TRR
JOURNAL OF THE TRANSPORTATION RESEARCH BOARD

People

Transportation Research Record
2019, Vol. 2673(2) 743–757

DOI: 10.1177/0361198119825834
journals.sagepub.com/home/trr
SAGE

Effects of Signalization at Rural Intersections Considering the Elderly Driving Population

Lishengsa Yue[1], Mohamed Abdel-Aty[1], Jaeyoung Lee[1], and Ahmed Farid[1]

Abstract

The main objective of this study is to quantify the safety impacts of signalization at Florida's rural three-leg and four-leg stop-controlled intersections by estimating crash modification factors. The intersections are those in which stop signs are provided for the minor approaches or all-way stop-controlled intersections. The crash modification factors (CMF) are estimated using the cross-sectional method. Generalized linear models (GLM) and multivariate adaptive regression spline models (MARS) are employed with four years of Florida crash data. The *K*-nearest neighbor (KNN) and *K*-means clustering algorithms are implemented to identify the comparison sites which are sites having similar characteristics as those of the converted intersections. Furthermore, the quasi-induced exposure method is used to evaluate separately the safety effects of signalization for elderly and non-elderly drivers. According to the results, signalization contributes to an increase in property damage only (PDO) and rear-end crashes. In addition, elderly drivers are more at risk of being involved in such crashes than non-elderly drivers. In particular, at rural four-leg two-way stop-controlled intersections, signalization decreases crash severity, and a greater percentage of the decrease is observed for the elderly drivers than non-elderly especially when the intersection has a high level of major-road average annual daily traffic (AADT) and elderly driver proportion. This study also demonstrates that the MARS model shows a better model fit than the GLM model due to its strength in capturing nonlinear relationships and interaction effects among variables. This study's findings have implications for both practitioners and researchers.

A crash modification factor (CMF) is a multiplicative factor used to estimate the change in crash counts after implementing a road safety countermeasure at a specific site. The percentage reduction in crash counts could be expressed as $(1 - CMF) \times 100\%$. Part D of the Highway Safety Manual (HSM) (*1*) provides a series of CMFs for converting a stop-controlled intersection to a signalized intersection based on the previous road safety evaluations (*2*, *3*). However, these CMFs, particularly those of rural intersections, are irrespective of the number of legs. In addition, no CMFs are available for all-way stop-control intersections in the HSM. The CMFs should be developed with care because, on the one hand, different rural intersections may be designed based on different considerations and warrants, while, on the other hand, latent spatial and temporal features vary among intersections. Therefore, the changes in the safety effects due to signalization may also be different for each intersection undergoing the signalization. Thus, it is necessary to distinguish the CMFs between the different types of intersections.

Some research has shown the need to estimate CMFs at different sites separately, not simply to propose an overall CMF for all sites. McGee et all (*4–6*) found that the effect of signalization at an intersection depends on the crash history at the intersection, the number of approaches (three- or four-leg), the signal timing, the operation speed and the total entering volume. Sacchi et al. (*7*) employed a full Bayes before-after methodology to account for spatial and temporal characteristics when evaluating the safety effects of signalization. In another study, Srinivasan et al (*8*) demonstrated that the safety effects of signalization with and without left-turn lanes on two-lane roads would make a difference. The differences in CMFs among sites may not always be significantly different. Harkey et al. (*9*) detected slight

[1]Department of Civil, Environmental, and Construction Engineering, University of Central Florida, Orlando, FL

Corresponding Author:
Address correspondence to Lishengsa Yue: 2017lishengsa@Knights.ucf.edu

variations in CMFs developed for different intersection types and crash severities.

As for signalization effects, most previous research supports the assertion that the count of rear-end crashes surges while that of angle crashes diminishes after signalization. Although the frequency of total crashes may increase, that of severe crashes is reduced (*1–4*, *9*, *10*).

The cross-sectional method is often used to calculate CMFs because data requirements are not extensive (*11*) and no data is required regarding the pre-signalization period. Thus, the method is chosen to be implemented in this study. Another advantage of the cross-sectional method is that the effects of the crash contributing factors, other than that which describes the deployment of the safety countermeasure (signalization), on crash frequency can be controlled (*12*). An essential step to be followed when applying the cross-sectional method is to develop safety performance functions (SPFs) to predict crashes. One of the modeling configurations widely used to develop SPFs is the generalized linear model (GLM) having the error term following the negative binomial distribution because of its ability to address over-dispersion. However, the effect of one independent variable on crash frequency is assumed to be fixed. That is, the predictor is multiplied by a fixed coefficient. Thus, the GLM may not be able to capture some nonlinear relationships and interaction effects among variables. Therefore, a machine learning method, namely the multivariate adaptive regression splines (MARS), is used to estimate the SPF. The MARS model uses flexible hinge functions or the product of the hinge functions to capture different effects from different regions of variable value space on the response variable (*13*). The MARS model may have the potential to demonstrate better performance than the traditional GLM model. Unlike other machine learning methods, the MARS model is not a black-box method and is able to output interpretable results. The results illustrate the relationships among variables based on fixed coefficients which makes it possible to estimate CMFs. Researchers have applied the MARS model to estimated CMFs for changes in the median width, the inside shoulder width and the outside shoulder width of urban freeway interchange influence areas. That is for both total and injury crash frequencies (*14*). Furthermore, Park and Abdel-Aty (*11*) computed CMFs for combined road-side treatments and demonstrated that the MARS model showed a better goodness-of-fit than the commonly used negative binomial model. However, few researchers applied the MARS model to evaluate the changes in crash trends after signalization. In this study, the MARS technique is employed for computing CMFs that quantify the safety effects of signalization.

Because the observed differences in crash frequencies may be due to factors other than signalization, the cross-sectional method requires that the crash counts of the sites being signalized be compared with those of similar sites (*15*). The known factors, other than signalization, are controlled during the SPF development process. These variables include geometric design factors (i.e., skew angle, street lighting, pedestrian crosswalk, number of lanes serving different movements, channelized lane design, and posted speed limit), traffic conditions (i.e., traffic volume, truck proportion) and other information (i.e., whether near a school zone). However, currently, there is no robust solution to control other unknown factors (*15*). In this study, the *K*-nearest neighbors (KNN) and *K*-means clustering algorithms are used in conjunction to select intersections with similar major- and minor-road traffic volumes; this indirectly accounts for unknown factors. To the best of our knowledge, this is the first time this combined selection method is used in estimating CMFs.

Another concern regarding Florida is whether the crash experiences of elderly drivers are different from those of non-elderly drivers. Florida has a large proportion of elderly drivers (65 years and older) and the safety issues related to elderly drivers need to be considered. It is not known whether elderly drivers are more vulnerable than non-elderly drivers at intersections. This poses the question of whether special considerations are needed at specific locations having considerable proportions of elderly drivers when signalizing stop-controlled intersections. This study is aimed at investigating whether elderly drivers are more at risk than non-elderly drivers at signalized intersections that were previously stop-controlled intersections. Stop-controlled intersections include those in which stop signs are deployed on the intersections' minor approaches only and all-way stop-controlled intersections. Since a high proportion of the elderly in a region does not necessarily indicate a high proportion of elderly drivers, the quasi-induced exposure method (*16–18*) is used to capture the real proportion of elderly drivers at an intersection. The quasi-induced exposure can aid in capturing the relationships between the crash frequency and crash contributing factors for the particular driving population examined.

The remainder of this paper is organized as follows. The analysis methodologies are explained in the following section. A description of the data used for this study is then provided followed by a presentation of the results. The paper concludes with a discussion of the results and a conclusion.

