Investigating the influence of segmentation in estimating safety performance functions for roadway sections

Salvatore Cafiso  
Department of Civil and Environmental Engineering  
School of Engineering, University of Catania  
Viale Andrea Doria 6, 1-95125 Catania, Italy  
Phone: +39 095 738 2213  
Fax: +39 095 738 2247  
E-mail: dcafiso@dica.unict.it

Carmelo D’Agostino  
Department of Civil and Environmental Engineering  
School of Engineering, University of Catania  
Viale Andrea Doria 6, 1-95125 Catania, Italy  
Phone: +39 095 738 2213  
Fax: +39 095 738 2247  
Email: dagosti@dica.unict.it

Bhagwant Persaud (Corresponding author)  
Ryerson University  
Department of Civil Engineering  
350 Victoria Street, Toronto, Canada M5B2K3  
Phone: 416-979-5345, Ext. 6464  
Fax: 416-979-5122  
E-mail: bpersaud@ryerson.ca

Word count: 4239 + 250×(4 figures + 4 tables) = 6239 words

Submission date: November 15, 2012


ABSTRACT

Safety performance functions (SPFs) are crucial to science-based road safety management. Success in developing and applying SPFs depends fundamentally on two key factors: the validity of the statistical inferences for the available data and on how well the data can be organized into distinct homogenous entities. The latter aspect plays a key role in the identification and treatment of road sections or corridors with problems related to safety. Indeed, the segmentation of a road network could be especially critical in the development of SPFs that could be used in safety management for roadway types, such as motorways (freeways in North America), which have a large number of variables that could result in very short segments if these are desired to be homogeneous. This consequence, from an analytical point of view, can be a problem when the location of crashes is not precise and when there is an over abundance of segments with zero crashes. Lengthening the segments for developing and applying SPFs can mitigate this problem, but at a sacrifice of homogeneity. This paper seeks to address this dilemma by investigating five approaches for segmentation for motorways, using sample data from Italy. The best results were obtained for the segmentation based on two curves and two tangents within a segment and the segmentation with fixed length. The segmentation characterized by a constant value of all original variables inside each segment was the poorest approach by all measures.

Keywords: road safety, rural motorways, safety performance functions, segmentation, crash prediction, general estimating equation.
INTRODUCTION
Safety performance functions (SPFs) are crucial to science-based road safety management using, e.g., the methods prescribed in the Highway Safety Manual (1). These functions are statistical models used to estimate the expected crash frequency for a facility (2) based on its characteristics, mainly traffic volume, which accounts for the majority of the variability in crash frequency, and geometric variables. These functions are developed from data for a number of similar sites. Success in development or application of an SPF for road segments depends strongly on how well the data can be organized into distinct homogenous entities, i.e., on the approach to segmentation.

Segmentation, when based on multiple variables, may lead to very short homogeneous segments (3). For example, when using the segmentation approach proposed by the HSM, the presence of very short segments does not allow proper statistical inference for several reasons. The most important are the non-perfect identification of the location of crashes, which is often taken from police reports, (4), and the fact that crashes are rare events resulting in a great number of segments with zero crashes. Lengthening segments to avoid these issues will sacrifice homogeneity.

In the literature there are a number of different approaches to segmentation. Miaou and Lum suggested that short sections, less than or equal to 80 m could create bias in the estimation of linear models, but not when using Poisson models (5). Similarly, Ogle et al. demonstrated that short segment lengths, less than 160 m, cause uncertain results in crash analysis (6). Cafiso et al. (7) showed that to increase performance in identifying correct positives as black spots, segment length should be related to AADT with lower AADT values requiring longer segment lengths. Qin et al. studied the relationship between segmentation and safety screening analysis (4) using different lengths of sliding windows to identify hazardous sites, and concluded that short segments as well as those that are too long create a bias in the identification of sites with safety problem.

Some studies focused on the relationship between crashes and road geometry in addressing segmentation. For example, Cenek et al. (8), who investigated this relationship, for rural roads data, used a fixed segment length of 200 m. A similar study was done by Cafiso et al. (9) using homogeneous section with different lengths on a sample of Italian two lane rural roads, aggregating variable related to curvature and roadside hazard. They concluded that model that contains geometry and design consistency variables are more reliable than others. Other studies suggested different ways to aggregate segment data to avoid lengths that are too short. For example, Koorey proposed the aggregation of curves and tangents when the radius of curves exceeds a predetermined threshold value (10).

