The Pennsylvania State University

The Graduate School

College of Engineering

SAFETY MODELING VIA SEGMENTATION
OF TRANSPORTATION NETWORKS

A Dissertation in

Civil Engineering

by

Jun Seok Oh

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

December 2010

The dissertation of Jun Seok Oh was reviewed and approved* by the following:

Venky N. Shankar
Associate Professor of Civil and Environmental Engineering
Dissertation Advisor
Chair of Committee

Evelyn Thomchick
Associate Professor of Supply Chain and Information Systems

Suzanne M. Shontz
Assistant Professor of Computer Science and Engineering

Thorsten Wagener
Associate Professor of Civil and Environmental Engineering

Peggy A. Johnson
Professor of Civil and Environmental Engineering
Head of the Department of Civil and Environmental Engineering

*Signatures are on file in the Graduate School

ABSTRACT

Safety Modeling via Segmentation of Transportation Networks

Jun Seok Oh

Dissertation Advisor: Associate Professor Venky N. Shankar

This dissertation proposes a methodology to address a long-standing question in traffic safety relating to the evaluation of safety risk and the benefits associated with safety interventions. Traditionally, safety risk has been assessed at the corridor level, with corridors being evaluated in terms of accident rates, as in accidents per million vehicle miles. This measure allowed safety planners and engineers to look for correlations at the aggregate level using economic and sociodemographic data from counties and cities. As roadway geometric data became more widely available, both in terms of general access to public agencies and in terms of measurement detail, statistical models for safety were developed incorporating correlations between safety outcomes at the roadway segment level and roadway geometrics. This approach avoids the problem of ecological correlation that is likely to occur in modeling using county or city level independent variables. The problem of ecological correlation occurs when correlations between safety outcomes in corridors are evaluated using mean accident rates and county or city level means for independent variables. This approach assumes corridor and regional means reflect segment means accurately, an assumption that is not tenable,

especially when segmental heterogeneity is significant, as has been shown to be, in the safety context. Heterogeneity refers to the deviations in patterns of accident occurrences at individual roadway segments, and how these deviations are referred to an "average site." This average site can be a virtual site represented by the group mean. Heterogeneity results in overdispersion of accidents, meaning that the variance of the accident distribution exceeds the mean. This is due to the fact that the probability of accident occurrence is not uniformly distributed in space and time. Therefore, one can expect accidents to cluster at various locations on the transportation network, such as at intersections, interchanges, lane drops or lane additions, horizontal or vertical curves, or at locations where decisions have to be made by drivers regarding lane changing, braking, speed reduction or acceleration, or route change.

Given this background, the problem of evaluating safety interventions is compounded by the challenge of modeling the effect of heterogeneity simultaneously alongside the modeling of the marginal effect of an intervention. There are two primary contributors to this challenge. The first contributor is selection bias, which arises when locations for safety interventions are not randomly chosen. The second contributor is the scale of measurement of this bias. It may be that selection bias at aggregate scales (for example, in instances where corridor length treatments are applied) is influenced by heterogeneity in a different manner compared to bias at smaller scales (for example, spot interventions). The impact of this variation is that the assessment of safety interventions can be varied depending on the scale at which the evaluation is conducted. Hence, the methodological problem of simultaneously addressing heterogeneity and selection bias is the objective of this dissertation. This dissertation attempts to provide some perspective to this problem via multiple scales, by proposing a joint model of heterogeneity and selection bias using a discrete-count approach, and using this framework to address the following research questions:

a) What is the impact of selection bias on safety intervention due to scale? In other words, if safety interventions are applied at locations where accident patterns are severe and frequent, how does one account for the lack of intervention at less problematic locations? And how does a statistical methodology derived for selection bias provide inference across scales, as segments are scaled up from very small lengths to lengths of the order of corridors?
b) How does one represent insights into the policy implications of selection bias in a manner that integrates context (i.e., roadway location and characteristics) and scale?

I use freeway roadway lighting as an example safety intervention to make these evaluations in this dissertation. Roadway lighting is installed in order to improve traffic flow, thereby also contributing to improved roadway safety. Roadway lighting is installed in various forms – as in median-side lighting, versus right-side lighting, versus tunnel lighting, versus ramp-mainline merge points, versus, installations on both sides of the traveled way. This dissertation involved data collection on all 1,528 centerline miles

of interstate freeway in Washington State and analyzed the correlation of accident frequencies with roadway lighting installation, after accounting for roadway geometrics and traffic flow levels. It was determined that certain installations are more effective than others, when selection bias is taken into account. For example, right-side lighting installation is found to be effective in reducing accident frequencies compared to other types of lighting installation, indicating a 30% reduction in accident frequencies compared to segments where there is no roadway lighting at the lighting segmentation scale. Such a result appears to justify the installation of right-side lighting at critical locations such as ramp merge points or departure points. The key phrase is “appears to justify”. This dissertation explores the extent to which scale affects inferences such as the above. With different scales of segmentation, such as interchange and non-interchange segments, one mile uniform length segments, or accident-cluster length segments, right-side installation has a smaller reduction of accident frequencies compared to accident reduction at the lighting segmentation scale. In the case of accident-cluster level segmentation, right-side lighting installation is associated with an increase in accident frequency. This example result demonstrates that the scale of data plays a very important role in safety inferences, especially when heterogeneity and selectivity bias are accounted for.

While roadway lighting is used as an example for application of this dissertation’s analytical framework, it is expected that the full-purpose self-contained computational framework for analyzing safety outcomes will be of substantial interest to the safety community at large. One can use this framework for the analysis of any safety intervention at any scale. The framework incorporates the typical geometric design decisions used in practice, and therefore, analysts can use this framework to address selectivity bias arising from roadway improvement projects involving all geometric types. In particular, the framework developed in this dissertation can also aid decision makers to conduct scenario testing. One example of scenario testing would be to examine the impact of energy-conservation efforts on traffic safety patterns on urban and rural freeways. Another would be to explore the design contexts associated with high levels of unobserved heterogeneity, where the discussion on the measurement of factors that do not currently exist in highway databases can be motivated. Example factors relating to heterogeneity could involve measures of segment-level kinematics such as speed, speed dispersion, and headway following distances. Or, they could involve microclimatic measurements such as pavement temperatures, determination of icing likelihoods, wind gust speeds and sun angles.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1

INTRODUCTION

Since the development of the American Association for State Highway Transportation Officials (AASHTO) Strategic Highway Safety Plan (SHSP), several states have adopted a similar approach in developing their own highway safety plan. The Washington Department of Transportation (WSDOT) is a recognized leader in this area. The WSDOT conducts statistical modeling and visualization analyses as part of their accident research initiative in order to establish systematic bases in its strategic safety plan. This dissertation is an in-depth, original look at statistical approaches appropriate for public agency decision making; hence, the focus of this dissertation is empirical. The main motivations for this dissertation are drawn from current and prior research conducted by the author for the Washington State Department of Transportation.

The WSDOT research effort was begun in 2006, with the author leading research activities in the area of data collection methods for freeway accident data systems. Significant goals of the effort were: a) exchangeability of data in multiple formats, b) usability of data for the development of statistical models, c) post-processing of model outputs for visualization, and d) usability of the above components for integrated prioritization of freeway corridors. The author demonstrated the viability of components "a" and "b" via two bodies of work, namely his MS thesis (Oh 2006) and a research report published for the WSDOT in June 2008 (Shankar et al, 2008).

The earlier work in 2006 reproduced accident, geometric, and traffic flow data by direction for 124 centerline miles in a consistent, complete record format. The latter work in 2008 extended this reproduction to the entire interstate system consisting of 1,528 centerline miles. To the author's knowledge, consistent and complete database compilation involving over 100 accident related variables, traffic flow, and geometrics at a statewide scale has not been done in the nation. While the lack of such efforts may sound surprising, potential reasons do exist. Some challenges occurred when creating record consistency and completeness in the proper format. For example, some highway log information such as number of lanes, shoulder widths, and presence of median barrier by type, were available as text documents in their original form. Traffic flow data such as annual average daily traffic (AADT) was available in electronic form at 0.1-mile or 1-mile intervals. The 0.1-mile data were interpolations of AADT measured using loop detectors which are not necessarily regularly placed at 0.1-mile or 1-mile intervals on the state interstate system. Interpolations were produced by WSDOT in-house through a feedback algorithm that ensures consistency with neighboring AADT computations. Accident data was available for multiple years in the form of detailed accident reports that contained information by severity type (property damage only, possible injury, evident injury, disabling injury and fatality), type of collision such as entering at angle, sideswipe, same direction, fixed object, overturn or headon, vehicle involvement, driver related factors such as alcohol or drug involvement, seat belt use, age, gender, occupant information including factors similar to that for the driver and in addition, occupant position in vehicle, and environmental factors such as occurrence of snowy, icy, rainy or dry driving conditions, as well as presence or absence of roadway lighting. It should be noted here that this database is event-specific. In order to construct segment-level decision frameworks, which is the primary objective of this dissertation, event-specific information needs to be aggregated to appropriate scales. The appropriateness of scale depends on the level and nature of the research questions being asked. In this dissertation, the following research questions are asked:

a) What is the impact of selection bias on safety intervention due to scale? In other words, if safety interventions are applied at locations where accident patterns are severe and frequent, how does one account for the lack of intervention at less problematic locations? And how does a statistical methodology derived for selection bias provide inference across scales, as segments are scaled up from very small lengths to lengths of the order of corridors?

b) How does one represent insights into the policy implications of selection bias in a manner that integrates context (i.e., roadway location and characteristics) and scale?

Given this background and objectives, the remainder of this dissertation is organized as follows. I review literature of direct relevance to the dissertation and in addition provide a bibliography relevant to the dissertation itself. The review includes segmentation and selection bias research related to transportation applications. In this sense, the technical benefit of the work will be to provide flexibility, and as a result, scalability in modeling safety.

It was noted that the marginal effects of key infrastructure variables are of interest. From a policy standpoint, this is definitely a substantial motivation, because cost effectiveness is a major factor driving prioritization schemes for safety, mobility, and accessibility related infrastructure improvements. Herein lies an issue of statistical significance; typically, transportation improvements are applied at locations where a need is determined to exist. Therefore, the application of improvements is not random; rather, it follows a selection rule. In the transportation case, a selection rule may be based on ordering of need. A decision making framework that is empirically based uses accident data observed at either the selection locations alone, or also at locations where improvements are not applied. In either case, some accounting needs to occur for the selection bias associated with the marginal effect of the improvement in question. For example, if I consider roadway lighting as the variable of interest, then it can be argued that the marginal effect of roadway lighting can be expected to decrease accident propensity at locations where lighting is installed. A policy based on the examination of just lighting-only locations may estimate the effectiveness of roadway lighting with bias. Statistical and econometric methods involving the treatment of selectivity bias are documented, and the literature review in this dissertation addresses that.

Following the literature review, I present the methods employed in this dissertation. A description of data collection and segmentation methods to obtain structured datasets at multiple scales for accident analyses is provided. Statistical modeling designs for this research are also presented. I discuss the results to demonstrate the viability of the methods proposed and conclude with major findings and recommendations for future work.

Chapter 2

RELATED WORKS AND RESEARCH QUESTIONS

2.1 Segmentation Methods

Location based analysis of transportation accidents plays an important role in safety prioritization. This is primarily because estimation of safety risk must be conducted using location-specific attributes such as geometrics, traffic volumes and environmental conditions. For the remainder of this dissertation, the words location and segment are used interchangeably. Alternative risk estimation approaches involving regional socio-economic characteristics do not provide accurate estimates of segment level risk, although they do provide trend estimates at the regional level for the system as a whole. Regional models have other shortcomings in their employment for segment-level risk estimation. First is the issue of ecological correlation. Ecological correlation is an effect that arises out of the use of regional means rather than location-specific values thereby introducing bias where location specific values deviate significantly from regional means. This is particularly true for the accident context since location specific frequencies of accidents can be significantly different from a regional mean. A regional mean in accident context can refer to a district or region of administration within a state. Typically a region or district of administration is an area delineated on the basis of topography, balance of geographic coverage and annual differences in environmental conditions. Washington State, which serves as the empirical context for this dissertation, has six regions encompassing in total 7,100 centerline miles of state highway. As a consequence of location specific deviations from a representative mean, the problem of heterogeneity arises in the modeling of safety risk. Heterogeneity can occur due to various contributing factors – geometric changes, environmental changes, traffic flow changes, and changes in driver behavior due to driving context such as urban versus rural. A consequence of heterogeneity is the concentration of accidents spatially that contributes to the problem of overdispersion in the estimation of accident counts. Overdispersion means the variance of the count is significantly greater than the mean. This problem of overdispersion is nonlinear in accident contexts – the variance-mean relationship is quadratic (see for example Shankar et al 1995). Nonlinearities compound the overdispersion problem when locations are scaled up to corridors or regions – another reason for bottom-up estimation. A final note relating to segmentation relates to the problem of aggregation bias. Aggregation bias occurs when event occurrence probabilities are estimated using mean values of the independent variables – a problem sure to occur when regional values are used. For the above mentioned reasons, I consider segmentation literature and methods that deal directly with the dependent variable of interest, frequency of accidents, or deal with independent variables in a manner that minimizes heterogeneity and aggregation bias.

Segmentation methods commonly used in the contemporary statistical context of traffic safety are independent variable based (see for example Shankar 1995; Lord 2000).

Since the early work of Shankar 1995, where fixed-length segments were employed, the current state of the art involves homogeneous segmentation where segment lengths are defined on the basis of *homogeneity of all independent variables*. Lord (2000) is a good example of this application. The rule here is for an X vector of dimension K, with Var[X_k]=0 for all X for any segment j. The main idea here is that by constraining the within-segment variance of X_k to be zero, all inference is conditioned on between-segment variance. However, the downside of this approach is one can end up with too small a segment length (of the order of 0.01 miles or 52.80 feet). This creates an artificially induced estimation and prediction problem – with numerous small segments, the sample of segments can have a predominance of zero frequency values and hence, may over-represent the concentration of accident risk at few locations. In contrast to the homogeneous length approach, the fixed-length approach can be viewed as arbitrary – for example, if one were to use one-mile segment lengths, the only justification is a practical one in the sense that the segmentation decision is usually driven by a variable of particular interest for the study. For example, Shankar et al (1995) proposed to study the interaction between roadway geometrics and environmental conditions such as precipitation and the effects these interactions had on accident occurrences on rural freeway sections in Washington State. The segment lengths are fixed to be consistent with the spatial interval of measurement for precipitation data. A secondary justification offered in the Shankar study was that the error distributions were roughly independently and identically distributed (IID) at that scale. It is not clear whether several segmentations were compared, but Shankar et al suggest that the inclusion of environmental data in the analysis of accident risk can pose substantial segmentation challenges. I interpret this to be a generalized challenge for the field of safety as a whole – the current literature has surprisingly not dealt with this issue at depth. I can speculate two reasons – the first relating to the availability of environmental data, and the second to the plethora of methodological issues that can complicate the segmentation problem. Environmental data for accident analysis are difficult to measure and maintain on a consistent basis, for the reason that weather station data is not readily useful for pavement level inferences – which is in fact the level at which roadway accidents occur. Weather station data are fixed station measurements that can contain considerable altitude variations depending on the location of the weather station. For example, in Washington State, over 250 weather stations are used as permanent recording stations; however, less than 10 percent of these stations are located close enough to the roadway for even rough assumptions regarding pavement level environmental conditions. So, considerable post-processing of weather station information is required before segment level analyses can commence. Importantly, pavement icing probabilities and temperature variations, factors strongly associated with accident occurrence, are not measured by permanent weather stations. Such data can be estimated (see for example Senn 2005) but require a meteorological basis, a methodological issue beyond the domain of traffic safety estimation methods.

Roadway segmentation is fairly common in Geographic Information System (GIS) applications in transportation. Nyerges (1990) introduced a locational referencing and roadway segmentation method on the basis of simple transportation referencing schema.

Most transportation organizations use a combination of three schemes for transportation information referencing. They are road name and milepost, control section, and link and node. A control section is defined as a segment which has homogeneous types of information. Link schemes are used for the connection of two or more roadways, with nodes being connection points on these roadways. While research based on the link-node concept is well established in graph theory and well-known transportation problems relating to traffic assignment, little integration has occurred in terms of relating this concept to network safety, especially via segmentation. Segmentation is an area that can be fruitful in terms of benefits for field integration – that is, bringing together concepts in traffic demand and assignment, traffic flow, and safety.

Dueker (2000) and Butler (2001) are other prominent examples of segmentation applications using GIS attributes for network links, nodes, route, traversal segments of routes and geographic coordinates. In an uncommon example, Quiroga (2000) uses dependent variables such as travel time and speed to segment transportation systems. However, the segmentation purpose in this paper is quite basic – the network is segmented so to provide the ability to retrieve particular component information; not necessarily for the purpose of evaluating model behavior across scales, which is a central issue in this dissertation. By far, the literature on dependent variable based assessments of traffic safety is limited to clustering analyses (see for example Tarko and Karlaftis 1998). In this thought-provoking paper, the authors look at the assessment of heterogeneity using clustering methods; however, their focus is strictly on the evaluation of similar heterogeneities across the network, a conclusion drawn on the basis of clustering. Locations within a cluster are estimated to share heterogeneities of similar magnitudes. I find this to be an interesting objective; while there appears to be statistical support, via automatic clustering, the method is not informed by an evaluation of its viability for statistical model estimation. That is, how does one construct a model with network segment clusters that are geographically dispersed? Research by Anagnostopoulos (2006) raises similar questions in his address of segmentation using time scales in addition to geographic scales.

In short, the literature on segmentation is scant and surprisingly limited in terms of guidance on issues relating to either linear or grid network components. For example, the following questions are not addressed:

a) How does one segment in order to address multidimensional heterogeneity? The common answer appears to be the very short segment approach – which I have discussed at some length as being problematic due to the problem of excess zeros. I refresh the reader that the problem of very short segments arises because the focus is on independent variables alone.

b) How does one segment in order to address longitudinal heterogeneity? A well established method is not apparent in the literature.

An integrative approach would involve a justifiable segmentation approach that serves as the foundation of multi-scale analysis of traffic safety. By “justifiable,” I mean in a somewhat narrow sense quantifiable objectives, such as consistency of outcomes in their ability to serve as key indicators of network safety, while at the same time, providing guidance on definitions for measures of mobility and accessibility as well. For example, if network safety evaluation can be consistently measured across scales via an objective measure such as total number of accidents, or severity index, or cost, then one can ask the question: Can the same scales be useful for simultaneous analysis of mobility and accessibility? In this sense, segmentation theory is fundamental. I discuss in the following chapter two approaches, one that is exogenous segmentation, and one that is endogenous segmentation.

There are two aspects of consistency here that motivate the rest of this dissertation. The first aspect as I just discussed is a general theory for segmentation – should it be purely endogenous, or some hybrid that accounts for exogenous variables as well. Incidentally, Anagnostopoulos (2006) provides some guidance here through the field of dynamic programming. The second aspect of consistency on an empirical basis is the development of understanding the variation in heterogeneity across scales, and how this can affect inferences. I acknowledge that this is purely statistical, but I argue that without a solid statistical basis, guidance on causal mechanism investigation can be misleading. For example, if we find that policy effects associated with lighting installation can be consistently inferred across scales through the sign of coefficients as well as the magnitude of impact, then the remaining important issue statistically speaking relates to uncertainty. How uncertainty varies across scales can be decomposed into model uncertainty and parameter uncertainty. This can provide further guidance for targeted research on causal mechanisms and extend the domain of methods from purely statistical to possibly statistical-physical or beyond.

2.2 Selectivity Bias and Two Step Process

The infrastructure policy assessment problem is essentially a selection bias problem. Research in the area of selection bias correction in traffic safety is nonexistent. This is not surprising since the field has spent a majority of its focus on evaluation techniques for before-after scenarios involving the application of policy variables. Green et al (2003) present the results on a before and after analysis for the impact of lighting at a rural intersection in District 12 in Kentucky State. They conclude the mean crash rate per year is reduced by about 45% after the lighting installation. Isebrands et al (2004) conducted a before and after study at rural intersections in Iowa State. They used linear regression models to get the mean ratios of night crashes to total crashes at the 10% level of significance. The ratio of night crashes to total crashes is reduced by 15% in after-lighting installation period. Also, Poisson regression models were used for comparing the mean crash rates during the before and after installation periods. Isebrands et al say the expected night crash rate in the before installation period is 54% higher than the after installation period at the 10% significant level. The day time crash rate is increased by 24% from the before to after period, but it is not found to be statistically significant. Washington et al (2006) present a method to evaluate the effectiveness of left turn lanes on traffic safety. They present the evaluation problem as an endogeneity issue. Endogeneity arises when a bi-directional relationship exists between the dependent variable and an independent variable. Strictly speaking, this is an effect that can bias parameter estimates for the policy relevant independent variable due to nonzero covariance between the policy variable and the error term in the estimation equation.

Washington et al use this interpretation in their evaluation on the effect of left turn lanes on traffic safety. They argue that locations with high accident profiles may have a higher probability of left turn lane installation, and proceed to estimate a statistical model that corrects for this dependence. In this dissertation I encounter a similar problem, but it is not interpreted in the same way, although the statistical underpinnings for problem resolution are very similar. The problem is motivated by the selection of roadway lighting as the policy variable of interest in this dissertation. I examine the issue of roadway lighting since it is emerging as a policy variable of great interest motivated by energy efficiency and sustainability issues. Agencies are beginning to ask whether roadway lighting has substantial safety benefits, and whether there are design situations where non-installation would provide substantial energy savings, but not result in increased societal costs due to safety problems. The roadway lighting installation decision is not set by firm rules that depend on traffic volume, or curvature, or restricted sight distance – common measures that would indicate a safety problem. Rather, it appears somewhat ad-hoc that lighting installation is found to be more frequent in urban than rural settings. One can argue that urban volumes are higher; in addition, the density of interchanges and overpasses is greater, implying that situations which require decision distances for drivers making lane changes or route choice changes may be driving lighting installation choice. I find only part of this to be true in my examination of lighting installation choice. There is substantial departure from these afore-mentioned conditions in cases where lighting has been installed either on the median or right hand side. Some of this departure is attributable to transition zone effects, i.e., when roadway character changes from rural to urban. In some other cases, it is attributable to the presence of short curves, whether they are located on bridges, stream crossings, or culverts. The main point here is that lighting installation choice does not appear to be restricted to safety motivating situations if one were to measure safety through observed outcomes such as accident counts and severity. *However, the situations where lighting installation occurs may have been perceived to be dangerous by engineers. It is the nature of the heterogeneity behind this perception that motivates the lighting installation choice problem as a selection bias problem.* In simple words, the problem can be characterized as follows:

The choice to install lighting is selective due to the presence of observable segment attributes such as traffic volume, curvature, interchange and overpass density, and lane capacity. In addition, the choice is influenced by unobserved heterogeneity. The unobserved heterogeneity may also affect the magnitude of the count outcome due to the fact that the outcome is conditioned on selectivity; hence, heterogeneity motivates the selection problem. The key to examining empirical evidence of heterogeneity rests in large part on scale. How does heterogeneity influence selection bias across scales? The larger the degree of heterogeneity, the potentially larger the selection bias. One can then expect that as heterogeneity increases, the estimation of lighting impacts on safety becomes more uncertain. The nominal way to estimate lighting impacts is to estimate average treatment effects across the sample.

Much of the work on selection bias correction was pioneered by James Heckman who won the Nobel Prize in Economics in 2000 along with Daniel McFadden. In fact, the two share a common methodological interest in the treatment of the selection bias problem. McFadden's methods (see for example Durbin and McFadden 1984) are applicable under specific types of selection bias, whereas Heckman's selection correction methods have been applied across a variety of policy scenarios (see for example, Heckman 1979, 1990). The essential idea Heckman posts is that the bias from using non-randomly selected samples arises because of a missing data problem. In other words, non-random selection of locations only provides us with before and after effects of a policy for that subsample. We do not observe the policy effects for observations that were not selected for the policy application. In the lighting case, the same question can be asked. In essence, this is a counterfactual approach, i.e., what would the policy effect be if it were applied randomly. To deal with this missing data problem, Heckman's selectivity bias correction involves two stages. In the first stage a probit or logit analysis is conducted to predict the probability of policy application for any observation. An observation here refers to the unit of analysis (in our case, a segment). In the second stage, a linear regression is conducted with the predicted probit/logit probability as the independent variable in lieu of the original policy variable. If the predicted probability coefficient is significant in the second stage, then the policy variable is estimated to statistically influence the outcome of the policy. In our case, the outcome is the number of accidents per segment, and the policy variable is the decision to install lighting in a given segment. Generally, the regression model has the following form:

$$y_i = \alpha + \beta' x_i + \varepsilon_i \qquad (1)$$

where α is an intercept term, β are estimable coefficients, x is the observed variable, y is the dependent variable, and ε is the random unobserved error term. Because the general regression model cannot capture selection bias in independent variables, the model is corrected by adding the Inverse Mill's ratio term for explaining the seemingly chosen portion.