Method

Cross-Sectional Method

The cross-sectional method is used to estimate the CMF of a treatment when it is difficult to directly analyze the change in crash frequency before and after the implementation of the treatment. This is usually the case if 1) the

date of the treatment implementation is unknown or 2) the crash data for the period before and after the treatment implementation are not available (*12*). The cross-sectional method requires SPFs to predict crashes at a specific site. The SPFs account for the factors related to crashes including geometric design, traffic volume and other factors. A commonly used SPF structure is the GLM with an error term that follows the negative binomial distribution. It is employed for this study with the functional form as follows:

$$N_{predictedi} = \exp(\beta_0 + \beta_1 \times \ln(AADT_i) + \beta_2 \times (control\,type_i) + \ \ldots \ + \beta_k x_{ki}) \quad (1)$$

Note that $N_{predictedi}$ is the predicted crash frequency at a specific site, i, while the term, $AADT_i$, is the annual average daily traffic at the site. It may be categorized as minor-road AADT and major-road AADT. The predictor, $control\,type_i$, is the control type of the site. It may be classified as either two-way stop control, all-way stop control or signal control. The term, x_{ki}, is the kth predictor of the site. The number of crash years is often used as the offset term or part of the offset term in the model.

The CMF of one type of treatment can then be calculated by taking the ratio of the predicted crash frequency at the site with the treatment to the predicted crash frequency at the site without the treatment using the following function:

$$CMF_i = \exp(\beta_k(x_{kt} - x_{kb})) \quad (2)$$

The term, x_{kt}, is the predictor, k, of the site with the treatment and $x_{kb}$is the predictor k of the site without the treatment., that is, the base condition.

In this study, the control type of an intersection is a binary variable. For signal control, the control type predictor is set as 1 while for stop control the control type is set as 0 and the CMF is simply $\exp(\beta_k)$.

For the cross-sectional method to be robust it is important that all locations are similar to each other in all factors that affect crash risk except for signalization (*15*). In this study, this requires that the selected stop-controlled intersections and signalized intersections should be comparable with each other.

MARS

The MARS are extensions to linear models which may be used to model complex relationships (e.g., nonlinear and interaction effects) among variables (*19*). The MARS partitions the independent variable space into several regions using "knots" and fits different spline models (i.e., "basis functions") to each region. This process is also known as a multivariate piecewise regression (*20*) and is appropriate when independent variables belonging to different clusters have different relationships (*21*). The basis MARS model is defined as follows (*22*):

$$\widehat{y_0} = a_0 + \sum_{m=1}^{M} a_m B_m(x) \quad (3)$$

The term, $\widehat{y_0}$, is the predicted response while a_0 is the constant of the basis function and a_m is the coefficient of the mth basis function. The term, $B_m(x)$, is the mth basis function which may be a hinge function or a product of two or more hinge functions. The hinge function has the form of $\max(0, x - constant)$ or $\max(0, constant - x)$. In this study, the natural log form of the MARS model, that is, $\hat{y} = \exp(\widehat{y_0})$, is used to develop the CMFs.

Three steps are needed to fit a MARS model (*22*). The first step is a constructive phase and a forward pass is used to introduce basis functions into several regions of the space of independent variables. This step usually introduces more basis functions than needed which may overfit the data. The second step is the pruning phase. Its goal is to find a subset of those basis functions in the first step which produces the least generalized cross-validation error ($GCVE$). The $GCVE$ is estimated using

$$GCVE(M) = \frac{1}{n}\frac{\sum_{i=1}^{n}(y_i - \hat{y})^2}{\left(\frac{1-C(M)}{n}\right)^2}$$
$$C(M) = M + dM \quad (4)$$

The final step is to select the optimum model from a sequence of recommended models based on the fitting results. In Equation 4, n is the number of observations while y_i is the response for observation i and $\hat{y}$ is the predicted response for the observation. The function $C(M)$ is the complexity penalty function while M is the number of non-constant basis functions. The parameter, d, is the user-defined cost of each basis function used for optimizing the MARS model.

KNN and K-Means Clustering

The KNN algorithm is a non-parametric method which is used to identify K most similar observations to a given observation. The similarity of two observations is measured according to the distance between them by considering all of their attributes (*23*). A commonly used distance is the Euclidean distance given by the following:

$$d(x, x') = \sqrt{(x_1 - x_1')^2 + (x_2 - x_2')^2 + \ \ldots \ + (x_n - x_n')^2} \quad (5)$$

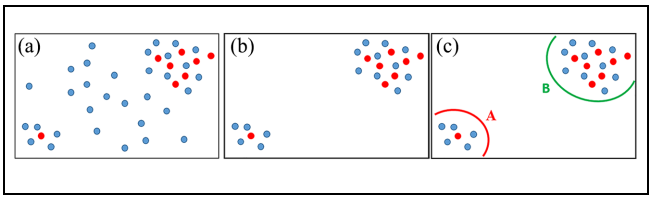

Figure 1. Using *K*-nearest neighbors (KNN) and *K*-means clustering to select similar stop-controlled intersections and signalized intersections: (*a*) original intersection sets; (*b*) intersections selected by KNN; and (*c*) intersections selected by *K*-mean.

Note: The red points represent the signalized intersections and the blue point represent the stop-controlled intersections.

The terms, x_n and $x_n^{'}$, are the n_{th} attributes of the two observations.

The *K*-means clustering algorithm is aimed at partitioning N observations into K ($K \leqslant N$) groups such that the summation of the variance in each cluster is minimized (*24*). Formally, the objective function is the following:

$$\min J = \min \sum_{j=1}^{K} \sum_{i=1}^{n} x_i^j - c_j{}^2 \tag{6}$$

where x_i^j is a data point, i, in cluster j and $x_i^j - c_j{}^2$ is the distance (usually the Euclidean distance) between the data point x_i^j and the cluster center c_j.

Both the KNN and *K*-means clustering algorithms require standardizing the predictors. Formulas (5) and (6) show that if an input feature has a variance that is significantly larger than other features it may have a strong effect on the objective function and render the algorithm unable to learn from other features correctly as expected (*25*).

The use of the method involving the combination of the KNN and *K*-means clustering algorithms produces a selection process depicted in Figure 1. In Figure 1*a*, an assumption was made that we have a set of locations which are stop-controlled intersections and signalized intersections. The intersections' attributes vary significantly as shown (here a two-dimensional space is used for convenience). Figure 1*b* shows that with the use of the KNN algorithm, only those stop-controlled intersections and signalized intersections which have similar features are selected. While the KNN algorithm only guarantees that, while there always exist pairs of similar stop-controlled intersections and signalized intersections, there may still be a significant difference among these selected pairs. As shown in Figure 1*b*, it may be more reasonable to analyze only the intersections in the upper right corner rather than all intersections depicted in Figure 1*b*. Thus, the *K*-means algorithm is then used to identify possible patterns in the dataset and to select only specific groups. In Figure 1*c*, the *K*-means algorithm distinguishes the attribute space into two parts, A and B. This makes it possible to filter out intersections belonging to part B.

In this study, the major-road AADT and the ratio of minor-road AADT to major-road AADT are used as control variables to identify similar sites, whether treated or untreated locations. The AADT was used as a reasonable exposure variable of an intersection. Intersections with similar AADT are expected to have similar crash frequencies.

Quasi-Induced Exposure Method

The quasi-induced exposure method is used to estimate the increase in the risk of being involved in a crash associated with driver-related or vehicle-related characteristics when there is no direct way to measure the intensity of exposure of these characteristics (*16*). The concept of the quasi-induced method is that the not-at-fault driver, involved in a two-vehicle collision (in these crashes only one of the two drivers was considered responsible for the crash) may be considered an approximately random sample of the road-user population (*16–18*). In this study, the ratio of the not-at-fault elderly drivers (age $\geqslant$ 65) to all not-at-fault drivers is used as the proportion of elderly drivers using an intersection. In this study, the proportion of elderly drivers was calculated at the county subdivision scale, that is, the intersections in the same county subdivision have the same elderly driver proportion. The county subdivision scale defines an area which has the size between the county and the census tract. Theoretically, a smaller scale such as the census tract or block group may better represent the heterogeneity of elderly drivers between intersections. However, these scales have inadequate counts of crashes which involve not-at-fault elderly drivers in the analysis period and this may lead to estimation bias.

