REGRESSION TO THE MEAN CORRECTION FOR COLLISION MODIFICATION FACTORS

by

Bernard James

Bachelor of Engineering, Ryerson University, Toronto, 2008

A thesis

presented to Ryerson University

in partial fulfillment of the

requirements for the degree of

Master of Applied Science

in the Program of

Civil Engineering

Toronto, Ontario, Canada, 2010

DECLARATION

I hereby declare that I am the sole author of this thesis or dissertation.

I authorize Ryerson University to lend this thesis or dissertation to other institutions or individuals for the purpose of scholarly research.

I further authorize Ryerson University to reproduce this thesis or dissertation by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

[redacted]

REGRESSION TO THE MEAN CORRECTION FOR COLLISION MODIFICATION FACTORS

By

Bernard James

Master of Applied Science in Civil Engineering, 2010

Department of Civil Engineering, Ryerson University

ABSTRACT

Collision Modification Factors (CMFs) are a simple method of representing the effectiveness of road safety treatments. With the release of the Highway Safety Manual (HSM) and the recent launching of a CMF Clearinghouse website, CMFs are likely to become more widely used for estimating the effects of potential road safety treatments. The presence of regression to the mean (RTM) bias has long been shown to affect the accuracy of CMFs that did not account for the RTM in their development. The purpose of this research was to study how the RTM depends on the number of years of data used for selecting high collision sites for treatment and on the relative number of sites selected. From this analysis, a function based on the number of years, percentage of high collision sites selected, and the mean and standard deviation of the site population from which the treated sites are drawn was developed to more accurately estimate the magnitude of the RTM effect. This function can be used to adjust CMFs that do not account for RTM, complementing the procedure developed and used to correct CMFs included in the HSM.

ACKNOWLEDGEMENTS

I would like to thank my supervisor, Dr. Bhagwant Persaud for his guidance and support. I would also like to thank the other members of the examining committee, Dr. Medhat Shehata and Dr. Jinyuan Liu for taking the time to review the thesis, as well as Dr. Darko Joksimovic for chairing the examination.

In addition, I would like to thank Transport Canada for partial funding of my research under the project, "Adoption of Highway Safety Manual Methodologies for Safety Assessment of Canadian Roads" for which the project manager was Deborah deGrasse. I would also like to thank the Natural Sciences and Engineering Research Council of Canada for support from a discovery grant to Dr. Persaud.

I am also grateful to the Turner-Fairbank Highway Research Center for providing me with the collision data from their Highway Safety Information System (HSIS) that was used for my research. I would also like to thank Craig Lyon and Yongsheng Chen for helping me with their thoughts and comments for my research as well as providing me with the data necessary to do so. Finally, I would like to thank my wife, Trista James, for her ongoing love and support while I worked on this thesis.

DEDICATION

I would like to dedicate this thesis to my parents, for their continued support over the years which has made it possible for me to attain this major milestone.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

LIST OF ABBREVIATIONS

AADT	Annual Average Daily Traffic
AMF	Accident Modification Factor
EB	Empirical Bayes
EB-MoM	Empirical Bayes Method of Moments
EB-SPF	Empirical Bayes Safety Performance Function Method
CMF	Collision Modification Factor or Crash Modification Factor
CRF	Crash Reduction Factor
HSM	Highway Safety Manual
RTM	Regression to the Mean
SPF	Safety Performance Function

1. INTRODUCTION

Collision (or Crash) Modification Factors (CMFs) are a simple method of representing the effectiveness of road safety treatments. As road agencies work towards making their roadways safer for their users, CMFs are becoming more important in assisting with the decision making for the specific design of all the various road features. These road features can encompass anything to do with road design that can potentially affect the number of collisions/accidents/crashes that occur on that section of highway. ("Accidents", "crashes", and "collisions" are terms used interchangeable in the literature, and in this thesis.) They can include features from cross section elements or intersection design and can range from small changes, such as additional signage or lighting, to larger changes such as grade changes or conversion of intersections to roundabouts. With the release of the Highway Safety Manual (HSM) in 2010 (AASHTO, 2010) and the launching of the CMF Clearinghouse website in 2009, both of which include databases of relevant CMFs, a major focus is being placed on the use of these CMFs and the accuracy of the predictions. The accuracy of these predictions depends largely on the methodology by which they were developed. To ensure accurate results, the development of these CMFs needs to take into consideration sources of error such as regression to the mean. Regression to the mean (RTM) is the phenomenon where the number of collisions at a location fluctuates from year to year, but ultimately returns to a long term average. This fluctuation is caused by the random nature by which collisions occur. Before-and-after studies that do not take this into account, would overestimate the safety effect of a treatment due to the natural reduction in the collisions that would occur in the after period even if the treatment was not applied.

