previously discussed exception of *VerticalCurve*.

The assumption of monotonicity for *AADT* is strongly supported by the estimated function shown in Figures 2c and 3c. There is an interesting non-linearity in this relationship, with a dampening of effect (change in slope) around 2,000 vehicles per day. Such a function cannot be easily modeled using the parametric functions typically employed in practice. Keeping in mind that *AADT* is a good proxy for (the inverse of) inter-vehicle spacing (if *VMT* is held constant), or equivalently, for the density of vehicles on the road segment, the estimated function in Figure 3c

implies that the expected number of crashes increases much faster for low vehicle density segments than for high density segments.

The relationship between the expected number of crashes and *HorizontalCurveDegree* shown in Figure 3b appears somewhat sinusoidal, with an inflection point very near the average value of 6.27 degrees, where the average is computed for segments containing horizontal curves. In increasing the degree of curvature from 4 to 12 degrees (and holding the other variables constant at their median values), the expected number of crashes is expected to increase by 0.06 crashes (a substantial increase, in practical terms).

The third variable that is related nonlinearly to the expected number of crashes is *HorizontalCurveLength*, although its relationship appears to be weaker than either *AADT* or *HorizontalCurveDegree*. Figure 3a indicates the expected number of crashes increases approximately 50% from 0.041 to 0.06, as the length of the curve increases from 0 to 2,000 feet. The rate of increase is greatest near zero, although this is the portion of the estimated function where the uncertainty in the estimate is greatest.

The fact that *ShoulderWidth* does not show up as an important variable in the fully linear and the nonlinear model is a bit surprising given the empirical results in some previous analyses. There are several possible reasons for this, including an incorrectly specified model, missing explanatory variables, estimation error due to the natural random variability in the data, or the possibility that for our data set *ShoulderWidth* does not contribute any explanatory power to the model. Most likely, it is simply reflecting the fact that wider lanes and shoulders can encourage driver inattention and higher speeds, resulting in more and more severe collisions, especially on two-lane roads, where driveways and crossings lead to interactions of vehicles at very different speeds. In other words, a more "forgiving roadside" can actually reduce safety (as described in Dumbaugh [2005]).

From a statistical perspective, it seems reasonable to rule out an incorrectly specified model because the Poisson-gamma model with a heterogeneity term included is a very flexible model (and is often used in these types of studies). The flexible functional forms used to model the

covariate relationships significantly reduces the possibility that incorrect parametric assumptions are made regarding these relationships. Also, various alternative specifications for incorporating *VMT* and *AADT* into the model were considered (although not shown here for brevity); but the estimated coefficients are similar, and the same coefficients are more than two standard errors from zero in each specification. The possibility of missing variables cannot be ruled out, but the variables included in our model are similar to those used in analogous studies. Also, the heterogeneity term $\phi_i$ is included to account for variability in the crash counts due to unobservable factors, such as travel speeds, impaired driving and weather conditions.

## 4.3 Out-of-sample forecasting

This section discusses forecasting and out-of-sample validation in a Poisson-gamma model. Out-of-sample validation is the "gold standard" in model fitting diagnostics, so we focus on this measure of fit. However, the same discussion applies to in-sample measures of fit. We begin with a brief discussion of forecasting in a Bayesian context. We then discuss the difficulty in obtaining accurate predictions in a Poisson-gamma model due to the substantial amount of unexplainable variability inherent to such models. The section concludes with the out-of-sample validation results obtained when the parametric and nonparametric procedures are applied to the crash data.

For the model in (1), the prediction of a future value $y_{n+1}$ is $E(y_{n+1} \mid y)$. Likewise, $Var(y_{n+1} \mid y)$ is a measure of predictive uncertainty. Analytically, based on standard Bayesian predictive inference (Bernardo and Smith, 1994), we have

$$E(y_{n+1} \mid y) = \sum_{y_{n+1}=0}^{\infty} \int \cdots \int y_{n+1} h_1(y_{n+1} \mid \phi_{n+1}, \alpha, \gamma, f_{n+1}, \eta) h_2(\phi_{n+1} \mid \eta) h_3(\alpha, \gamma, f_{n+1}, \eta \mid y) d\phi_{n+1} d\alpha \, d\gamma \, df_{n+1} d\eta$$

where $\gamma = (\gamma_1, \ldots, \gamma_7)$ and $f_{n+1} = [f(\log(x_{1,n+1})), \ldots, f(\log(x_{8,n+1}))]$. The values of $E(y_{n+1} \mid y)$ and $Var(y_{n+1} \mid y)$ cannot be computed analytically but they are straightforward to obtain from the MCMC sampling algorithm used for estimation. The reason is that both quantities are functions of the random parameters $\alpha$, $\gamma$, $f_{n+1}$ and $\eta$. Since samples from the posterior distributions of these

parameters are available from the MCMC algorithm, we can use them to estimate $E(y_{n+1} \mid y)$ and $Var(y_{n+1} \mid y)$.

To discuss issues related to the difficulty in obtaining accurate predictions, the density function $h_1$ in the integral represents the uncertainty in $y_{n+1}$ given $\phi_{n+1}$, $\alpha$, $\gamma$, $f_{n+1}$ and $\eta$. This uncertainty is due to the variability in $y_{n+1}$ generated from a Poisson-gamma distribution when the mean parameter is known, i.e., when $\mu(x_{n+1}, z_{n+1}) = \phi_{n+1} g(x_{n+1}, z_{n+1})$ is known. The density functions $h_2$ and $h_3$ taken together represent the uncertainty in the mean parameter $\mu(x_{n+1}, z_{n+1})$ given the data $y$ in the estimation sample (because $\mu(x_{n+1}, z_{n+1})$ is a function of $\phi_{n+1}$, $\alpha$, $\gamma$ and $f_{n+1}$). More specifically, the density function $h_2$ represents the variability in the future observation $\phi_{n+1} \sim \Gamma(\eta, \eta)$ given $\eta$. The density function $h_3$ represents the uncertainty regarding $\alpha$, $\gamma$, $f_{n+1}$ and $\eta$ given the data in the estimation sample (i.e., $h_3$ represents the uncertainty due to the imperfect information about $\alpha$, $\gamma$, $f_{n+1}$ and $\eta$ captured by the estimation methodology). We note that the variability captured by $h_1$ and $h_2$ cannot be controlled by the analyst. On the other hand, the uncertainty captured by $h_3$ can be reduced using a good estimation methodology.