Data

In this study, there are 438 rural four-leg intersections and 520 rural three-leg intersections. That includes 121 signalized intersections and 837 stop-controlled intersections. The crash, geometric design and traffic data of these intersections were identified for four years (2011–2014) from multiple sources. The crash records were collected from the Crash Analysis Reporting System (CARS) database maintained by the Florida Department of Transportation (FDOT). The geometric design data were collected from Google Maps. The traffic data were obtained from the Highway Performance Monitoring System (HPMS) and the FDOT. For the application of the cross-sectional method, typically three to five years of

crash records are recommended. Furthermore, between 100 and 1,000 intersection samples are required (*26*). The KNN and *K*-means methods were applied to select similar intersections. The selection process selected 140 rural four-leg intersections and 79 rural three-leg intersections for estimating crash prediction models. The distributions of the attributes of the selected intersections are summarized in Table 1.

Results

Development of GLM and MARS Models

Both GLM and MARS models of property damage only (PDO) crashes, severe (KAB as designated in the HSM) crashes, rear-end crashes and angle crashes were developed for rural four-leg intersections and rural three-leg intersections. The results are shown in Table 2 (GLM results) and Table 3 (MARS model results). Rear-end crashes involving elderly drivers at rural four-leg two-way stop-controlled intersections are used as an example to demonstrate the MARS model results. The crash trends after signalization of two-way stop-controlled intersections and of all-way stop-controlled intersections were modeled separately.

For the GLM, correlations between intersection attributes were examined first and only one of two highly correlated variables was kept in the final model depending on which variable could produce a better model fit. The results show that the signalization effect on crashes varies with different AADTs, crash types, and locations. For MARS modeling, no assumptions are made regarding the variables' distributions and thus meaningful standard errors of the coefficients of each selected variable are not presented. However, if the terms were not critical, the MARS algorithm would not have included them in the model (*27*). In this study, the two-way maximum order of interactions was considered for the MARS algorithm. It was found that increasing the maximum order of interactions (greater than 2) did not produce a significant difference in the final results, and the two-way maximum order of interactions is sufficient to explain the results. In addition, both GLM and MARS models were developed for crashes involving elderly drivers. Also, crashes not involving elderly drivers were modeled using the GLM and the MARS techniques since the impact of signalization may vary between drivers of different age groups. Accordingly, the major road AADT was adjusted by the elderly driver proportion and non-elderly driver proportion. The adjusted AADT represents the AADT generated by a specific group of drivers.

Table 4 compares the goodness-of-fit results of the GLM and MARS models. The measures are the mean absolute deviation (MAD) and root mean squared error (RMSE). It is worth noting that due to limited crash frequency, the control type is not invariably statistically significant in the GLM and MARS models of some types of crashes at specific locations. The models with insignificant control type parameters are not presented in Table 4. The results show that the MARS model always performs better than the GLM model on all the common crash types modeled using regression techniques. This may be because the MARS model can account for a more complex interaction effect between independent variables. Regarding the MARS model, each variable's effect is not consistent in its value space and the value space is separated by several knot values. In each region, a different effects function is mapped.

Estimation of CMFs

Table 5 shows the summary of CMF functions related to control type (i.e., signal and stop control) at different locations for different crash types. Because there are interaction terms involving the control type variable in the GLM and MARS models, the CMF is not a fixed value but a function involving all variables (as well as their coefficients) associated with the control type in the exponential model. These variables may be the control type itself or interaction terms between the control type variable and other variables.

Figure 2 shows that the signalization effect varies between locations. Signalization increases rear-end crashes and PDO crashes at rural intersections. Yet, at rural four-leg two-way stop-controlled intersections, signalization decreases severe crashes (i.e., KAB crashes). In addition, angle crashes not involving elderly drivers are reduced at rural four-leg two-way stop-controlled intersections after signalization. These conclusions are generally consistent with previous studies (*1–4*, *9*, *10*). In addition, Figure 2 shows that the actual crash trend depends on the major-road AADT and the elderly driver proportion. Generally, for a specific group of drivers (elderly or non-elderly), higher major road AADTs and large proportions of the driver group lead to surges in specific crash types given that all else is fixed. The crash types are PDO and rear-end crashes. On the other hand, more KAB crashes reduce.

Signalization is likely to increase PDO crashes and rear-end crashes involving elderly drivers more than those not involving elderly drivers especially when the major road AADT and the elderly driver proportion are high. In Figure 2, the green plots of PDO and rear-end crashes (representing crashes involving elderly drivers) are predominantly above the blue plots (representing crashes not involving elderly drivers) except where the major road AADT and the elderly driver proportion are low. Furthermore, at rural four-leg minor-road stop-controlled intersections, the signalization becomes better

Table 1. Descriptive Statistics of Attributes of Sampled Intersections

		Rural three-leg intersections				Rural four-leg intersections					
		Two-way stop-control		Signalization		All-way stop-control		Two-way stop-control		Signalization	
Variable		Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.
Crash (2011–2014)	Crash frequency	4.667	4.320	9.368	4.621	2.357	2.959	3.954	4.388	8.000	6.494
Geometric design (constant in 2011–2014)	Skew angle	9.417	13.375	5.789	11.336	2.500	7.638	7.241	11.148	4.400	11.843
	Street lighting	7% have		64% have		11% have		21% have		58% have	
	Pedestrian crosswalk	3.3% have		29% have		14% have		10% have		72% have	
	No. of exclusive right turning lane	0.567	0.767	1.824	0.636	0.107	0.416	0.241	0.570	0.720	1.173
	No. of exclusive left turning lane	0.717	0.783	2.235	0.562	0.500	1.291	0.621	1.026	2.040	1.541
	No. of exclusive through lane	0.767	0.945	2.529	1.281	0.321	1.056	0.138	0.462	1.160	2.014
	No. of through & right turning lane	0.400	0.494	0.235	0.437	0.464	1.261	0.563	0.872	1.560	1.502
	No. of through & left turning lane	0.717	0.640	0.118	0.332	0.071	0.378	0.184	0.518	0.160	0.473
	No. of right & left turning lane	0.717	0.490	0.059	0.243	Do not have		Do not have		Do not have	
	No. of through & left & right turning lane	0.317	0.504	0.059	0.243	3.429	1.317	3.195	1.150	1.760	1.451
	Channelized right-turn lane	10% have		42% have		Do not have		Do not have		Do not have	
	Channelized left-turn lane	Do not have		Do not have		Do not have		Do not have		Do not have	
	Speed limit on major road	50.875	7.850	48.421	6.021	46.058	7.006	48.373	8.699	42.400	8.675
	Speed limit on minor road	45.660	7.908	41.667	8.225	40.288	9.443	44.103	9.918	37.300	8.658
Traffic (averaged from 2011–2014)	Major-road AADT	7721	5164	12640	4593	3890	3647	4216	3134	6267	3064
	Minor-road AADT	2673	2041	6457	3237	1447	1010	1622	850	3192	1749
	Truck proportion	0.084	0.055	0.087	0.050	0.092	0.063	0.085	0.054	0.106	0.037
	Elderly driver proportion county subdivision	0.133	0.053	0.129	0.042	0.143	0.066	0.134	0.046	0.130	0.050
Other	Adjacent to school zone	No		No		No		No		No	
	Having approaches serving as ramp	No		No		No		No		No	
Total	Sample size	60		19		28		87		25	