1.1. Background Information

The CMF estimates the new number of collisions to be expected after implementing the safety treatment by multiplying the CMF by the number of collision that would have occurred without the treatment. Collision Modification Factors (CMFs), as they are referred to in this thesis, are also termed Accident Modification Factors (AMFs) or Crash Modification Factors (CMFs), all of which function in exactly the same way. With a similar functionality, many studies also refer to Crash Reduction Factors (CRFs), which represents the safety effect as a percentage reduction in the expected number of collisions. The advantage with representing the factor as a modification factor rather than as a reduction factor is that this allows the modification to be clearly indicated as either an increase or decrease in the number of collisions. (A CMF greater that 1 indicates an increase, while a CMF less than 1 indicates a decrease, and unlike the case of crash reduction factors, the sign is always positive).

An example of a CMF listed in the Highway Safety Manual's "Knowledge" document, conversion of stop controlled intersections in rural areas (with Annual Average Daily Traffic (AADT) volumes of 7185 to 17220) to single lane roundabouts has a CMF of 0.42 (NCHRP 17-27 Project Team, iTrans Consulting Inc., 2009) based on research by Persaud et al. (2001). Put simply, this would imply that if an intersection has the characteristics identified by the CMF were to be converted to a single lane roundabout, and if it is estimated to experience 10.0 collisions per year without conversion, then it would be estimated to have 10 x 0.42 = 4.2 collisions per year after the conversion.

1.1.1. Regression to the Mean

The regression to the mean phenomenon or bias is a problem that affects most before and after studies. It is sometimes referred to as selection bias because a site is typically selected for treatment based on having an abnormally high accident count (Hauer, 1997). Regression to the mean is simply explained as a statistical phenomenon whereby the number of accidents at a particular site fluctuates up or down around a long term average (Bahar, 2009). This phenomenon can be shown by the example in Figure 1-1 where the number of accidents moves up or down about the long term mean (Shen & Gan, 2003). From this example we can see that if the site is treated based on one of the high points, very probably there will have been an immediate reduction in the number of collisions in the following year regardless of whether or not a safety treatment was implemented. Therefore, a before and after study conducted without taking into consideration regression to the mean will produce an exaggerated safety treatment effect. Even if a site with high accident counts is selected not because it has a high accident count but through some other selection process, while one may assume there is no longer selection bias, such a site may still become subject to the regression to the mean phenomenon, distorting the safety effect estimates in a simple before-after evaluation study (Hauer, 1997).

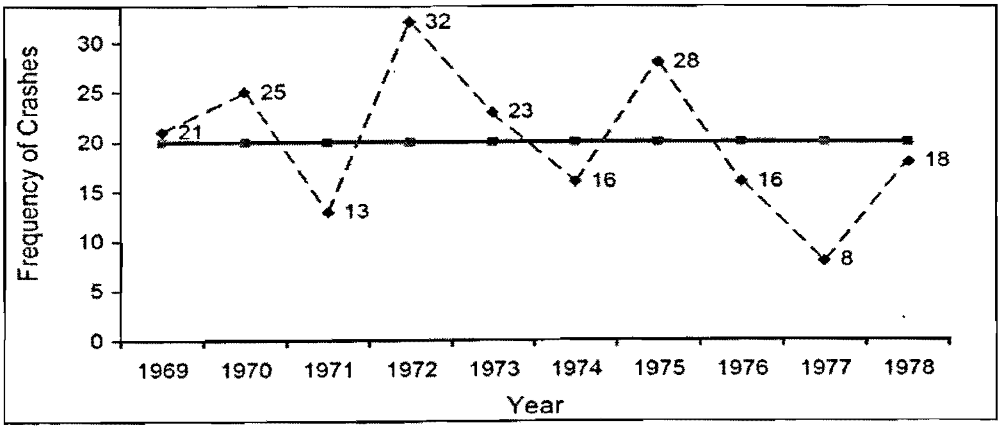

Figure 1-1: Regression to the Mean Example (Shen & Gan, 2003)