If the total variation in a data set is dominated by unexplained variability (i.e., the variability captured by $h_1$ and $h_2$), then, regardless of whether one employs a parametric or nonparametric model, the out-of-sample forecasts will be poor. For the model and data set considered in this paper, the unexplained variability is very high given the inherent variability in Poisson-gamma data and the substantial variability in $\phi_{n+1} \sim \Gamma(\eta, \eta)$ when $\eta$ is less than one. These two sources of variability dominate the uncertainty due to the estimation methodology. This unfortunately means the model will provide poor forecasts no matter how good the estimation methodology is. This is supported by the simulation results discussed in section 5. In fact, we can show via simulation that even if the true values of $\alpha$, $\gamma$, $f_{n+1}$ and $\eta$ are known (i.e., there is no estimation error), the out-of-sample predictions for individual segments will still be very poor.

To obtain out-of-sample validation results for the crash data, the sample of 7710 observations is divided into an estimation sample with 5140 observations (two-thirds of the full sample) and a

validation sample of 2570 observations. The quality of the out-of-sample forecasts is measured using

$$RMSE = \sqrt{\sum_{i=1}^{2570}[y_i - \hat{g}(x_i, z_i)]^2}$$

where $\hat{g}(x_i, z_i)$ depends on the estimated values of $\alpha$, $\gamma$ and $f$ obtained from the estimation sample and the sum is over the 2570 observations in the validation sample. Note that the expected value of future observations given the data is $\phi_i \hat{g}(x_i, z_i)$. However, forecasts of future values of $\phi_i$ are one, since $\phi_i \sim \Gamma(\eta, \eta)$ with $E(\phi_i) = 1$ and there is no information in the *y*-values in the estimation sample about the $\phi_i$-values in the validation sample. For the nonparametric model the RMSE is 0.507, while for the fully linear model the RMSE is 0.489. As expected, the out-of-sample results for both models are very poor and nearly the same because the unexplained variability in the validation sample dominates the controllable estimation error.

## 5. Quality of the estimation procedures and out-of-sample forecasts

This section reports the results of three simulation experiments that show for three different sets of functions (linear, nonlinear and flat): (1) the relative performance of the three estimation procedures (parametric, nonparametric with monotonicity constraints and nonparametric without monotonicity constraints), and (2) the out-of-sample forecasting performance for the three estimation procedures. The results show the quality of the function estimates depends crucially on the estimation methodology that is used. They also show that the inherent variability in count data obtained from a Poisson-gamma model makes accurate out-of-sample forecasting very difficult – no matter how well the unknown functions are estimated.

The three simulations use the model in equation (1) to generate the *y*-data with

$$\log[g(x_i, z_i)] = \alpha + \gamma_1 Collector_i + f_1(\log(VMT_i)) + f_2(\log(AADT_i))$$
$$+ HorizCurve_i \times [f_3(\log(HorizontalCurveLength_i)) + f_4(\log(HorizontalCurveDegree_i))]. \quad (10)$$

For each simulation, $n$ = 7710 observations (as in the actual data set) are generated to form an estimation sample, and a second independent set of 7710 observations are generated to form a forecasting (validation) sample. The estimated coefficients and functions from the estimation sample are used to forecast the $y$-values in the forecasting sample. Both samples are generated using the same $x$, $z$, $\alpha$, $\gamma_1$, $\eta$ and function values.

Our goal is to create estimation and forecasting samples with properties similar to those of the actual crash data under various scenarios for the "true" coefficients and functions in (10). To accomplish this, the $z$-variable included in the model is the indicator variable for *Collector*. This is the significant $z$-variable from the actual data. Also, $\gamma_1$ is set to the estimated value reported in section 4.

The four $x$-variables included in the model are *VMT*, *AADT*, *HorizontalCurveLength* and *HorizontalCurveDegree*. These are $x$-variables that have estimated functions with a range greater than 0.5 (see Figure 2) and therefore contribute to the variability in the actual crash count data. The functions associated with the $x$-variables vary across the three simulations. For the first two simulations, the range of the functions associated with each variable are set to give approximately the same function ranges as the corresponding estimated functions obtained using the actual crash data (the third simulation sets $f_j(\log(x_j)) = 0$ for all four functions). For example, for the variable *HorizontalCurveDegree*, the range for $f_4$ is set to 0.89, which is the range of the estimated function for *HorizontalCurveDegree* shown in Figure 2b. The functions are discussed in more detail below.

The value of $\eta$ used to generate the $\phi_i$ values is set to 0.65. The value of $\alpha$ in each simulation is set so that the sample mean and standard deviation of the simulated $y$-data are similar to the sample mean and standard deviation of the $y$-data in the actual crash data. 10 runs are done for each simulation.

The first simulation sets the four functions in (10) to the linear functions:

$$f_j(\log(x_j)) = r_j \log(x_j), \qquad j = 1, \ldots, 4,$$

where $r_j$ is specified to give the function the appropriate range, as discussed above. The second simulation sets the four functions in (10) to the nonlinear functions:

$$f_j(\log(x_j)) = r_j \exp\left\{\left([\log(x_j) - a_j]/b_j\right)^3\right\}, \qquad j = 1, \ldots, 4$$

where $a_j$ is the minimum of the $\log(x_j)$ values across the 7710 observations, $b_j$ is the maximum of the $\log(x_j) - a_j$ values (so $0 \le [\log(x_j) - a_j] / b_j \le 1$ for all $x_j$), and $r_j$ is set as in the first simulation. The third simulation sets the four functions to be flat functions $f_j(\log(x_j)) = 0$. This represents the case where there is no relationship between $y$ and any of the $x$-variables.

The numerical measure we use to measure the quality of the three estimation procedures is the root-mean-squared-error (RMSE):

$$RMSE = \sqrt{\sum_{i=1}^{7710} [g(x_i, z_i) - \hat{g}(x_i, z_i)]^2}$$

where $g(x_i, z_i)$ is defined in (10) and $\hat{g}(x_i, z_i)$ represents the estimate of $g(x_i, z_i)$. The quantity $g(x_i, z_i)$ in the above expression is the portion of the mean function that depends on the unknown functions $f_j$ and the unknown regression coefficients $\gamma_j$. The RMSE values for three simulation scenarios are reported in Table 3.