Table 2. Estimated Parameters of GLMs for Different Crash Types

Parameter	Total PDO crashes		Total KAB crashes		Total rear-end crashes		Total angle crashes	
	Without elderly drivers	Involving elderly drivers	Without elderly drivers	Involving elderly drivers	Without elderly drivers	Involving elderly drivers	Without elderly drivers	Involving elderly drivers
(a) Rural four-leg signalized intersection vs rural four-leg two-way stop-controlled intersection								
Constant	**–4.6444**	**–8.9000**	**–11.3233**	**–8.9010**	**–14.6089**	**–9.3326**	**–10.4273**	——
	(1.2706)	**(1.4406)**	**(2.3099)**	**(1.2316)**	**(2.8455)**	**(1.6744)**	**(1.9866)**	
ln(AADT)	**0.4513**	**0.8630**	**1.0898**	**0.9687**	**1.3999**	**0.8275**	**1.0249**	——
	(0.1576)[a]	**(0.2154)**[b]	**(0.2554)**[a]	**(0.1828)**[b]	**(0.3126)**[a]	**(0.2503)**[b]	**(0.2175)**[a]	
ln(AADT)*	**0.0700**	**0.1204**	–0.0693	–0.0959	**0.1314**	**0.2173**	–0.0635	——
control type	**(0.0287)**[a]	**(0.0573)**[b]	(0.038)[a]	(0.0521)[b]	**(0.0352)**[a]	**(0.069)**[b]	(0.0333)[a]	
Ratio	——	**1.886**	**2.5171**	**1.6264**	**1.6769**	**1.7733**	**2.4872**	——
		(0.6156)	**(0.6656)**	**(0.5253)**	**(0.728)**	**(0.7042)**	**(0.6126)**	
Street lighting	——	——	——	——	——	–0.6514(0.3497)	——	——
Dispersion	0.6057	1.8512	0.6391	0.9093	0.4049	0.9198	0.5687	——
MAD	1.4277	0.6463	1.1001	0.5307	0.7471	0.3562	1.3681	——
RMSE	2.0497	1.4674	1.4811	0.8356	1.1860	0.7865	1.9210	——
(b) Rural four-leg signalized intersection vs rural four-leg all-way stop-controlled intersection								
Constant	**–8.8755**	**–5.9350**	——	——	**–10.3577**	**–7.3139**	——	——
	(1.6576)	**(1.5000)**			**(2.1001)**	**(1.7669)**		
ln(AADT)	**0.9338**	**0.5860**	——	——	**1.0176**	**0.5869**	——	——
	(0.1987)[a]	**(0.2519)**[b]			**(0.2465)**[a]	**(0.2965)**[b]		
ln(AADT)*	**0.078**	0.1366	——	——	**0.1255**	**0.2781**	——	——
control type	**(0.0317)**[a]	(0.0749)[b]			**(0.0374)**[a]	**(0.1084)**[b]		
Dispersion	0.2389	1.1382	——	——	0.1234	0.8102	——	——
MAD	1.2812	1.2758	——	——	0.9534	0.8010	——	——
RMSE	1.9464	2.2005	——	——	1.3426	1.2783	——	——
(c) Rural three-leg signalized intersection vs rural three-leg two-way stop sign control intersection								
Constant	——	**–5.7979**	——	——	**–9.622**	**–7.6601**	——	——
		(2.0701)			**(1.6925)**	**(2.3397)**		
ln(AADT)	——	0.4978	——	——	**0.961**	**0.7249**	——	——
		(0.3011)[b]			**(0.1888)**[a]	**(0.3338)**[b]		
ln(AADT) *	——	**0.1404**	——	——	**0.0534**	**0.1533**	——	——
control type		**(0.0500)**[b]			**(0.0220)**[a]	**(0.0552)**[b]		
Dispersion	——	0.3221	——	——	0.0879	0.5084	——	——
MAD	——	0.6181	——	——	1.0736	0.5558	——	——
RMSE	——	0.8418	——	——	1.4695	0.8432	——	——

Note: —— = the control type variable is insignificant. Control type = 1 represents signalized intersections while control type = 0 represents stop-controlled intersections. Numbers in parentheses are standard errors. Bold coefficients are significant at the 95% level while non-bold coefficients are significant at the 90% level.

[a]The non-elderly driver-related AADT, which is equal to the major-road AADT multiplied by (1-elderly driver proportion at county subdivision level).

[b]The elderly driver-related AADT, which is equal to the major-road AADT multiplied by the elderly driver proportion at county subdivision level.

Table 3. Estimated Parameters of MARS for Signalization Effect at Rural Four-Leg Two-Way Stop-Controlled Intersections

Significant basis function	Function information	Coefficient
Basis0	Intercept	−1.9115
Basis1	NOT(MISSING(lnAADT[a]))	−17.5012
Basis3	Basis1*MAX(lnAADT - 7.0579,0)	−0.868
Basis4	Basis1*MAX(7.0579 - lnAADT,0)	−1.0089
Basis5	Control type[b] = 0	0.0509
Basis7	Basis1*NOT(MISSING(truck proportion))	18.9801
Basis9	MAX(truck proportion - 0.1154,0)	−22.8769
Basis10	MAX(0.1154- truck proportion,0)	−813.26
Basis11	MAX(ratio - 0.9048,0) Basis5	−1047.18
Basis12	MAX(0.9048 - ratio,0) Basis5	−3.3863
Basis13	Basis3*NOT(MISSING(street_lighting))	2.3218
Basis15	Basis13*(street lighting = 1)	−3.5903
Basis17	(pedestrian crosswalk[c] = 1)	−3.3728
Basis19	Basis10*NOT(MISSING(major))	802.6
Basis21	Basis19*MAX(major[d] -40,0)	−2.9086
Basis22	Basis19*MAX(40-major,0)	−1113.5
Basis23	Basis10*MISSING(major)*MAX (skew angle - 35,0)	185.98
Basis24	Basis10*MISSING(major)*MAX (35-skew angle,0)	−32.395
Basis25	Basis10*MAX(N_exclusive_left[e] - 5,0)	117.72
Basis26	Basis10*MAX(5 - N_exclusive_left[e],0)	5.0955
Basis27	Basis3*MAX(ratio[f] - 0.8485,0)	482.76
Basis28	Basis3*MAX(0.8485 - ratio,0)	0.283
Basis29	Basis17*MAX(skew angle - 30,0)	0.4457
Basis30	Basis17*MAX(30 - skew angle,0)	0.1105

Note: The estimation results are for rear-end crashes involving elderly drivers.

[a]Elderly driver-related AADT, which is equal to the major-road AADT multiplied by the elderly driver proportion at county subdivision level.

[b]Control type = 0 represents stop control; control type = 1 represents signalized intersections.

[c]There is a pedestrian crosswalk.

[d]Major means speed limit on major road.

[e]Number of exclusive left-turn lanes.

[f]Ratio of minor-road AADT to major-road AADT.

at decreasing more severe crashes (KAB) for elderly drivers than for non-elderly drivers, with the increase of the major road AADT and the elderly driver proportion. Two possible reasons may contribute to this: (1) as shown in Figure 3, signalization increases KAB rear-end crashes for both elderly drivers and non-elderly drivers. However, the increase is not as large for elderly drivers as it is for non-elderly drivers; and (2) the reduction in KAB angle crashes involving elderly drivers is larger than that of KAB angle crashes involving non-elderly drivers after signalization especially when the major road AADT and elderly driver proportion are high.

Also, from Figure 2, it may be inferred that the elderly drivers are more sensitive than the non-elderly drivers. A larger slope of the estimated CMF profile in Figure 2 indicates that CMFs change substantially for elderly drivers by slight changes in major road AADTs and driver proportions.