1.1.2. Development of CMFs

Collision Modification Factors can be developed through several different methods that can be used to evaluate the effects of a safety treatment. The popular methods are typically applied either through some form of a before-and-after study or a cross-sectional study (Forbes, 2003). The before-and-after study methods involve comparing the number of collisions expected to occur without the implementation of the treatment to the number of collisions that actually occur after implementation. The methodologies for developing the CMFs using before-and-after methods are documented by Shen and Glen (2003) and these various methods are listed below. The cross-sectional studies on the other hand do not require the treatment to be installed at the observation sites, but instead compare sites that have the treatment-related feature with those that do not have the feature being studied. The methodology for developing CMFs using cross-sectional data is documented by Bonneson and Pratt (2008). The most common methods for developing CMFs are:

1. **The simple (or naïve) before-and-after study method**: The simple before-and-after study, often referred to as the naïve before-and-after study, is a simple comparison between the number of accidents in the before period against the number of accidents in the after period. The CMF is calculated by dividing the total number of after crashes by the total number of before crashes for periods of equal duration. It is therefore considered a naïve method because it assumes that the number of crashes before a treatment is a good estimate of the number expected to occur without the treatment and does not take into account any other factors that can affect this estimate such as changes in traffic volume and external causal factors. These external factors could include weather conditions, economic conditions, changes in traffic policies

and other similar changes that could cause changes in the total number of collisions. More importantly, when this method is used, sites are treated based on having a high accident record which introduces a large regression to the mean error where, without any treatment, the total number of collisions would have naturally declined in the after period (Shen & Gan, 2003).

2. **The before-and-after study with comparison group method**: The before-and-after study with comparison group method is similar to that of the simple before-and-after study but it goes further by attempting to compensate for the external causal factors by using a comparison group of untreated sites. It assumes that any changes in accident patterns that would have occurred in the comparison group would have also occurred in the treatment sites if no treatments were implemented. The CMF from a simple before-study is simply adjusted by multiplying it by the ratio of before to after collisions in the comparison group. While this does account for external factors that could affect collision patterns, it still does not account for regression to the mean (Shen & Gan, 2003).

3. **The before-and-after study with the Empirical Bayes (EB) method**: The Empirical Bayes method for before-and-after studies goes further by introducing the use of a model estimate for the mean crash frequency of similar sites as well as using the crash record of the site. The mean crash frequency of similar sites is usually estimated from a Safety Performance Function (SPF) that is a model estimate of the expected number of crashes at an untreated "reference" site based on the AADT, and sometimes other characteristics of the site. By using this combined method with mathematical techniques the effects experienced from unrelated factors and

regression to the mean are minimized in order to determine the true estimate of crashes expected without the treatment, and ultimately, a true safety effect of the treatment (Shen & Gan, 2003).

4. **Cross-sectional study**: Cross-sectional studies estimate the safety effect of a feature by comparing crashes at sites with that feature to crashes at sites that are similar on all counts with the exception of that particular feature. Thus, it may be inferred that the difference in crashes would represent the reduction in collisions due to the feature and therefore can be used to estimate a collision modification factor. While it is recognized that before-and-after studies are better at estimating the safety effects of a treatment, cross-sectional studies are sometimes employed when it is not possible to do a before-and-after study. Also, it is important to note that one of the difficulties with the cross-sectional study is finding sites that have exactly the same features with the exception of the one feature being studied (Bonneson & Pratt, 2008).

1.1.3. Reliability of CMFs

Given the large number of studies being conducted worldwide to develop CMFs for various safety treatments, there is a pressing need to assess the reliability of the CMFs produced from these studies. Based on the method by which the study is conducted, one is able to determine whether the study accounted for all the potential sources of error. For a CMF to be considered reliable it must be both precise and accurate. Precision and accuracy are illustrated by the bull's eye target in Figure 1-2. As indicated, if the results from a safety treatment cluster at the same off bull's eye value they would be considered precise but not accurate, while if they scatter around the target then they are considered to be neither precise nor accurate (NCHRP 17-

27 Project Team, iTrans, 2007). It is therefore necessary for the CMF to produce consistent results on target and as predicted for the CMF to be considered reliable and therefore safe for road agencies to use. This precision and accuracy can be measured by the standard error of the CMF estimate, which is the methodology used for rating CMFs in the Highway Safety Manual (HSM).