The RMSE used to measure the quality of the out-of-sample forecasts is:

$$RMSE = \sqrt{\sum_{i=1}^{7710} [y_i - \hat{g}(x_i, z_i)]^2}$$

where the sum is over the 7710 observations in the forecasting sample and the coefficients and function values used to compute $\hat{g}(x_i, z_i)$ are from the estimation sample. As in section 4.3 the forecasts of $\phi_i$ are one. The forecasting sample is large, in order to remove the effects of random variability across observations on the forecasting *RMSE* values.

**Table 3: Root-mean-squared-errors (RMSE)**
**for the three simulation experiments and three estimation procedures**

| | Estimation method | | | |
|---|---|---|---|---|
| | True | Nonparametric with monotonicity | Nonparametric w/out monotonicity | Linear |
| | Part 1: True functions are linear functions | | | |
| RMSE for the function estimates | --- | 0.019 | 0.023 | 0.014 |
| RMSE for out-of-sample forecasts | 0.374 | 0.374 | 0.375 | 0.374 |
| | Part 2: True functions are nonlinear functions | | | |
| RMSE for the function estimates | --- | 0.025 | 0.028 | 0.053 |
| RMSE for out-of-sample forecasts | 0.387 | 0.389 | 0.390 | 0.392 |
| | Part 3: True functions are flat functions | | | |
| RMSE for the function estimates | --- | 0.011 | 0.018 | 0.011 |
| RMSE for out-of-sample forecasts | 0.394 | 0.394 | 0.394 | 0.394 |

## 5.1 Quality of the estimation procedures

The simulation results reported in Part 1 of the table, where the true functions are linear, show just a 0.005 increase in the RMSE value when the nonparametric monotone procedure is used rather than the linear (parametric) procedure. This increase is expected given the information in the linearity assumption. However, the increase is small and the nonparametric procedure does nearly as well as the linear procedure – even when the linearity assumption is satisfied. The results in Part 2 of the table, where the true functions are nonlinear, show an RMSE increase of 0.028 when the linear procedure is used rather than the nonparametric monotone procedure (a 112% increase). This is substantially larger than the difference reported in Part 1. Finally, the results in Part 3 show there is no difference between the linear and nonparametric monotone procedure when the true functions are flat functions. The three sets of simulation results imply that little is lost using the nonparametric monotone procedure when the linearity assumptions are satisfied but there is a substantial gain if the linearity assumptions are violated.

Comparing the RMSE values for the two nonparametric procedures in columns 3 and 4 shows the gain from incorporating the monotonicity constraint into the nonparametric procedure.

The biggest percentage gains are for the linear and flat functions, 21% and 64%, respectively. The exponential functions in Part 2 are rapidly increasing functions, which dissipates some of the impact of imposing monotonicity conditions on the function estimates in this case. Even so, there is a 12% increase in the quality of the function estimate. Monotonicity information is the most valuable when the function is slowly increasing or flat. Also, the importance and impact of the monotonicity information is greater in smaller sample sizes and/or when there are a large number of functions to be estimated. In other words, the less "information per function" there is in the data, the more important the monotonicity information becomes.

## 5.2 Quality of the out-of-sample forecasts

The simulation results in Table 2 show that, even if the estimate of $g(x, z)$ is very good, this is not reflected in significantly smaller forecast errors. For example, the results from the first simulation shown in Part 1 of the table (when the true functions are linear) indicate essentially no difference in the RMSE for the forecasted values obtained using the three estimation procedures. In fact, even if the true function values are known (i.e., the best possible scenario), the RMSE remains unchanged – as per the second column in the table, labeled "True". Similar results hold for the other two simulations as well. In other words, in a forecasting context the gain from better function estimates is dominated by the uncontrollable variability.

## 6. Conclusion

This paper uses a Bayesian semi-parametric estimation procedure for monotonic function estimation in a Poisson-gamma model of crash counts. The methodology uses quadratic regression splines with a Bayesian variable selection technique for choosing the knot points. In addition, monotonicity constraints are imposed on function estimates through prior distributions for unknown parameters. The model is a compromise between a fully parametric analysis and a fully nonparametric analysis. The monotonicity constraints, if appropriate, incorporate valuable information and structure into the model that often results in better estimates while still allowing for functional flexibility of the relationships. Using a semi-parametric procedure is particularly

important if the standard parametric assumptions in a Poisson-gamma model (typically linearity assumptions) are violated.

An important benefit of such procedures is that nonlinear relationships can be detected that are not observable when parametric functions are forced onto the model. These nonlinear relationships often have important subject matter implications. In terms of roadway safety, we find strong nonlinear relationships between the number of crashes and the degree of horizontal curvature and traffic intensity (*AADT*), *ceteris paribus*. When horizontal curvature is present, we find only a weak relationship to crash rates until curvature reaches approximately four degrees (of subtended angle) per 100 feet of curve (or a radius of roughly 1,400 feet), at which point the expected number of crashes begins to increase substantially. Similarly, for low levels of congestion there is a strong increase in the expected number of crashes as traffic levels rise (while holding *VMT* constant, by reducing segment length) but with an eventual reduction in the rate of increase, as congestion worsens. Neither of these relationships can be modeled using typical functions and are likely to be overlooked using a standard parametric analysis.

The monotonicity constraint incorporates important information into the model. The additional information provides better function estimates, as indicated by section 5's simulation results, as well as smoother estimates without hard-to-interpret "wiggles" that unconstrained nonparametric procedures often produce. The Bayesian methodology also gives meaningful confidence bands that provide important measures of uncertainty regarding function estimates. In addition, the confidence bands can be used to determine if standard parametric assumptions are satisfied (in which case they can be safely incorporated into the model) and to determine which specific explanatory variables are actually related to crash counts.

## 6.1 Future research

Total crashes (i.e., the sum of fatal, disabling injury, non-disabling injury, possible injury and no injury crashes) were analyzed here – rather than only fatal crashes, or fatal and disabling injury crashes. This is because the roadway sections are very short, on average, resulting in very

low (and often zero) crash counts. In general, higher crash counts provide more information, and thus more reliable function estimates. Nevertheless, the issue of distinguishing the five crash types is an important one. Modeling the different types of crashes will make for more valuable inferences regarding the impact and import of various covariates. Given the sparseness present in most crash data sets that allow for control of site-specific design attributes, one meaningful way of "borrowing strength" across crash types is to pursue a nonparametric multivariate analysis, where the five types of crash counts are treated as a vector and the "Poisson count vector" is analyzed nonparametrically. Park and Lord (2007) and Ma et al. (2008) have used parametric multivariate analysis with linear relationships to model multiple categories of crash data. Nonparametric function estimation in non-Gaussian (e.g., Poisson-gamma) multivariate models is an interesting direction for future research in both the statistics and transportation literature. At this point in time, such methods are not available, but extension of the current modeling methods to these more complex contexts is quite feasible.