The MARS model successfully captured more nonlinear relationships between the control type and other variables. In the GLM model, the interaction term is between the major road AADT and the control type. In the MARS model, more interaction terms between the control type and other factors except AADT show that the crash trends after signalization may also vary between the range of other factors. These factors include the ratio of minor-road AADT to major-road AADT, number of lanes of specific movements, speed limit on the major road, and skew angle. This indicates the advantage of the MARS model when it comes to capturing the nonlinear relationship between variables.

Since the CMF results estimated from the MARS model depend heavily on the original intersection attributes including geometric designs and traffic conditions than the GLM model, a base condition was set according to the average level of those attributes across the sampled intersections. The CMF is estimated under the base condition using the functional forms provided in Table 5.

The MARS models' results have similar trends to those of the GLM models. Signalized rural intersections would bring about more rear-end crashes and PDO crashes (as depicted in Figure 4) while at rural four-leg two-way stop-controlled intersections, signalization decreases KAB crashes by 32% (i.e., CMF = 0.68). Signalization increases rear-end crashes involving elderly drivers by 202% (i.e., CMF = 3.02) at rural three-leg two-way stop-controlled intersections.

In addition, the MARS model also leads to the same conclusion which is that elderly drivers are more vulnerable than non-elderly drivers when it comes to signalization. While there is a slight difference between the MARS and the GLM model results, the MARS model better distinguishes the trends of crashes involving elderly drivers and those not involving elderly drivers after signalization. Regarding elderly-driver-involved PDO and rear-end crashes, signalization contributes to crashes at rural four-leg all-way stop-controlled intersections more than it does at rural four-leg two-way stop-controlled intersections. For elderly-driver-not-involved PDO and rear-end crashes, the conclusion is contrary.

Discussion and Conclusion

In this study, GLM and MARS models are employed to analyze the changes in crash frequencies after signalization of different types of rural stop-controlled intersections. The joint KNN and *K*-means clustering algorithm is proposed to help select comparison sites, which was seldom attempted in previous research. In this study, the

Table 4. Comparison of the Goodness-of-Fit Results of GLM and MARS Models

Model	Measurements	Total PDO crashes		Total KAB crashes		Total rear-end crashes		Total angle crashes	
		Without elderly	Involving elderly	Without elderly	Involving elderly	Without elderly	Involving elderly	Without elderly	Involving elderly
(a) Rural four-leg signalized intersection vs two-way stop-controlled intersection									
GLM	MAD	1.4277	0.6463	1.1001	0.5307	0.7471	0.3562	1.3681	N/A
	RMSE	2.0497	1.4674	1.4811	0.8356	1.1860	0.7865	1.9210	N/A
MARS	MAD	1.1866	0.4561	0.9184	N/A	0.5418	0.208	N/A	N/A
	RMSE	1.6709	0.8823	1.2677	N/A	0.8786	0.4534	N/A	N/A
(b) Rural four-leg signalized intersection vs all-way stop-controlled intersection									
GLM	MAD	1.2812	1.2758	N/A	N/A	0.9534	0.8010	N/A	N/A
	RMSE	1.9464	2.2005	N/A	N/A	1.3426	1.2783	N/A	N/A
MARS	MAD	N/A	0.4557	N/A	N/A	0.8878	0.5817	N/A	N/A
	RMSE	N/A	0.6197	N/A	N/A	1.2776	0.9706	N/A	N/A
(c) Rural three-leg signalized intersection vs two-way stop-controlled intersection									
GLM	MAD	N/A	0.6181	N/A	N/A	1.0736	0.5558	N/A	N/A
	RMSE	N/A	0.8418	N/A	N/A	1.4695	0.8432	N/A	N/A
MARS	MAD	1.489	0.4908	N/A	N/A	N/A	0.4557	N/A	N/A
	RMSE	1.8748	0.6126	N/A	N/A	N/A	0.6197	N/A	N/A

Table 5. Summary of CMF Functions for Different Crash Types

Crash type	Elderly driver	GLM	MARS
(a) CMF for rural four-leg signalized intersection vs two-way stop-controlled intersection			
PDO	Not involved	Exp(0.07*ln(major road AADT*(1-elderly driver proportion)))	Exp(3.5874*MAX(0.6378 - ratio,0))
	Involved	Exp(0.1204*ln(major road AADT*elderly driver proportion))	Exp(2.3624*MAX(0.7895 - ratio,0))
KAB	Not involved	Exp(–0.0693*ln(major road AADT*(1-elderly driver proportion)))	Exp(1.2848*MAX(8.9994-ln(major road AADT*(1-elderly driver proportion)),0)-3.3103*MAX(ratio - 0.2836,0)-0.9983*MAX(number of through & left-turn lane - 1,0))
	Involved	Exp(–0.0959*ln(major road AADT*elderly driver proportion))	N/A
Rear-end	Not involved	Exp(0.1314*ln(major road AADT*(1-elderly driver proportion)))	Exp(–9.3774*MAX(ln(major road AADT*(1-elderly driver proportion))- 8.9994,0) + 1.2303*MAX(8.9994 - ln(major road AADT*(1-elderly driver proportion)),0) + 5.7162*MAX(0.6935 - ratio,0)-0.3531*MAX(skew angle-30,0) + 2.4684*MAX(Number of exclusive right-turn lane-1,0))
	Involved	Exp(0.2173*ln(major road AADT*elderly driver proportion))	Exp(–0.0509 + 1047.18*MAX(ratio-0.9048,0) + 3.3863* MAX(0.9048 - ratio,0))
Angle	Not involved	Exp(–0.0635*ln(major road AADT*(1-elderly driver proportion)))	N/A
Angle	Involved	N/A	N/A
Severe(KAB) rear end	Not involved	Exp(0.2594* ln(major road AADT*(1-elderly driver proportion)))	Not-Provided
	Involved	Exp(0.2339* ln(major road AADT*elderly driver proportion))	Not-Provided
Severe(KAB) angle	Not involved	Exp(–0.1462* ln(major road AADT*(1-elderly driver proportion)))	Not-Provided
	Involved	Exp(–0.0985* ln(major road AADT*elderly driver proportion))	Not-Provided
(b) CMF for rural four-leg signalized intersection vs all-way stop-controlled intersection			
PDO	Not involved	Exp(0.078*ln(major road AADT*(1-elderly driver proportion)))	1.53
	Involved	Exp(0.1366*ln(major road AADT*elderly driver proportion))	4.89
Rear end	Not involved	Exp(0.1255*ln(major road AADT*(1-elderly driver proportion)))	Exp(0.3070*MAX(major[a] - 42.5,0))
	Involved	Exp(0.2781*ln(major road AADT*elderly driver proportion))	Exp(1.757-0.8979*MAX(35 - major,0))
(c) CMF for rural 3-leg signalized intersection vs two-way stop-controlled intersection			
PDO	Not involved	N/A	1.79
	Involved	Exp(0.1404*ln(major road AADT*elderly driver proportion)	9.48
Rear end	Not involved	Exp(0.0534*ln(major road AADT*(1-elderly driver proportion)))	N/A
	Involved	Exp(0.1533*ln(major road AADT*elderly driver proportion)	3.02

Note: N/A = the signalization effect is not significant in the model; Not-Provided = the model is not applied to the crash; ratio = the ratio of minor-road AADT to major-road AADT; major = speed limit on the major road. For rural three-leg two-way stop-controlled intersections and rural four-leg all-way stop-controlled intersections, the control type is insignificant in the model for angle crashes and KAB crashes. Accordingly, the CMF is not available.