The Highway Safety Manual (HSM) (1) recommends the use of homogeneous segments with respect to AADT, number of lanes, curvature, presence of ramp at the interchange, lane width, outside and inside shoulder widths, median width and clear zone width. There is no prescribed minimum segment length for application of the predictive models, but there is a suggestion of a segment length not less than 0.10 miles.

Given the variety of approaches and the fact that there is no apparent preferred one, this paper seeks to investigate alternative methodologies for segmentation, including the HSM procedures, using sample data from Italian motorways. All but one of these methods aggregate and redefine variables over longer segments while seeking to retain the geometric and exposure characteristics of the segment as best as possible. SPFs calibrated for different segmentations are compared in terms of goodness of fit and the variables captured. Stepwise regression was used for each of five different segmentation concepts to select the best combination of variables. In addition, for each segmentation concept, two simpler models were estimated and compared, a base model and curvature-based model that is described later.
DATA DESCRIPTION

The data used for this investigation pertain to Italian rural motorway, the “A18” Messina-Catania, which is approximately 76 km (47.2 miles) long. The cross section is made up from 4 lanes, 2 in each direction, divided by a median with barriers. The analysis period is for the 8 years from 2002 until 2009, during which 887 severe (fatal plus injury) crashes according to the official statistics on motor vehicle collisions provided by the Italian National Institute of Statistics (ISTAT) (11). Table 1 shows basic statistics for the dataset used for analysis.

In this study, only the road segments were analyzed; interchange data and the part of segment directly influenced by the presence of intersection were discarded. Every segment contiguous to an intersection starts from a distance of 50 m (164 ft) from the bevel for the insertion of the service lanes for exit from, and entry into the main flow. The available data, in addition to AADT (Figure 1), were: radius of curvature, vertical gradient, type of section, and roadside features (presence and typology of the lateral and median barriers). With this wide variety of variables it was impossible to achieve segmentation with perfectly homogeneous sections; thus, it was necessary to use variables as appropriate to characterize the segment features.

Table 1. Details of the database used to estimate models.

<table>
<thead>
<tr>
<th>Year</th>
<th>Range AADT</th>
<th>Crash (Fatal + Injury)</th>
<th>Length (km)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2002</td>
<td>8696 - 24904</td>
<td>94</td>
<td></td>
</tr>
<tr>
<td>2003</td>
<td>9082 - 26123</td>
<td>95</td>
<td></td>
</tr>
<tr>
<td>2004</td>
<td>9423 - 26947</td>
<td>100</td>
<td></td>
</tr>
<tr>
<td>2005</td>
<td>10944 - 26882</td>
<td>104</td>
<td></td>
</tr>
<tr>
<td>2006</td>
<td>7792 - 26414</td>
<td>113</td>
<td></td>
</tr>
<tr>
<td>2007</td>
<td>7917 - 27001</td>
<td>119</td>
<td></td>
</tr>
<tr>
<td>2008</td>
<td>7651 - 26783</td>
<td>113</td>
<td></td>
</tr>
<tr>
<td>2009</td>
<td>9066 - 26743</td>
<td>93</td>
<td></td>
</tr>
<tr>
<td>Total Crash</td>
<td></td>
<td>831</td>
<td>145.08 (two directions)</td>
</tr>
</tbody>
</table>

Figure 1. Details of annual traffic (millions of vehicles) during the analysis period.
ANALYSIS
This section describes the variables used, the segmentation approaches evaluated, and the models investigated, before presenting the results.

Variables
In order to divide the sample into homogeneous segments, it was necessary to combine all the variables into a usable form, paying attention to the final form of the equation used for developing the SPF’s for each segmentation approach (11). The Average Annual Daily Traffic (AADT), which describes the exposure to crash risk, was not constant across some segments for two segmentation concepts; a length-weighted AADT was used as an approximation when this situation occurred. This approximation is more appropriate where a linear relationship between the dependent variable and AADT exists. (For the SPF’s developed in this study, the crash-AADT relationship was approximately linear.) Other variables were similarly aggregated over segments in which they may not be constant for one or more segmentation approaches. The variables considered, apart from AADT, are related to geometry. The original data were: curvature, gradient of each segment and barrier condition. The aggregation of these data is described below:

- **Curvature treated as curvature change rate (CCR)** (12) of the segment, calculated as follows:
  \[
  CCR = \frac{\sum |\gamma_i|}{L} \text{[gon/m]}
  \]
  where \(\gamma_i\) is the deflection angle for a contiguous element (curve or tangent) \(i\) within a section of length \(L\);

- **Slope Change Rate (SCR)** for the vertical profile of the road segment, which represents the variation of the slope inside a single segment, calculated as follows:
  \[
  SCR = \frac{\sum |\delta_i|}{L} \text{[gon/m]}
  \]
  where \(\delta_i\) is the deflection angle for a slope related to the horizontal alignment within a section of length \(L\);

- **I**, defined as the weighted average of the vertical gradient (up and down) with the reference length within each segment;

- **I_d**, defined as the weighted average of the vertical gradient (down) with the reference length within each segment.

- **Roadside Hazard (RSH)** along a motorway, which is based both on type of section (trench, embankment, viaduct) and on the type of barrier with reference to the European standard EN 1317-1 1998 (13) (14) (15) (16). RSH assumes 6 possible values (from 1 to 6, in increasing order of potential hazard), defined as follows: first, we considered only the conditions of the outer margin, assigning 1 to the trench, 2 to embankment with appropriate barriers (complying with EN 1317-1), 3 to the viaduct with adequate lateral barrier, 4 to embankment where the side dam is not adequate, 5 to the viaduct with inadequate lateral barrier. A value of 1 was added if the median barrier is not adequate (Table 2). For this variable, the percentage of the length of a segment in which the RSH value was 6 (RSH6) or 5 and 6 (RSH56) was used;

- **TUN**, which indicates the percentage of the length of segment that is a tunnel;
Table 2. Value of RSH by type of section and condition of lateral and median barriers (LA and MA indicate adequate lateral and median barriers; LN and MN indicate inadequate lateral and median barriers.)

<table>
<thead>
<tr>
<th></th>
<th>RSH for Lateral Barrier only</th>
<th>RSH for Median Barrier only</th>
<th>RSH for Lateral and Median Barrier combinations</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>LA</td>
<td>LN</td>
<td>MA</td>
</tr>
<tr>
<td>Trench</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Embankment</td>
<td>2</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>Viaduct</td>
<td>3</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>Tunnel</td>
<td>4</td>
<td>4</td>
<td></td>
</tr>
</tbody>
</table>

Segmentation Approaches
In order to assess the influence of the organization of the data into segments on the goodness of fit of an SPF, five different segmentation approaches were assessed, taking as reference, the traffic (AADT) and the curvature. Specifically, these are:

- **Segmentation 1**: Homogeneous segments with respect to AADT and curvature, as suggested by HSM, using AADT and curvature as explanatory variables;
- **Segmentation 2**: Data organized to have within each segment 2 curves and 2 tangents, avoiding having short segments when using a single curve;
- **Segmentation 3**: Segments have constant AADT; other variables may not be constant.
- **Segmentation 4**: Segments have a constant length. Specifically, a length of 650 m was chosen, coinciding with the maximum length of an interchange area, and selected to be just longer than the longest horizontal curve. This length was chosen to minimize the problem of incorrect location of crashes on Italian motorways.
- **Segmentation 5**: All the variables used in the stepwise procedure are constant within each segment with their original value.

For the segmentation based on curvature and AADT, very short segments were eliminated in order to have segments with length more than 100 m. Using different segmentation approaches also changes the range of variation of the variables used to estimate the model. Table 3 shows the range of the variables used for each segmentation approach. For Segmentation 5, characterized by the value of the original data constant inside each segment, it is not possible to use an aggregated variable for RSH and TUN, so these are used as categorical variables with their original value.

Table 3. Range (min-max) of variables for segmentation approaches investigated.