The selection bias correction model is presented as:

$$\mathrm{Y_i} = \alpha + \beta' \mathrm{x_i} + \sigma\lambda_{\mathrm{i}} + \varepsilon_{\mathrm{i}} \tag{2}$$

where λ is the Inverse Mill's ratio and σ is the estimated coefficient. The Inverse Mill's ratio is a measure of the "selection hazard," meaning that it provides the instantaneous probability of a segment being selected for lighting given that it has not been chosen. In other words, the Inverse Mill's ratio takes into account the probability of a non-selected location being selected, thereby accounting for non-randomness. Some problems exist with this basic Heckman approach, especially when the error distribution is not normal. In this case, Lee (1984) has provided several options through non-normal distributions. In a more general sense, for our application, using generalized extreme value distributions to estimate lighting installation choice is reasonable. The second aspect of the Heckman two-step process that is restrictive relates to the nature of the choice. So far I have only discussed a binary choice followed by a regression function. In the lighting choice situation, the choice is polytomous, meaning, we have more than two choices (lighting or no lighting) and up to seven distinct and mutually exclusive choices. These relate to the location on the roadside where installation is chosen. For example, the location choices can include a) median side only, b) median and right side, c) right side only, d) lighting in tunnels, e) median side point, and f) right side point. Lighting at a single point such as a merge point between an on-ramp and the mainline, is different from continuous lighting, where a series of luminaires are installed. Choices “a” through “d” reflect continuous lighting types, whereas “e” and “f” reflect point installations; the baseline choice is no lighting. For a polytomous model with an extreme value distribution, I propose to use the mixed logit approach, which accounts for choice heterogeneity in the selection problem very flexibly. It is flexible in the sense that no apriori structure is required for estimating say, a nested logit type model. One can directly estimate a single-level mixed logit model, compute the predicted probabilities, and then proceed to incorporate them in the second stage regression model. Given these preliminaries, some methodological issues of note are to be presented. First, in the sense of Heckman, the selection bias problem is similar to an omitted variables bias problem (see Puhani 2000 for details). Second, the problem of identification is a significant one in the estimation of the regression equation for the outcome. As collinearity between the vector of regressors in the choice equation and the outcome increases, identification becomes increasingly challenging. At a minimum, it is recommended that at least one variable in the choice equation be omitted from the outcome regression equation. A more generalized view of this is based on the fact that the Inverse Mill’s ratio is quasi-linear, meaning that only in very extreme samples where choice selection as opposed to no selection approaches unity does the Inverse Mill’s ratio become non-linear, which in turn means that identical regressors in both equations may still make estimation feasible in those extreme ranges (Puhani 2000). This is not usually the case in the lighting choice problem; the non-selection probability is non-trivial. This probability will vary depending on the scale (length of segment) at which lighting choice is evaluated. Hence, a good empirical procedure is to estimate Inverse Mill's ratios from the choice equation and regress those on the regressors in the outcome equation. If the adjusted R-squared of this regression is less than 0.8, then proceeding with the outcome regression is admissible. Otherwise, it is desirable to find more instruments in the choice equation, namely, regressors that are not in the outcome equation. A final note relates to the consistency of the variance-covariance matrix of the regression outcome equation. This may not be consistent, and it may be necessary to adjust for inconsistency using robust approaches such as the White estimator (White 1980) or an estimator that accounts for variability arising from the first-stage prediction (Murphy and Topel 1985).

To summarize, selection bias correction in transportation safety policy is practically non-existent; it appears there is one article in the published literature that addresses the topic as an endogeneity issue, which is not necessarily an accurate characterization in most policy application situations. The lighting choice problem serves as an example. The second limitation with the selection bias correction problem relates to extending methods from the Heckman two-step process with a binary choice selection equation to a polytomous selection choice equation. The second variant I will address relates to the use of nonlinear regression models for the outcome equation in the second step, whereas the Heckman procedure uses the classical ordinary least squares (OLS) regression equation. The need for a nonlinear regression in the second step arises from the fact that I will be dealing with counts of accidents in years, for any given segment.

Using a linear OLS model will produce inconsistent parameter estimates. A nonlinear regression model such as Poisson or negative binomial model has been found to be suitable for count regressions where heterogeneity is plausible. If heterogeneity is significant, a negative binomial (NB) regression usually suffices. The complication in my empirical case arises from the fact that the NB regression is a second-stage regression and not an independent regression. Hence, I must factor in the heterogeneity effect common to both the selection equation and the count outcome equation. This makes the estimation process somewhat challenging since the estimation now requires a single-step procedure as opposed to a two-step procedure in order to be efficient.

Given these preliminaries, I discuss the extension of the Heckman approach to our empirical case by using a richer lighting choice set in the first step, and accident frequencies as the outcome in the second step. This procedure, if feasible, can be conducted from any subset of frequencies, such as fatalities, injuries, or non-injuries. It can also be conducted for collision types, such as rear ends, overturns, fixed object hits, or other types. In this sense, a procedure accommodating a polytomous choice along with a count regression outcome that accounts for heterogeneity and overdispersion is a useful method. *This is a central and main scientific contribution of my dissertation, the other being a justifiable theory for network segmentation.* It helps answer the research questions I initially raised in the introduction section of this dissertation. These questions are revisited for quick reference:

a) What is the impact of selection bias on safety intervention due to scale? In other words, if safety interventions are applied at locations where accident patterns are severe and frequent, how does one account for the lack of intervention at less problematic locations? And how does a statistical methodology derived for selection bias provide inference across scales, as segments are scaled up from very small lengths to lengths of the order of corridors?

b) How does one represent insights into the policy implications of selection bias in a manner that integrates context (i.e., roadway location and characteristics) and scale?

2.3 Mixed Logit Model

Since the lighting choice equation involves seven choices as previously described, the multinomial logit model (MNL) developed by McFadden (1984) is the most popular discrete choice model and can be used for the choice estimation equation in a Heckman-type two stage selectivity bias correction process. MNL operates on the basis of independent and identical distribution of random components in the utilities assumption. Generally, the utility function (U) of MNL has the following form:

$$U_{ij} = \alpha_j + \beta_j^{'} x_{ij} + \varepsilon_{ij} \tag{3}$$

where, *α* is a constant term for alternative *j*, *x* is the observed variable in individual *i* for alternative *j*, *β* are coefficients, and *ε* is a randomly distributed unobserved utility.

With the utility function, the probability that individual *i* selects alternative *j* is given by:

$$\Pr(j_i \mid x_i) = \frac{e^{U_{ij}}}{\sum_{p=1}^{J} e^{U_{ip}}}$$

However, the MNL model is based on the IID assumption and cannot capture preference heterogeneity in individual characteristics. The mixed logit model provides features for model analysis, including observed and unobserved heterogeneity, from a variety of sources. The individual specific random parameter is introduced in the mixed logit model, so that the parameters are randomly distributed over individuals with unique means and variances in each individual. With unique parameters in each individual, the utility function of the mixed logit model can be considered as follows:

$$U_{ij} = \alpha_j + (\overline{\beta}_j^{'} + \sigma_{ij}^{'})x_{ij} + \varepsilon_{ij} \tag{4}$$

where *σ'* are the observed heterogeneity term to capture preference heterogeneity for individual *i* with alternative *j*, and the other parameters are the same as those in the utility function of the MNL model.

Deb et al (2006) introduced the treatment-effects model, which can be used when one treatment is chosen from more than two choices. This model uses a mixed multinomial logit (MMNL) structure to capture the effects of unobserved factors as well as observed factors. I begin my main analysis by applying the treatment-effects model of Deb et al (2006). In their paper, they use a shared heterogeneity term to motivate the selection bias problem influencing treatment-effects. They analyze the choice of health insurance plan type, and the outcome measures medical care usage by the treatment-effects model. For the model development, the indirect utility function is defined as follows:

$$U_{ij} = \alpha_j + \beta_j^{'} x_{ij} + (\delta_j^{'} l_{ij} + \varepsilon_{ij}) \tag{5}$$

where *x* denotes the exogenous variable, *l* is an unobserved factor for individual *i* and treatment *j*, and *ε* is the independently and identically distributed error term.

With the above utility function, the probability of treatment can be described with a mixed multinomial logit structure as follows:

$$\Pr(d_i \mid x_i, l_i) = \frac{e^{U_{ij}}}{1 + \sum_{p=1}^{J} e^{U_{ip}}}$$

where d is the binary variable representing the observed treatment choice.

So, the expected outcome equation for individual *i* can be defined as:

$$E(Y_i \mid d_i, z_i, l_i) = \alpha + \beta^{'} z_i + \sum_{j=1}^{J} \omega_j d_{ij} + \sum_{j=1}^{J} \lambda_j l_{ij} \tag{6}$$

where *z* is a set of exogenous variable associated with the treatment effects parameters, *d* is the choice variable, and *l* is the shared latent variable.

As noted, the shared heterogeneity makes it a single-step estimation procedure which is likely to be more efficient than the traditional Heckman type approach. In summary, the Deb approach allows for evaluating treatment selection effects on a count outcome through observed and unobserved factors.

Chapter 3

SAFETY ANALYSIS PROCESS AND EMPRIRCAL SETTINGS

3.1 Safety Analysis Process

The focal point of this dissertation is to demonstrate the effect of data scale on selection bias in safety analysis. For the analysis on the basis of the data scale effect, various segmentation approaches were employed. Figure 1 shows the taxonomy of segmentation types used in this dissertation. In the top half of the figure above the dotted line, exogenous segmentation types are shown, along with the modeling method employed for that type of segmentation. In the bottom half of the figure, endogenous segmentation types are shown, along with modeling method used for analyzing that type of segmentation. Exogenous segmentation includes segmentation on the basis of independent variables such as lighting type, exposure offsets such as length of segment, and geometric network classifiers such as interchange and noninterchange segments. Endogenous segmentation is based on the outcome's distribution in space – in this case the total annual frequency of accidents. Accident clusters are identified using the method of medoid based clustering, which in turn provides for the opportunity to directly link lighting presence with accident occurrence in space.

The modeling method used for analyzing exogenous segmentation datasets involves the joint model of lighting type or sequence choice and the accident count outcome. In essence, this is analogous to a full information simulated maximum likelihood method. This method becomes computationally burdensome and impractical in datasets where endogenous segmentation yields a large number of segments. Therefore, for the endogenous segmentation dataset, a two-step Heckman-type approach is used involving predicted probabilities from a discrete choice model of lighting type sequence in the outcome equation as an independent variable.

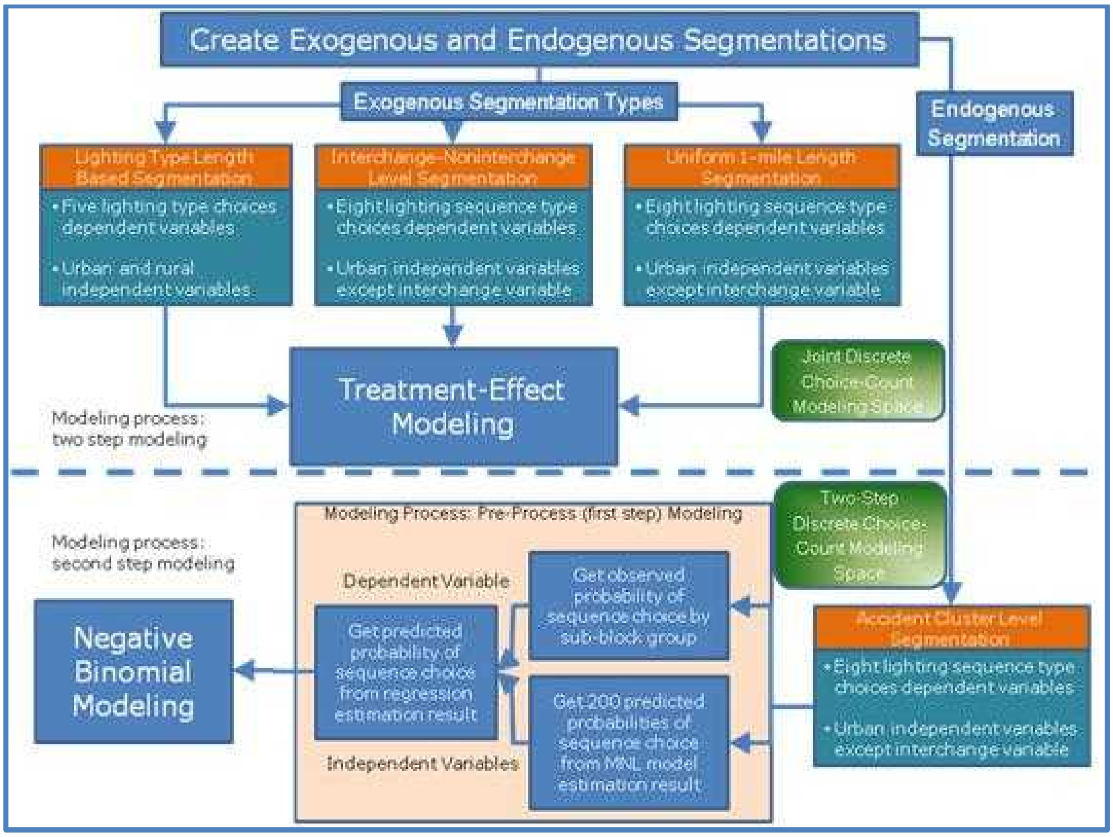

Create Exogenous and Endogenous Segmentations
Exogenous Segmentation Types
Endogenous Segmentation
Lighting Type Length Based Segmentation
•Five lighting type choices dependent variables
•Urban and rural independent variables
Interchange-Noninterchange Level Segmentation
•Eight lighting sequence type choices dependent variables
•Urban independent variables except interchange variable
Uniform 1-mile Length Segmentation
•Eight lighting sequence type choices dependent variables
•Urban independent variables except interchange variable
Joint Discrete Choice-Count Modeling Space
Modeling process: two step modeling
Treatment-Effect Modeling
Two-Step Discrete Choice-Count Modeling Space
Modeling Process: Pre-Process (first step) Modeling
Modeling process: second step modeling
Dependent Variable
Get observed probability of sequence choice by sub-block group
Negative Binomial Modeling
Get predicted probability of sequence choice from regression estimation result
Get 200 predicted probabilities of sequence choice from MNL model estimation result
Independent Variables
Accident Cluster Level Segmentation
•Eight lighting sequence type choices dependent variables
•Urban independent variables except interchange variable

Figure 1: Safety Analysis Procedure.

Safety analysis will be achieved following the process in Figure 1. Four types of segmentation datasets are created for safety analysis. Segmentation datasets by lighting installation type, interchange existence, and one mile section are prepared as exogenous segmentation datasets while accident clustering is used for the case of endogenous segmentation. Definitions of segmentation will be presented in the next section.

Two types of models are estimated for safety analysis. The key variable for selection bias is the lighting installation variable, which allows the predicted lighting choice values to become independent variables at the second step of model estimation. A treatment-effect model is used for the exogenous segmentation datasets. As described in the previous section, the treatment-effect model is the one step estimation model that condenses Heckman's two step selection bias estimation into one operation.

Both the treatment-effect model and the mixed multinomial logit model cannot handle the large dataset containing small scale segments. The size of data can be one problem because both models are based on simulation, in which the allocation of data into memory at the random draw step for computation can halt the modeling. This may cause the models to create inappropriate coefficients by the problem even though they are successfully estimated. Due to the main problem of regular selection bias estimation, alternate two step estimation methods are used for the small scale dataset in this dissertation. The predicted lighting choice probabilities are estimated by regression with observed probabilities and random predicted probabilities, as dependent variables and independent variables, respectively, at the pre-process modeling step. The regression model can be defined as:

$$p_{ij} = \alpha_j + \sum_{k=1}^{200} \beta_{jk} x_{ijk} + \varepsilon_{ij} \tag{7}$$

where p is predicted probabilities of sequence choice by regression for individual i and alternative j, α is the constant, β are coefficients of random choice predicted probabilities, and x are the 200 random predicted probabilities of sequence choice.

Table 1: Highway Capacity Manual Design Criteria for Geometric Sub-Block.

Six traffic flow level (q)	Six shoulder width combinations	Four curve combinations
q ≤ 1440 vphpl	(Left SW ≤ 2) AND (Right SW ≤ 2)	(Horizontal Curve = 0) AND (Vertical Curve = 0)
1400 < q ≤ 1650 vphpl	(Left SW ≤ 2) AND (2 < Right SW ≤ 10)	(Horizontal Curve ≠ 0) AND (Vertical Curve = 0)
1400 < q ≤ 1900 vphpl	(Left SW ≤ 2) AND (Right SW > 10)	(Horizontal Curve = 0) AND (Vertical Curve ≠ 0)
1900 < q ≤ 2150 vphpl	(Left SW > 2) AND (Right SW ≤ 2)	(Horizontal Curve ≠ 0) AND (Vertical Curve ≠ 0)
2150 < q ≤ 2400 vphpl	(Left SW > 2) AND (2 < Right SW ≤ 10)	
q > 2400 vphpl	(Left SW > 2) AND (Right SW > 10)	

For the simulation method, random choice probabilities are obtained by generating two hundred random numbers on the basis of a multivariate random distribution. The inputs for random number generation are mean and standard deviation numbers from the multinomial logit model estimation results. The predicted probabilities are calibrated against the observed probabilities for a sequence type (calculated by dividing the number of observations in the sequence choice within the sub-block by the total number of observations in the sub-block). The sub-block for computation of observed probabilities is defined based on the Highway Capacity Manual (HCM) level design criteria. This process is carried out for the various sub-blocks defined in Table 1. The total interstate network is divided into 144 sub-blocks by the criteria shown in Table 1.

The negative binomial model is estimated with predicted probabilities from pre-processing and other exogenous variables at the second step. This is similar to the second step estimation in Heckman's two selection correction model; the model equation is:

$$\ln(y_i) = \alpha + \beta' z_i + \sigma_j p_{ij} + \varepsilon_i \tag{8}$$

where y is the accident count for individual *i*, *α* is the constant, *β* are coefficients of exogenous variables, *z* are exogenous dependent variables, *σ* are coefficients of predicted sequence choice probabilities, and *p* are predicted sequence choice probabilities by regression.

3.2 Segmentation Data Setting

Exogenous segmentation refers to segmentation using independent variables alone as previously discussed, whereas endogenous segmentation refers to dependent variable based segmentation. The motivations for exogenous segmentation are purely statistical – that is, the *segmentation is based on the type of variation in the X vector and how that variation is associated with the variation in the outcome. For exogenous segmentation, three types of variations are used, such as lighting installation type, interchange and non-interchange section, and one mile section as example cases.* In contrast, *endogenous segmentation is based on the nature of occurrence of the outcome, in the case of this dissertation, for example, frequency of accidents.* In a sense, frequency of accidents can be time-invariant, that is, if accidents are clustering around specific locations on the network, there must be an underlying causal mechanism, and this causal mechanism does not vary substantially over time, unless interventions occur. Investigating the causal mechanism requires approaches beyond pure statistics or associative modeling. Kinematic underpinnings need to be explored, as in heterogeneities that can occur due to environmental interactions and driver to driver interactions. While this aspect of research is beyond the scope of this dissertation, the goal of this dissertation is to set the table for this type of unifying discussion. In order to set the table, a consistent empirical template is required.

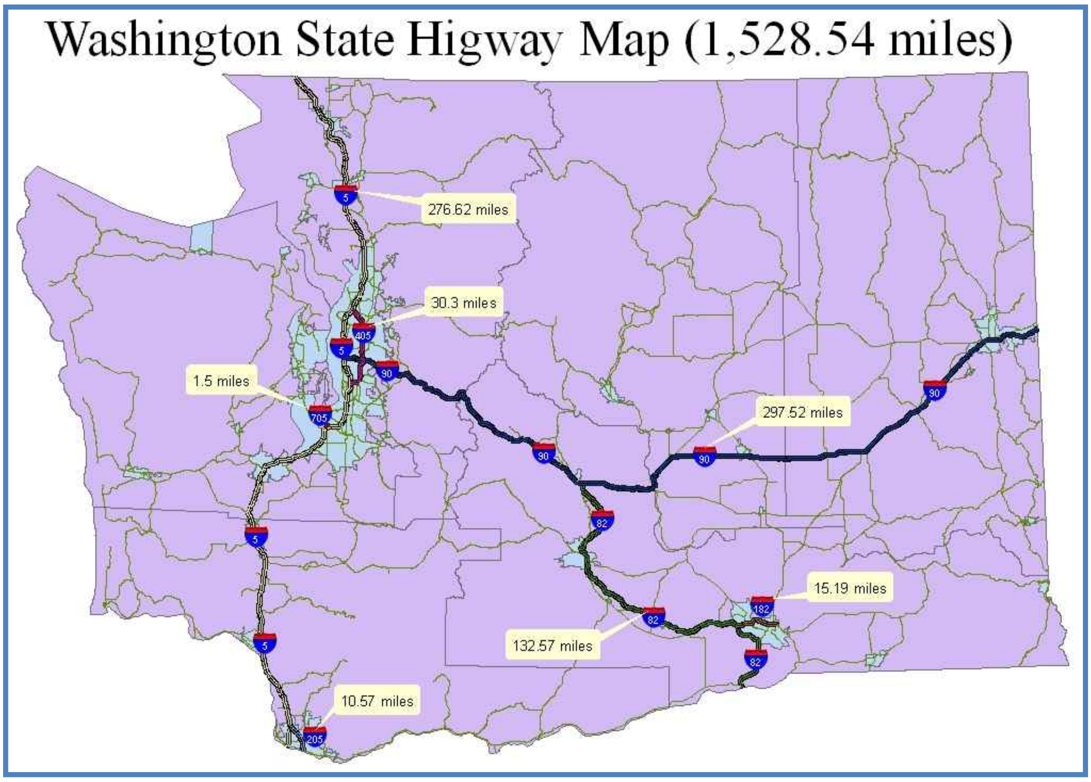


Figure 2: Seven Interstates in Washington State.

A total of 1,528 centerline miles were scanned, by direction, covering interstates in Washington State for input into the segmentation datasets. As seen in Figure 2, Washington State covers 7 interstates including: I-5, I-82, I-90, I-182, I-205, I-405, and I-705. Lighting pole installation, interstate, and overpass information were collected by scanning the WSDOT interstate driving view images provided by the SRWeb. The lighting data is aggregated by lighting type definition as shown in Table 2.

Table 2: Definitions of Lighting Types for Lighting Segmentation.

	Description of lighting
No lighting	Lighting pole does not exist
Median continuous only	Continuous lighting poles present at median side
Median point only	Point lighting pole presents at median side
Right continuous only	Continuous lighting poles present at shoulder side
Right point only	Point lighting pole present at shoulder side
Both lighting	Lighting poles are installed at both side
Tunnel lighting	Lighting installed in a tunnel

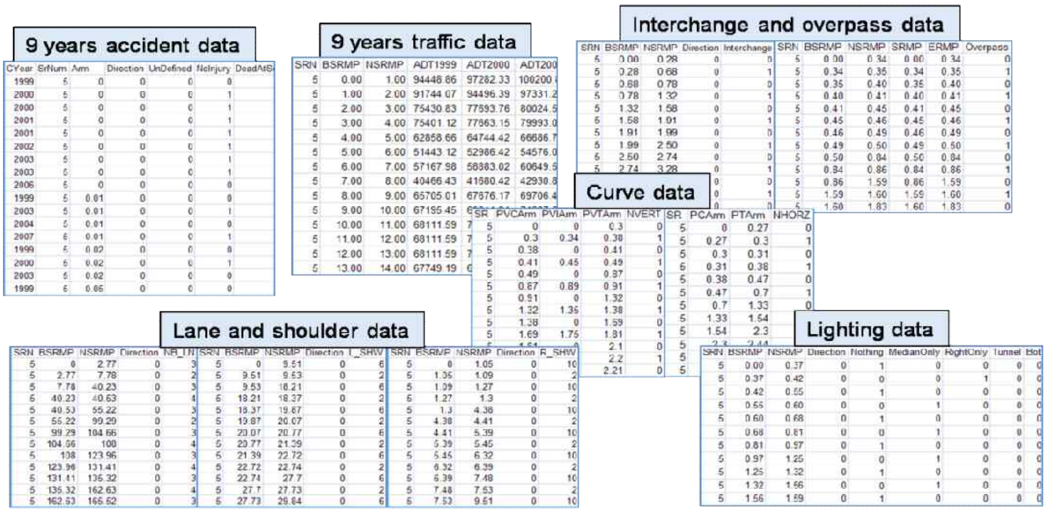
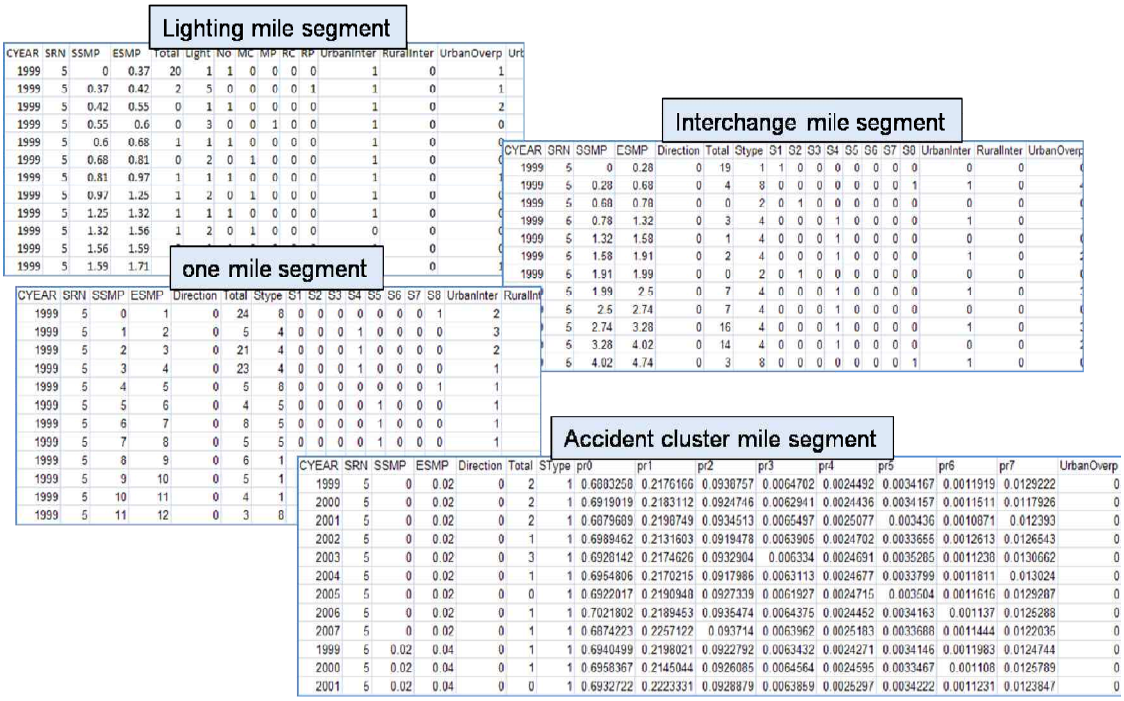

9 years accident data
CYear SrNum Arm Direction UnDefined NoInjury DeadAtS
9 years traffic data
SRN BSRMP NSRMP ADT1999 ADT2000 ADT200
Interchange and overpass data
SRN BSRMP NSRMP Direction Interchange
SRN BSRMP NSRMP SRMP ERMP Overpass
Curve data
SR PVCArm PVIArm PVTArm NVERT
SR PCArm PTArm NHORZ
Lane and shoulder data
SRN BSRMP NSRMP Direction NB_LN
SRN BSRMP NSRMP Direction L_SHW
SRN BSRMP NSRMP Direction R_SHW
Lighting data
SRN BSRMP NSRMP Direction Nothing MedianOnly RightOnly Tunnel Bot

Figure 3: Input Transportation Data for Segmentation.