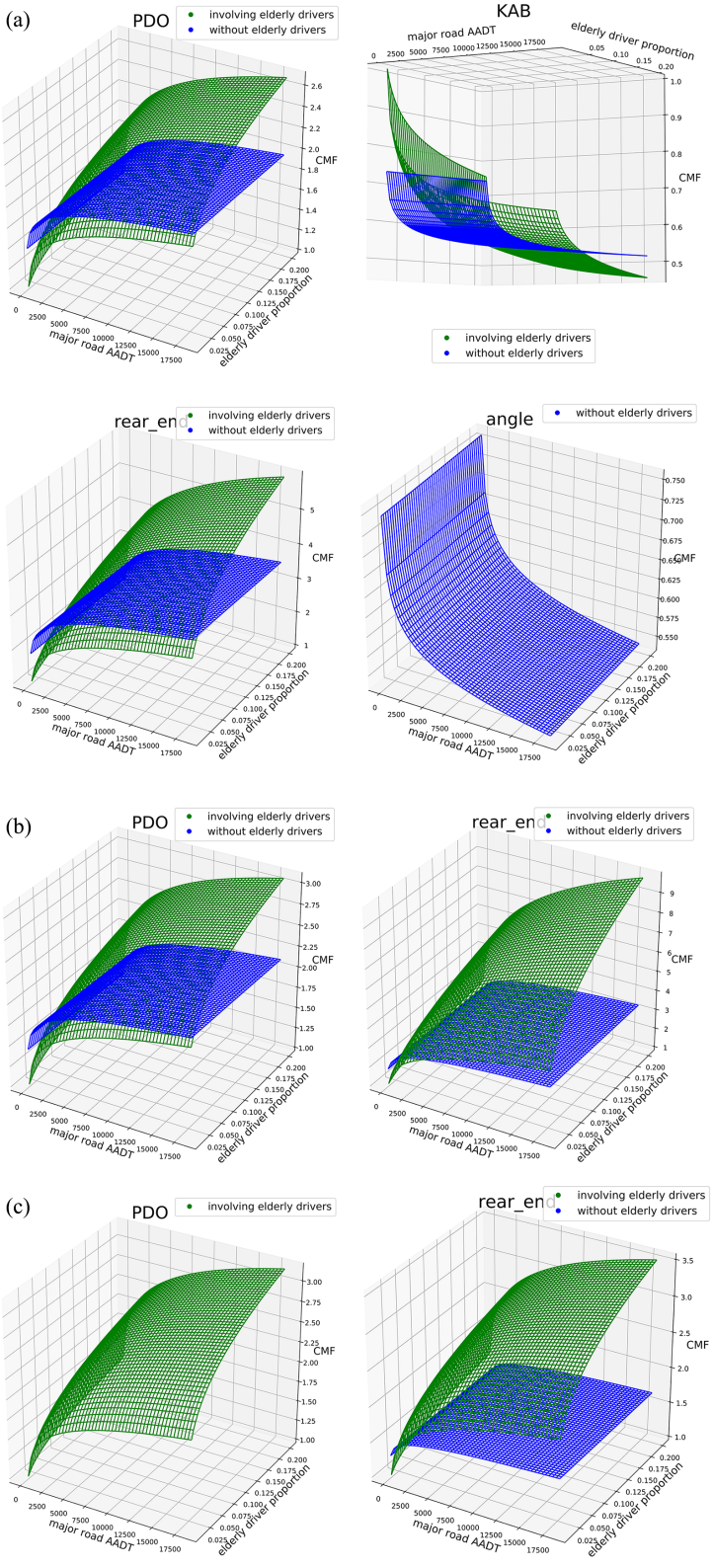

Figure 2. CMF estimated for crashes involving elderly drivers and crashes not involving elderly drivers by the GLM model: (*a*) rural four-leg intersections, signalized vs two-way stop-controlled; (*b*) rural four-leg intersections, signalized vs all-way stop-controlled; and (*c*) rural three-leg intersections, signalized vs two-way stop-controlled.

Note: For rural four-leg two-way stop-controlled intersection.

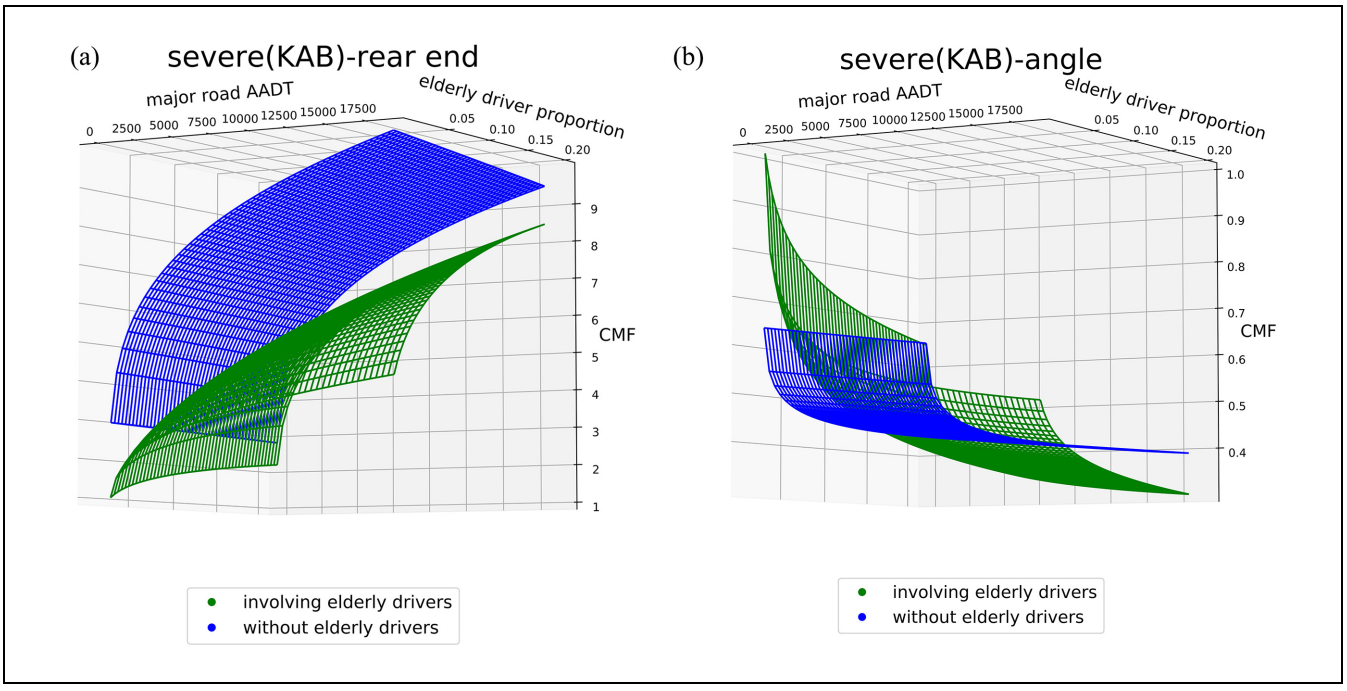

Figure 3. CMF estimated for severe crashes involving elderly drivers and crashes not involving elderly drivers by GLM model.
Note: For rural four-leg two-way stop-controlled intersection.