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Length (m)</strong></td>
<td>100.1 - 1563.4</td>
<td>234.7 - 3307.6</td>
<td>4882.3 - 21856.3</td>
<td>650.0</td>
<td>12 - 979.1</td>
</tr>
<tr>
<td><strong>SCR (gon/m)</strong></td>
<td>0 - 0.31</td>
<td>0 - 0.10</td>
<td>0.036 - 0.086</td>
<td>0 - 0.28</td>
<td>0 - 0.35</td>
</tr>
<tr>
<td><strong>CCR (gon/m)</strong></td>
<td>0 - 0.031</td>
<td>0 - 0.014</td>
<td>0.034 - 0.068</td>
<td>0 - 0.024</td>
<td>0 - 0.33</td>
</tr>
<tr>
<td><strong>RSH6 (%)</strong></td>
<td>0 - 70.23</td>
<td>0 - 55</td>
<td>0 - 10.61</td>
<td>0 - 66.03</td>
<td>-</td>
</tr>
<tr>
<td><strong>RSH56 (%)</strong></td>
<td>0 - 100</td>
<td>0 - 100</td>
<td>0 - 12.03</td>
<td>0 - 100</td>
<td>-</td>
</tr>
<tr>
<td><strong>RSH</strong></td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1 - 6</td>
</tr>
<tr>
<td><strong>TUN (%)</strong></td>
<td>0 - 100</td>
<td>0 - 75.4</td>
<td>0 - 49.1</td>
<td>0 - 100</td>
<td>0 - 1</td>
</tr>
<tr>
<td><strong>I (Gon)</strong></td>
<td>-0.042 - 0.045</td>
<td>-0.031 - 0.031</td>
<td>-0.0086 - 0.0088</td>
<td>-0.038 - 0.038</td>
<td>-0.042 - 0.045</td>
</tr>
<tr>
<td><strong>I₄ (Gon)</strong></td>
<td>0 - 0.043</td>
<td>0 - 0.031</td>
<td>0.0023 - 0.014</td>
<td>0 - 0.038</td>
<td>0 - 0.043</td>
</tr>
</tbody>
</table>
Models
The Generalized Estimating Equation (GEE) (17) method was used to estimate model coefficients, using the Statistical Analysis System (SAS) software package. The GEE procedure is classified as a multinomial analogue of a quasi-likelihood function, which allows the consideration of time trend in the models. Consistent with the state of research in developing these models, the negative binomial error distribution was assumed for the count of observed crashes (2).

The analysis of the road network using the predictive models includes the choice of a period of analysis. In general, this period of analysis depends on the availability of both traffic and crash data, but, in the literature, numerous studies have shown that periods longer than 5 years could introduce bias into the mathematical model for the variables linked to any physical changes of the network, or to the natural time trend, which, without the use of GEE, cannot be taken into account. The GEE procedure incorporates time trend, so is well suited to modeling data for long time periods. Specifically, it accounts for the temporal correlation that results when data for long periods are disaggregated into separate observations for each year.

To evaluate the goodness of fit (g.o.f.) of the models, two different methodologies were applied: the Quasilikelihood under the Independence model Criterion (QIC) (18) (19) and the Cumulative Residuals (CURE) plot (20). The QIC statistic is analogous to Akaike’s Information Criterion (AIC) statistic used for comparing models fit with likelihood-based methods. Since the generalized estimating equations (GEE) method is not a likelihood-based method, the AIC statistic is not applicable. The QIC has the following form:

\[ QIC = -2Q(\hat{\mu}; I) \]

where \( I \) represents the independent covariance structure used to calculate the quasi-likelihood and \( \hat{\mu} = g^{-1}(x^T \beta) \) where \( g^{-1} \) is the inverse of link function.

When using QIC to compare two structures, or two models, the model with the smaller statistic is preferred (18) (19) (21). The smaller the value of QIC, the better is fit of the model to the data. Therefore QIC can be used to compare and rank different models. For the present study, another advantage of QIC is that the g.o.f. of models with different numbers of parameters can be compared.

The CURE method to evaluate the goodness of fit is based on the study of residuals, i.e., the difference between the number of crashes observed at a site and the expected value at the same site and in the same year. Assuming that residuals are normally distributed with expected value equal to 0 and a variance equal to \( \sigma^2 \) (20), it is possible to calculate the variance of the expected value as the square of the cumulative residuals. The trend in the residuals with respect to AADT (or other variables) can be evaluated relative to the variance to qualitatively assess goodness of fit. The CURE plot is used for the examination of residuals after the estimation of the SPFs. It is used to examine whether the chosen functional form indeed fits each explanatory variable along the entire range of its values represented in the data. An upward or downward drift is a sign that the SPF consistently predicts fewer or more crashes, respectively, than were counted. Thus, it is desirable that the plot of cumulative residuals should stay flat or at least oscillate between over and under prediction and not stray beyond the \( \pm 2\sigma^* \) boundaries.