WSDOT’s database was used for collecting information on the number of lanes, shoulder widths, number of horizontal curves, and number of vertical curves. Safety analysis in this dissertation spans accidents over a recent nine year period. The WSDOT Transportation Data Office (TDO) provides accident location and accident type information from 1999 to 2007. Also, annual average daily traffic (AADT) flow for every one mile is provided as traffic information. Sample transportation datasets for safety analysis are shown in Figure 3.

Four data templates are prepared for the segmentation data on the basis of segment length for each segmentation type. Raw data in Figure 3 are aggregated into data templates by counting and weighting methods; the segmented datasets are shown in Figure 4.



Lighting mile segment

CYEAR	SRN	SSMP	ESMP	Total	Light	No	MC	MP	RC	RP	UrbanInter	RuralInter	UrbanOverp	Urb
1999	5	0	0.37	20	1	1	0	0	0	0	1	0	1	
1999	5	0.37	0.42	2	5	0	0	0	0	1	1	0	1	
1999	5	0.42	0.55	0	1	1	0	0	0	0	1	0	2	
1999	5	0.55	0.6	0	3	0	0	1	0	0	1	0	0	
1999	5	0.6	0.68	1	1	1	0	0	0	0	1	0	0	
1999	5	0.68	0.81	0	2	0	1	0	0	0	1	0	0	
1999	5	0.81	0.97	1	1	1	0	0	0	0	1	0	1	
1999	5	0.97	1.25	1	2	0	1	0	0	0	1	0	0	
1999	5	1.25	1.32	1	1	1	0	0	0	0	1	0	0	
1999	5	1.32	1.56	1	2	0	1	0	0	0	0	0	0	
1999	5	1.56	1.59									0	0	
1999	5	1.59	1.71									0	1	

Interchange mile segment

CYEAR	SRN	SSMP	ESMP	Direction	Total	Stype	S1	S2	S3	S4	S5	S6	S7	S8	UrbanInter	RuralInter	UrbanOverp
1999	5	0	0.28	0	19	1	1	0	0	0	0	0	0	0	0	0	0
1999	5	0.28	0.68	0	4	8	0	0	0	0	0	0	0	1	1	0	4
1999	5	0.68	0.78	0	0	2	0	1	0	0	0	0	0	0	0	0	0
1999	6	0.78	1.32	0	3	4	0	0	0	1	0	0	0	0	1	0	1
1999	5	1.32	1.58	0	1	4	0	0	0	1	0	0	0	0	0	0	0
1999	5	1.58	1.91	0	2	4	0	0	0	1	0	0	0	0	1	0	2
1999	6	1.91	1.99	0	0	2	0	1	0	0	0	0	0	0	0	0	0
	5	1.99	2.5	0	7	4	0	0	0	1	0	0	0	0	1	0	1
	5	2.5	2.74	0	7	4	0	0	0	1	0	0	0	0	0	0	0
	5	2.74	3.28	0	16	4	0	0	0	1	0	0	0	0	1	0	3
	5	3.28	4.02	0	14	4	0	0	0	1	0	0	0	0	0	0	2
	5	4.02	4.74	0	3	8	0	0	0	0	0	0	0	1	1	0	0

one mile segment

CYEAR	SRN	SSMP	ESMP	Direction	Total	Stype	S1	S2	S3	S4	S5	S6	S7	S8	UrbanInter	RuralInt
1999	5	0	1	0	24	8	0	0	0	0	0	0	0	1	2	
1999	5	1	2	0	5	4	0	0	0	1	0	0	0	0	3	
1999	5	2	3	0	21	4	0	0	0	1	0	0	0	0	2	
1999	5	3	4	0	23	4	0	0	0	1	0	0	0	0	1	
1999	5	4	5	0	5	8	0	0	0	0	0	0	0	1	1	
1999	5	5	6	0	4	5	0	0	0	0	1	0	0	0	1	
1999	5	6	7	0	8	5	0	0	0	0	1	0	0	0	1	
1999	5	7	8	0	5	5	0	0	0	0	1	0	0	0	1	
1999	5	8	9	0	6	1										
1999	5	9	10	0	5	1										
1999	5	10	11	0	4	1										
1999	5	11	12	0	3	8										

Accident cluster mile segment

CYEAR	SRN	SSMP	ESMP	Direction	Total	SType	pr0	pr1	pr2	pr3	pr4	pr5	pr6	pr7	UrbanOverp
1999	5	0	0.02	0	2	1	0.6803258	0.2176166	0.0938757	0.0064702	0.0024492	0.0034167	0.0011919	0.0129222	0
2000	5	0	0.02	0	2	1	0.6919019	0.2183112	0.0924746	0.0062941	0.0024436	0.0034157	0.0011511	0.0117926	0
2001	5	0	0.02	0	2	1	0.6879689	0.2198749	0.0934513	0.0065497	0.0025077	0.003436	0.0010871	0.012393	0
2002	5	0	0.02	0	1	1	0.6989462	0.2131603	0.0919478	0.0063905	0.0024702	0.0033655	0.0012613	0.0126543	0
2003	5	0	0.02	0	3	1	0.6928142	0.2174626	0.0932904	0.006334	0.0024691	0.0035285	0.0011238	0.0130662	0
2004	5	0	0.02	0	1	1	0.6954806	0.2170215	0.0917986	0.0063113	0.0024677	0.0033799	0.0011811	0.013024	0
2005	5	0	0.02	0	0	1	0.6922017	0.2190948	0.0927339	0.0061927	0.0024715	0.003504	0.0011616	0.0129287	0
2006	5	0	0.02	0	1	1	0.7021802	0.2189453	0.0935474	0.0064375	0.0024452	0.0034163	0.001137	0.0125288	0
2007	5	0	0.02	0	1	1	0.6874223	0.2257122	0.093714	0.0063962	0.0025183	0.0033688	0.0011444	0.0122035	0
1999	5	0.02	0.04	0	1	1	0.6940499	0.2198021	0.0922792	0.0063432	0.0024271	0.0034146	0.0011983	0.0124744	0
2000	5	0.02	0.04	0	1	1	0.6958367	0.2145044	0.0926085	0.0064564	0.0024595	0.0033467	0.001108	0.0125789	0
2001	5	0.02	0.04	0	0	1	0.6932722	0.2223331	0.0928879	0.0063859	0.0025297	0.0034222	0.0011231	0.0123847	0

Figure 4: Four Types of Segmentation Results for Accident Analysis.

The input for choice variables in the random parameter model or treatment-effect model must be dummy information. In the case of lighting segmentation, lighting information is a logical value for each lighting choice type, but in other segmentation cases, lighting choice values represent the proportion of luminary cover.

Table 3: Definitions of Lighting Sequence Types for Other Segmentation.

	Description of sequence
Sequence 1	No lighting presence
Sequence 2	Median continuous lighting presence
Sequence 3	Right continuous lighting presence
Sequence 4	Median continuous with no lighting presence
Sequence 5	Right continuous with no lighting presence
Sequence 6	Median point with no lighting presence
Sequence 7	Right point with no lighting presence
Sequence 8	other lighting types presence

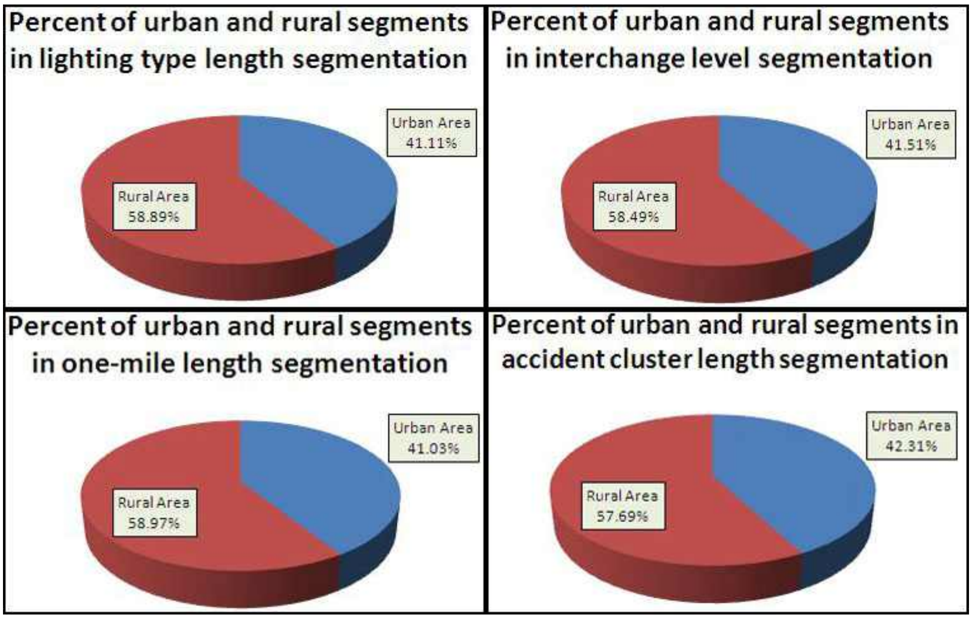


Figure 5: Proportions of Segment Length in Urban and Rural Areas.

Due to the variation in definition, lighting variables are replaced by lighting sequence choice variables in the other segmentation cases. Lighting sequence types for model estimation are shown in Table 3. Here, it is not considered the order of lighting type for lighting sequence. For instance, median continuous lighting presence is followed by no lighting presence as well as no lighting presence is followed by median continuous lighting presence in case of median continuous with no lighting presence.

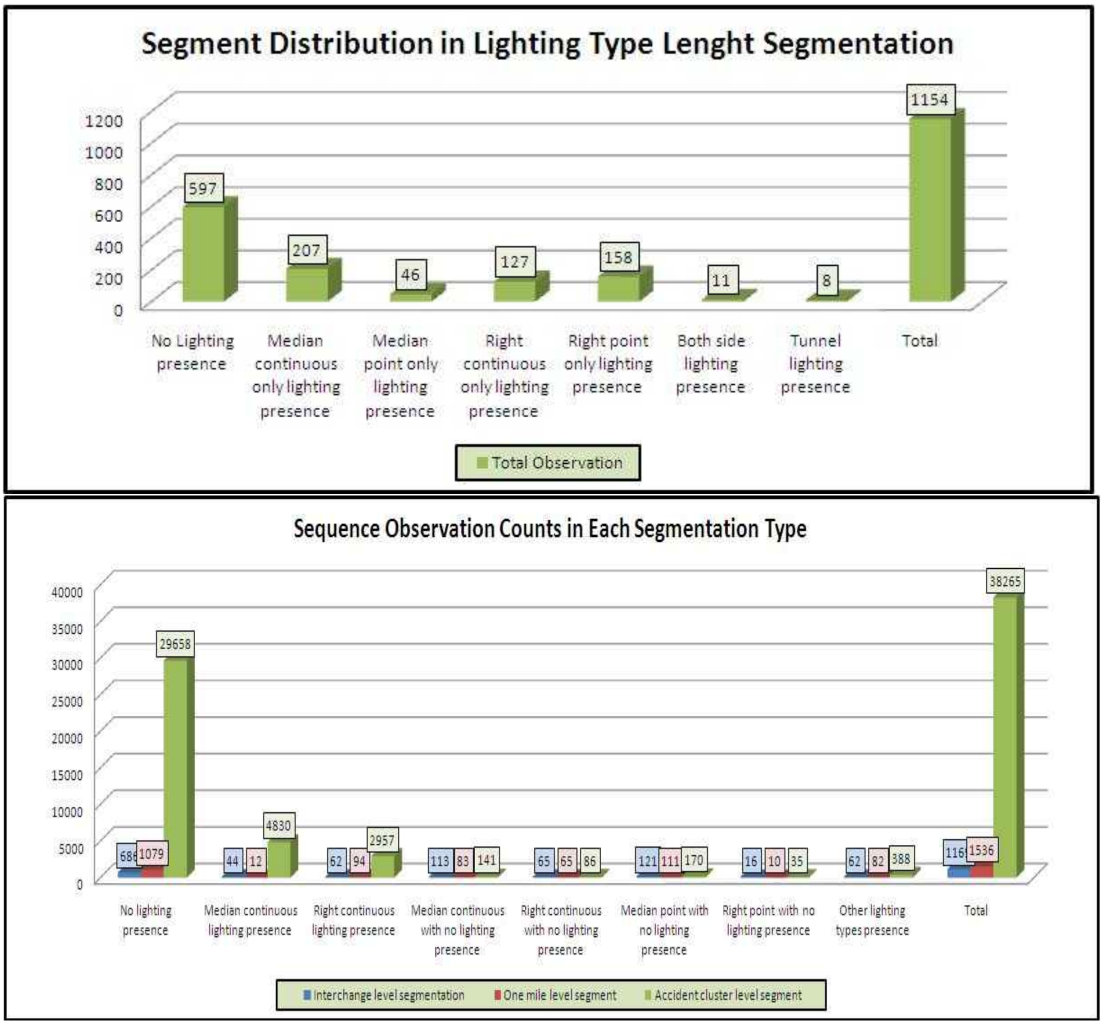
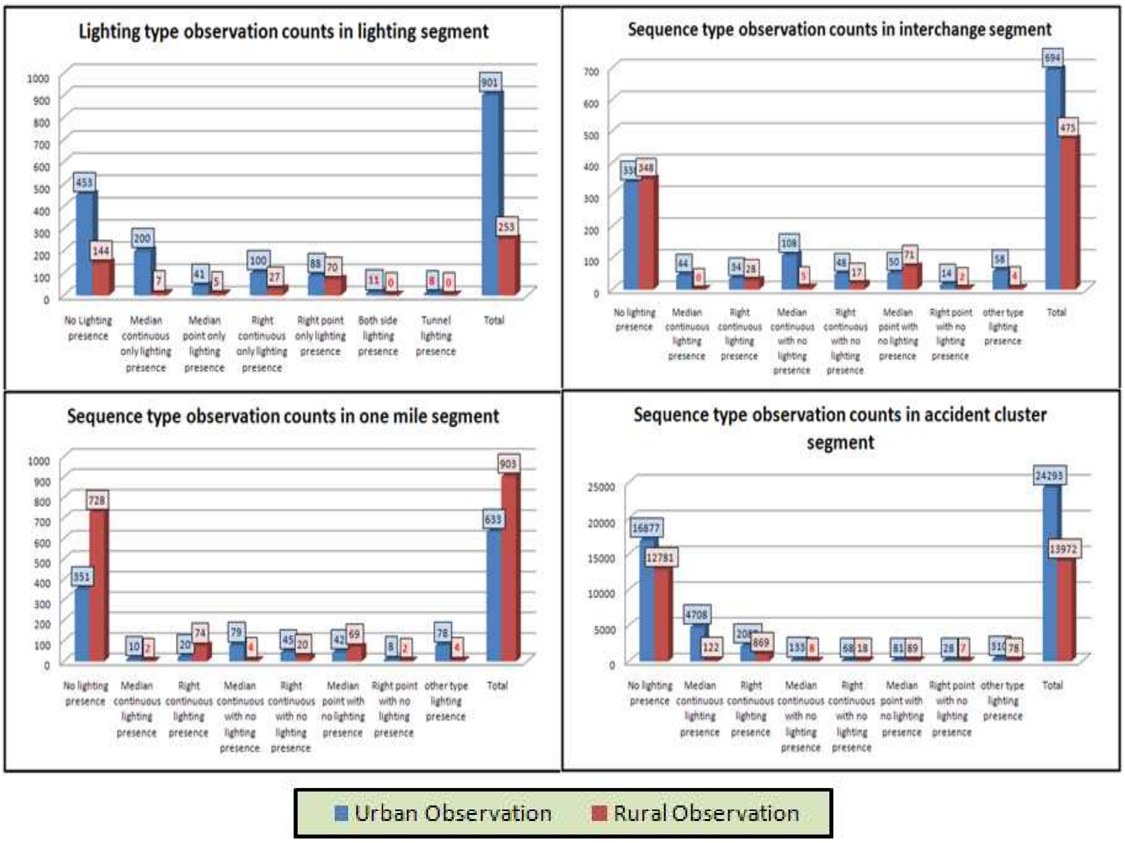


Figure 6: Lighting Observation Counts in Lighting Segmentation and Sequence Observation Counts for Each Segmentation Type.

As shown in Figure 2, interstates 182, 205, 405, and 705 exist in urban areas, while interstates 5, 82, and 90 cover both urban and rural areas. The total percentage of segment length in the rural area is greater than the percentage of urban interstate length. Proportions of segment length in the two different areas are shown in Figure 5. The rural area comprises around 58 percent of all segmentation.


Lighting type observation counts in lighting segment

Sequence type observation counts in interchange segment

Sequence type observation counts in one mile segment

Sequence type observation counts in accident cluster segment

Figure 7: Segment Observation Counts in Urban and Rural Area.

Figure 6 presents lighting and sequence observation counts for each segmentation type. The proportion of no lighting presence is more than 50 percent for all segmentation cases. In the case of lighting segmentation, no lighting presence covers about 52 percent of all observations. No lighting presence occupies roughly 59 percent in interchange segmentation, 70 percent in one mile segmentation, and 78 percent in accident-cluster segmentation respectively. When the segment length size decreases, the percentage of no lighting presence increases. This can affect choice model estimation negatively because of the lack of observations in the other lighting types or sequences types.

Although the rural proportion of the interstate system is greater than its urban counterpart, the lighting observations and sequence observations can appear in different ways. One can expect that more lighting poles are installed in the urban area because of safety impact factors such as traffic flow; this is statistically shown in Figure 7. In most cases, the lighting installed segment count in the urban area is greater than those in the rural area. Only the median lighting with no lighting sequence type has more observations in the rural area than in the urban area. This may be because no lighting observations affect sequence more than median lighting installation. As seen in Figure 7, too few observations exist for both-side lighting and tunnel lighting in the lighting segmentation for both the urban and the rural area. Due to negative effects of lack of observations in the model estimation process, two lighting cases are excluded. In the sequence segmentation cases, too many zeros also exist in the rural area, which dramatically inflates the coefficients for the estimation of the choice model; therefore, all rural independent variables will be excluded from model estimation.

3.3 Descriptive Statistics Results

Figures 8 through 11, and Tables 4 through 7 show the descriptive statistics for key variables in the four types of segment datasets for total accidents. Lighting segments are identified based on the segmentation method in the previous section. Descriptive statistics for lighting segmentation are presented in Figure 8 and Table 4. The mean length of the lighting segments is approximately 1.42 miles. Minimum segment length is 0.01 miles and maximum segment length is 62.27 miles. The percent of no lighting segments is 85.4. Point lighting segments constitute 5.57 percent of the network, continuous lighting segments accounts for 8.03 percent, and other type segments comprise less than 1 percent.

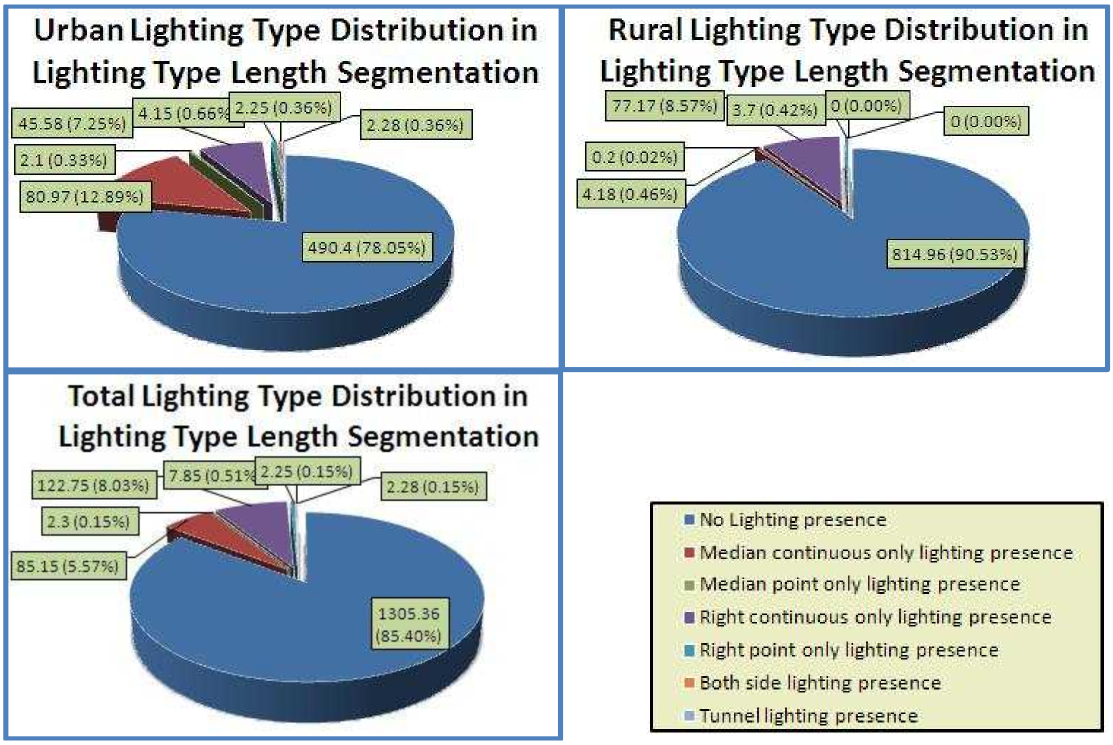


Figure 8: Lighting Type Segment Observation in Lighting Segmentation.

Table 4: Descriptive Statistics for Key Variables in the Interstate Lighting Segment Dataset for Washington State.

Variable	Mean	Minimum	Maximum
Segment length in miles	1.42	0.01	62.27
Percent of segments increasing direction of travel	50.00%	0	1
Number of interchanges in urban segment	1.12	0	10
Number of interchanges in rural segment	1.78	0	13
Number of overpasses in urban segment	0.72	0	12
Number of vertical curves in urban segment	2.04	0	55
Number of vertical curves in rural segment	8.48	0	81
Number of horizontal curves in urban segment	1.32	0	46
Left shoulder width in feet in urban segment	6.83	2	14
Left shoulder width in feet in rural segment	7.18	2	10.21
Right shoulder width in feet in urban segment	6.61	2	24
Right shoulder width in feet in rural segment	7.30	2	16
Log average daily traffic per lane in urban	10.53	8.61	11.40
Log average daily traffic per lane in rural	9.42	7.51	10.61
Number of total accidents in segment	11.42	0	247

On the average, 1.12 interchanges exist in each urban segment with a maximum of 10, while the mean number of interchanges in rural segments is 1.78 with maximum of 13. The average number of overpasses in each segment is 0.72 with a maximum of 12. The average number of vertical curves per urban segment is 2.04 with a maximum of 55, while the mean number of vertical curves per rural segment is 8.48 with a maximum of 81. Each segment has 1.32 horizontal curves, on the average, with a maximum of 46.