It also would be helpful to incorporate the effects of weather conditions, other design features (such as sight distances, driveway frequency, population density, and clear zone width), and other factors (e.g., distance to the nearest hospital) into the analysis. Such data are not readily available in most settings, but they may be quite meaningful in terms of crash frequencies and outcomes. In general, the Bayesian approach employed here enables substantial specification flexibility for more appropriate modeling – and interpretation – of count-based relationships.

## References


Abdel-Aty, M. A., Radwan, A. E., 2000. Modeling traffic accident occurrence and involvement. *Accident Analysis and Prevention* 32(5), 633-642.

Bernardo, J., Smith A. F. M., 1994. *Bayesian Theory*. John Wiley and Sons, London, England.

Biller, C., 2000. Adaptive Bayesian regression splines in semiparametric generalized linear model. *Journal of Graphical and Computational Statistics* 9 (1), 122-140.

Biller, C., Fahrmeir, L., 2001. Bayesian varying-coefficient models using adaptive regression splines. *Statistical Modeling* 1, 195-211.

Casella, G., George, E. I., 1992. Explaining the Gibbs sampler. *American Statistician*, 46, 167-174.

Chin, H. C. C., Quddus, M. A., 2003. Applying the random effect negative binomial model to examine traffic accident occurrence at signalized intersections. *Accident Analysis and Prevention* 35(2), 253-259.

Dumbaugh, E., 2005. Safe Streets, Livable Streets. *Journal of the American Planning Association* 71(3): 283-300.

Dunson, D. B., 2005. Bayesian semiparametric isotonic regression for count data. *Journal of the American Statistical Association* 100 (470), 618-627.

Fahrmeir, L., Osuna, L., 2007. Structured count data regression. Working paper, Department of Statistics, Ludwig Maximilian Universitt Munchen.

Friedman, J., Tibshirani, R., 1984. The monotone smoothing of scatterplots. *Technometrics* 26 (3), 243-250.

Gelfand, A. E., Smith, A. F. M., 1990. Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association* 85 (410), 398-409.

Gurmu, S., Rilstone, P., Stern, S., 1999. Semiparametric estimation of count regression models. *Journal of Econometrics* 88(1), 123-150.

Kockelman, K., Bottom, J., Kweon, Y., Ma, J., Wang, X., 2006. "Safety Impacts and Other Implications of Raised Speed Limits on High-Speed Roads." Final Report for National Cooperative Highway Research Program Project #17-23. Digest 303 available at http://www.ce.utexas.edu/prof/kockelman/public_html/NCHRPSpeedLimits17-23.pdf.

Kumara, S. P., Chin, H. C., 2003. Modeling accident occurrence at signalized T intersections with special emphasis on excess zeros. *Traffic Injury Prevention* 4(1), 53-57.

Kweon, Y.J., Kockelman, K., 2005. The safety effects of speed limit changes: use of panel models, including speed, use, and design variables. *Transportation Research Record 1908*, 148-158.

Lord, D., Washington, S. P., Ivan, J. N., 2005a. Poisson, Poisson-gamma and zero-inflated regression models of motor vehicle crashes: balancing statistical fit and theory. *Accident Analysis and Prevention* 37(1), 35-46.

Lord, D., Manar, A., Vizioli, A., 2005b. Modeling Crash-Flow-Density and Crash-Flow-V/C Ratio for Rural and Urban Freeway Segments. *Accident Analysis and Prevention* 37 (1), 185-199.

Ma, J., Kockelman, K., Damien, P., 2008. A Multivariate Poisson-Lognormal Regression Model for Prediction of Crash Counts by Severity, using Bayesian Methods. *Accident Analysis and Prevention* 40 (3), 964-975.

Maddala, G. S., 1977. *Econometrics*, McGraw Hill, New York.

Mahamassani, H. S., Baaj, M. H., Tong, C. C., 1988. Characterization and evolution of spatial density patterns in urban areas. *Transportation* 15 (3), 233-256.

Maher, M. J., 1990. A bivariate negative binomial model to explain traffic accident migration. *Accident Analysis and Prevention* 22 (5), 487–498

Mammen, E., 1991. Estimating a smooth monotone regression function. *Annals of Statistics* 19, 724-740.

Miaou, S. P., 1994. The relationship between truck accidents and geometric design of road sections: Poisson versus negative binomial regressions. *Accident Analysis and Prevention* 26 (4), 471-482.

Neelon, B., Dunson, D. B., 2004. Bayesian isotonic regression and trend analysis. *Biometrics* 60, 398-406.

Park, E. S., Lord, D., 2007. Multivariate Poisson-Lognormal Models for Jointly Modeling Crash Frequency by Severity. *Transportation Research Record* No. 2019: 1-6.

Ramsay, J. O., 1998. Estimating smooth monotone functions. *Journal of the Royal Statistical Society*, *Series B* 60 (2), 365-375.

Schipper, M., Taylor, J. M. G., Lin, X., 2007. Bayesian generalized monotonic functional mixed models for the effects of radiation dose histograms on normal tissue complications. *Statistics in Medicine* 26 (25), 4643-4656.

Shively, T. S., Sager, T. W., Walker, S. G., 2009. A Bayesian approach to nonparametric monotone function estimation. *Journal of the Royal Statistical Society*, *Series B* 71 (1), 159-175.

Shively, T. S., Walker, S. G., Damien, P., 2009. Monotone nonparametric function estimation in non-Gaussian models, under review.

Smith, M., Kohn, R., 1996. Nonparametric regression using Bayesian variable selection. *Journal of Econometrics* 75 (2), 317-343.

Ulfarsson, G. F., Shankar, V. N., 2003. Accident count model based on multiyear cross-sectional roadway data with serial correlation. *Transportation Research Record 1840*, 193-197.

Wahba, G., 1978. Improper priors, spline smoothing and the problem of guarding against model errors in regression. *Journal of the Royal Statistical Society, Series B* 40 (3), 364-372.

Wong, C., Kohn, R., 1996. A Bayesian approach to additive semiparametric regression. *Journal of Econometrics*, 74 (2). 209-235.

Wright, I., Wegman, E., 1980. Isotonic, convex, and related splines. *Annals of Statistics* 8 (5), 1023-1035.

Xie, Y., Zhang, Y., 2008. Crash frequency analysis with generalized additive models. *Transportation Research Record 2061*, 39-45.