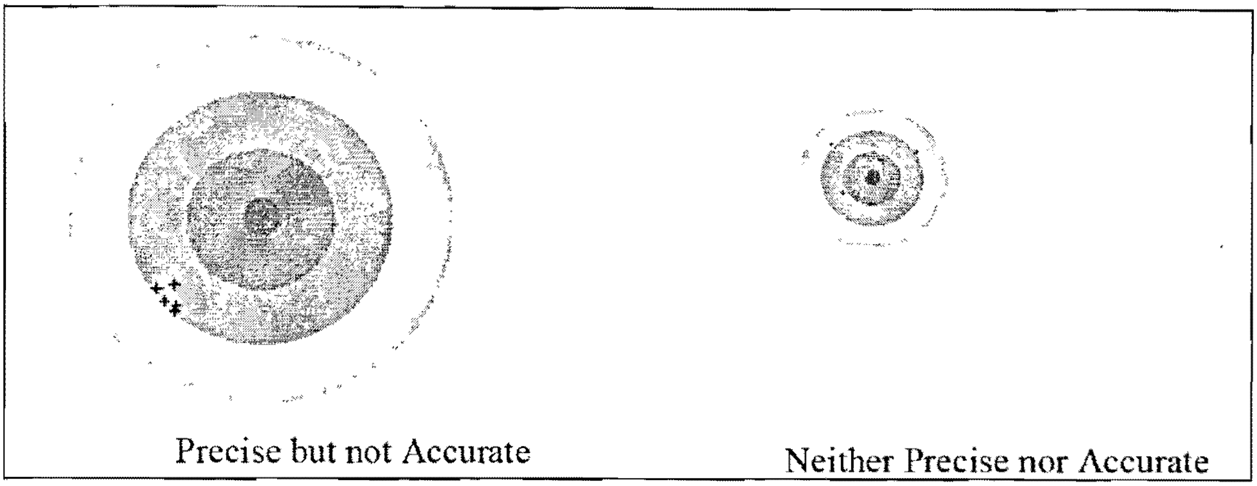

Figure 1-2: Illustration of precision and accuracy (NCHRP 17-27 Project Team, iTrans, 2007)

The Crash Modification Factor Clearinghouse (http://www.cmfclearinghouse.org), which is a web-based collection of CMFs that will be constantly updated with CMFs from new studies, has also identified the need to rate the reliability of the CMFs. To do this, a 5-star quality rating system has been developed to indicate the quality or confidence in the results of the study that have been determined by the review committee (University of North Carolina Highway Safety Research Center, 2010). The rating system is based on Study Design, Sample Size, Standard Error, Potential Bias, and Data Source of the study. Points are given for those categories of either 0, 1 or 2 and based on double weight for study design and sample size. A maximum number of 14 points can be achieved, which would result in a 5-star rating.

1.2. Problem Statement

While there are now significant resources available for accessing existing CMFs, the origins of these CMFs usually come from individual studies and assessments that are derived using varying methodologies. Due to the limited resources of many of the road agencies or individuals who conduct these studies, it is often not possible for them to be conducted in such a comprehensive manner as to account for all sources of error, specifically, regression to the mean. Regression to the mean, as noted earlier, occurs as a result of the number of collisions at a given location fluctuating up and down each year around a long term average. This average can be defined as the normal number of collisions at a location that can theoretically be determined from the average of data collected over an adequate number of years. If insufficient years of data are used to develop the CMF, it is possible that the resulting safety treatment effect determined is simply a result of the fluctuation in the yearly number of collisions rather than a real reduction in the number of collisions. Regression to the mean errors are also associated with selecting a small proportion of the highest collision sites from a population for treatment, as these few abnormally high crash sites will have much lower means in any other period. The main concern as a result of the regression to the mean phenomenon is based on using an incorrect measure of the true mean for the number of crashes in the before period and using that incorrect mean as the estimate of crashes in the after period. However, many CMFs are often published based on too few years of data, and without using the empirical Bayes method to correct for regression to the mean bias, which produces inflated results for the CMFs. This immediately raises the issue of the reliability of the effectiveness of the published CMFs. Road agencies want to be sure that the treatments they decide to use will indeed yield the results expected on the basis of the CMFs, so as not to waste money on treatments that would not work.