The mean left shoulder width is 6.83 feet per urban segment with a minimum of 2 feet and a maximum of 14 feet. The average left shoulder width is 7.18 feet per rural segment with a minimum of 2 feet and a maximum of 10.21 feet. The average right shoulder width per urban segment is 6.13 feet with a minimum of 2 feet and a maximum of 24 feet, while the mean right shoulder width per rural segment 7.30 feet with a minimum of 2 feet and a maximum of 16 feet.

Mean log AADT per lane in each urban segment is 10.23 with a minimum of 8.61 vehicles per day and a maximum of 11.40 vehicles per day, while average log AADT per lane in each rural segment is 9.42 with a minimum of 7.54 vehicles per day and a maximum of 10.61 vehicles per day. The mean number of total accidents is 11.42 per segment with a maximum of 247.

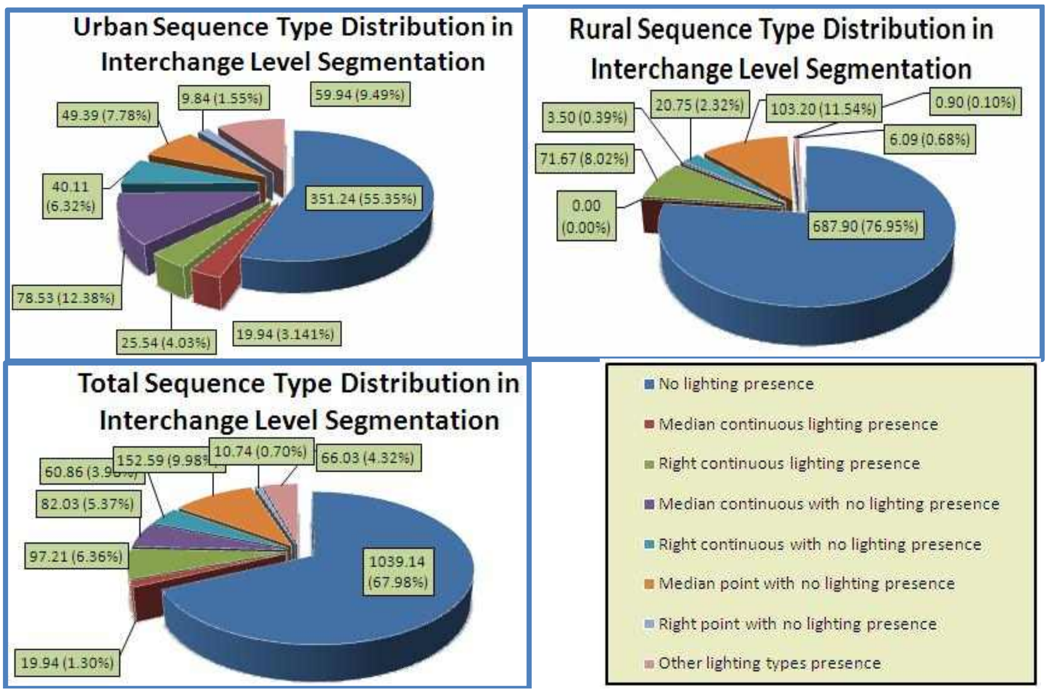


Figure 9: Sequence Type Segment Observation in Interchange Segmentation.

Figure 9 and Table 5 show descriptive statistics for the interchange segmentation dataset. The mean length of the lighting segments is 1.31 miles, while the minimum segment length is 0.01 miles and the maximum segment length is 20.38 miles. The percent of no lighting segments is 67.98, while the segments with median point with no lighting encompasses 9.98 percent of the total segments. Only right continuous lighting segments constitute 6.36 percent of the network, median continuous with no lighting segments maintains 5.37 percent of the network, and other sequence segments make up less than 5 percent.

Table 5: Descriptive Statistics for Key Variables in Interchange Segment Dataset for Washington State.

Variable	Mean	Minimum	Maximum
Segment length in miles	1.31	0.01	20.38
Percent of segments increasing direction of travel	50.00%	0	1
Number of overpasses in urban segment	0.88	0	16
Number of vertical curves in urban segment	2.45	0	30
Number of horizontal curves in urban segment	1.56	0	37
Left shoulder widths in feet in urban segment	7.02	2	18
Right shoulder widths in feet in urban segment	6.75	2	18
Log average daily traffic per lanes in urban	10.31	8.16	11.39
Number of total accidents in segment	10.69	0	388

The average number of overpasses in each urban segment is 0.88 with a maximum of 16. On the average, 2.45 vertical curves exist in each urban segment with a maximum of 30, while the mean number of horizontal curves in each urban segment is 1.56 with a maximum of 37. Each segment has a left shoulder width of 7.02 feet, on the average, with a minimum of 2 feet and a maximum of 18 feet, while the average right shoulder width per urban segment 6.75 feet with a minimum of 2 feet and a maximum of 18 feet. Mean log AADT per lane in each urban segment is 10.31 with a minimum of 8.16 vehicles per day and a maximum of 11.39 vehicles per day, while the mean number of total accidents is 10.69 per segment with a maximum of 388.

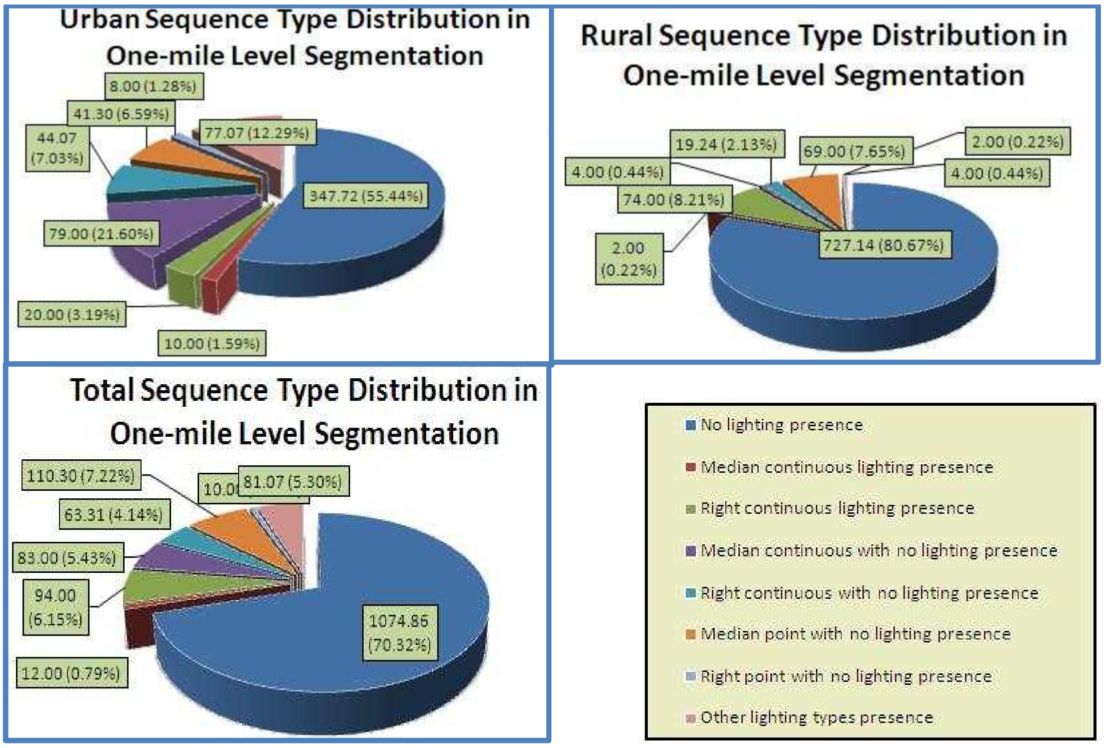


Figure 10: Sequence Type Segment Observation in One Mile Segmentation.

Figure 10 and Table 6 show the descriptive statistics for the one mile segmentation. The mean length of the lighting segments is 1.00 mile with a minimum segment length of 0.01 miles and a maximum segment length of 1.00 mile. The percent of no lighting segments is 70.32, while the percent of segments with median point with no lighting is 7.72. Only right continuous lighting segments constitute 6.15 percent of the network, median continuous with no lighting segments accounts for 5.43 percent of the network, other type lighting segments comprises 5.30 percent, and other sequence segments make up less than 5 percent of the network.

On the average, 0.97 overpasses exist for each urban segment with a maximum of 7. The average number of vertical curves for urban segments is 2.68 with a maximum of 9, while the mean number of horizontal curves for urban segments is 1.66 with a maximum of 5. Each segment maintains a 7.16 feet left shoulder width, on the average, with a minimum of 2 feet and a maximum of 14 feet, while the average right shoulder width per urban segment is 6.89 feet with a minimum of 2 feet and a maximum of 15.08 feet. Mean log AADT per lane in each urban segment is 10.14 with a minimum of 8.16 vehicles per day and a maximum of 11.40 vehicles per day. The mean number of total accidents is 8.14 per segment with a maximum of 172.

Table 6: Descriptive Statistics for Key Variables in One Mile Segment Dataset for Washington State.

Variable	Mean	Minimum	Maximum
Segment length in miles	1.00	0.01	1.00
Percent of segments increasing direction of travel	50.00%	0	1
Number of overpasses in urban segment	0.97	0	7
Number of vertical curves in urban segment	2.68	0	9
Number of horizontal curves in urban segment	1.66	0	5
Left shoulder widths in feet in urban segment	7.16	2	14
Right shoulder widths in feet in urban segment	6.89	2	15.08
Log average daily traffic per lane in urban	10.14	8.16	11.40
Number of total accidents in segment	8.14	0	172

Descriptive statistics for accident-cluster segmentation is shown in Figure 11 and Table 7. The mean length of the lighting segments is 0.04 miles approximately with a minimum segment length of 0.00 miles and maximum segment length of 2.18 miles. The percent of no lighting segments is 83.70, while the percent of segments with only right continuous is 7.92. Only median continuous lighting segments constitute 5.31 percent of the network, median continuous with no lighting segments is 1.38 percent of the network, and other sequence segments make up less than 1 percent of the network.

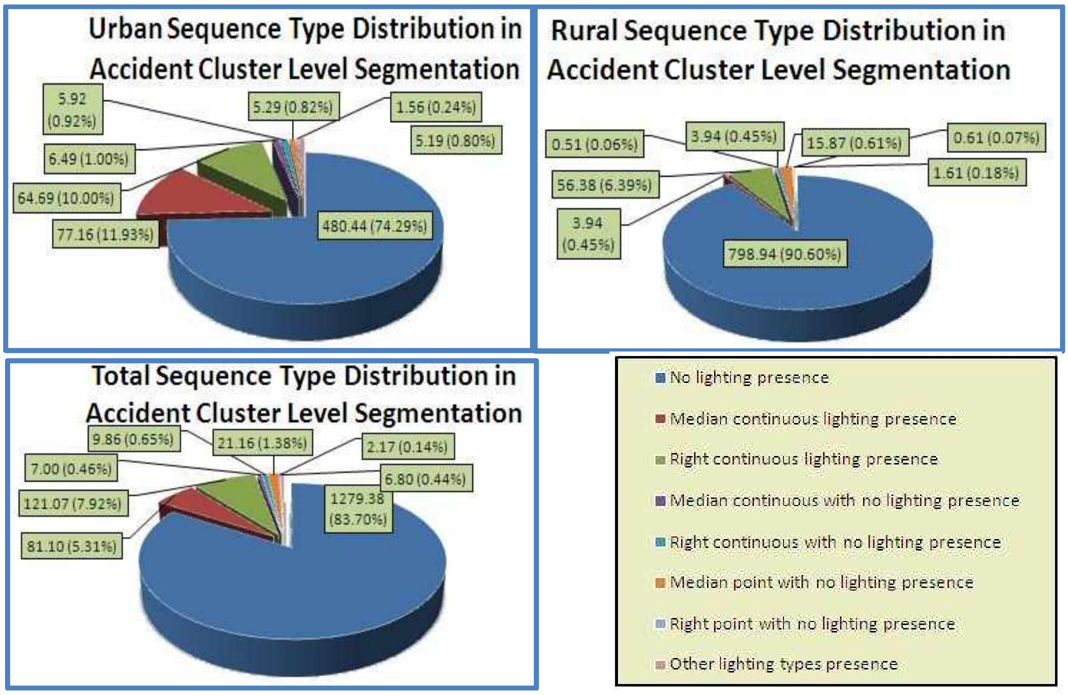


Figure 11: Sequence Type Segment Observation in Accident-Cluster Segmentation.

The average number of overpasses in each urban segment is 0.03 with a maximum of 3. On the average, 0.38 vertical curves exist in urban segments with a maximum of 5, while the mean number of horizontal curves in urban segments is 0.33 with a maximum of 4. Each segment has a 6.98 left shoulder width, on the average, with a minimum of 2 feet and a maximum of 26 feet, while the average right shoulder width per urban segment is 6.83 feet with a minimum of 2 feet and a maximum of 24 feet. Mean log AADT per lane in each urban segment is 10.46 with a minimum of 8.16 vehicles per day and a maximum of 11.63 vehicles per day. The mean number of total accidents is 0.42 per segment with a maximum 26.

Table 7: Descriptive Statistics for Key Variables in Accident-Cluster Segment Dataset for Washington State.

Variable	Mean	Minimum	Maximum
Segment length in miles	0.04	0.00	2.18
Percent of segments increasing direction of travel	50.00%	0	1
Number of overpasses in urban segment	0.03	0	3
Number of vertical curves in urban segment	0.38	0	5
Number of horizontal curves in urban segment	0.33	0	4
Left shoulder widths in feet in urban segment	6.98	2	26
Right shoulder widths in feet in urban segment	6.84	2	24
Log average daily traffic per lane in urban	10.46	8.16	11.63
Number of total accidents in segment	0.42	0	26

Chapter 4

STATISTICAL MODELING RESULTS FOR INTERSTATE LIGHTING SEGMENTATION

I will now discuss the results from the modeling trials relating lighting choice to accident count outcomes. Prior to describing the model results and the implications for the modeling alternatives, I presented the descriptive statistics of the dataset I used from Washington State in the previous chapter. This refined dataset is based on the raw data that was obtained from the Washington State Department of Transportation.

4.1 Negative Binomial Model Results

I will begin with the baseline model results based on lighting type presence in the lighting segmentation dataset. Lighting type is distinguished as only median continuous, only right continuous, only median point, only right point, or none. I do not use both lighting, tunnel lighting data, number of rural overpasses, and number of horizontal curves in the rural area because the segment lengths are very small, and hence, can contribute to convergence problems. Table 8 shows the negative binomial model of accident frequencies with lighting choice variables on all Washington interstates. The median-side continuous lighting variable appears to have counter-productive effects in the model. Most geometric infrastructures, such as shoulders, median barriers, and guardrails, are installed to improve safety on interstates. As a part of roadway infrastructure, lighting poles are installed to decrease accident frequencies to improve driver vision at night or in adverse weather conditions. This is the productive or positive effects of lighting installations on interstate safety. As seen in the negative binomial model of accident frequencies in the case of lighting segmentation, the increase in the median continuous lighting installation appears to increase accident frequencies, and thus, produces counter-productive safety effects on interstates.

Table 8: Negative Binomial Model of Accident Frequencies with Lighting Choice Variables in case of Lighting Segmentation.

	Coefficient	Std. Err.	t-statistics	P Value	95% Conf. Interval	
Constant	-10.410	0.308	-33.84	0.00	-11.013	-9.807
Only median continuous lighting	0.312	0.033	9.48	0.00	0.247	0.377
Only median point lighting	-1.196	0.070	-17.17	0.00	-1.332	-1.059
Only right continuous lighting	-0.080	0.042	-1.91	0.06	-0.161	0.002
Only right point lighting	-1.729	0.048	-36.16	0.00	-1.823	-1.635
Number of urban interchanges	0.281	0.024	11.89	0.00	0.235	0.328
Number of rural interchanges	-0.052	0.033	-1.57	0.12	-0.118	0.013
Number of urban overpasses	0.207	0.014	15.11	0.00	0.180	0.234
Number of urban vertical curves	0.063	0.008	8.19	0.00	0.048	0.079
Number of rural vertical curves	0.086	0.004	19.69	0.00	0.078	0.095
Number of urban horizontal curves	0.131	0.012	10.59	0.00	0.107	0.156
Urban left shoulder widths	-0.019	0.004	-4.65	0.00	-0.027	-0.011
Rural left shoulder widths	0.121	0.014	8.52	0.00	0.093	0.149
Urban right shoulder widths	-0.015	0.004	-3.95	0.00	-0.023	-0.008
Rural right shoulder widths	0.131	0.014	9.27	0.00	0.103	0.158
Urban log AADT per number of lanes	1.131	0.028	39.85	0.00	1.076	1.187
Rural log AADT per number of lanes	1.012	0.034	29.93	0.00	0.945	1.078
Overdispersion	1.157	0.021	56.32	0.00	1.118	1.198
log likelihood at constant						-30982.55
log likelihood at convergence						-27427.18

If a coefficient of a choice variable has a positive sign, it contributes to increasing the value of a dependent variable compared to a baseline variable. However, if the sign of the coefficient is negative, it contributes more to decreasing the value in relation to the baseline variable. The coefficient of median continuous lighting variable has a positive sign, while the other lighting variables have negative signs. Therefore, with traffic and geometry controls, more accident occurrences are expected with median continuous lighting presence than without lighting, but lighting presence contributes more to reduced accident frequencies than no lighting presence in the other three lighting cases.

Baseline model results with lighting sequence variables in interchange and one mile segmentation are shown in Table 9 and Table 10. Here, all variables are limited in the urban area because the lack of lighting presence in the rural area contributes to convergence problems.

Table 9: Negative Binomial Model of Urban Accident Frequencies with Lighting Sequence Choice Variables in case of Interchange Segmentation.

	Coefficient	Std. Err.	t-statistics	P Value	95% Conf. Interval	
Constant	1.581	0.017	90.49	0.00	1.547	1.615
Only median continuous lighting	0.418	0.057	7.32	0.00	0.306	0.530
Only right continuous lighting	-0.330	0.049	-6.75	0.00	-0.426	-0.234
Median continuous with no lighting	0.495	0.039	12.87	0.00	0.420	0.571
Right continuous with no lighting	0.131	0.047	2.80	0.01	0.039	0.223
Median point with no lighting	0.085	0.035	2.43	0.02	0.017	0.154
Right point with no lighting	0.395	0.088	4.47	0.00	0.222	0.568
Other sequences	0.751	0.048	15.61	0.00	0.657	0.845
Number of overpasses	0.087	0.011	7.86	0.00	0.065	0.108
Number of vertical curves	0.023	0.008	2.75	0.01	0.007	0.040
Number of horizontal curves	0.183	0.012	14.73	0.00	0.159	0.208
Left shoulder widths	-0.051	0.004	-11.87	0.00	-0.060	-0.043
Right shoulder widths	-0.030	0.004	-7.04	0.00	-0.038	-0.021
Log AADT per number of lanes	0.085	0.006	15.26	0.00	0.074	0.096
Overdispersion	0.992	0.015	64.582	0.000	0.962	1.022
log likelihood at constant						-35535.45
log likelihood at convergence						-33582.91

Table 10: Negative Binomial Model of Urban Accident Frequencies with Lighting Sequence Choice Variables in case of One Mile Segmentation.

	Coefficient	Std. Err.	t-statistics	P Value	95% Conf. Interval	
Constant	0.926	0.013	72.66	.00	0.901	0.951
Only median continuous lighting	1.049	0.091	11.49	0.00	0.870	1.228
Only right continuous lighting	-0.457	0.041	-11.03	0.00	-0.53	-0.376
Median continuous with no lighting	0.919	0.039	23.60	0.00	0.842	0.995
Right continuous with no lighting	0.489	0.042	11.55	0.00	0.406	0.572
Median point with no lighting	0.459	0.033	13.90	0.00	0.394	0.524
Right point with no lighting	0.784	0.099	7.90	0.00	0.589	0.978
Other sequences	0.948	0.040	23.46	0.00	0.869	1.027
Number of overpasses	0.098	0.011	8.67	0.00	0.076	0.121
Number of vertical curves	-0.026	0.009	-2.78	0.01	-0.044	-0.008
Number of horizontal curves	0.136	0.013	10.73	0.00	0.111	0.160
Left shoulder widths	-0.058	0.004	-13.04	0.00	-0.066	-0.049
Right shoulder widths	-0.033	0.004	-7.72	0.00	-0.042	-0.025
Log AADT per number of lanes	0.165	0.007	24.13	0.00	0.151	0.178
Overdispersion	0.267	0.051	5.19	0.00	0.183	0.389
log likelihood at constant						-42346.821
log likelihood at convergence						-37447.936

The only right-side continuous lighting presence variables for the two models have negative coefficient signs; they do not have counter-intuitive effects in both models. Here, the term counter-intuitive effect is used in similar manner as the term counter-productive. Most people intuitively expect infrastructures are installed to improve safety on the roadway, but if accident frequencies are increased in the segment where the infrastructure is installed compared to locations where it is not installed, the infrastructure does not contribute to improving safety; this is the opposite concept of the lighting infrastructure to conventional intuition. The only right-side continuous lighting presence variables for the two models seems to contribute to decreasing the accident frequencies, and it is only these lighting variables that do not have counter-intuitive effects on safety.

The median-side continuous lighting variable and any type of lighting with no lighting presence variables have positive coefficient signs, so they have greater effects on accident frequencies than the no lighting presence variable. The lighting presence variables, except the median continuous lighting variable in Table **10**, have negative effects on the increase of accident frequencies, but all lighting sequence variables contribute to increasing accident frequencies. This is because the no lighting portion in the lighting sequence segments appear to contribute more to increasing accident frequencies than the contribution of decreasing accident frequencies for the lighting installed portion.

Table **11**: Negative Binomial Model of Urban Accident Frequencies with Lighting Sequence Choice Variables in case of Accident-Cluster Segmentation.

	Coefficient	Std. Err.	t-statistics	P Value	95% Conf. Interval	
Constant	-9.726	0.114	-85.41	0.00	-9.949	-9.503
Only median continuous lighting	0.267	0.010	25.63	0.00	0.247	0.288
Only right continuous lighting	0.217	0.016	13.69	0.00	0.186	0.248
Median continuous with no lighting	0.530	0.050	10.67	0.00	0.433	0.628
Right continuous with no lighting	0.602	0.068	8.80	0.00	0.468	0.736
Median point with no lighting	0.332	0.072	4.62	0.00	0.191	0.472
Right point with no lighting	0.227	0.126	1.79	0.07	-0.021	0.474
Other sequences	0.361	0.033	10.79	0.00	0.295	0.426
Number of overpasses	0.599	0.021	28.94	0.00	0.558	0.640
Number of vertical curves	0.012	0.009	1.42	0.16	-0.005	0.029
Number of horizontal curves	-0.029	0.009	-3.26	0.00	-0.046	-0.012
Left shoulder widths	-0.018	0.001	-15.00	0.00	-0.020	-0.015
Right shoulder widths	-0.017	0.001	-14.49	0.00	-0.019	-0.015
Log AADT per number of lanes	0.849	0.011	80.75	0.00	0.829	0.870
Overdispersion	1.331	0.015	88.33	0.00	1.302	1.361
log likelihood at constant						-186199.12
log likelihood at convergence						-180676.97

The one mile segmentation model results have larger coefficients than the interchange segmentation model in the sense that lighting sequence has greater effects on accident frequencies when compared to no lighting presence.

Table 11 shows the baseline model results with lighting sequence variables in the accident-cluster segmentation. This model is estimated in the urban area as well. Each lighting sequence presence variable has a positive sign and counter-intuitive effect on accident frequencies in the model. Interstates with lighting pole installation have a greater expectation of accident frequencies than areas without lighting presence.

4.2 The Mixed Multinomial-Selection Negative Binomial Count Treatment Effects Model and Negative Binomial Model with Pre-Processing

This subsection presents the results based on a polytomous selection schema for lighting choice or lighting sequence choice incorporated with the negative binomial count outcome. In the cases of lighting, interchange, and one mile segmentation, the effects of geometry and traffic flow on the choice of lighting installation are estimated by the mixed multinomial logit model. The multinomial logit model is used for estimating the effects of traffic and geometry on lighting sequence choice in the case of accident-cluster segmentation. The baseline choice is no lighting presence in all choice models. The four types of lighting presence referenced in a previous chapter are used as alternative choices for the mixed multinomial logit model in the lighting segmentation case. Alternative choices of sequence for the logit models are the seven types of lighting sequence in the other three segmentation cases.

Table 12: Mixed Multinomial Logit Model of the Lighting Type Installation Probability in Lighting Segmentation.