## Acknowledgement

The authors wish to thank two anonymous reviewers for their comments and Dr. Jianming Ma for his assistance in data assembly.

## Appendix

This appendix outlines the MCMC sampling algorithm used to implement the nonpamametric monotone estimation procedure employed in the paper. Full details of the algorithm as it applies to estimating monotonic functions in the class of models that have log-concave likelihood functions, of which the Poisson-gamma is a member, can be found in Shively, Walker and Damien (2009). The algorithm discussed below is for the Poisson-gamma model in equation (1) with a single continuous explanatory variable $x$ with unknown function $f(x)$ and a single linear regressor $z$ with coefficient $\gamma$ but it is straightforward to generalize to multiple functions and multiple linear regressors.

Let $\pi$ represent the likelihood function

$$\pi(\alpha, f(x_i), \gamma, \phi_i \mid y) = exp\{-\phi_i \exp[\alpha + f(x_i) + \gamma z_i]\} \frac{\{\phi_i \exp[\alpha + f(x_i) + \gamma z_i]\}^{y_i}}{y_i!}. \qquad \text{(A1)}$$

Given $J = (J_1, \ldots, J_{m+2})$, the quadratic regression spline function $f_m$ in (3) that approximates $f(x)$ can be written in matrix notation as $f_m = X_J \beta_J$ where $f_m = (f_m(x_1), \ldots, f_m(x_n))'$, and $X_J$ and $\beta_J$ are defined as in section 2.1. The constraints on the $\beta_j$ values to ensure that the resulting function is

non-decreasing depend on the $J_j$ values. For example, if $J_j = 1$ for all $j$ then the constraints are $\beta_1 \geq 0$, $\beta_1 + 2\tilde{x}_1\beta_2 \geq 0$, and $\beta_1 + 2\tilde{x}_{j+1}\beta_2 + 2\sum_{k=1}^{j}(\tilde{x}_{j+1} - \tilde{x}_k)\beta_{k+2} \geq 0, j = 1, \ldots, m$ (with $\tilde{x}_{m+1} = 1$). In general, the linear restrictions on the elements of $\beta_J$ required to ensure the function is non-decreasing can be written as $\delta_J = L_J\beta_J$, where $L_J$ is a lower triangular matrix that depends on $J$ and the $\tilde{x}_j$ values, and each element of $\delta_J$ must be greater than or equal to zero. The portion of the $\delta_J$-parameter space that guarantees a non-decreasing function is the multi-dimensional generalization of the first quadrant. Setting the variance matrix in the prior for $\beta_J$ discussed in section 2.2 to $\Omega_J = L_J^{-1}(L_J')^{-1}$ gives a distribution for $\delta_J$ that is a $N(0, cI)$ distribution constrained to the multi-dimensional generalization of the first quadrant, where $I$ is the identity matrix with appropriate dimensions.

To make the model analytically tractable we re-parameterize to give $f_m = W_J\,\delta_J$ where $W_J = X_J L_J^{-1}$. Using this parameterization with the likelihood function in (A1) gives

$$\pi(y \mid \alpha, J, \delta_J, \gamma, \phi) = \prod_{i=1}^{n}\pi(y_i \mid \alpha, J, \delta_J, \gamma, \phi)$$

$$= \exp\{-s(y, \alpha, J, \delta_J, \gamma, \phi)\}$$

where

$$s(y, \alpha, J, \delta_J, \gamma, \phi) = \sum_{i=1}^{n}\left[\phi_i \exp\{\alpha + w_{Ji}\delta_J + \gamma z_i\} - y_i\{\log(\phi_i) + \alpha + w_{Ji}\delta_J + \gamma z_i\}\right]$$

and $w_{Ji}$ represents the $i$-th row of $W_J$. The corresponding posterior distribution is

$$\pi(\alpha, J, \delta_J, \gamma, \phi, \eta \mid y) \propto \exp\{-s(y, \alpha, J, \delta_J, \gamma, \phi)\}\,\pi(\alpha)\pi(J, \delta_J)\pi(\gamma)\pi(\phi \mid \eta)\pi(\eta)$$

where $\pi(\alpha)$, $\pi(J, \delta_J)$, $\pi(\gamma)$, $\pi(\phi \mid \eta)$ and $\pi(\eta)$ represent prior distributions.

The key idea in the sampling algorithm is to introduce a latent variable $v$ such that

$$\pi(v, \alpha, J, \delta_J, \gamma, \phi, \eta \mid y) \propto e^{-v} I(v > s(y, \alpha, J, \delta_J, \gamma, \phi))$$

$$\times\, \pi(\alpha)\pi(J, \delta_J)\pi(\gamma)\pi(\phi \mid \eta)\pi(\eta)$$

where $I$ is the indicator function with $I(v > s(y, \alpha, J, \delta_J, \gamma, \phi)) = 1$ if $v > s(y, \alpha, J, \delta_J, \gamma, \phi)$ and =0 otherwise.

For notational purposes, let $J_{(-j)} = J$ without the $j$-th element, $\delta_{(-j)} = \delta$ without the $j$-th element, and $\phi_{(-i)} = \phi$ without the $i$-th element. Using this notation, the MCMC sampling algorithm described below is used to carry out function estimation. For excellent discussions of Bayesian inference using MCMC methods, see Gelfand and Smith (1990) and Casella and George (1992).

(0) Start with some initial values $v^{[0]}$, $\alpha^{[0]}$, $J^{[0]}$, $\delta^{[0]}$, $\gamma^{[0]}$, $\phi^{[0]}$ and $\eta^{[0]}$;

(1) Generate $v$ conditional on $\alpha$, $J$, $\delta_J$, $\gamma$, $\phi$, $\eta$, $y$;

(2) Generate $(J_j, \delta_j)$ conditional on $v$, $\alpha$, $J_{(-j)}$, $\delta_{(-j)}$, $\gamma$, $\phi$, $\eta$, $y$; $j = 1, \ldots, m + 2$; $(J_j, \delta_j)$ will be generated as a block;

(3) Generate $\alpha$ conditional on $v$, $J$, $\delta_J$, $\gamma$, $\phi$, $\eta$, $y$;

(4) Generate $\gamma$ conditional on $v$, $\alpha$, $J$, $\delta_J$, $\phi$, $\eta$, $y$;

(5) Generate $\phi_i$ conditional on $v$, $\alpha$, $J$, $\delta_J$, $\gamma$, $\phi_{(-i)}$, $\eta$, $y$; $i = 1, \ldots, n$;

(6) Generate $\eta$ conditional on $v$, $\alpha$, $J$, $\delta_J$, $\gamma$, $\phi$, $y$;

Let $\delta^{[l]}$ and $J^{[l]}$ be the iterates of $\delta$ and $J$ in the sampling period. Then an estimate of the posterior mean of the $i$-th element of $f_m$, and therefore an estimate of $f_m(x_i)$ is $\frac{1}{L}\sum_{l=1}^{L} w_{J^{[l]},i}\,\delta^{[l]}_{J^{[l]}}$ .