1.3. Objective

Qualitative methods have been developed to adjust for regression to the mean bias in CMFs from published studies that ignore this bias; however, there is no quantitative method that verifies those processes. It is to be expected that the fewer years of data that have been used, the greater the regression to the mean bias would be. It is also to be expected that if smaller proportions of the high accident sites are used, the regression to the mean bias would be greater. The purpose of this research is to evaluate the methods for estimating the expected number of collisions at a site without treatment, and correspondingly, for correcting for regression to the mean. The research will also empirically explore how regression to the mean depends on the number of years of observed data and on the proportion of high accident sites selected. This will help to determine whether it is possible to develop a specific process to correct for regression to the mean in published collision modification factors that are suspected of having regression to the mean bias.

1.4. Thesis Outline

This thesis is divided into 8 main chapters as follows:

- **Chapter 1 - Introduction**: This chapter introduces Collision Modification Factors (CMFs) and Regression to the Mean (RTM) and outlines the objective of the thesis.
- **Chapter 2 - Literature Review**: This chapter reviews material related to the development of Collision Modification Factors, the effects of regression to the mean and the procedures developed to account for regression to the mean.

- **Chapter 3 - Analysis Data**: This chapter identifies the datasets that were used for the analysis, which include the real data as well as the methodology for creation of the simulated data.

- **Chapter 4 – Comparison of Methods for Estimating Expected Collisions**: This chapter identifies the various methods for estimating the expected number of accidents in the presence of regression to the mean, and compares the results of each method.

- **Chapter 5 - Empirical Estimation of Regression to the Mean Effect**: This chapter addresses the methodology for the selected method of estimating the regression to the mean using the datasets and gives the results of the analysis.

- **Chapter 6 – Collision Modification Factor Corrections for Regression to the Mean:** This chapter applies the methods used for estimating the number of expected accidents to devise a formula for correcting collision modification factors that are suspected of having regression to the mean bias.

- **Chapter 7 - Summary, Conclusions and Recommendations**: This chapter summarizes the results from the research and provides the conclusions and recommendations for the methodology used for the estimation of regression to the mean for correction of collision modification factors.

- **Chapter 8 - Further Study**: This chapter summarizes the limitations of the research completed in order to identify areas for future study to build upon the findings, methodology and results of this research.

2. LITERATURE REVIEW

A literature review was conducted on the existing methods for using and developing collision modification and for correcting for regression to the mean. Online journals, text books and project reports were reviewed for this purpose. These sources have been grouped into several main categories and are summarized in this section.

2.1. Collision Modification Factors

Lord and Bonnenson (2006) analyzed the role and application of accident modification factors within highway design process, specifically due to the forthcoming release of the Highway Safety Manual and the expected increase in the use of accident modification factors. It was identified by the authors that many road agencies still use crash reduction factors (CRFs) instead of accident/collision modification factors (AMFs or CMFs), which is very limiting, as this excludes the instances where there is an increase in collisions, and which is why all new studies use AMFs or CMFs, which can reflect both increases and decreases in collisions. The relationship between the CRF and AMF or CMF is given in Equation 2-1.

$$\text{AMF} = 1 - \text{CRF} \qquad \textbf{Equation 2-1}$$

The authors then explained how these AMFs could not only be used for countermeasures to treat existing road segments, but also how they can be used to evaluate various design alternatives where the existing collision history does not yet exist. This is accomplished through the use of safety performance functions that would give an expected number of collisions such that the design alternatives can be compared. Furthermore, it is also possible to combine the AMFs from several design alternatives in order to evaluate more complex alternatives (Lord & Bonneson, 2006).