	Only median continuous lighting			Only median point lighting		
	Coefficient	Std. Err.	t-statistics	Coefficient	Std. Err.	t-statistics
Constant	-7.752	0.893	.68	-2.021	1.430	-1.41
Number of urban interchanges	-0.070	0.066	-1.05	-0.607	0.140	-4.32
Numb r of rural interchanges	0.220	0.330	0.67	-0.772	0.433	-1.78
Number of urban overpasses	0.157	0.037	4.30	-0.373	0.134	-2.77
Number of urban vertical curves	-0.139	0.023	-5.92	-1.227	0.111	-11.08
Number of rural vertical curves	-0.781	0.146	-5.34	-1.943	0.326	-5.97
Number of urban horizontal curves	-0.021	0.037	-0.57	-0.465	0.104	-4.46
Urban left shoulder widths	-0.007	0.010	-0.69	0.058	0.018	3.22
Rural left shoulder widths	-0.074	0.063	-1.16	0.116	0.101	1.14
Urban right shoulder widths	0.022	0.010	2.32	0.039	0.016	2.43
Rural right shoulder widths	-0.139	0.063	-2.20	-0.138	0.092	-1.50
Urban log AADT per number of lanes	0.654	0.082	7.97	0.045	0.132	0.34
Rural log AADT per number of lanes	0.810	0.120	6.76	0.180	0.181	0.99
	Only right continuous lighting			Only right point lighting		
Constant	3.257	0.805	4.05	3.366	0.857	3.93
Number of urban interchanges	-0.749	0.079	-9.42	-0.014	0.115	-0.12
Number of rural interchanges	-1.467	0.150	-9.80	-0.876	0.179	-4.90
Number of urban overpasses	0.242	0.042	5.73	-1.075	0.132	-8.11
Number of urban vertical curves	-0.021	0.027	-0.79	-0.696	0.074	-9.38
Number of rural vertical curves	0.094	0.016	5.97	-1.648	0.104	-15.90
Number of urban horizontal curves	-0.019	0.041	-0.48	-1.253	0.093	-13.50
Urban left shoulder widths	-0.063	0.012	-5.13	0.085	0.014	5.93
Rural left shoulder widths	0.223	0.050	4.43	0.199	0.037	5.32
Urban right shoulder widths	-0.074	0.012	-6.03	0.026	0.013	2.02
Rural right shoulder widths	0.027	0.045	0.61	0.275	0.039	7.09
Urban log AADT per number of lanes	-0.335	0.075	-4.47	-0.430	0.081	-5.33
Rural log AADT per number of lanes	-0.593	0.103	-5.78	-0.368	0.095	-3.87
log likelihood at constant						-10668.76
log likelihood at convergence						-10583.66

Table 13: Mixed Multinomial Logit Model of the Lighting Sequence Probability in Interstate Segmentation.

	Only median continuous			Only right continuous		
	Coefficient	Std. Err.	t-statistics	Coefficient	Std. Err.	t-statistics
Constant	-20.992	1.58	-13.28	-2.887	0.068	-42.18
Number of urban overpasses	-0.170	0.086	-1.97	0.305	0.073	4.17
Number of urban ver ical curves	-0.350	0.048	-7.28	-0.094	0.041	-2.30
Number of urban horizontal curves	0.118	0.064	1.83	-0.014	0.060	-0.24
Urban left shoulder widths	-0.088	0.021	-4.27	0.099	0.026	3.84
Urban right shoulder widths	-0.201	0.020	-9.91	-0.185	0.023	-7.94
Urban log AADT per number of lanes	2.023	0.145	13.91	0.047	0.030	1.55
	Median continuous with no lighting			Right continuous with no lighting		
Constant	-5.299	0.194	-27.30	-3.485	0.086	-40.36
Number of urban overpasses	0.637	0.042	15.11	0.647	0.051	12.64
Number of urban vertical curves	-0.217	0.028	-7.76	-0.130	0.037	-3.53
Number of urban horizontal curves	0.101	0.034	2.96	-0.030	0.055	-0.55
Urban left shoulder widths	-0.104	0.014	-7.28	-0.200	0.019	-10.61
Urban right shoulder widths	-0.090	0.014	-6.35	-0.182	0.019	-9.67
Urban log AADT per number of lanes	0.488	0.024	20.04	0.340	0.023	14.80
	Median point with no lighting			Right point with no lighting		
Constant	-5.844	0.258	-22.64	-1.922	0.049	-39.29
Number of urban overpasses	0.151	0.115	1.31	0.407	0.059	6.92
Number of urban vertical curves	-0.338	0.073	-4.64	-0.052	0.033	-1.58
Number of urban horizontal curves	0.088	0.101	0.88	-0.009	0.040	-0.23
Urban left shoulder widths	-0.113	0.030	-3.75	0.047	0.019	2.54
Urban right shoulder widths	0.100	0.034	2.92	0.006	0.018	0.34
Urban log AADT per number of lanes	0.257	0.046	5.59	-0.091	0.023	-3.91
	Other sequences					
Constant	-5.366	0.202	-26.53			
Number of urban overpasses	0.783	0.045	17.59			
Number of urban vertical curves	-0.068	0.031	-2.18			
Number of urban horizontal curves	0.075	0.029	2.56			
Urban left shoulder widths	-0.061	0.018	-3.43			
Urban right shoulder widths	-0.105	0.017	-6.01			
Urban log AADT per number of lanes	0.359	0.028	13.02			
log likelihood at constant						-13361.55
log likelihood at convergence						-13218.09

Table 14: Mixed Multinomial Logit Model of the Lighting Sequence Probability in One Mile Segmentation.

	Only median continuous			Only right continuous		
	Coefficient	Std. Err.	t-statistics	Coefficient	Std. Err.	t-statistics
Constant	-6.528	0.251	-25.96	-2.671	0.043	-61.91
Number of urban overpasses	0.850	0.100	8.51	0.368	0.110	3.33
Number of urban vertical curves	0.124	0.085	1.46	0.349	0.060	5.84
Number of urban horizontal curves	-0.433	0.129	-3.37	0.160	0.095	1.68
Urban left shoulder widths	-0.425	0.040	-10.70	1.077	0.082	13.08
Urban right shoulder widths	-0.614	0.046	-13.44	-0.375	0.044	-8.51
Urban log AADT per number of lanes	0.843	0.053	16.06	-1.045	0.090	-11.67
	Median continuous with no lighting			Right continuous with no lighting		
Constant	-6.279	0.216	-29.09	-4.092	0.079	-51.98
Number of urban overpasses	1.050	0.048	21.66	1.072	0.055	19.47
Number of urban vertical curves	-0.025	0.035	-0.71	-0.003	0.043	-0.07
Number of urban horizontal curves	0.357	0.049	7.32	-0.058	0.063	-0.91
Urban left shoulder widths	-0.133	0.019	-7.16	-0.299	0.023	-13.26
Urban right shoulder widths	-0.176	0.018	-9.66	-0.318	0.023	-14.07
Urban log AADT per number of lanes	0.509	0.033	15.33	0.490	0.032	15.25
	Median point with no lighting			Right point with no lighting		
Constant	-6.572	0.255	-25.74	-2.766	0.046	-60.67
Number of urban overpasses	0.717	0.114	6.31	0.567	0.064	8.83
Number of urban vertical curves	-0.123	0.091	-1.35	-0.068	0.043	-1.59
Number of urban horizontal curves	0.531	0.110	4.82	0.121	0.059	2.06
Urban left shoulder widths	-0.046	0.044	-1.05	-0.080	0.023	-3.51
Urban right shoulder widths	-0.036	0.044	-0.83	-0.152	0.022	-6.91
Urban log AADT per number of lanes	0.179	0.070	2.56	0.150	0.033	4.59
	Other sequences					
Constant	-6.194	0.208	-29.74			
Number of urban overpasses	1.134	0.049	23.08			
Number of urban vertical curves	0.021	0.037	0.58			
Number of urban horizontal curves	0.626	0.050	12.42			
Urban left shoulder widths	-0.179	0.019	-9.49			
Urban right shoulder widths	-0.216	0.019	-11.51			
Urban log AADT per number of lanes	0.476	0.033	14.28			
log likelihood at constant						-128789.87
log likelihood at convergence						-12721.13

Table 15: Mixed Multinomial Logit Model of the Lighting Sequence Probability in Accident-Cluster Segmentation.

	Only median continuous			Only right continuous		
	Coefficient	Std. Err.	t-statistics	Coefficient	Std. Err.	t-statistics
Constant	-12.328	0.163	-75.50	8.226	0.140	58.70
Number of urban overpasses	0.223	0.033	6.79	0.321	0.040	8.08
Number of urban vertical curves	-0.104	0.012	-8.89	-0.007	0.015	-0.47
Number of urban horizontal curves	0.116	0.012	9.74	0.004	0.017	0.25
Urban left shoulder widths	-0.042	0.002	-26.61	-0.061	0.002	-26.78
Urban right shoulder widths	-0.121	0.002	-76.60	-0.202	0.002	-83.36
Urban log AADT per number of lanes	1.148	0.015	76.17	-0.839	0.013	-64.73
	Median continuous with no lighting			Right continuous with no lighting		
Constant	-9.877	0.717	-13.77	-6.053	0.842	-7.19
Number of urban overpasses	0.547	0.136	4.02	1.387	0.103	13.47
Number of urban vertical curves	-0.153	0.060	-2.57	0.291	0.072	4.05
Number of urban horizontal curves	0.334	0.058	5.74	0.677	0.074	9.19
Urban left shoulder widths	0.010	0.008	1.19	-0.026	0.012	-2.27
Urban right shoulder widths	-0.043	0.008	-5.27	-0.043	0.012	-3.75
Urban log AADT per number of lanes	0.494	0.066	7.44	0.050	0.077	0.64
	Median point with no lighting			Right point with no lighting		
Constant	-4.747	1.181	-4.02	-1.740	0.689	-2.53
Number of urban overpasses	-27.961	459391.80	0.00	1.003	0.122	8.23
Number of urban vertical curves	-0.500	0.138	-3.63	0.203	0.069	2.94
Number of urban horizontal curves	1.069	0.120	8.89	-0.023	0.077	-0.30
Urban left shoulder widths	0.117	0.019	6.15	0.058	0.011	5.07
Urban right shoulder widths	0.047	0.018	2.67	0.013	0.011	1.19
Urban log AADT per number of lanes	-0.309	0.112	-2.75	-0.410	0.065	-6.31
	Other sequences					
Constant	-16.104	0.578	-27.87			
Number of urban overpasses	0.518	0.095	5.48			
Number of urban vertical curves	0.003	0.039	0.09			
Number of urban horizontal curves	-0.557	0.046	-12.16			
Urban left shoulder widths	0.047	0.006	8.43			
Urban right shoulder widths	-0.037	0.005	-7.05			
Urban log AADT per number of lanes	1.153	0.053	21.65			
log likelihood at constant						-198831.60
log likelihood at convergence						-186604.31

These models use traffic and geometry factors such as number of interchanges, number of overpasses, number of horizontal and vertical curves, number of lanes, and shoulder width, as independent variables. For the mixed multinomial logit model, in the case of lighting segmentation, all urban and rural factors are used except number of overpasses and horizontal curves in the rural area. Other logit models use urban geometry without interstate and urban traffic factors as independent variables.

Tables **12** through **14** show the mixed multinomial logit model results of installation probability for each lighting type and lighting sequence type. Table **15** also shows the multinomial logit model results for sequence probabilities. The predicted outcomes of this lighting choice model are included in the negative binomial model as additional independent variables for selectivity bias correction.

The selectivity bias correction model results with predicted outcomes in the case of lighting segmentation are shown in Table **16**. Most independent variables are significant in the selectivity correction model estimations. If the t-statistic value of the variable is greater than or equal to 2, or the p-value is less than or equal to 0.05, the variable is significant at the 95 percent confidence level. If the variable is significant, the coefficient is highly different from zero and it will significantly increase or decrease the dependent variable at the confidence level. Most geometry variables have a positive relationship with the number of accidents, except for urban shoulder widths, in the correction model. The traffic flow variable also has a positive coefficient sign; this trend is applicable to the uncorrected model in Table **8** as well.

Table 16: Negative Binomial Results with Selectivity Bias Correction in Lighting Segmentation.

	Coefficient	Std. Er .	t-statistics	P Value	95% Conf. Interval	
Constant	-10.574	0.329	-32.13	0.00	-11.219	-9.929
Only median continuous lighting	0.324	0.054	6.03	0.00	0.219	0.430
Only median point lighting	-1.352	0.085	-15.82	0.00	-1.519	-1.185
Only right continuous lighting	-0.060	0.060	-1.00	0.32	-0.178	0.058
Only right point lighting	-2.178	0.060	-36.06	0.00	-2.296	-2.059
Number of urban interchanges	0.287	0.026	11.04	0.00	0.236	0.338
Number of rural interchanges	-0.030	0.036	-0.85	0.40	-0.100	0.040
Number of urban overpasses	0.233	0.015	15.32	0.00	0.203	0.262
Number of urban vertical curves	0.042	0.009	4.89	0.00	0.025	0.059
Number of rural vertical curves	0.079	0.005	16.32	0.00	0.069	0.088
Number of urban horizontal curves	0.098	0.014	6.85	0.00	0.070	0.126
Urban left shoulder widths	-0.013	0.004	-3.07	0.00	-0.022	-0.005
Rural left shoulder widths	0.127	0.015	8.27	0.00	0.097	0.157
Urban right shoulder widths	-0.011	0.004	-2.53	0.01	-0.019	-0.002
Rural right shoulder widths	0.138	0.015	9.23	0.00	0.109	0.167
Urban log AADT per number of lanes	1.134	0.030	37.41	0.00	1.075	1.194
Rural log AADT per number of lanes	1.020	0.036	27.95	0.00	0.948	1.091
Heterogeneity of only median continuous	0.014	0.048	0.28	0.78	-0.081	0.108
Heterogeneity of only median point	0.173	0.045	3.84	0.00	0.085	0.261
Heterogeneity of only right continuous	-0.061	0.046	-1.31	0.19	-0.151	0.030
Heterogeneity of only right point	0.591	0.039	15.18	0.00	0.514	0.667
Overdispersion	0.806	0.041	19.60	0.00	0.729	0.890
log likelihood at constant						-38010.84
log likelihood at convergence						-37884.24

The number of interchanges in the rural area has a negative sign in both models, but the p-value is greater than 0.05 and the variable is not significant at the 95 percent confidence level. Because the variable is not significant, there is no significant evidence that suggests the number of interchanges decrease accident frequency on interstates. The right continuous lighting presence variable is significant in the uncorrected model, but it is not significant in the selectivity correction model. Median continuous lighting has a positive effect on accident frequency and it is significant in both models. Median point lighting, right continuous lighting, and right point lighting variables are significant in both estimations and are associated with decreased accident frequencies.

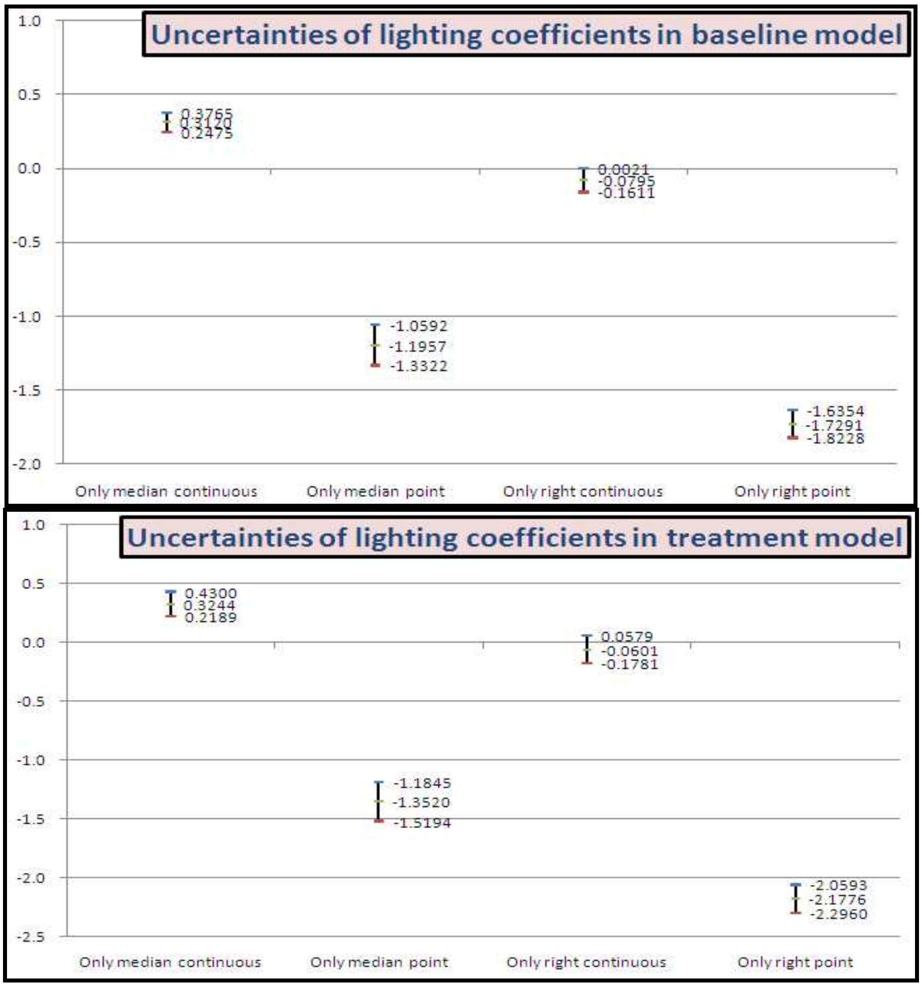


Figure 12: Uncertainty of Lighting Choice Variables in Lighting Segmentation.

The standard errors after selectivity correction are slightly larger than the standard errors of the initial models. This is due to the fact that unobserved heterogeneity is accounted for in their shared lambda variable. The coefficients for the right continuous and right point lighting variables are changed by roughly 25 percent. This implies that the unobserved heterogeneity has a greater influence on model estimation in the selectivity correction model. The coefficient for the number of vertical curves and horizontal curves in urban area are changed by 34 percent and 25 percent, while the change in coefficient for left and right shoulder width in the urban area are changed by 29 percent and 31 percent. The coefficient change for the number of rural interchange variable is 42 percent, and the coefficient change for the median point lighting and urban overpass variables are roughly 10 percent, while the changes in coefficient for other variables is less than 10 percent.

The confidence interval of parameter is the useful method to know the reliability of an estimate because it can measure the uncertainty of estimated parameters. The mean value of the estimated parameter, its standard error, and sample size are associated with the computation of the confidence interval. The 95 percent confidence interval is generally used for the confidence interval. Figure **12** shows the 95 percent confidence intervals for key variables in lighting segmentation. The selectivity correction model has more confidence intervals than the baseline model for all lighting choice variables because of the heterogeneity effect. The confidence interval for the median continuous lighting variable is increased by 64 percent. In the case of the right continuous lighting variable, the confidence interval is increased by 45 percent. For the two types of point lighting variables, the confidence interval was found to increase by about 20 percent. The increases in uncertainty are because the selectivity bias model captures unobserved heterogeneity in parameter estimation.

Table 17: Negative Binomial Results with Selectivity Bias Correction in Interchange Segmentation.

	Coefficient	Std. Err.	t-statistics	P Value	95% Conf. Interval	
Constant	1.286	0.035	36.40	0.00	1.216	1.355
Only median continuous	1.074	0.098	10.99	0.00	0.883	1.266
Only right continuous	-0.431	0.107	-4.04	0.00	-0.640	-0.222
Median continuous with no lighting	0.822	0.119	6.89	0.00	0.588	1.056
Right continuous with no lighting	-0.099	0.130	-0.77	0.44	-0.354	0.155
Median point with no lighting	0.382	0.123	3.10	0.00	0.141	0.624
Right point with no lighting	0.013	0.098	0.13	0.90	-0.180	0.206
Other sequences of lighting	0.937	0.089	10.51	0.00	0.762	1.111
Number of overpasses	0.113	0.012	9.34	0.00	0.090	0.137
Number of vertical curves	0.049	0.009	5.60	0.00	0.032	0.067
Number of horizontal curves	0.089	0.013	6.73	0.00	0.063	0.114
Left shoulder widths	-0.057	0.005	-11.88	0.00	-0.066	-0.047
Right shoulder widths	-0.037	0.005	-7.94	0.00	-0.047	-0.028
Log AADT per number of lanes	0.095	0.006	14.91	0.00	0.083	0.108
Heterogeneity of only median continuous	-0.666	0.093	-7.16	0.00	-0.849	-0.484
Heterogeneity of only right continuous	0.077	0.104	0.74	0.46	-0.126	0.281
Heterogeneity of median continuous with no lighting	-0.244	0.128	-1.92	0.06	-0.494	0.006
Heterogeneity of right continuous with no lighting	0.294	0.140	2.10	0.04	0.020	0.567
Heterogeneity of median point with no lighting	-0.029	0.101	-0.29	0.78	-0.227	0.169
Heterogeneity of right point with no lighting	0.129	0.104	1.25	0.21	-0.074	0.333
Heterogeneity of other sequences	-0.064	0.082	-0.78	0.44	-0.224	0.096
Overdispersion	0.434	0.060	7.27	0.00	0.332	0.569
log likelihood at constant	-46801.00					
log likelihood at convergence	-46629.62					

Table 17 shows the selectivity bias correction model results with predicted outcomes in the case of interchange segmentation. Most independent variables have positive effects on accident occurrences in the correction model and the baseline model in Table 9. Traffic and geometry variables are significant because the p-values are less than 0.05 in both models. The shoulder width variables have negative effects on accident frequency. The right point lighting with no lighting variable has a negative sign in the selectivity correction model, but it is not significant. The right continuous lighting variable has a negative effect on accident frequency in both models. The median continuous lighting, median continuous with no lighting, median point with no lighting, and right point with no lighting variables are significant in both estimations and are associated with increased accident frequencies.

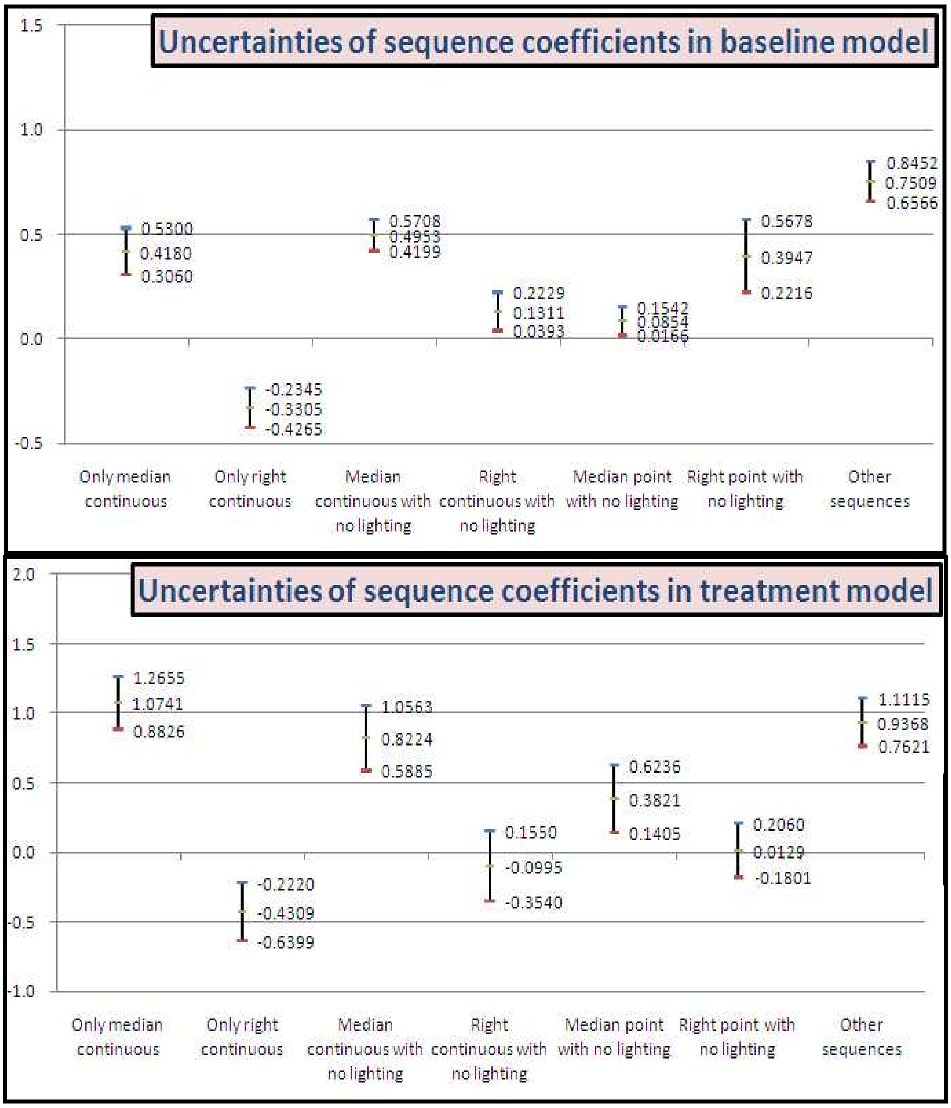


Figure 13: Uncertainty of Sequence Choice Variables in Interchange Segmentation.