We briefly outline the generation of $v$ and $(J_j, \delta_j)$ in steps 1 and 2. As shown below, $\delta_j \mid J_j = 1$ is generated in step 2 from a constrained normal distribution that depends on the roots of the function

$$\tilde{s}(\delta_j) = s(\delta_j; y, \alpha, J_j = 1, J_{(-j)}, \delta_{(-j)}, \gamma, \phi) - v$$

in the $\delta_j$-space, where $\tilde{s}(\delta_j)$ is a convex function in $\delta_j$, and $\alpha$, $J_{(-j)}, \delta_{(-j)}$, $\gamma$, $\phi$ and $v$ are held constant at their previously generated values. $\alpha$, $\gamma$ and $\phi_i$, $i = 1, \ldots, n$, are generated similarly to $\delta_j \mid J_j = 1$ in that they reduce to generating random variates from a constrained normal distribution for $\alpha$ and $\gamma$, and from a constrained gamma distribution for $\phi_i$, with the constraints in each case

depending on the roots of $s(y,\alpha,J,\delta_J,\gamma,\phi)-v$ in the $\alpha$, $\gamma$ and $\phi_i$ spaces, respectively. The parameter $\eta$ is straightforward to generate (although tedious) using a Metropolis-Hasting step. Details of the entire algorithm and its implementation are available from the authors on request.

1. Generate $v$:

Generate $v^*$ from an $Exp(1)$ distribution and compute $v = v^* + s(y, \alpha, J, \delta_J, \gamma, \phi)$.

2. Generate $(J_j, \delta_j)$; $j = 1, \ldots, m+2$:

$(J_j, \delta_j)$ are generated as a block. To generate these values, we have

$$\pi(J_j, \delta_j | \cdots) \propto I[v > s(y, \alpha, J, \delta_J, \gamma, \phi)]\pi(\delta_j | J_j)\pi(J_j) \tag{A2}$$

where "$\cdots$" represents $(y, v, \alpha, J_{(-j)}, \delta_{(-j)}, \gamma, \phi, \eta)$. For $J_j = 0$, this yields

$$\pi(J_j = 0 | \cdots) \propto I\left(v > s(y,\alpha,J_j = 0, J_{(-j)},\delta_{(-j)},\gamma,\phi)\right)\pi(J_j = 0)$$

To find $\pi(J_j = 1 | \cdots)$, we integrate $\delta_j$ out of the density function in (A2) with $J_j$ set to one. To accomplish this, note that

$$\pi(J_j = 1, \gamma_j | \cdots) \propto I[\tilde{s}(\delta_j) < 0]\pi(\delta_j | J_j = 1)\pi(J_j = 1)$$

where $\pi(\delta_j | J_j = 1)$ is the prior distribution for $\delta_j$ given $J_j = 1$ and $\tilde{s}(\delta_j)$ is defined above. If $\tilde{s}(\delta_j)$ is greater than zero for all $\delta_j \geq 0$, then $\pi(J_j = 1 | \cdots) = 0$. Otherwise, let $a^*_{\min}$ and $a_{max}$ represent the roots of this function. Noting that the monotonicity restriction is $\delta_j \geq 0$, let $a_{min} = \min\{0, a^*_{\min}\}$. Then

$$\pi(J_j = 1, \delta_j | \cdots) \propto I\left(a_{\min} < \delta_j < a_{\max}\right)\pi(\delta_j | J_j = 1)\pi(J_j = 1). \tag{A3}$$

$\delta_j$ can now be integrated out analytically to give $\pi(J_j = 1 | \cdots)$. Given $\pi(J_j = 0 | \cdots)$ and $\pi(J_j = 1 | \cdots)$, $J_j$ is generated from a Bernoulli distribution.

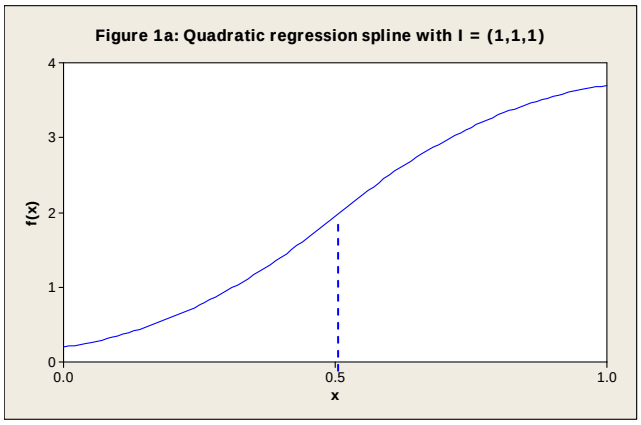
If $J_j = 0$, $\delta_j$ does not need to be generated. If $J_j = 1$, then $\delta_j \mid J_j = 1$ has a $N(0, c)$ distribution constrained to the interval $(a_{min}, a_{max})$. Well-known algorithms exist for generating from univariate constrained normal distributions.