The selection of the explanatory variables to be included in the model was made using a stepwise methodology inserting at first all the variables available, and testing for each of the five segmentations in order to keep only the variables that were significant. This method was applied using different set of variables, and avoiding problems due to correlation of variables. In the end, one model was calibrated with different combinations of variables for each segmentation approach (Model form A). Two other models were calibrated, one using only
curvature (CCR) and AADT (Model form B), and one as base model for each approach, using AADT (Model form B) as the only explanatory variable. For all the models the segment length is included as an offset variable. In Table 4 the estimated models are presented with the value of QIC, and standard error and level of significance of variables. Models A, B and C assume, respectively, the following form:

\[
\begin{align*}
\text{Model A: } & E(Y) = e^{\alpha_0 + \alpha_5 \times L \times AADT + \sum \beta_i Var_{i+S}} \\
\text{Model B: } & E(Y) = e^{\alpha_0 + \alpha_5 \times L \times AADT + \sum \beta_i Var_{i+S}} \\
\text{Model C: } & E(Y) = e^{\alpha_0 + \alpha_5 \times L \times AADT + \sum \beta_i Var_{i+S}}
\end{align*}
\]

where:
- \( E(Y) \): expected annual crash frequency of random variable \( Y \);
- \( L \): length of road segment [m];
- \( AADT \): average annual daily traffic [veh/day];
- \( \alpha_0 \), \( \alpha_5 \): exponent of constant term of the model, and time trend, where the subscript \( S \) indicates the segmentation approach number;
- \( \alpha_i \): exponent of AADT, where the subscript \( S \) indicates the segmentation approach number;
- \( \beta_i \): set of parameters of the regression for different set of variables, with \( S \) indicating the segmentation approach number, and \( i (1=1, 2,...,7) \) the variable;
- \( Var_{i+S} \): set of variables resulting from the stepwise procedure, for each segmentation approach \( S \);
- \( CCR \): Curvature change rate [gon/m].

**Results**

The model calibration results are presented in Table 4, while the plots of the cumulative residuals are presented in Figures 2-4. As is evident from Table 4, Segmentation 4, with constant segment length, allows a greater number of variables to be fit than the other segmentations for the primary Model form (A). The segmentation based only on AADT (Segmentation 3), allows the estimation of a model with five of eight variables considered in the stepwise procedure. Similarly, for the model estimated for Segmentation 2, which is achieved by inserting 2 curves and 2 tangents in each segment, five variables were also significant, but the value of QIC is lower than for the Segmentation 3 SPF. The segmentation that gives the worst results in terms of number of variables that could be included in the model is Segmentation 5 in which all variables are constant within each segment. Besides, the model for Segmentation 5 has the highest value of QIC.

For Segmentation 3, which is based on AADT, the variables selected with the stepwise procedure have in general greater standard errors than those selected for other segmentation approaches. This is likely because, in motorways, AADT changes only at interchanges, so this segmentation approach can yield very long segments, with considerable within-segment variation in the other variables that cannot be adequately modeled.

Results reported in Table 4 show not only differences in the g.o.f. and number of explanatory variables, but also, sometimes, differences in the sign of the coefficients. This indicates, depending on the segmentation, an opposite influence of the variable on the expected number of crashes estimated by the SPF.
In general, Segmentation 2 gives the best results for the primary model form, based on both QIC and the CURE plots. The cumulative residual curves in Figure 2 oscillate closer to the value of zero than for the other segmentations. For model form B, in terms of the CURE plots, the best model is estimated from Segmentation 3, as is shown in Figure 3. For model form (C), based only on AADT, the model that oscillates closest to the value of zero is based on Segmentation 4. Only one model exceeds the ±2σ boundary -- the AADT-based model for Segmentation 3 (Figure 4 for the value of AADT close to 15,000 vehicle-day). Segmentation 5, characterized by constant value of variables inside each segment, gives the poorest results, similar to the earlier conclusion based on QIC.

In general, for all models, the CURE plots reveal that the models tend to underestimate the number of crashes for low values of AADT, and to overestimate crashes for higher values of AADT.