The standard errors of the selectivity correction model are slightly larger than the standard errors of the uncorrected models because of unobserved heterogeneity. The coefficient for the median continuous with no lighting variable is changed by 66 percent, and the coefficient for the right point with no lighting variable is changed by 97 percent. The coefficient for the number of horizontal curves variable is changed by 52 percent. The median continuous lighting, right continuous with no lighting, median point with no lighting, and the number of vertical curves coefficients are changed by more than 100 percent, while the percent change for other variables is less than 50.

Figure **13** shows the 95 percent confidence intervals for key variables in interchange segmentation. The increase in the confidence interval for median continuous lighting variable is 71 percent, while uncertainties of the other sequence variables are increased by over 100 percent. The other lighting sequence variable has an 85 percent increase in parameter uncertainty.

Table **18** shows the selectivity bias correction model results with predicted outcomes in the case of one mile segmentation. The number of vertical curves variable has a negative sign in the baseline model. This variable has a positive sign in the selectivity correction model, but it is not significant in the 95 percent confidence interval. Both the left and right shoulder width variables have negative effects on accident frequency. Both model results show that the right continuous lighting variable contributes to a greater decrease of accident frequency than no lighting presence, but the other sequence type variables have positive effects on accident occurrences.

Table 18: Negative Binomial Results with Selectivity Bias Correction in One Mile Segmentation.

	Coefficient	Std. Err.	t-statistics	P Value	95% Conf. Interval	
Constant	0.734	0.025	28.96	0.00	0.684	0.784
Only median continuous	1.341	0.122	10 97	0.00	1.101	1.580
Only right continuous	-1.026	0.057	-18.16	0.00	-1.137	-0.916
Median continuous with n lighting	1.181	0.072	16.39	0.00	1.040	1.322
Right continuous with no lighting	0.582	0.070	8.37	0.00	0.446	0.719
Median point with no lighting	0.790	0.117	6.75	0.00	0.560	1.019
Right point with no lighting	0.479	0.071	6.71	0.00	0.339	0.619
Other sequences of lighting	1.144	0.067	17.00	0.00	1.012	1.276
Number of overpasses	0.086	0.012	6.86	0.00	0.061	0.110
Number of vertical curves	0.003	0.009	0.30	0.76	-0.016	0.021
Number of horizontal curves	0.104	0.013	7.91	0.00	0.078	0.130
Left shoulder widths	-0.062	0.005	-12.73	0.00	-0.072	-0.053
Right shoulder widths	-0.045	0.005	-9.17	0.00	-0.055	-0.036
Log AADT per number of lanes	0.161	0.007	21.56	0.00	0.146	0.175
Heterogeneity of only median continuous	-0.099	0.080	-1.24	0.21	-0.255	0.057
Heterogeneity of only right continuous	0.717	0.049	14.69	0.00	0.622	0.813
Heterogeneity of median continuous with no lighting	-0.101	0.070	-1.44	0.15	-0.239	0.036
Heterogeneity of right continuous with no lighting	-0.006	0.061	-0.10	0.92	-0.126	0.114
Heterogeneity of median point with no lighting	0.065	0.062	1.05	0.29	-0.056	0.186
Heterogeneity of right point with no lighting	0.016	0.068	0.24	0.81	-0.118	0.150
Heterogeneity of other sequences	-0.112	0.063	-1.79	0.07	-0.235	0.011
Overdispersion	0.267	0.051	5.19	0.00	0.183	0.389
log likelihood at constant						-50269.07
log likelihood at convergence						-50095.44

For most variables, except for right point with no lighting, number of vertical curves, and left shoulder width variables, the standard errors of the selectivity correction model are slightly larger than the standard errors of the uncorrected models. The standard error is slightly decreased after selectivity correction in the case of the right point with no lighting variable. The coefficients for the right continuous lighting and the number of vertical curves variables are changed by over 100 percent. The coefficient for median point with no lighting and right point with no lighting are changed by 72 percent and 39 percent, and the coefficient for right shoulder width is changed by 36 percent, while the other variables are changed between 2 percent and 29 percent.

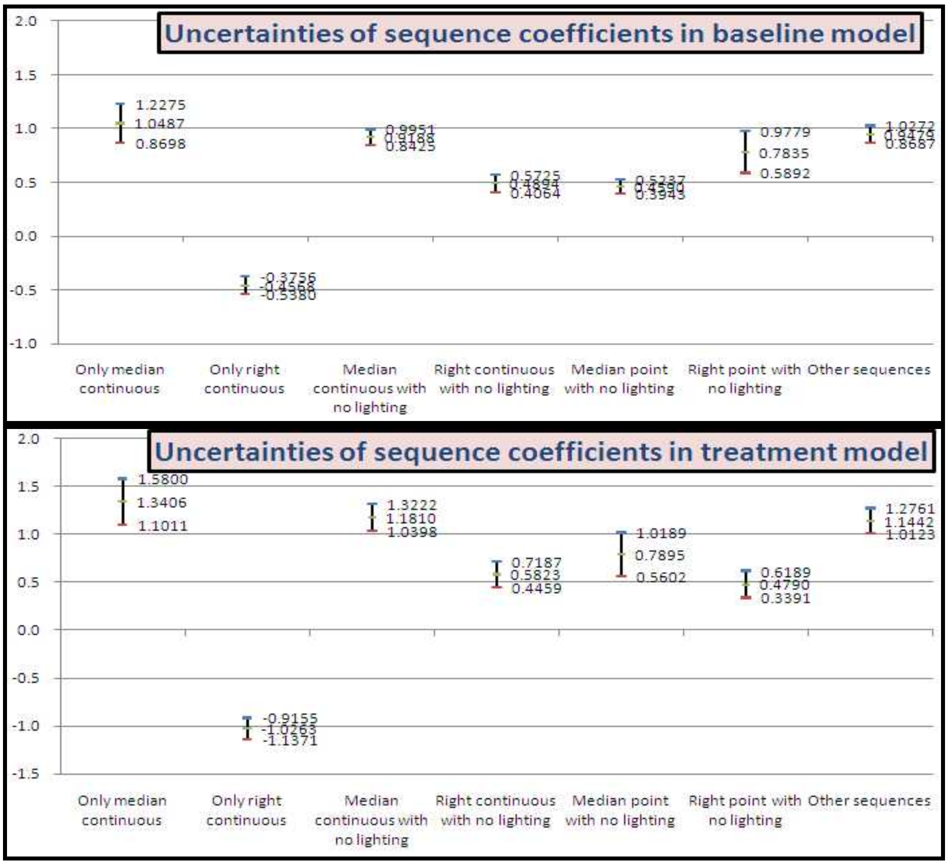


Figure 14: Uncertainty of Sequence Choice Variables in One Mile Segmentation.

The 95 percent confidence intervals for key variables in one mile segmentation are shown in Figure 14. Uncertainties of parameter estimation for the median continuous and the right continuous variables have 34 percent and 36 percent increases after selectivity correction. The confidence interval for the median continuous with no lighting variable is increased by 85 percent. About a 65 percent increase of uncertainty is found in the case of the right continuous with no lighting and other lighting sequence variables. The 95 percent confidence interval for the median point with no lighting variable is increased by over 200 percent, while uncertainty for the right point with no lighting variable is decreased by 28 percent after selectivity correction estimation.

Table 19: Negative Binomial Results with Random Sequence Choice Probability in Accident-Cluster Segmentation.

	Coefficient	Std. Err.	t-statistics	P Value	95% Conf. Interval	
Constant	-11 473	0.509	-22.52	0.00	-12.471	-10.474
Only median continuous lighting	2.388	0.867	2.75	0.01	0.688	4.089
Only right continuous lighting	1.055	2.752	0.38	0.70	-4 339	6.449
Median continuous with no lighting	-41.300	25.684	-1.61	0.11	-91.641	9.041
Right continuous with no lighting	-67.612	71.261	-0.95	0.34	-207.280	72.056
Median point with no lighting	252.773	52.448	4.82	0.00	149.977	355.568
Right point with no lighting	209.563	134.157	1.56	0.12	-53.380	472.507
Other sequences of lighting	32.587	12.727	2.56	0.01	7.642	57.532
Number of overpasses	0.627	0.021	30.31	0.00	0.587	0.668
Number of vertical curves	0.010	0.009	1.13	0.26	-0.007	0.026
Number of horizontal curves	-0.022	0.009	-2.51	0.01	-0.040	-0.005
Left shoulder widths	-0.019	0.001	-15.41	0.00	-0.021	-0.016
Right shoulder widths	-0.026	0.001	-21.75	0.00	-0.028	-0.024
Log AADT per number of lanes	0.874	0.011	82.70	0.00	0.854	0.895
Overdispersion	1.363	0.015	88.84	0.00	1.333	1.394
log likelihood at constant						-184095.81
log likelihood at convergence						-179084.78

Table 19 shows the results of the negative binomial with random sequence choice variables model in the case of accident-cluster segmentation. All independent variables are significant in the initial model shown in Table 11, but any variables associated with lighting installation with no lighting are not significant in this model because the increase in variance by random draws affects the standard errors and t-statistic values. Traffic and geometry variables, except vertical curves, seem to be significant in both models. The increase in the number of horizontal curves and the shoulder width variables contribute to the decrease in accident frequency on interstates.

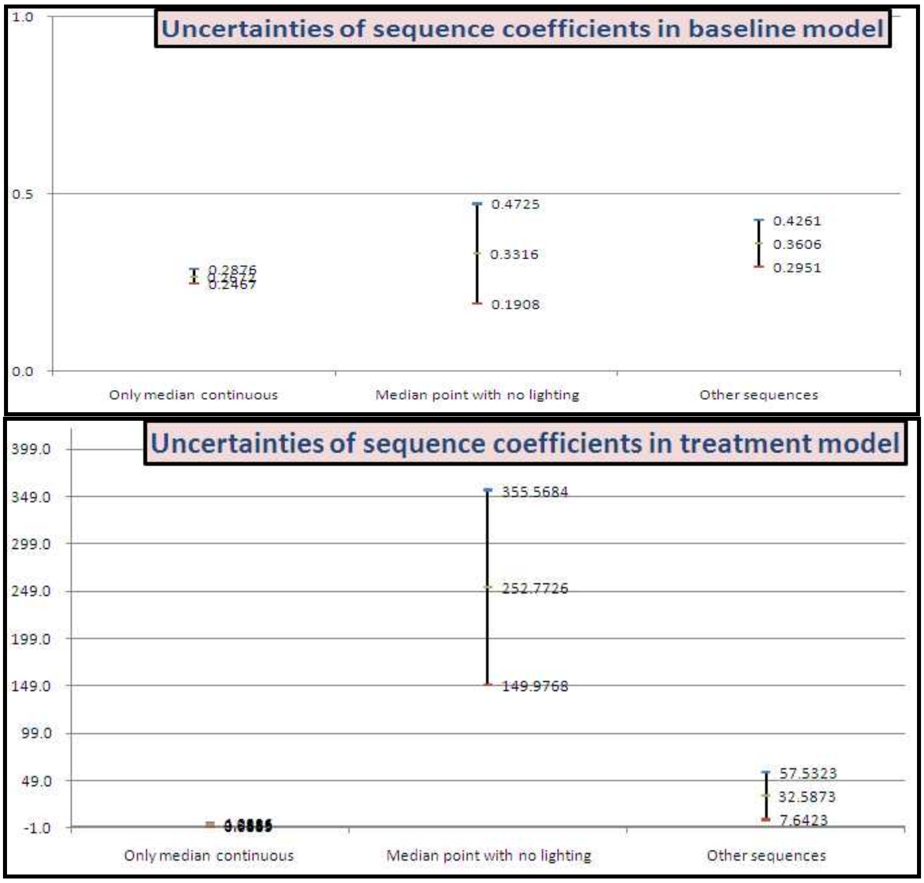


Figure 15: Uncertainty of Sequence Choice Variables in Accident-Cluster Segmentation.

The standard errors for lighting sequence variables in the pre-processing model are larger than the standard errors of those variables in the initial models. Too few observation findings affect the insignificance of the coefficients and the large standard errors of all other sequence variables except the median continuous and the right continuous lighting variables. The standard errors for other geometry variables and the traffic variable do not change after pre-processing. The coefficients for median continuous lighting, median point with no lighting, and other lighting sequence variables experience significant change. The coefficient for the number of horizontal curves is changed by 23 percent, and the coefficient for right shoulder widths is changed by 51 percent, while the other variables encounter between 2 percent and 6 percent coefficient changes.

The 95 percent confidence intervals for key variables in accident-cluster segmentation are shown in Figure **15**. The only median continuous variable, median point with no lighting variable, and other lighting sequence variable are used for uncertainty comparison between the two models because the estimated parameters of the other variables are not significant. In the case of the three lighting sequence variables, although they are significant in the pre-processing model, the coefficients and standard errors are too large to compare to those in the uncorrected model. Because the input of the lighting sequence variable by pre-processing has random probabilities in a normal distribution, random numbers significantly impact the heterogeneity of lighting sequence parameters in count estimation.

4.3 Elasticity Estimation Result of Model Outputs

This section describes the elasticity of model estimation. Elasticity describes the ratio of the percent change in one variable to the percent change in another variable. Elasticity is defined in this dissertation as the percent change in the dependent variable by the percent change of each independent variable. Accident frequency is the dependent variable, while lighting choice or sequence, geometry, and traffic variables are the independent variables. All count models have a non-log form accident frequency variable and non-log form geometry variables. So, the elasticity of geometry to accident frequency can be shown as:

$$E_x^y = \hat{\beta} x \tag{9}$$

where y is the non-log form dependent variable, x consists of variables in non-log forms, and β is the coefficient of the independent variable. Since log AADT numbers are used for the traffic variable in the models, the elasticity of the log form variable will be defined by following:

$$E_x^y = \hat{\beta} \tag{10}$$

where y is the non-log form dependent variable, x consists of variables in log forms, and β is the coefficient of the independent variable.

Table **20** shows the elasticity of the negative binomial model and the selectivity correction model in lighting segmentation. A 1 percent change in the median continuous lighting variable will increase about a 0.06 percent change in accident occurrences for both estimations. When median point lighting is increased by 1 percent, accident frequency will be decreased by 0.05 percent in the baseline model and 0.06 percent in the selectivity correction model, respectively. A 1 percent change in the right continuous lighting variable results in a less than 0.01 percent change in accidents for both models. The right point lighting variable decreases by 0.26 percent and 0.32 percent in accident frequency when it is increased by 1 percent. All other geometry variables affect less than a 0.3 percent change on accident frequency, but a 1 percent increase in traffic increases accident occurrence by about 1 percent.

Table 20: Elasticity of Negative Binomial Model and Selectivity Correction Model in Lighting Segmentation.

	Negative binomial estimation				Selectivity correction estimation			
Variable	Elasticity	Std.Err.	95% Conf. Interval		Elasticity	Std.Err.	95% Conf. Interval	
Only median continuous lighting	0.060	0.006	0.047	0.073	0.06	0.010	0.042	0.083
Only median point lighting	-0.051	0.004	-0.059	-0.044	-0.058	0.005	-0.067	-0.049
Only right continuous lighting	-0.009	0.005	-0.018	0.000	-0.007	0.007	-0.020	0.006
Only right point lighting	-0.255	0.009	-0.274	-0.237	-0.322	0.012	-0.345	-0.298
Number of urban interchanges	0.249	0.021	0.207	0.290	0.254	0.023	0.208	0.299
Number of rural interchanges	-0.019	0.012	-0.044	0.005	-0.011	0.013	-0.037	0.015
Number of urban overpasses	0.117	0.008	0.101	0.134	0.132	0.009	0.114	0.150
Number of urban vertical curves	0.102	0.013	0.077	0.128	0.068	0.014	0.041	0.096
Number of rural vertical curves	0.153	0.011	0.133	0.174	0.140	0.011	0.119	0.161
Number of urban horizontal curves	0.137	0.013	0.111	0.163	0.102	0.015	0.073	0.132
Urban left shoulder widths	-0.101	0.022	-0.144	-0.058	-0.072	0.023	-0.118	-0.026
Rural left shoulder widths	0.183	0.022	0.140	0.225	0.191	0.023	0.145	0.237
Urban right shoulder widths	-0.081	0.020	-0.121	-0.041	-0.056	0.022	-0.099	-0.012
Rural right shoulder widths	0.200	0.022	0.157	0.243	0.211	0.023	0.165	0.256
Urban log AADT per number of lanes	1.131	0.241	0.658	1.605	1.134	0.257	0.630	1.639
Rural log AADT per number of lanes	1.012	0.078	0.860	1.164	1.020	0.082	0.858	1.181

The 95% confidence intervals of elasticity for key variables in lighting segmentation are shown in Figure 16. Uncertainties of elasticity for median continuous and right continuous variables have a 62 percent and a 19 percent increase after selectivity correction. The confidence interval of elasticity for median point lighting is increased by 44 percent, while the uncertainty of elasticity for right point lighting has a 26 percent increase.

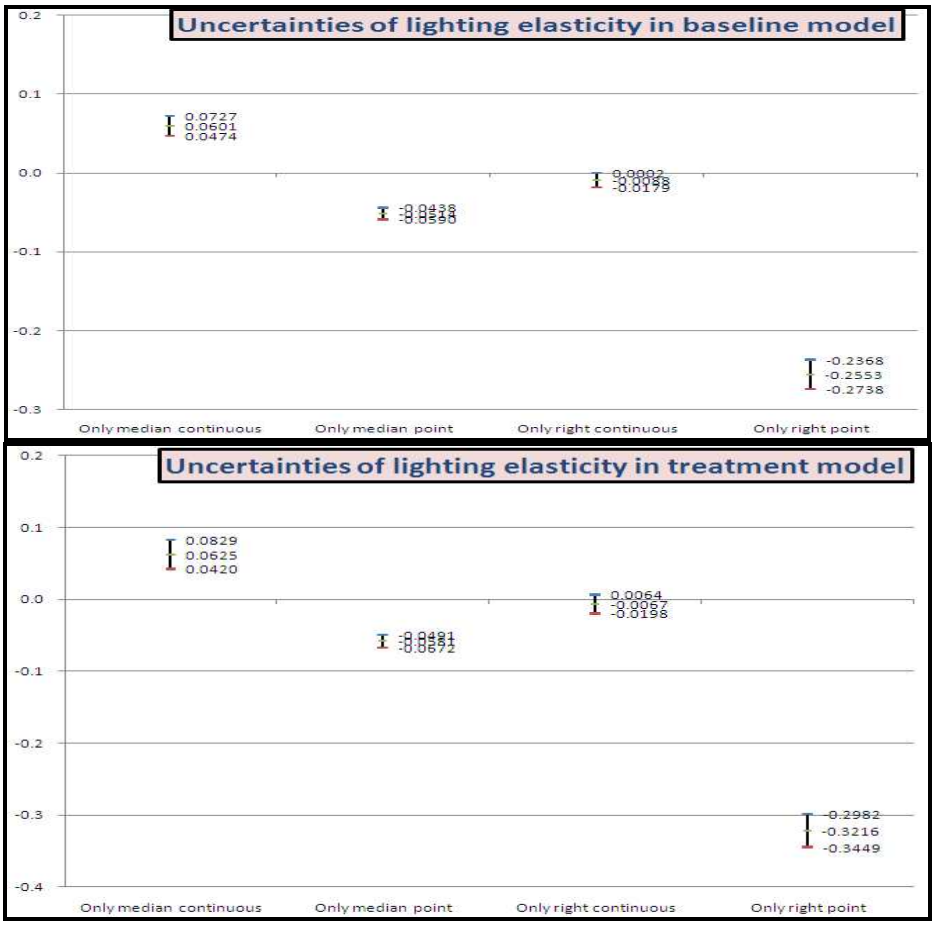


Figure 16: Uncertainty of Lighting Choice Variables Elasticity in Lighting Segmentation.

Table 21 shows the elasticity of the negative binomial model and the selectivity correction model in interchange segmentation. A 1 percent change of all variables contributes to a less than 1 percent change in accident occurrences. The right continuous lighting variable decreases by 0.02 percent in both models, and the right continuous with no lighting variable slightly increases accident frequency in the initial model but decreases accidents in the correction model. After selectivity correction, both shoulder width variables decrease in accident frequency by around 0.03 percent. A 1 percent change in the number of horizontal curves results in a 0.17 percent increase of accidents in the baseline model. The other variables affect less than a 0.1 percent change in accidents.

Table 21: Elasticity of Negative Binomial Model and Selectivity Correction Model in Interchange Segmentation.

	Negative binomial estimation				Selectivity correction estimation			
Variable	Elasticity	Std.Err.	95% Conf. Interval		Elasticity	Std.Err.	95% Conf. Interval	
Only median continuous lighting	0.016	0.002	0.011	0.020	0.040	0.004	0.032	0.049
Only right continuous lighting	-0.018	0.003	-0.023	-0.012	-0.023	0.006	-0.034	-0.012
Median continuous with no lighting	0.048	0.004	0.040	0.056	0.079	0.012	0.056	0.103
Right continuous with no lighting	0.007	0.003	0.002	0.012	-0.006	0.007	-0.020	0.009
Median point with no lighting	0.001	0.000	0.000	0.002	0.005	0.002	0.002	0.009
Right point with no lighting	0.041	0.009	0.023	0.059	0.001	0.010	-0.019	0.021
Other lighting sequences	0.040	0.003	0.034	0.046	0.050	0.005	0.040	0.060
Number of overpasses	0.045	0.006	0.034	0.057	0.060	0.006	0.047	0.072
Number of vertical curves	0.034	0.012	0.010	0.058	0.072	0.013	0.047	0.097
Number of horizontal curves	0.170	0.012	0.146	0.194	0.082	0.012	0.058	0.106
Left shoulder widths	-0.214	0.018	-0.250	-0.178	-0.237	0.020	-0.276	-0.197
Right shoulder widths	-0.119	0.017	-0.152	-0.086	-0.150	0.019	-0.187	-0.113
Log AADT per number of lanes	0.085	0.034	0.018	0.153	0.095	0.039	0.018	0.172

The 95% confidence intervals of elasticity for key variables in interchange segmentation are shown in Figure 17. The confidence interval of elasticity for the median continuous variable is increased by 83 percent, while the right point with no lighting variable has an 11 percent increase of uncertainty after selectivity correction. Selectivity correction also incurs a 70 percent increase in the confidence interval for the other lighting sequence variable. Uncertainties of elasticity for the right continuous lighting variable, the median continuous with no lighting variable, the right continuous with no lighting variable are increased by over 100 percent, while the elasticity of median point with no lighting variable has more than a 200 percent increase in the confidence interval.

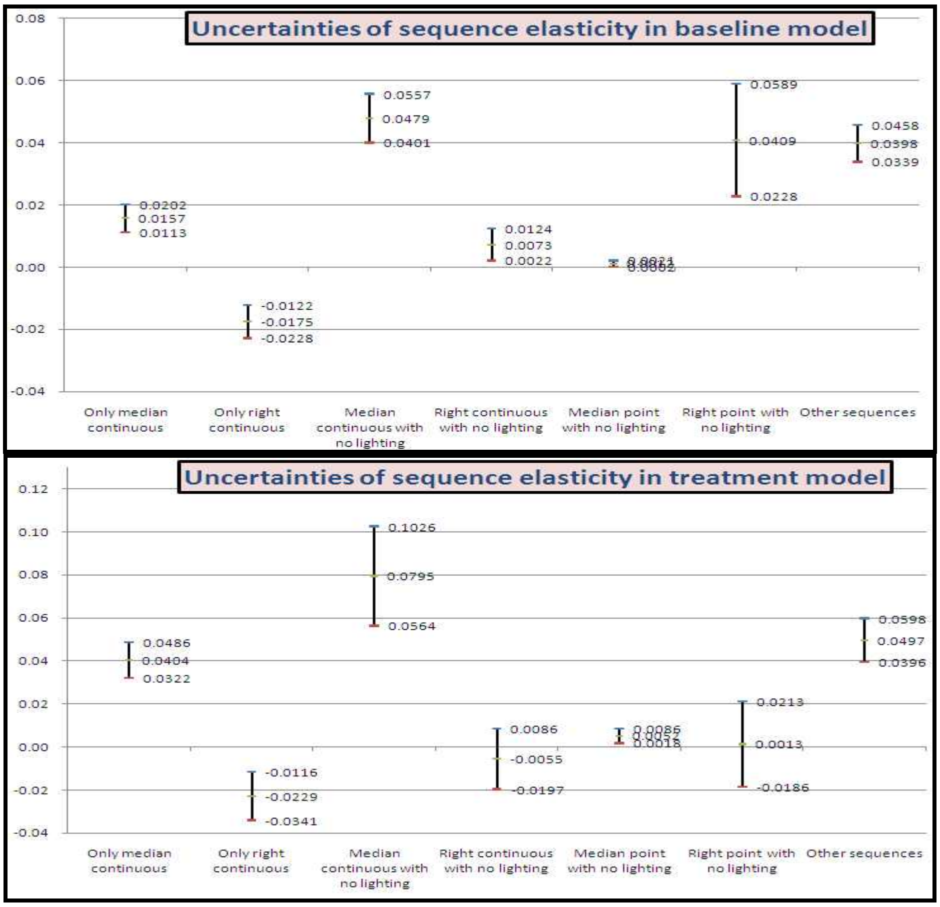


Figure 17: Uncertainty of Sequence Choice Variables Elasticity in Interchange Segmentation.