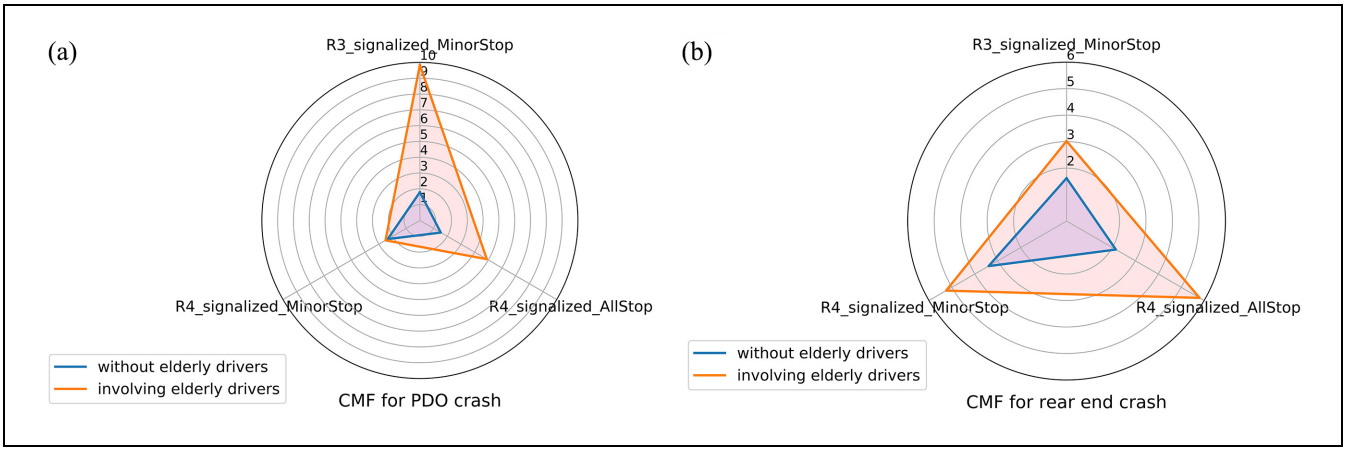

Figure 4. CMF estimated for crashes involving elderly drivers and crashes not involving elderly drivers by MARS model.
Note: (1) The base condition is set as follows: the major-road AADT is 10,000 vpd, the ratio of the minor-road AADT to the major-road AADT is 0.4, the elderly driver proportion is 0.15, the speed limit on the major road is 45 mph, the skew angle is 0 and there is no exclusive right-turn lane or shared through and left-turn lane. The treatment is signalization. By using the MARS model functional forms in Table 5, the CMFs could be calculated. (2) "R3_signalized_MinorStop" denotes that the CMF is estimated by comparing signalization and two-way stop control at rural three-leg intersections; that is the same for "R4_signalized_MinorStop" and "R4_signalized_AllStop". (3) The CMF for rear-end crashes not involving elderly drivers is not computed using the MARS technique but is estimated by the GLM technique using the same base condition.

AADT is used as the control variable to identify similar treated and untreated locations. Theoretically, incorporating more control variables such as geometric designs will aid in identifying similar sites. However, excessive control variables may lead to under-dispersion. It is the condition at which the variance of the crash frequencies is less than the mean. The variables of interest may become insignificant in the SPF model due to the lack of variance in crash frequencies. In addition, the roadway characteristics such as the number of lanes and speed limit are highly correlated with the AADT. Also, other traffic information such as truck proportion is reflected by the AADT. Thus, only the AADT is selected as the control variable. It is standardized with a mean of zero

and a unit variance before it is inputted into the joint KNN and *K*-means algorithm. The variables other than AADT are tested when developing the SPFs.

Furthermore, applications of the quasi-induced exposure methodology and the MARS model also improve the estimation of CMFs of signalization. This study demonstrates that the safety effects of signalization vary among sites. Signalization increases rear-end crashes and PDO crashes at rural intersections (i.e., three-leg/four-leg intersections, two-way stop-controlled/all-way stop-controlled). In addition, the GLM technique's results also demonstrate that at rural four-leg two-way stop-controlled intersections, signalization decreases KAB crashes and angle crashes not involving elderly drivers.

Both the GLM and MARS models show that signalization increases PDO crashes and rear-end crashes involving elderly drivers more than it increases such crashes not involving elderly drivers. That is particularly so when the intersection has a high major-road AADT and a high elderly driver proportion. Under the base conditions mentioned in Figure 4, when signalizing a rural four-leg two-way stop-controlled intersection, the GLM model estimates the CMFs for PDO crashes involving and not involving elderly drivers to be 2.41 and 1.66, respectively. For rear-end crashes involving and not involving elderly drivers, the CMFs are 4.89 and 3.28 respectively. On the other hand, with the use of the MARS model, the CMFs for PDO crashes involving and not involving elderly drivers are 2.51 and 2.35, respectively. The CMFs of rear-end crashes involving and not involving elderly drivers are 5.25 and 3.40 respectively. A possible reason for this may be that the reaction times of elderly drivers are longer than those of non-elderly drivers.

Signalizing rural four-leg two-way stop-controlled intersections is beneficial for elderly drivers. KAB crashes involving elderly drivers decrease more substantially than those not involving elderly drivers after signalization. That is observed with the increase of major-road AADT and elderly driver proportion. The GLM model demonstrates that KAB crashes involving elderly drivers decrease by 51% and those not involving elderly drivers by 47%. Two possible reasons may contribute to this difference: (1) severe rear-end crashes involving elderly drivers increase considerably less than those not involving elderly drivers after signalization; and (2) the reduction in KAB angle crashes involving elderly drivers is larger than that of KAB angle crashes not involving elderly drivers particularly when the major road AADT is high and when the proportion of elderly drivers is large.

Thus, at rural four-leg two-way stop-controlled intersections with relatively high major-road AADTs and high elderly driver proportions, signalization may be recommended because it would significantly reduce the severity of crashes involving elderly drivers. However, other supplemental countermeasures should also be considered. To reduce PDO and rear-end crashes involving elderly drivers, the countermeasures include but are not limited to lower posted speed limits and redundant signs to reduce the risk of failure to comply.

The comparative analysis revealed that the MARS model outperforms the GLM model. The MARS model uses multiple knots to separate the value space of an independent variable into several regions, and is able to locate the regions that describe the safety effects of signalization. For the identified significant regions, the MARS model can even fit the data with different basis functions. This is why the MARS model outperforms the GLM model in capturing complex and nonlinear relationships among variables. In this study, the GLM model only captures the interaction effect between the major road AADT and the control type while the MARS model captures the impacts of several other additional variables on safety due to signalization.

However, the advantage of the MARS model would be annulled if the value space of an independent variable lacked variation and only contained several discrete values which are close to each other. If this is the case, developing the MARS to further divide the narrow value space is futile. A severe case may happen, that the portions of these discrete values in total samples are far away from the uniform distribution, that is, the majority of the observations have the variable with the same value. In this case, the estimated knots related to those less-observed values may be biased. In this study, the MARS model identified several variables which are associated with the control type, including the major-road AADT, the ratio of the minor-road AADT to the major-road AADT, the number of lanes of specific functions, the speed limit on the major road and the skew angle. However, among these variables, only the AADT and the ratio can be treated as continuous variables which have various values the other variables are more similar to categorical variables which are only observed for several discrete values. Some of these variables also lack variation. For example, theoretically, the number of exclusive right-turn lanes can be at most four at intersections, however, in the study, 81% of the sampled intersections do not have exclusive right-turn lanes and only 8% of sampled intersections have two or more exclusive right-turn lanes. This may bias the results. A possible way to obtain more reliable results from the MARS model is to use the average level of independent variables as the model input and analyze the model results in a general base condition which was conducted in this study.

This study is not without limitations. The study uses the quasi-induced exposure method to estimate the

proportion of elderly drivers at an intersection. Although many researchers consent to the benefit of this method (*16–18*, *28*), it is recommended that its assumption be validated regarding the not-at-fault drivers when applying it to a new data set. A possible technique is to compare the relative exposure estimated via the quasi-induced exposure method with the exposure "truth" from external data sources such as the vehicle miles traveled (VMT) statistics (*29*). However, in this study such data sources are not available. Thus, to a limited extent, there may be a deviation between the estimated proportion and the actual proportion of elderly drivers.

It is worth elaborating on the shortcomings of the models employed. The relationship between the crash frequency and the signalization effects on specific crashes (severe and angle crashes) at some intersections was not captured by the GLM and MARS models because of the limited crash data. The MARS model is more sensitive to the sample size than the GLM model. The MARS model failed to capture the effects of signalization on the count of severe crashes and angle crashes more frequently than the GLM model.