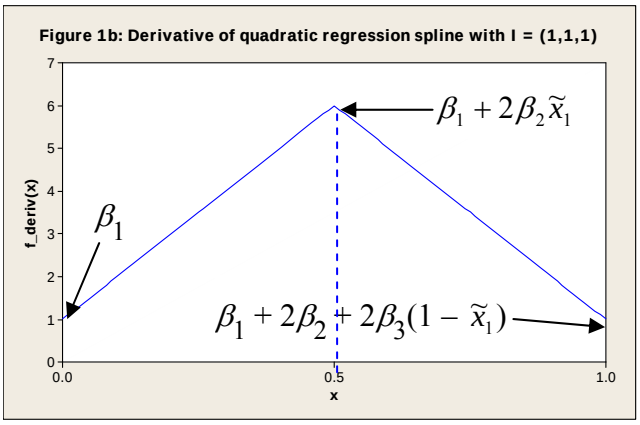
Shively, Walker and Damien (2009) give a rejection sampling algorithm with a high acceptance rate for generating $(J_j, \delta_j)$ that avoids the direct calculation of the roots $a_{min}$ and $a_{max}$. The rejection algorithm finds tight bounds $b_{min}$ and $b_{max}$ such that $b_{min} \leq a_{min}$ and $a_{max} \leq b_{max}$. Using these bounds, $\pi(J_j = 1, \delta_j \mid \cdots)$ in (A3) is approximated by

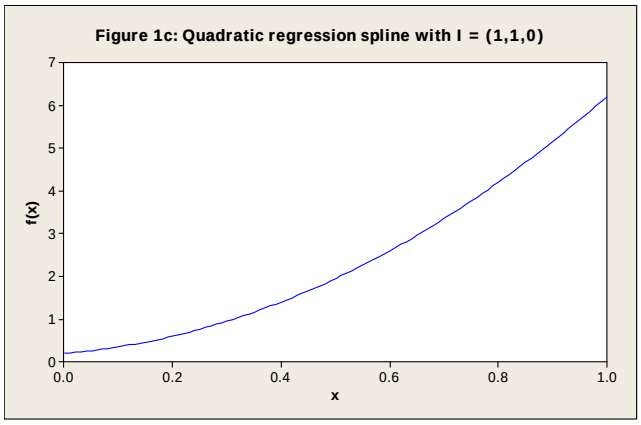
$$\pi_{Approx}(J_j = 1, \delta_j \mid \cdots) \propto I\left(b_{\min} < \delta_j < b_{\max}\right)\pi(\delta_j \mid J_j = 1)\pi(J_j = 1).$$

The approximation is exact for all $\delta_j$ values except $b_{min} < \delta_j < a_{min}$ and $a_{max} < \delta_j < b_{max}$. Thus, if the bounds $b_{min}$ and $b_{max}$ are tight (which they almost always are given the well-behaved function $\tilde{s}(\delta_j)$) the rejection sampling algorithm will have a high acceptance rate. This significantly increases the computationally efficiency of the MCMC algorithm and improves its numerical stability (numerical problems arise in the exact method if the roots $a_{min}$ and $a_{max}$ are not computed with sufficient accuracy – however, computing them with sufficient accuracy makes the sampling algorithm too computationally intensive to use in practice with large data sets). The rejection sampling algorithm can also be modified to apply to the generation of $\alpha$, $\gamma$ and $\phi_i$, $i = 1, \ldots, n$.

# Figure 1: Quadratic regression splines and derivatives



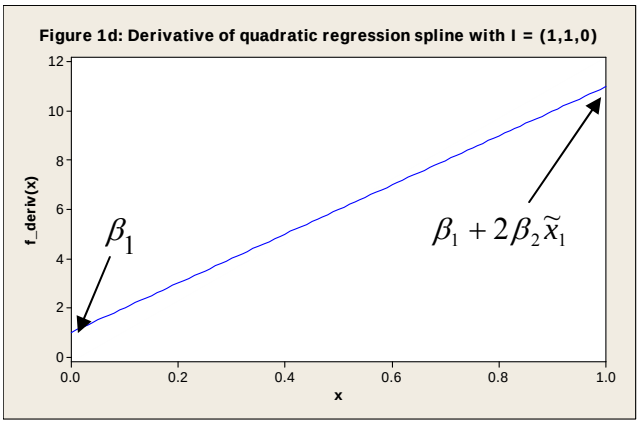
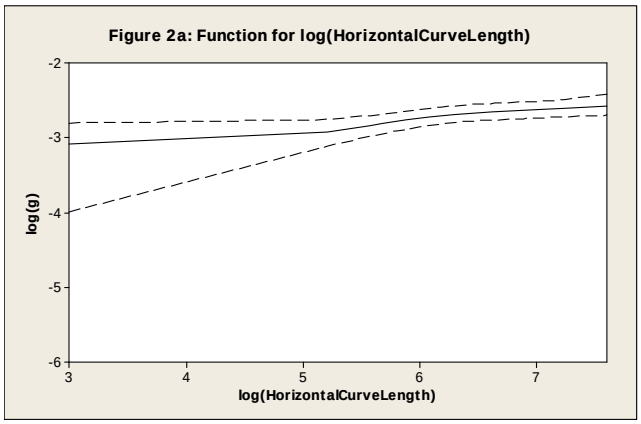
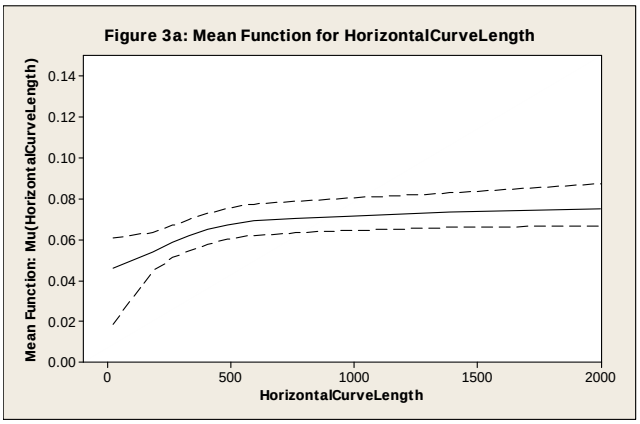



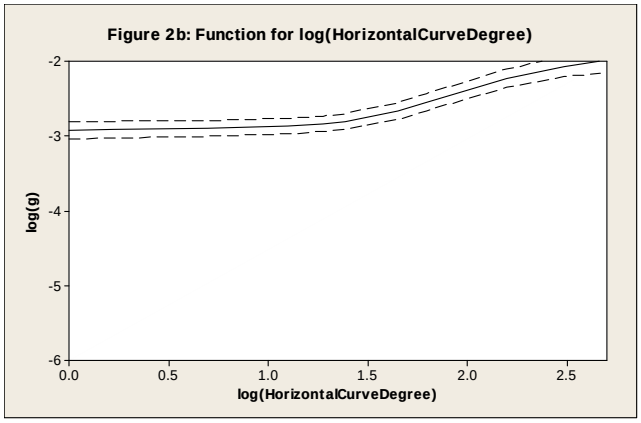
**Figure 2: Estimated functions on the log-log scale**