Elasticity of the initial model and the selectivity correction model in one mile segmentation is shown in Table 22. A 1 percent change of most variables incur less than a 0.1 percent change in accident frequency. The right continuous lighting variable decreases by 0.03 percent and 0.06 percent in each model. A 1 percent change in the number of vertical curves variable decreases accident frequency by 0.03 percent in the uncorrected model, but slightly increases in the correction model. A 1 percent increase in the shoulder width variables decreases accident frequency by about 0.1 percent in both models.

Table 22: Elasticity of Negative Binomial Model and Selectivity Correction model in One Mile Segmentation.

	Negative binomial estimation				Selectivity correction estimation			
Variable	Elasticity	Std.Err.	95% Conf. Interval		Elasticity	Std.Err.	95% C nf. Interval	
Only median continuous lighting	0.008	0.001	0.006	0.010	0.010	0.001	0.008	0.013
Only right continuou lighting	-0.028	0.003	-0.033	-0.023	-0.063	0.004	-0.071	-0.055
Median continuous with no lighting	0.050	0.003	0.044	0.055	0.064	0.005	0.055	0.073
Right continuous with no lighting	0.021	0.002	0.017	0.025	0.025	0.003	0.019	0.031
Median point with no lighting	0.003	0.000	0.002	0.004	0.005	0.001	0.003	0.007
Right point with no lighting	0.057	0.007	0.042	0.071	0.035	0.005	0.024	0.045
Other lighting sequences	0.051	0.003	0.045	0.056	0.061	0.004	0.053	0.069
Number of overpasses	0.039	0.005	0.030	0.048	0.034	0.005	0.024	0.044
Number of vertical curves	-0.029	0.010	-0.049	-0.008	0.003	0.010	-0.017	0.024
Number of horizontal curves	0.093	0.009	0.076	0.110	0.071	0.009	0.053	0.089
Left shoulder widths	-0.170	0.013	-0.196	-0.144	-0.184	0.015	-0.213	-0.156
Right shoulder widths	-0.094	0.012	-0.119	-0.070	-0.128	0.014	-0.156	-0.101
Log AADT per number of lanes	0.165	0.029	0.107	0.222	0.161	0.032	0.098	0.223

Figure 18 presents the 95% confidence intervals of elasticity for key variables in one mile segmentation. Uncertainties of elasticity for the median continuous lighting variable and the right point with no lighting variable are increased by 30 percent, while the right continuous lighting variable and the other lighting sequence variable have a 50 percent increase in uncertainty after selectivity correction. The confidence interval of elasticity for the median continuous with no lighting variable is increased by more than 64 percent, while the elasticity of the right continuous with no lighting variable has over a 57 percent increase in the confidence interval. Selectivity correction influences more than a 100 percent increase in the elasticity confidence interval for the median point with no lighting variable, but it decreases the uncertainty of elasticity for the right point with no lighting variable by 29 percent.

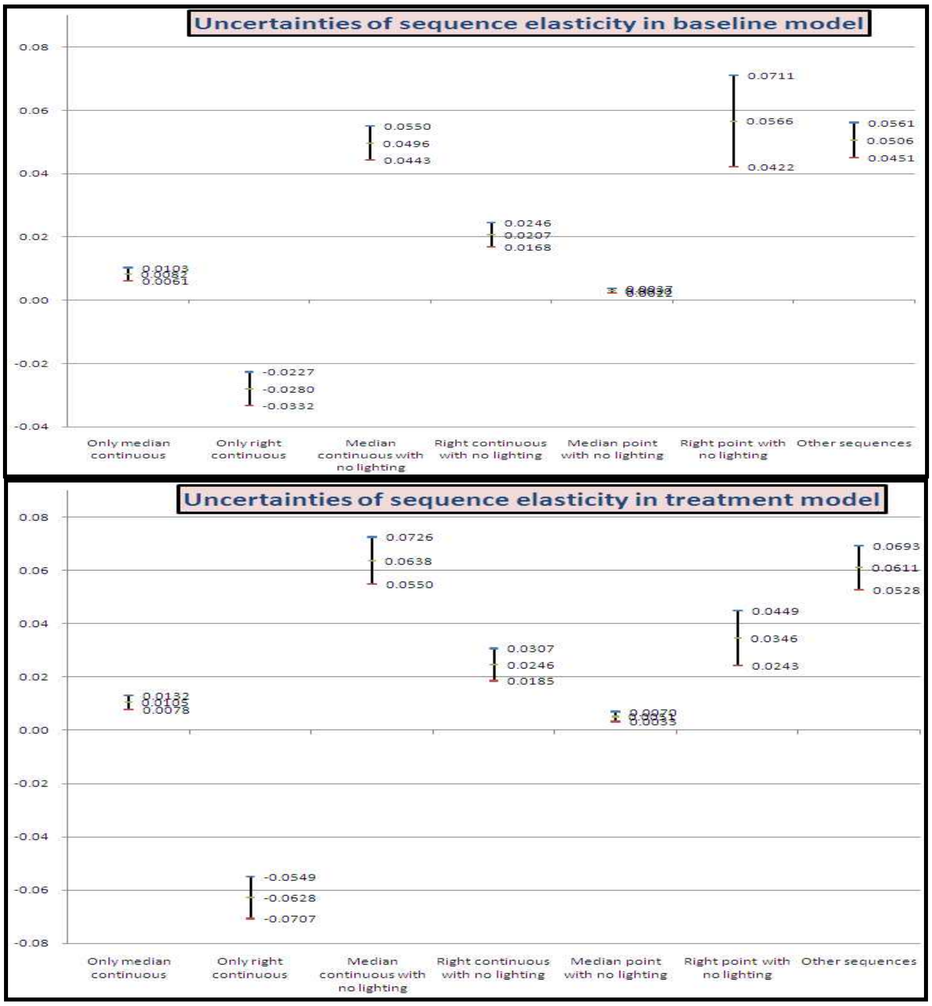


Figure 18: Uncertainty of Sequence Choice Variables Elasticity in One Mile Segmentation.

Table 23 shows the elasticity of the negative binomial model and the pre-process model in accident-cluster segmentation. A 1 percent change of all variables contributes to a less than 1 percent change in accident occurrences. A 1 percent change in the median continuous lighting variable, the median point with no lighting variable, and the other lighting sequence variable influences a less than 0.1 percent change in accident frequency in the initial model, but contributes more than a 0.4 percent change of accidents in the pre-process model. Other sequence variables also have greater contributions to accidents in the pre-process model than in the baseline model, but they are not significant. Both model results show that a 1 percent increase in traffic increases accident frequency by 0.8 percent on interstates.

Table 23: Elasticity of Negative Binomial Model and Pre-process Model in Accident-Cluster Segmentation.

	Ne ative binomial estimation				Selectivity correction estimation			
Variable	Elasticity	Std.Err.	95% Conf. Interval		Elasticity	Std.Err.	95% Conf. Interval	
Only median continuous lighting	0.0 2	0.002	0.048	0.056	0.463	0.168	0.133	0.792
Only right continuous lighting	0.019	0.001	0.016	0.021	0.091	0.237	-0.373	0.554
Median continuous with no lighting	0.003	0.000	0.002	0.003	-0.226	0.141	-0.502	0.050
Right continuous with no lighting	0.002	0.000	0.001	0.002	-0.189	0.199	-0.580	0.202
Median point with no lighting	0.001	0.000	0.001	0.002	0.843	0.175	0.500	1.186
Right point with no lighting	0.000	0.000	0.000	0.001	0.242	0.155	-0.062	0.545
Other lighting sequences	0.005	0.000	0.004	0.005	0.416	0.163	0.098	0.735
Number of overpasses	0.018	0.001	0.017	0.019	0.019	0.001	0.018	0.020
Number of vertical curves	0.005	0.003	-0.002	0.011	0.004	0.003	-0.003	0.010
Number of horizontal curves	-0.010	0.003	-0.015	-0.004	-0.007	0.003	-0.013	-0.002
Left shoulder widths	-0.124	0.008	-0.140	-0.107	-0.131	0.009	-0.148	-0.114
Right shoulder widths	-0.117	0.008	-0.133	-0.101	-0.177	0.008	-0.193	-0.161
Log AADT per number of lanes	0.849	0.110	0.634	1.065	0.874	0.111	0.658	1.091

Figure 19 shows the 95% confidence intervals of elasticity for significant choice variables in accident-cluster segmentation. The elasticities for significant key choice variables also capture large heterogeneity like parameters for these variables, and this causes enormous increases (more than 1000 percent) of uncertainty. Heterogeneity in the random selection probabilities at the pre-processing step derives large standard errors and uncertainties of parameters at the second step estimation; this becomes the cause of the immense change of elasticity uncertainty after pre-processing on the basis of random choice.

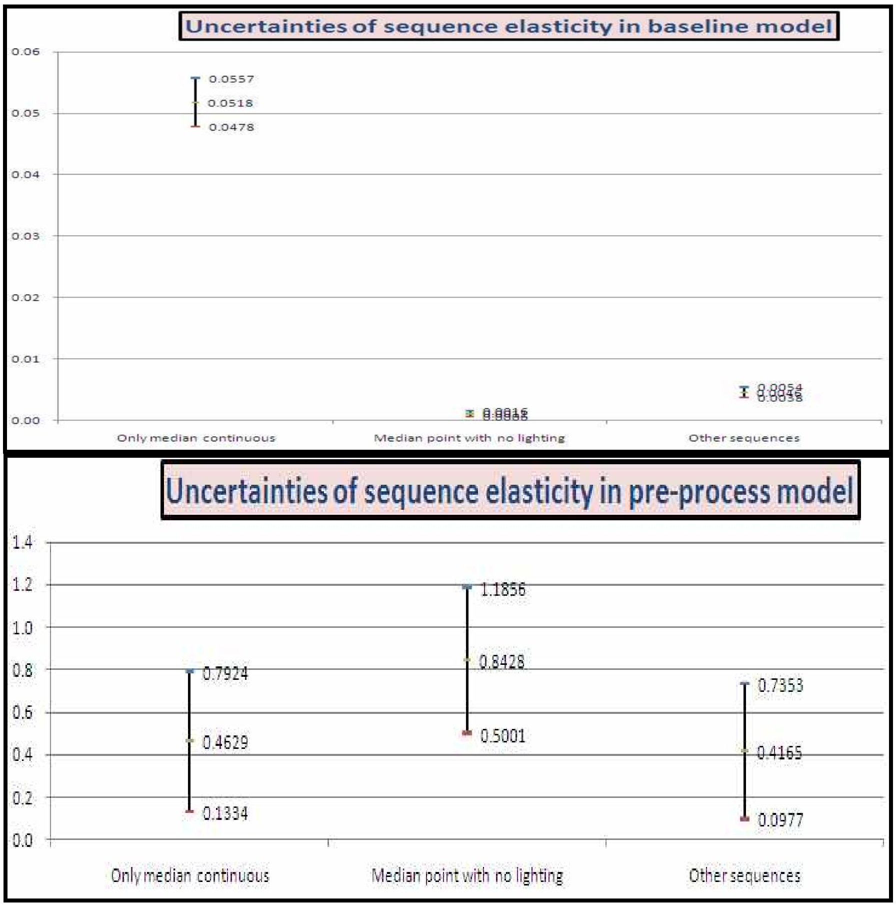


Figure 19: Uncertainty of Sequence Choice Variables Elasticity in Accident-Cluster Segmentation.

4.4 Summary of Findings in Model Results

This subsection describes the summary of findings for the modeling results. The median-side continuous lighting variable appears to have counter-productive effects on accident frequencies in the overall negative binomial models, while the median-side point lighting variable has negative signs except for the accident-clustering segmentation model. All other lighting sequence variables have positive signs in the overall models. Most geometric infrastructures, such as shoulders, median barriers, and guardrails, are installed to improve safety on interstates. It is shown that the shoulder width variables have negative signs in overall model results, and this implies installed geometric infrastructures have productive effects on accident frequencies.

Many independent variables have a positive relationship with accident frequencies in the overall selectivity bias correction models. The model results show that the right-side continuous lighting variable contributes to a greater decrease of accident frequency than no lighting presence, but the other sequence type variables have a positive relationship with accident occurrences. The right-side continuous lighting is associated with decreased accident frequencies in the lighting presence segmentation model and the lighting sequence segmentation models. The median-side continuous lighting variable seems to increase accident frequencies in the overall models. Most geometry variables have a positive relationship with the number of accidents, except for urban shoulder widths lighting segmentation in the overall models. The traffic flow variable also has positive coefficient signs in all selectivity bias correction models.

The selectivity bias correction model has more confidence intervals than the baseline model for all lighting choice variables because of the heterogeneity effect. It means the standard errors of the selectivity correction model are larger than the standard errors of the uncorrected models because of unobserved heterogeneity. The standard error is slightly decreased after selectivity bias correction in the case of the right-side point with no lighting variable in one mile segmentation. In case of accident cluster segmentation, too few observation findings affect the insignificance of the coefficients and the large standard errors of all other sequence variables except the median-side continuous and the right continuous lighting variables. The standard errors for other geometry variables and the traffic variable do not change after pre-processing. The coefficients for median-side continuous lighting, median-side point with no lighting, and other lighting sequence variables experience significant change. For elasticities of coefficients, all lighting and geometry variables affect small percent change on accident frequency in all baseline and selectivity bias correction models. A 1 percent increase in traffic increases accident occurrence by about 1 percent in case of lighting segmentation and accident cluster segmentation in both models.

Chapter 5

CONCLUSION AND DIRECTIONS FOR FUTURE RESEARCH

This chapter includes a discussion of major conclusions and methodological issues directing future research in the area of network segmentation and segmentation based model building for traffic safety inference. Major conclusions are discussed first, with a breakdown in terms of parameter inference conclusions and computation specific conclusions. These conclusions relate comprehensively to this dissertation's initial objective which is reiterated below:

To provide some perspective to the problem of modeling heterogeneity and selection bias via multiple scales, by proposing a joint model of heterogeneity and selection bias using a discrete-count approach, and using this framework to address the following research questions:

c) What is the impact of selection bias on safety intervention due to scale? In other words, if safety interventions are applied at locations where accident patterns are severe and frequent, how does one account for the lack of intervention at less problematic locations? And how does a statistical methodology derived for selection bias provide inference across scales, as segments are scaled up from very small lengths to lengths of the order of corridors?

d) How does one represent insights into the policy implications of selection bias in a manner that integrates context (i.e., roadway location and characteristics) and scale?

Major conclusions from this dissertation

I successfully addressed the problem of simultaneously addressing heterogeneity and selection bias by developing a framework that incorporates a discrete-count statistical model across multiple scales. I showed that the discrete-count model is estimable through simulation based inference. I also showed that incorporating the discrete-count model at multiple scales through endogenous and exogenous segmentation is feasible, albeit subject to computational barriers. Nevertheless, several significant conclusions in terms of parameter inference were derived, which I discuss next. I discuss the computational barriers in a forthcoming section.

I conclude that right-continuous lighting is associated with fewer accident occurrences compared to no-lighting in all corrected models. The right continuous lighting variables have negative signs in all baseline (uncorrected, non-selectivity-bias, non-heterogeneity-inclusive) models except those involving accident-cluster segmentation. In the case of selectivity correction model, the variables for sequence type that includes full or partial proportion of right continuous lighting have negative effects on accident frequencies across scale. The median continuous lighting variable has positive signs for both baseline and selectivity correction estimations across scales. Both median point lighting and right point lighting have negative signs in the case of lighting type segmentation, but the sign is positive in lighting sequence segments involving point lighting and no lighting combinations. The same trend is also shown in the selectivity-heterogeneity correction model. Other variables such as the number of urban overpasses in a segment, had a positive effect on accident frequencies across scale. An increase in the number of urban vertical curves increases accident frequencies. The number of urban horizontal curves has a negative sign only in the case of the accident-cluster segmentation. An increase in shoulder widths decreases accident occurrences while traffic flow increases accident frequencies.

The median point with no lighting presence parameter in accident-cluster level segmentation has the highest uncertainty band of 205.59 from 149.98 to 355.57. The uncertainty band width is 0.51 from -0.35 to 0.16 for the right continuous with no lighting parameter in interchange segmentation. This presents that the right continuous with no lighting variable in interchange segmentation has a potential possibility to decrease accident frequencies. The right continuous lighting parameter in interchange segmentation has consistently negative uncertainty from -0.64 to -0.22 with 0.42 band width. The highest uncertainty band widths of parameters are found in selectivity correction estimation results. This is because the simulation based estimation captures more heterogeneity by random draws.

The uncertainty band width is 0.04 from -0.02 to 0.02 for the right point with no lighting elasticity in interchange segmentation. The urban left shoulder width elasticity in lighting segmentation has consistently negative uncertainty from -0.12 to -0.03 with 0.09 band width. The AADT per lane parameters in lighting segmentation and accident-cluster segmentation have elasticity of greater than 1 or close to 1, and other parameters are significantly inelastic (less than 0.4). The elasticity of the urban log AADT per lane parameter is 1.13 with the uncertainty band width of 1.01 from 0.63 to 1.64, while the rural log AADT per lane has the elasticity of 1.02 with the band width of 0.32 from 0.86 to 1.18 in the case of lighting segmentation. The elasticity of the urban log AADT per lane parameter is 0.85 and it swings from 0.63 to 1.07 with the band width of 0.44. This implies that there is a significant increase in accident occurrences with a small increase in traffic flows. However, traffic flow parameters are significantly inelastic (less than 0.2) in the case of interchange segmentation and one mile segmentation. This presents how scale affects safety analysis in transportation as well. The elasticities of the lighting sequence variables are increased from less than 0.002 to the range between 0.009 and 0.843 after selectivity correction in the case of accident cluster segmentation. This implies that the simulation based modeling method increases the contribution of lighting presence to the safety on interstates. The increase of predicted probability for lighting sequence by random draws seems to increase the coefficients and elasticities of lighting sequence variables in the count estimation step of the model.

Computation specific conclusions

I used all interstates in Washington States data for model estimation in this dissertation. I experienced that the input matrix size is a significant factor for simulation based safety analysis in transportation field. Robust inference requires panels with longitudinal histories of six years or more usually, and in the statewide context, this can involve several thousand or hundreds of thousands of observations depending on scale. The simultaneous maximum likelihood estimation method is used for getting the proper parameters and optimizing the models. The initial parameters are randomly chosen at the first step of estimation. The probabilities of the dependent variables are obtained by the probability function of the model. New parameters are estimated by the probability and observed numbers. Until the likelihood function is maximized (close to zero), this process is repeated simultaneously for multiple parameters. A Newton-Raphson algorithm is generally used for the convergence of the maximum likelihood function; this is the optimization algorithm to find maximized function. This algorithm is repeated while the absolute value of the function's first derivate is greater than the tolerance, and the tolerance is generally 10^{-8}. The input value is replaced by subtracting the combination number of the first derivative and the second derivative from the old input value in this algorithm. In the maximum likelihood estimation method, parameters are input values and are simultaneously replaced by the Newton-Raphson algorithm. The maximum likelihood function and estimated parameters are obtained when the iteration is stopped by the first derivative of the likelihood function satisfying the tolerance. Although the simulation based approach offers a feasible method for full-information maximum likelihood modeling of heterogeneity and selectivity bias, model estimation is severely hampered by dataset size. For example, the sample size of endogenous segmentation is 38,265 observations (while other segmentation types such as lighting-type-specific, one-mile, or interchange-type) have fewer than 1,600 observations. The input matrix size is 612,240 cells with 16 variables in endogenous segmentation. The mixed logit model uses Halton draws to generate random numbers. Generally, a pseudo random number generation is used to create random numbers, and it is based on a uniform distribution. However, despite its fast speed to generate random numbers, it is an insufficient method to generate random numbers for model estimation because of its high discrepancy, which means the draws are far from uniform. Instead, the Halton sequence is used to simulate random draws with low discrepancy. The Halton sequence is also called a reverse radix-based sequence because it uses a radical inverse function to gain the point in the interval corresponding to a specific number (Kocis and Whiten 1997). The discrepancy is estimated by selecting a representative subinterval from the sequence draw, and then sliding the subinterval through the draw range. Since the Halton draw focuses on a uniform interval rather than equal randomization, it promises lower discrepancy than the pseudo random draw. With 200 quasi-random (based on the Halton sequence) draws, the mixed logit model treatment-effects model for endogenous segmentation was unable to handle this matrix size, and estimation fails to proceed iteratively. It is to be noted that the 200-draw procedure is done repeatedly at each iteration in order to evaluate the function (log-likelihood for the observed sample) and the gradient. The initial computations begin with converged parameter values from the multinomial logit baseline, using the well-known Broyden-Fletcher-Goldfarb-Shanno (BFGS) method. The line search sub-problem involving step sizes is rarely ever reached at the end of the first iteration. As a point of reference, it requires typically around 100 iterations for convergence of the treatment-effects model in the smaller segmentation datasets.

Matrix management in such cases had to be done manually. To reduce the size of matrix in the models, the interstate dataset was divided into seven individual, interstate-by-number datasets. However, this produced sample size problem of another kind. Some interstates, such as I-182, I-205, I-405, and I-705, had sample sizes that were too small as a result of few clusters. The lack of variability in these samples forces the analyst to make judgments on combining interstates, which in turn implies restrictive assumptions on heterogeneity across interstates. In essence, the computational problem comes with significant tradeoffs, one that should serve as a useful objective in future research.

The nine year panel on accidents used in this dissertation serves as a good empirical example of endogenous segmentation that can create its own type of matrix size problems. The cluster specification that leads to endogenous segmentation involves the method of medoids. If other clustering methods were to be used along with other functional classes of roads, such as divided highways, arterials and collectors, the clustering method itself would be subject to computational challenges, let alone the challenge of post-processing the clusters for joint modeling of heterogeneity and selection bias.

The matrix size problem also contributes to limitations in lighting-type variable definitions. The treatment-effects model handles the lighting-type variables at the mixed multinomial logit estimation step. The lighting choices are dummy numbers from the lighting-type segmentation, while they become proportions of lighting type presence in other segmentations. In the outcomes component, the lighting type probabilities are inserted as independent variables via the selection-bias with heterogeneity correction. The crucial part is the estimation of lighting type probabilities for the "second stage" estimation. The mixed logit model produces the proper probability output for small matrix sizes with up to roughly 1,000 rows, 3 columns, and 100 quasi-random draws, but it produces unexpected probabilities (sometimes nonsensical) with 13,824 rows, 7 columns, and 100 random coefficients in larger- scale segmentations. In purely exogenous segmentation, the problem of micronumerosity occurs. For example, tunnel and both side lighting are relatively minimal in observation size compared to right-side and median-side lighting samples in the lighting segmentation dataset. This can cause identification type problems since the vector of "1s" is small. I tested the convergence without tunnel and both-side lighting types incrementally. When compared with the uncorrected negative binomial model, the treatment effects models show substantial changes in the magnitude of coefficients associated with lighting type. The model estimation results show larger standard errors and uncertainties after selectivity correction by treatment as well. The selectivity correction model captures heterogeneity of lighting choice at the logistic estimation step, and this heterogeneity can cause the change of coefficients associated with lighting choice and cause greater uncertainties in parameters and elasticities.

Policy Implications of the Findings

In the field of transportation, the current policy for lighting installation is applied based on random area choice or darkness in the area on a roadway. It does not consider the impact of lighting presence on safety. The policy also does not consider removal or reinstallation of lighting presence based on the improvement of safety and energy consumption. So, decision makers must consider the policies for lighting installation based on the analysis of the effects of lighting installations on accident frequencies on the roadway. This dissertation provides the example policy implication of lighting installation by safety analysis on interstates; this can assist decision makers in determining future policies for managing luminaries on interstates. Several policy implications arise from the models estimated in this dissertation. For example, the findings on the right-continuous lighting type support current practice which favors that type of installation. The elasticities for median point lighting type and other sequences of lighting types from the accident-cluster model show that their installation produces counter-productive effects that are not negligible. Median point lighting type is estimated to produce an elasticity greater than one on the high end of the 95% confidence band, suggesting that lighting policy should consider abandoning this type of installation on freeways. The known installation types such as median continuous and right continuous lighting types produce productive safety effects (elasticities around 0.2) decreasing accident occurrences even at the high end of the 95% confidence band. With the modeling results, decision makers can consider the removal of median point lighting from segments where accident frequencies are high compared to no lighting areas. Also, they may consider installing more median lighting poles in these segments to convert median point lightings to median continuous lighting types because the median continuous lighting has produced effects that improve safety. If significant accident frequencies are found in no lighting segments, decision makers can consider the policy of median continuous or right continuous lighting in the segments.

The lighting policy can also be considered based on the elasticities of parameters as well. In the case of lighting segmentation, other lighting presences, except median continuous lighting, contribute to decreasing accident frequencies compared to no lighting presence. However, the elasticities of the parameters are significantly small, and this means that a large increase in these types of lighting installation will only have a small decrease in accident frequencies on interstates. Compared to the elasticities of other infrastructure parameters, the elasticities of lighting sequence parameters are significantly small in the interchange and one mile segmentation cases. The policy decision makers should consider more investment in right-continuous lighting, median-point lighting, right-point lighting installation, and re-installation of lighting poles in sequence segment cases, rather than investment in other geometry infrastructures to improve safety on interstates.