Table 4. Value of regression parameters, (p-value) and (Standard error) for different segmentations (1, 2, 3, 4, 5) and Model forms (A, B, C).

<table>
<thead>
<tr>
<th></th>
<th>Seg_1 (Curve Based)</th>
<th>Seg_2 (2 Curves, 2 Tangents)</th>
<th>Seg_3 (AADT Based)</th>
<th>Seg_4 (Fixed Length)</th>
<th>Seg_5 (Homogeneous)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interc. (β_{0,g,τ,g})</td>
<td>-20.4439 (&lt;.0001)</td>
<td>-21.7516 (&lt;.0001)</td>
<td>-13.8951 (0.0003)</td>
<td>-20.1429 (&lt;.0001)</td>
<td>-20.7288 (&lt;.0001)</td>
</tr>
<tr>
<td>AADT (β_{1,g})</td>
<td>1.3652 (&lt;.0001)</td>
<td>1.4797 (&lt;.0001)</td>
<td>0.8279 (0.0143)</td>
<td>1.3475 (&lt;.0001)</td>
<td>1.4273 (&lt;.0001)</td>
</tr>
<tr>
<td>CCR (β_{2,τ,g})</td>
<td>2508.331 (0.0054)</td>
<td>484.9824 (0.0042)</td>
<td>-2931.75 (0.0273)</td>
<td>262.6808 (0.0066)</td>
<td>0.2111 (0.0003)</td>
</tr>
<tr>
<td>I (β_{3,δ})</td>
<td>-</td>
<td>-</td>
<td>11.8788 (0.0172)</td>
<td>-14.3209 (&lt;.0001)</td>
<td>-6.0280 (0.0500)</td>
</tr>
<tr>
<td>i_δ (β_{4,δ})</td>
<td>5.1423 (0.0010)</td>
<td>-</td>
<td>11.8788 (0.0172)</td>
<td>-14.3209 (&lt;.0001)</td>
<td>-6.0280 (0.0500)</td>
</tr>
<tr>
<td>Tun (β_{5,τ,g})</td>
<td>0.0058 (0.0015)</td>
<td>0.0050 (0.0087)</td>
<td>0.0258 (0.0001)</td>
<td>0.0046 (0.0097)</td>
<td>-0.4540 (&lt;.0001)</td>
</tr>
<tr>
<td>RSH6 (β_{6,τ,g})</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>RSH56 (β_{7,τ,g})</td>
<td>-0.0031 (&lt;.0001)</td>
<td>-0.0623 (0.0001)</td>
<td>-0.0634 (0.0004)</td>
<td>-0.0037 (&lt;.0001)</td>
<td>-0.0009</td>
</tr>
<tr>
<td>SCR (β_{8,τ,g})</td>
<td>-2.3927 (&lt;.0001)</td>
<td>-2.3927 (&lt;.0001)</td>
<td>-2.3927 (&lt;.0001)</td>
<td>-2.3927 (&lt;.0001)</td>
<td>-2.3927 (&lt;.0001)</td>
</tr>
<tr>
<td>QIC</td>
<td>3322.00</td>
<td>1081.65</td>
<td>1761.16</td>
<td>2706.95</td>
<td>4510.73</td>
</tr>
</tbody>
</table>

Model with AADT and CCR (Model form B)

<table>
<thead>
<tr>
<th></th>
<th>Seg_1 (Curve Based)</th>
<th>Seg_2 (2 Curves, 2 Tangents)</th>
<th>Seg_3 (AADT Based)</th>
<th>Seg_4 (Fixed Length)</th>
<th>Seg_5 (Homogeneous)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interc. (β_{0,g,τ,g})</td>
<td>-19.1411 (&lt;.0001)</td>
<td>-20.7128 (&lt;.0001)</td>
<td>-26.3161 (&lt;.0001)</td>
<td>-20.7128 (&lt;.0001)</td>
<td>-19.783 (&lt;.0001)</td>
</tr>
<tr>
<td>AADT (β_{1,g})</td>
<td>1.2075 (&lt;.0001)</td>
<td>1.3713 (&lt;.0001)</td>
<td>1.8705 (&lt;.0001)</td>
<td>1.3713 (&lt;.0001)</td>
<td>1.2891 (&lt;.0001)</td>
</tr>
<tr>
<td>CCR (β_{2,τ,g})</td>
<td>23.7961 (0.0021)</td>
<td>489.8783 (0.0031)</td>
<td>2250.728 (0.005)</td>
<td>291.9741 (0.0020)</td>
<td>0.2022 (0.0006)</td>
</tr>
<tr>
<td>QIC</td>
<td>3325.57</td>
<td>1109.89</td>
<td>1462.87</td>
<td>2593.60</td>
<td>4580.81</td>
</tr>
</tbody>
</table>