In addition, the study does not address the issue of middle-aged drivers. Middle-aged drivers may be riskier than elderly drivers. They may make more unsafe driving decisions. Thus, the CMFs related to crashes involving middle-aged drivers may be different from those related to crashes involving elderly drivers. It is worth investigating the CMFs of crashes involving middle-aged drivers in the future.

Acknowledgments

The authors thank the Florida Department of Transportation (FDOT) for funding this research and allowing access to its databases. The authors are grateful to Lianet Peraza, Michelle Pruss and Morgan Morris for helping in collecting the data.

Author Contributions

The authors confirm their contributions to the paper as follows: study conception and design, Mohamed Abdel-Aty, Jaeyoung Lee; data collection, Lishengsa Yue, Ahmed Farid; analysis and interpretation of results, Lishengsa Yue, Jaeyoung Lee; draft manuscript preparation, Lishengsa Yue, Ahmed Farid, Jaeyoung Lee, Mohamed Abdel-Aty. All authors reviewed the results and approved the final version of the manuscript.

References

1. *Highway Safety Manual*, 1st ed. AASHTO, Washington, D.C., 2010.
2. Davis, G., and N. Aul. *Safety Effects of Left-Turn Phasing Schemes at High-Speed*. MN/RC-2007-03. Minnesota Department of Transportation, 2007.
3. Harwood, D. W., F. Council, E. Hauer, W. Hughes, and A. Vogt. *Prediction of the Expected Safety Performance of Rural Two-Lane Highways*. FHWA-RD-99-207. Federal Highway Administration, Washington, D.C., 2000.
4. McGee, H. W. *NCHRP Report No. 49l: Crash Experience Warrant for Traffic Signals*. Transportation Research Board of the National Academies, Washington, D.C., 2003.
5. Yuan, J., and M. Abdel-Aty. Approach-Level Real-Time Crash Risk Analysis for Signalized Intersections. *Accident Analysis & Prevention*, Vol. 119, 2018, pp. 274–289.
6. Yu, R., M. Quddus, X. Wang, and K. Yang. Impact of data aggregation approaches on the relationships between operating speed and traffic safety. *Accident Analysis & Prevention*, Vol. 120, 2018, pp. 304–310.
7. Sacchi, E., T. Sayed, and K. El-Basyouny. A Full Bayes before-after Study Accounting for Temporal and Spatial Effects: Evaluating the Safety Impact of New Signal Installations. *Accident Analysis & Prevention*, Vol. 94, 2016, pp. 52–58.
8. Srinivasan, R., B. Lan, D. L. Carter, and U. o. N. C. H. S. R. Center. Safety Evaluation of Signal Installation with and without Left Turn Lanes on Two Lane Roads in Rural and Suburban Areas. *Research and Analysis Group, North Carolina Department of Transportation*, 2014.
9. Harkey, D. L. *Accident Modification Factors for Traffic Engineering and ITS Improvements*. NCHRP Report 617. Transportation Research Board, Washington, D.C., 2008.
10. Wang, J. H., M. Abdel-Aty, and J. Lee. Examination of the Transferability of Safety Performance Functions for Developing Crash Modification Factors: Using the Empirical Bayes Method. *Transportation Research Record: Journal of the Transportation Research Board*, 2016. 2583: 73–80.
11. Park, J., and M. Abdel-Aty. Assessing the Safety Effects of Multiple Roadside Treatments Using Parametric and Nonparametric Approaches. *Accident Analysis & Prevention*, Vol. 83, 2015, pp. 203–213.
12. Abdel-Aty, M., C. Lee, J. Park, J. Wang, M. Abuzwidah, and S. Al-Arifi. *Validation and Application of Highway Safety Manual (Part D) in Florida*. Florida Department of Transportation, 2014.
13. Briand, L. C., B. Freimut, and F. Vollei. Using Multiple Adaptive Regression Splines to Support Decision Making in Code Inspections. *Journal of Systems and Software*, Vol. 73, No. 2, 2004, pp. 205–217.
14. Haleem, K., A. Gan, and J. Lu. Using Multivariate Adaptive Regression Splines (MARS) to Develop Crash Modification Factors for Urban Freeway Interchange Influence Areas. *Accident Analysis & Prevention*, Vol. 55, 2013, pp. 12–21.
15. Gross, F., B. Persaud, and C. Lyon. *A Guide to Developing Quality Crash Modification Factors*. FHWA-SA-10-032. U.S. Department of Transportation, Federal Highway Administration, Washington, D.C., 2010.
16. Martínez-Ruiz, V., P. Lardelli-Claret, E. Jiménez-Mejías, C. Amezcua-Prieto, J. J. Jimenez-Moleon, and J. D. D. L. del Castillo. Risk Factors for Causing Road Crashes Involving Cyclists: An Application of a Quasi-

Induced Exposure Method. *Accident Analysis & Prevention*, Vol. 51, 2013, pp. 228–237.
17. Keall, M., and S. Newstead. Induced Exposure Estimates of Rollover Risk for Different Types of Passenger Vehicles. *Traffic Injury Prevention*, Vol. 10, No. 1, 2009, pp. 30–36.
18. Stamatiadis, N., and J. A. Deacon. Quasi-induced Exposure: Methodology and Insight. *Accident Analysis & Prevention*, Vol. 29, No. 1, 1997, pp. 37–52.
19. Friedman, J. H. Multivariate Adaptive Regression Splines. *The Annals of Statistics*, Vol. 19, No. 1, 1991, pp. 1–67.
20. Abraham, A., D. Steinberg, and N. S. Philip. Rainfall Forecasting Using Soft Computing Models and Multivariate Adaptive Regression Splines. *IEEE SMC Transactions, Special Issue on Fusion of Soft Computing and Hard Computing in Industrial Applications*, Vol. 1, 2001, pp. 1–6.
21. Snedecor, G. W., and W. G. Cochran. *Statistical Methods*, 8th ed. Iowa State University Press, Ames, 1989.
22. Put, R., Q. Xu, D. Massart, and Y. Vander Heyden. Multivariate Adaptive Regression Splines (MARS) in Chromatographic Quantitative Structure–Retention Relationship Studies. *Journal of Chromatography A*, Vol. 1055, No. 1-2, 2004, pp. 11–19.
23. Zakka, K. *A Complete Guide to K-Nearest-Neighbors with Applications in Python and R*. https://kevinzakka.github.io/2016/07/13/k-nearest-neighbor/. Accessed March 15, 2018.
24. *A Tutorial on Clustering Algorithms*. https://home.deib.polimi.it/matteucc/Clustering/tutorial_html/index.html. Accessed April 4, 2018.
25. *scikit-learn: Machine Learning in Python*. http://scikit-learn.org/stable/modules/preprocessing.html. Accessed April 15, 2018.
26. Carter, D., R. Srinivasan, F. Gross, and F. Council. *NCHRP 20-7 (314) Final Report: Recommended Protocols for Developing Crash Modification Factors*. Transportation Research Board of the National Academies, Washington, D.C., 2012.
27. Milborrow, S. *Notes on the Earth Package*. http://www.milbo.org/doc/earth-notes.pdf. Accessed June 2, 2018.
28. Cerrelli, E. C. Driver Exposure: The Indirect Approach for Obtaining Relative Measures. *Accident Analysis & Prevention*, Vol. 5, No. 2, 1973, pp. 147–156.
29. Jiang, X., R. W. Lyles, and R. Guo. A Comprehensive Review on the Quasi-Induced Exposure Technique. *Accident Analysis & Prevention*, Vol. 65, 2014, pp. 36–46.

The Standing Committee on Highway Safety Performance (ANB25) peer-reviewed this paper (19-03826).

All opinions and results are solely those of the authors.