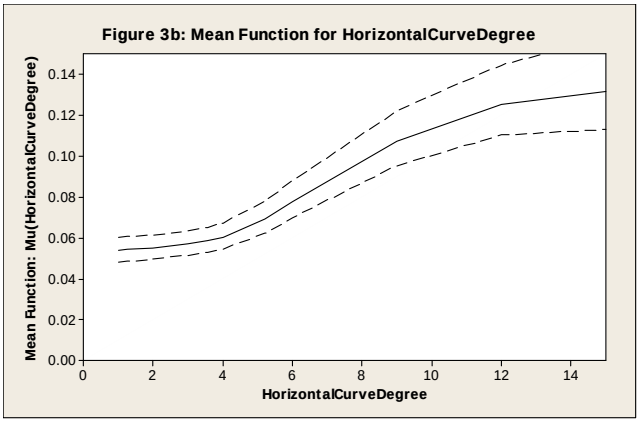
**Figure 3: Estimated mean functions on the original *x*-scale**

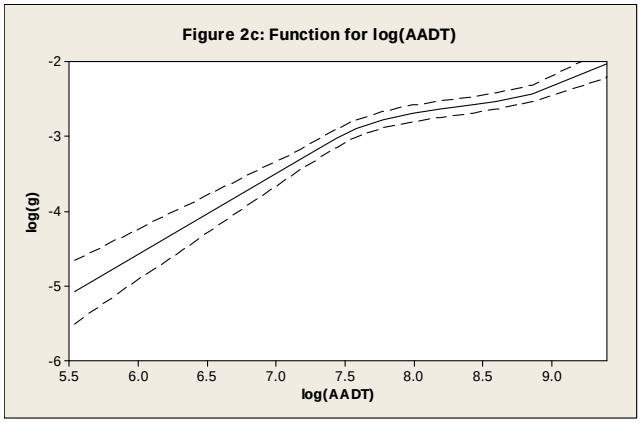
In both sets of figures, the solid curve is the function estimate and the dashed lines are 50% confidence bands.

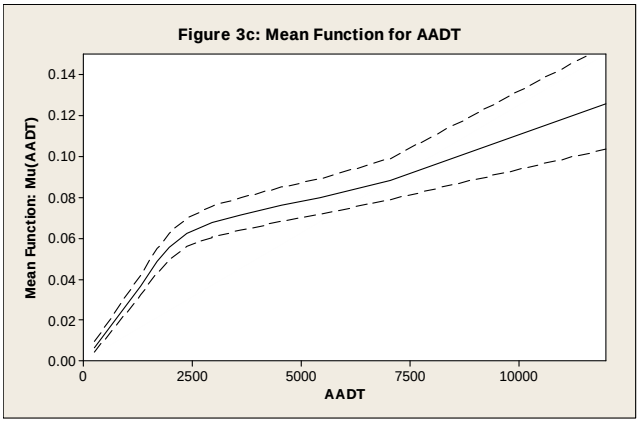


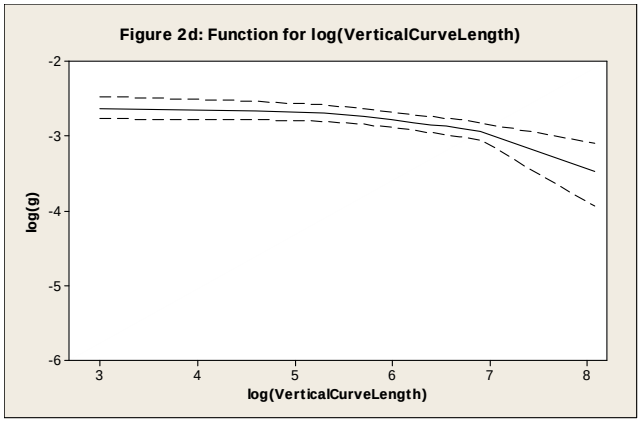



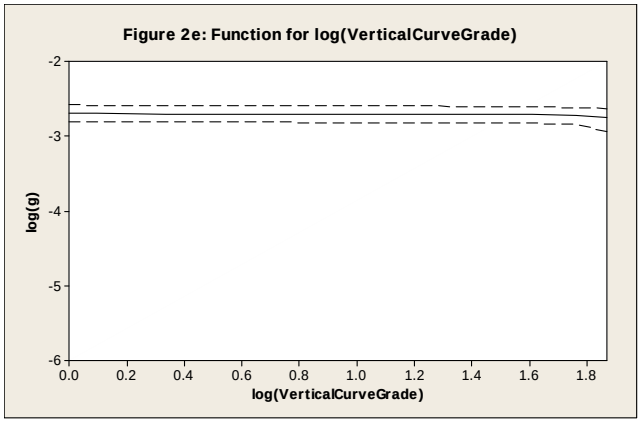



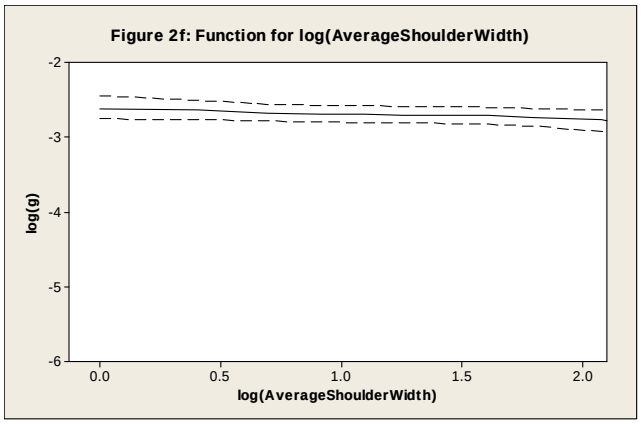



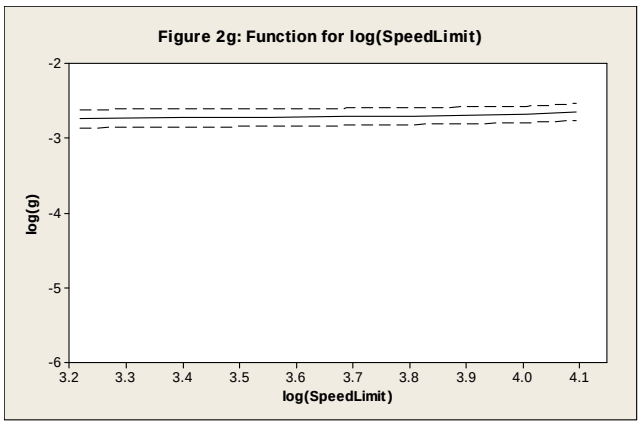



## Figure 2 (continued)




## Figure 3 (continued)