Base Model (Model form C)

<table>
<thead>
<tr>
<th></th>
<th>Seg_1 (Curve Based)</th>
<th>Seg_2 (2 Curves, 2 Tangents)</th>
<th>Seg_3 (AADT Based)</th>
<th>Seg_4 (Fixed Length)</th>
<th>Seg_5 (Homogeneous)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interc. (β_{0,g,τ,g})</td>
<td>-19.1467 (&lt;.0001)</td>
<td>-20.8182 (&lt;.0001)</td>
<td>-26.2944 (&lt;.0001)</td>
<td>-19.1993 (&lt;.0001)</td>
<td>-19.8970 (&lt;.0001)</td>
</tr>
<tr>
<td>AADT (β_{1,g})</td>
<td>1.2358 (&lt;.0001)</td>
<td>1.2000 (&lt;.0001)</td>
<td>1.1515 (&lt;.0001)</td>
<td>1.2403 (&lt;.0001)</td>
<td>1.2163 (&lt;.0001)</td>
</tr>
<tr>
<td>QIC</td>
<td>2947.62</td>
<td>1103.22</td>
<td>1363.27</td>
<td>2583.40</td>
<td>4410.09</td>
</tr>
</tbody>
</table>
Figure 2 CURE Plots with ±2σ for model form A.
Figure 3 CURE Plots with ± 2σ for model form B.
SUMMARY AND CONCLUSIONS
The purpose of this paper was to investigate the influence of segmentation on the performance of safety performance functions (SPFs), in terms of goodness of fit and the variables that could be modeled. To do this, it was necessary to sometimes aggregate variables into a usable form, to have a constant value of each modeled variable in each segment.

Figure 4 CURE Plots with ±2σ for model form C.
Five different segmentation approaches were evaluated with three different model forms. One approach is based on the Highway Safety Manual (HSM) method, using curvature and AADT as the basis, one has two curves and two tangents inside each segment, one is based on a constant AADT in each segment, one has a fixed length of each segment, and one has all the variables constant within each segment relative to their original value.

One model was calibrated with different combinations of variables for each segmentation approach. Two other models were calibrated for each approach, one using only curvature and AADT, and one a base model, using AADT as the only explanatory variable. The models were estimated for a sample of rural motorways segments in Italy, using data for the years 2002 through 2009. The Generalized Estimating Equation (GEE) procedure was applied to develop the SPFs, which were evaluated using cumulative residual (CURE) plots and the Quasilikelihood under the Independence Model Criterion (QIC) value.

The best results were obtained for the segmentation based on two curves and two tangents (Segmentation 2) and the segmentation with fixed length (Segmentation 4). Segmentation 5, characterized by a constant value of all original variables inside each segment, was the poorest approach by all measures. This is likely because it yields very short segments, resulting in non-perfect identification of the location of crashes and in a large number of segments with zero crashes. Both factors create difficulties in making sound statistical inference.

Fixed length segmentation may be the most flexible in practical applications. This is because the segment length can be determined by data availability and quality, and other factors to optimize the SPF calibration. The length chosen for this research was based on pragmatic reasoning. Given the promise shown by the results, further research can explore alternative considerations for determining the length of fixed length segments used for SPF development. Similarly, the segmentation approach (Segmentation 2) based on two curves and two tangents in each segment, which also showed promise, could be further explored by considering different numbers of curves and tangents in a segment.

ACKNOWLEDGMENT
The authors wish to express their gratitude to the East Sicily Highway Police Department and to the Management of Sicilian motorways (CAS) for their help in getting the data used for the analysis. The Ryerson University collaboration was made possible by a Discovery Grant from the Natural Sciences and Engineering Research Council of Canada (NSERC).

REFERENCES