Coupled with these safety insights, decision makers can utilize energy consumption models for various lighting types to determine the optimal installation lengths in terms of energy and safety, while promoting traffic flow without breakdowns. If a certain type of lighting does not have positive affects accident reduction, then these luminaries need to be turned off even at night or under bad weather for energy conservation. For example, from the modeling results, median point lighting has counter-productive effects on accident frequencies. So, decision makers can consider turning off the power to median point lighting for energy conservation on interstates. In rural areas, the sudden presence of point lighting after the a long period of darkness can alter the driver's behavior because of the sudden vision adjustment; this may be a contributing factor of crashes on the roadway as well. Median point lighting can also be considered for removal or turned off for energy saving as well as for safety. On the contrary, although median continuous or right continuous lighting consume a large amount of energy, the continuous lightings should be kept on for interstate safety. The safety cost can be inflated by vehicle maintenance, the treatment of the injured and killed people, crash handling, recovery of the roadway, and traffic flow recovery. Since the cost of safety includes many costly actions, it must be deliberated whether an infrastructure significantly decreases safety cost by reducing accidents, even though the energy cost may be very high. Based on the modeling results in this dissertation, two continuous lightings significantly contribute to decreasing accident frequencies. Because two continuous lightings seem to significantly reduce the safety cost compared to the cost of energy consumption by lighting poles, decision makers must consider whether to keep turning on or installing more continuous lightings on interstates. Instead of turning the luminary on or off, an adjustment to luminary density can be considered to get efficient energy saving while also reducing accident risk on interstates. To find the most efficient luminary strength, the lighting models should be considered to compromise between energy saving under Energy Policy Act 2005 and the cost of safety. This methodology is used for defining the commercial lighting power limits developed by the American Society of Heating, Refrigerating and Air-Conditioning / Illuminating Engineering Society of North America (ASHRAE/IESNA) 90.1-2004. The methodology calculates lighting power allowances for building spaces and whole buildings, and it uses available efficient lamp/ballast/fixture data, and illuminance values from IESNA illuminance recommendations. Energy-efficient design is promoted through the resulting lighting power densities (LPD) by this model.

Transferability Issues for Model Results

Accident analysis based on statistical modeling must consider roadway environmental factors such as number of curves, shoulder widths, and traffic volume. Since each state has different roadway environments and conditions, it is not easy to apply transportation accident analysis policies that can be used in one state to another state. In regards to data aggregation, the data by exogenous segmentation affects model transferability more for other states than endogenous segmentation data. The exogenous segmentation creates aggregated data based on independent variables that include geometrics and traffic information, while the endogenous segmentation is the data aggregation by accident cluster in this dissertation. Washington State has several unique environmental conditions influenced by a combination of factors such as mountainous terrains, frequent rains, and size of the urban area. The size of the mountainous area affects vertical and horizontal curves design, while the size of the urban area can influence the traffic flow. So, the model results in Washington State can to be applied to states having similar topographic features or urban size. As such, the results in this dissertation are difficult to be applied to states having mostly flat areas and small urban areas such as Iowa, Nebraska, or Kansas. Also, if weather condition is to be considered as independent variables, it is difficult to apply the Washington State modeling results to Arizona, New Mexico, Texas, or Nevada because the weather conditions are completely different in these states. The models estimated in this dissertation focus on the policy decision of lighting installation to decrease accident frequencies on interstates. The predicted probability of lighting choice reflects data aggregation at the discrete choice modeling step in the case of exogenous segmentation, such as lighting segmentation and interchange segmentation, while data aggregation only affects the predicted probability of accident frequencies in the endogenous segmentation case. The models are estimated with Washington State interstates data; since the data segmentation affects lighting installation choice in exogenous segmentation cases, it is very difficult to apply the policy implication of lighting installation to other states based on the Washington State results. However, it is possible to use the policy implication of lighting installation from the analysis results to other states or nationwide in the case of endogenous segmentation. The transferability issue demonstrates the importance of data aggregation in accident analysis; this is another contribution from this dissertation for policy decision-making by accident analysis in the field of transportation.

Methodological Issues Directing Future Research

Regarding segmentation theory, some methodological issues are bound to arise. I have presented an arbitrary segmentation approach while building my treatment effects model. This segmentation approach assumed that lighting segments are defined by the boundaries of existence of a particular type of lighting; so, it is a purely exogenous segmentation process. However, it is not so pure that all X variables are homogeneous in this definition; only the lighting definition is homogeneous. This introduces to some extent the "counting problem". As an example, I have computed the number of horizontal curves at the lighting segment level. The count of curves is not a complete count in the sense that some curves extend beyond the boundaries of the segment. So "counting" is done in a limited sense for the purpose of extracting variables for model estimation. I do not think this can be avoided, since the only recourse is to have purely homogeneous segments where all Xs are homogeneous. As discussed previously, this artificially induces the problem of excess zeros, and further, does not have potential to serve as a template for causal investigation since the scale can be very small. At the very least, scale should be defined by a minimum length – that is a length that can accommodate vehicle to vehicle interaction effects and environmental effects. This being said, I examined other scales, such as an accident clustering segment scale, and physical scales based on interchange and every one mile density. Both scales are very meaningful choices. The first is based on the outcome and hence is purely endogenous. I was able to draw insights into the causal nature of the accident occurrence process by accommodating minimal lengths of scale issues.

Here, instead of using lighting installation dummies, the eight types of lighting sequence variables are used for model estimation in purely endogenous and physical scales. A similar convergence problem was observed while I ran both models for the interstate segmentation and one mile scale datasets with all urban and rural geometry, and traffic variables. Since the problem stems from the lack of lighting sequence observations in the rural area, I only considered urban area data for model estimations. The change of parameters, standard errors, and uncertainties associated with lighting sequence are found after selectivity correction in these segmentation cases as well. However, the magnitude of the parameter changes, the standard errors, and the uncertainties are varied in each scale, and this shows how segmentation scale impacts selectivity bias and heterogeneity of choice. The negative binomial model results after pre-processing by random sequence choice probabilities in case of endogenous scale is described in the previous chapter as well. Accident-cluster segmentation creates a very small scale dataset. Due to small scale segmentation, the dataset has too many 0 values in all independent variables, which creates a convergence problem for treatment effect model estimation. Instead of a treatment effect model, I estimated a negative binomial model with pre-processing for capturing heterogeneity in the lighting sequence choice variables. I simply applied random sequence probabilities for heterogeneity, but this creates too much overdispersion and more problems in the parameter estimation of some sequence choices. Although some sequence choice variables have insignificant problems, this model result also presents the effects of heterogeneity on selectivity bias and uncertainties. More meaningful random methodology should be considered for heterogeneity controls in small scale datasets for further research.

The visualization of modeling results can be conducted with Internet map service technologies for future research as well. The visualization of statistical model results will allow decision makers to visually inspect severe heterogeneity associated with lighting type. The visualization template will permit people to see the heterogeneity along the centerline and explore which locations have similar magnitudes.

References

Anagnostopoulos, A. Vlachos, M. Hadjieleftheriou, M. Keogh, E. Yu, P. S., 2006. Global distance-based segmentation of trajectories. *Conference on Knowledge Discovery in Data, Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, 34 – 43.

Butler, J. A. Dueker K. J., 2001. Implementing the Enterprise GIS in Transportation Database Design. *Urban and Regional Information system Association Jornal*, Vol. 13, No. 1.

Deb, P. Trivedi, P. K., 2006, Maximum simulated likelihood estimation of a negative binomial regression model with multinomial endogenous treatment. The Stata Journal, Vol. 6, No. 2, 246-255

Dueker, K. J. Butler, J. A., 2000. A geographic information system framework for transportation data sharing. *Transportation Research Part C: Emerging Technologies*, Vol 8, Issue 1-6, 13-36.

Dubin, J. A. McFadden, D. L., 1984. An Econometric Analysis of Residential Electric Appliance Holdings and Consumption. *Econometrica*, Vol. 52, No. 2, 345-362

Green, E. R. Agent, K. R. Barrett, M. L. Pigman, J. G., 2003. Roadway Lighting and Driver Safety. *Kentucky Transportation Cabinet Project Report*.

Heckman, J. J., 1979. Sample Selection Bias as a Specification Error, Econometrica. *The Econometric Society*, Vol. 47, No. 1, 153-161.

Heckman, J. J., 1990. Varieties of Selection Bias. *The American Economic Review, Papers and Proceedings of the Hundred and Second Annual Meeting of the American Economic Association*, Vol. 80, No. 2, 313-318.

Hausman, J. McFadden, D., 1984. Specification Tests for the Multinomial Logit Model. *Econometrica*, Vol. 52, No. 5, 1219-1240

Isebrands, H. Hallmark, S. Hans, Z. McDonald, T. Preston, H. Storm, R., 2004. Safety Impacts of Street Lighting at Isolated Rural Intersections – Part II, *Minnesota Department of Transportation and Minnesota Local Road Research Board Project Report*.

Karlaftis M . G. Tarko A., 1998, Heterogeneity Considerations in Accident Modeling, *Accident Prevention and Analysis*, Vol. 30, Issue 4, 425-433.

Kim, D. G. Washington, S., 2006. The significance of endogeneity problems in crash models: An examination of left-turn lanes in intersection crash models. *Accident Analysis & Prevention*, Vol. 38, Issue 6, 1094-1100.

Kocis, L. and Whiten, W. J., 1997. Computational investigations of lowdiscrepancy sequences. ACM Transactions on Mathematical Software, 23(2):266-294

Lee, L. F., 1984. Tests for the Bivariate Normal Distribution in Econometric Models with Selectivity. *Econometrica*, Vol. 52, No. 4, 843-863.

Lord, D., 2000. The Prediction of Accidents on Digital Networks: Characteristics and Issues Related to the Application of Accident Prediction Models. *PhD Dissertation, Department of Civil Engineering, University of Toronto, Toronto*.

Murphy, K. M. Topel, R. H., 1985. Estimation and inference in two-step econometric models. *Journal of Business and Economic Statistics*, Vol. 3, Issue. 4, 370-379

Nyerges, T. L., 1990. Locational referencing and highway segmentation in a geographic information system. *Institute of Transportation Engineers Journal*, 27-31.

Oh, J., 2006. Text Extraction and Post Processing for Statistical Analysis of Interstate Accidents: A Case Study. *Master of Science Thesis, The Pennsylvania State University, USA*.

Puhani, P. A., 2000. The Heckman Correction for Sample Selection and its Critique. Journal of Economic Survey, Vol. 14, No. 1, 53-68

Quiroga, C. A., 2000. Performance measures and data requirements for congestion management systems. *Transportation Research Part C: Emerging Technologies*, Vol 8, Issue 1-6, 287-306.

Senn, L., 2005. Summary Report: Washington State Road Weather Information Systems. *Washington State Department of Transportation Project Report*.

Shankar, V. N. Mannering, F. L. Barfield, W., 1995. Effect of roadway geometrics and environmental conditions on rural accident frequencies, *Accident Analysis and Prevention,* 27(3), 371-389.

White, H., 1980. A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity. *Econometrica*, Vol. 48, No. 4, 817-838.

Bibliography

Carson, J. Mannering, F. L., 2001. The effect of Ice Warning Signs on Ice-Accident Frequencies and Severities. *Accident Analysis and Prevention*, Vol. 33, 99-109.

Carvalho, D. F. Estrella, J. C. Freire, A. P. Goularte, R. Santana, RH. C. Fortes, RP. M., 2007. Functional and device interoperability in an architectural model of geographic information system. *Proceedings of the 25th annual ACM international conference on Design of communication*, 127-133.

Dubin, J. A. Rivers, D., 1989. Selection Bias in Linear Regression, Logit and Probit Models. *Sociological Methods & Research, SAGE Journal*, Vol. 18, No. 2-3, 360-390.

Econometric Software. 2003. NLogit Version 3.0. *Econometric Software Inc.*

Giles, M. J., 2001. Heckman's Methodology for Correcting Selectivity Bias: An Application to Road Crash Costs. *Edith Cowan University Working Paper*, No. 01.11.

Greene, W. H., 2003. Econometric Analysis, Fifth edition. *Prentice Hall Inc.*

Greene, W. H. Hensher, D. A., 2007. Heteroscedastic control for random coefficients and error components in mixed logit. *Transportation Research Part E: Logistics and Transportation Review*, Vol 43, Issue 5, 610-623.

Hensher, D. A. Rose, J. M. Green, W. H., 2005. Applied Choice Analysis: A Primer, *Cambridge University Press.*

Jang, S. G. Kim, T. J., 2006. Modeling an interoperable multimodal travel guide system using the ISO 19100 series of international standards**.** *Proceedings of the 14th annual ACM international symposium on Advances in geographic information systems*, 115-122.

Kan, M. Y. Klavans, J. L. Mckeown, K. R., 1998. Linear Segmentation and Segment Significance. *Proceedings of 6th International Workshop of Very Large Corpora*, 197-205.

Kennedy, P., 2003. A Guide to Econometrics, 5th Edition. *The MIT Press*.

Kutner, H. M. Nachtsheim, J. C. Neter, J., 2004. Applied Linear Regression Models fourth edition, *McGraw-Hill/Irwin.*

Little, RJ. A., 1985. A Note About Models for Selectivity Bias, *Econometrica, The Econometric Society* , Vol. 53, No. 6, 1469-1474.

Mannering, F. L., 1986. Selectivity Bias in Models of Discrete and Continuous Choice: An Empirical analysis. *Transportation Research Record: Journal of the Transportation Research Board, Transportation Research Board of the National Academies,* 58-62.

Mannering, F. L. Hamed, M. M., 1990. Occurrence, Frequency, and Duration of Commuters' Work-To-Home Departure Delay. *Transportation Research Record: Journal of the Transportation Research Board, Transportation Research Board of the National Academies*, Vol. 24B, No. 2, 99-109.

Marshall, D., 2005. Programming Microsoft Visual C# 2005: The Language. *Microsoft Press.*

McFadden, D. Train, K., 2000. Mixed MNL Models for Discrete Response. *Journal of Applied Econometrics*, Vol. 15, No. 5, 447-470.

Milton, J. C. Mannering, F. L., 1996. The relationship between highway geometrics, traffic related elements and motor vehicle accidents, *Final Research Report, Washington State Department of Transportation, Washington*, WA-RD 403.1.

Mohammadian, A. Doherty, S. T., 2005. Mixed Logit Model of Activity-Scheduling Time horizon Incorporating Spatial-Temporal Flexibility Variables. *Transportation Research Record: Journal of the Transportation Research Board*, *Transportation Research Board of the National Academies,* Vol. 1926, 33-40.

National Highway Traffic Safety Administration, 2004. *Traffic Safety Facts.*

Poch, M. Mannering, F. L., 1996. Negative binomial analysis of intersection-accident frequencies. *Journal of Transportation Engineering,* 122(3), 105-113.

Revelt, D. Train, K., 1998. Mixed Logit with Repeated Choices: Households' Choices of Appliance Efficiency Level. *The Review of Economics and Statistics, The MIT Press,* Vol. 80, No. 4, 647-657.

Rosen, D. Shankar, V. N. Ulfarsson, 2007. Relationship of Shopping Activity Duration and Travel Time with Planning-Level Network and Socioeconomic Factors. *Transportation Research Record: Journal of the Transportation Research Board, Transportation Research Board of the National Academies*, Vol 1894, 160-169.

Shah, L., 1968, A Methodology to Related Traffic Accidents to Highway Design Characteristics, *Ohio University*.

Shankar, V. Milton, J. Mannering, F., 1997. Modeling accident frequencies as zero-altered probability processes: An empirical inquiry*, Accident Analysis and Prevention*, 29(6), 829-837.

Shankar V.N. Park M. Oh J. Sathyanarayan S. 2008, An Assessment of Interstate Safety Investment Priorities in Washington State. *Washington State Project Report*.

Skibo, C. Young, M. Johnson, B., 2005. Working With Microsoft Visual Studio 2005. *Microsoft Press*.

Train, E. K., 2003. Discrete Choice Methods with Simulation, *Cambridge University Press*.

Ulfarsson, F. G. Shankar, N. V. Vu, P., 2005. The Effect of Variable Message and Speed Limit Signs on Mean Speeds and Speed Deviations. *International Journal of Vehicle Information and Communication Systems*, Vol. 1, No. 1/2.

Vella, F., 1998. Estimating Models with Sample Selection Bias: A Survey. *The Journal of Human Resources, University of Wisconsin Press*, Vol. 33, No. 1, 127-169.

Walker, J., 2002. Mixed Logit (Or Logit Kernel) Model: Dispelling Misconceptions of Identification. *Transportation Research Record: Journal of the Transportation Research Board, Transportation Research Board of the National Academies*, Vol. 1805, 86-98.

Washington, S. P. Karlaftis, M. G. Mannering, F. L., 2003. Statistical AND Econometic Method FOR Transportation Data Analysis. *Chapman and Hall/CRC.*

Washington State Department of Transportation, 2000. Local Agency Safety Management System, *WSDOT*.

Ye, K. Myers, H. R. Walpole, E. R. Myers, L. S., 2006. Probability and Statistics for Engineers and Scientists, Eighth edition, *Prentice Hall Inc*.

Glossary

Endogenous: Parameters are endogenous by correlation between parameters. The measurement errors, simultaneity, omitted variables, and sample selection errors can be causes of endogeneity.

Exogenous: Exogeneity occurs when a parameter comes from outside of the equation without correlation with any other variables. It arises when a bi-directional relationship exists between the dependent variable and an independent variable; the opposite meaning of endogeneity.

Heckman's correction: The statistical method to correct selectivity bias of choice variables. This has two stage processes. The binary choice model is estimated in the first stage, and the choice probability is used for the independent variables in the second stage count estimation.

Heterogeneity: This term is used when a variable has a large number of structural variations in the model estimation. The randomness can be described by the same terminology.

Inverse Mill's ratio: Terminology to explain the probability of a non-selected proportion being selected in the selection bias correction model. It accounts for non-randomness.

Mixed multinomial logit regression: The multinomial logit regression capturing the randomness of parameters in the estimation.

Multinomial logit regression: The regression model in which the generalized logistic regression accounts for two or more discrete choice out comes.

Negative binomial regression: The regression analysis for the count estimation. This model is used for count estimation overdispersion exists within the data.

Overdispersion: Overdispersion occurs when the variance is greater than the mean value.

Poisson regression: The regression analysis for count estimation. The dependent variable is based on Poisson distribution.

Selection bias: This is a statistical bias by an error in choosing the individuals or groups to take part in a study. This occurs mostly by errors from the method of collecting samples.

Selection bias correction model: Count model that minimizes selection bias by capturing the seemingly selected portion. The Inverse Mill's ratio term is added in the general model in order to account for the probability of a non-selected portion being selected.

Uncorrected model: Negative binomial model without selectivity bias correction. This model is used for comparing selection bias in choice variables. This is also called the initial model or the baseline model in the dissertation.

Appendix

COMMANDS FOR MODEL ESTIMASTIMATION

NEGATIVE BINOMIAL AND TREATMENT-EFFECT ESTIMATION

1. Lighting segmentation

```
. insheet using "E:\LightingMileUrbanRuralFullInteraction.csv"
(23 vars, 9630 obs)

. nbreg total mc mp rc rp urbaninter ruralinter urbanoverp
urbanvcurve ruralvcurve urbanhcurve  urba nlftshw rurallftshw
 urbanrighshw ruralrighshw urbanlnadtpl rurallnadtpl,
dispersion(mean)

. mtreatnb total urbaninter ruralinter urbanoverp urbanvcurve ruralvcurve urbanhcurve urbanlftshw
rurallftshw urbanrighshw ruralrighshw urbanlnadtpl rurallnadtpl, mtreatment (light urbaninter ruralinter
urbanoverp urbanvcurve ruralvcurve urbanhcurve urbanlftshw rurallftshw urbanrighshw ruralrighshw
urbanlnadtpl rurallnadtpl) simulationdraws(100)
```

2. Interchange segmentation

```
. insheet using "E:\SequenceOfInterMileUrbanruralADTPLWithoutTunnelandBoth.csv"
(31 vars, 10521 obs)

. nbreg total sequence2 sequence3 sequence4 sequence5 sequence6 sequence7 sequence8 urbanoverp
urban vcurve urbanhcurve
urbanlftshw urbanrighshw urbanlnadtpl,  dispersion(mean)

. mtreatnb total urbanoverp urbanvcurve urbanhcurve urbanlftshw urbanrighshw urbanlnadtpl,
mtreatment (stype urbanoverp urbanvcurve urbanhcurve urbanlftshw urbanrighshw urbanlnadtpl)
simulationdraws(100)
```

3. One mile segmentation

```
. insheet using "E:\SequenceOfOneMileUrbanruralADTPLWithoutTunnelandBoth.csv"
(31 vars, 13824 obs)

. nbreg total sequence2 sequence3 sequence4 sequence5 sequence6 sequence7 sequence8 urbanoverp
```

```
urban vcurve urbanhcurve
urbanlftshw urbanrighshw urbanlnadtpl,  dispersion(mean)

. mtreatnb total urbanoverp urbanvcurve urbanhcurve urbanlftshw urbanrighshw urbanlnadtpl,
mtreatment (stype urbanoverp urbanvcurve urbanhcurve urbanlftshw urbanrighshw urbanlnadtpl)
simulationdraws(100)
```

4. Accident-cluster segmentation

```
. insheet using "E:\DataSet_For_NB_With_Obs_Model.csv"
(16 vars, 218637 obs)

. nbreg total s2 s3 s4 s5 s6 s7 s8 urbanoverp urbanvcurve urbanhcurve urbanlftshw urbanrshw
urbanlnadtpl , dispersion(mean)
```

MULTINOMIAL LOGIT ESTIMATION AND PRE-PROCESSING

1. Multinomial Logit Estimation

```
. insheet using "E:\DataSet_For_MNL_Model.csv"
(8 vars, 218637 obs)

. mlogit stype urbanoverp urbanvcurve urbanhcurve urbanlftshw urbanrshw urbanlnadtpl
```

2. Get Random Mean and Standard Deviation

```
. ereturn list

. matrix list e(b)

. matrix list e(V)
```

3. Regression Estimation by Predicted Probabilities

```
. regress obchoice rchoice1 rchoice2 rchoice3 rchoice4 rchoice5 rchoice6 rchoice7 rchoice8 rchoice9
rchoice10 rchoice11 rchoice12 rchoice13 rchoice14 rchoice15 rchoice16 rchoice17 rchoice18 rchoice19
rchoice20 … rchoice191 rchoice192 rchoice193 rchoice194 rchoice195 rchoice196 rchoice197
rchoice198 rchoice199 rchoice200
```

4. Negative Binomial Estimation with Random Sequence Probabilities

```
. set memory 570m
(583680k)

. insheet using "E:\DataSet_For_NB_Model.csv"
(24 vars, 218637 obs)

. nbreg total pr1 pr2 pr3 pr4 pr5 pr6 pr7  urbanoverp urbanvcurve urbanhcurve urbanlftshw urbanrshw
urbanlnadtpl , dispersion(mean)
```

PREDICTION AND ELASTICITY

```
. predict mean, xb

. predict stderr, stdp

. summarize

. mfx, eyex

. mfx, eydx
```

VITA

Jun Seok Oh

EDUCATION

Dec 2010 **Doctor of Philosophy, Civil Engineering, The Pennsylvania State University**
Pennsylvania, U.S.A *Major : Transportation Systems*

Dec 2006 **Master of Science, Civil Engineering, The Pennsylvania State University**
Pennsylvania, U.S.A *Major : Transportation Systems*

Feb 2004 **Master of Science, Computer Science, Chungbuk National University,**
Chungbuk, Republic of Korea *Major : Database Systems*

Feb 2002 **Bachelor of Engineering, Information and Computer Engineering,**
Hansung University, Seoul, Republic of Kore *Major : Information Engineering*

EXPERIENCE

2010-2010 **Schreyer Institute for Teaching Excellence,**
2007-2008 **The Pennsylvania State University, Pennsylvania, U.S.A**
Educational Researcher and Statistical Analyst and System Developer

2005-2010 **Pennsylvania Transportation Institute,**
Transportation Econometrics Application Laboratory,
The Pennsylvania State University, Pennsylvania, U.S.A
Assistant Researcher

2001-2003 **Department of Computer Science, Database Laoratory**
Chungbuk National University, Chungbuk, Republic of Korea
Assistant Instructor, Assistant Researcher

2001-2001 **N-guru Company, Seoul, Republic of Korea**
Computer Programmer

1999-2001 **GIS/ITS Institute, Hansung University, Seoul, Republic of Korea**
Assistant Researcher

SELECTED PUBLICATIONS

- Oh J. S., Ahn Y. E., Jang S. Y., Lee B. G., Ryu K. H., “Design of Vehicle Location Tracking System using Mobile Interface”, The Korea Information Processing Society Journal, Vol. 9-D, No 6 , pp 1071~1081, 2002
- Oh J. S., Jang S. Y., Ahn Y. E., Ryu K. H., “Real-time Vehicle Position Monitoring System using PDA”, Proceedings of the 18th Korean Information Processing Society Fall Conference, Vol. 9, No 2, pp 1685~1688, 2002