Belluz and Forbes (2003) in a paper on the Synthesis of Safety for Traffic Operations assessed the various methods of measuring road safety, which include Motor Vehicle Crashes (MVCs), Safety Performance Functions (SPFs) and Collision Modification Factors (CMFs). The Synthesis of Safety for Traffic Operations included CMFs for transportation practitioners in Canada to use, as theses were identified as being an important tool. It was noted that due to the small number of studies done in Canada, there is an information gap which makes it difficult for practitioners in Canada to practice Evidence Based Road Safety (EBRS) which is the preferred method for road safety treatment. However, in the absence of this, it was stated by the authors that reliable CMFs developed in other jurisdictions can be used as long as they are applied carefully. In concluding, the authors recommended that a uniform process be used for reporting safety effects and suggested that, with training and additional research, it would be possible to overcome this information gap (Belluz & Forbes, 2003).

2.2. Effect of Regression to the Mean

Hauer (1998) analyzes how bias by selection often results in an over-estimation of effectiveness. It is explained that the effectiveness of a countermeasure is often derived from the comparison of accidents before and after implementation of the treatment. Using a numerical example, it is demonstrated that there is a reduction in accidents in the after period simply due regression to the mean. The author explains that using a Poisson probability distribution an estimate for the regression to the mean is determined to allow for elimination of this bias (Hauer, 1980).

Elvik (2004) discusses the extent of bias in the selection of sites for road safety treatment in Norway. It is noted that the site selection process in Norway is a complex one that takes into account many factors in addition to the accident record. For the sites that were treated it was found that the percentage of those sites that had a higher than normal accident rate was the same as the percentage of those sites that had a lower than normal accident rate. This suggests that there exists very little bias in the selection of sites for treatment in Norway. This would ultimately remove any regression to the mean errors for the before-and-after studies that would result from selecting sites purely based on bad accident records. However, the whole purpose of treating sites is to improve safety by reducing the total number of collisions. One would question the relevance of treating so many sites that have lower than normal accident rates simply for the purpose of attaining greater statistical accuracy where bias is considered to be a bad thing. With the development of the Empirical Bayes (EB) approach, which can be used to correct for regression to the mean, the question arises as to whether we should avoid bias in site selection or simply ensure that we can account for it accurately. The author then concludes by noting that Norway could make their selection for road safety treatment more effective by selecting fewer safer than normal sites. However, based on the data available, the paper was unable to quantify the extent to which it can be made more effective (Elvik, 2004).

Maher and Mountain (2009) address the sensitivity of estimates of regression to the mean. It is noted that methods of accounting for regression to the mean (RTM) require some type of assumption regarding the distribution of the true mean. The EB method assumes a gamma distribution, as this works well for the mathematics of the Bayes Theorem. However, the authors noted that with the advances in computational techniques for Bayes Theorem, using Markov Chain Monte Carlo methods, it is possible to use other distributions. It was concluded that it is possible to get good RTM estimates using various distributions. While the RTM estimates varied by up to 20% based on the different distributions it was noted that this variation becomes significantly smaller with better predictive models that are used in the EB method. Based on the results of the analysis, no firm conclusion could be reached as to which distribution is best for the EB method nor as to if there is evidence to show that any of them would always be better than the traditionally used gamma distribution. The authors suggested that the distributions work better on a case by case basis, and different trials should be done to determine the best fit. It is important to note that this research supports using the EB method as being the preferred method for estimation of RTM regardless of the distribution used (Maher & Mountain, 2009).

2.3. Methods to Account for Regression to the Mean in Before-After Studies

Abbess et al. (1981) estimate the effectiveness of remedial treatment with special reference to the regression to the mean effect. In this paper it is explained that the Bayesian approach can be used to analyze blackspot data for collisions and to determine the effectiveness of the treatment. This was identified as necessary, given that other methods of determining the effectiveness of the treatment often ignore the main source of the problem and the methodology for the Bayesian approach automatically accounts for over-estimation of the treatment effect due to regression to the mean. At the time the paper was published, the author noted that methods and data were not yet available to assess the importance of regression to the mean. Graphs and data are presented to show the presence of regression to the mean with the collision data and to demonstrate a good fit to the accident mean of the gamma distribution whose parameters can then be estimated. As such, the paper identifies a formula to estimate the regression to the mean. In conclusion, the authors stated that Bayesian methods can successfully be used to analyze accident blackspots by properly estimating the expected number of accidents in the presence of regression to the mean (Abbess, Jarrett, & Wright, 1981).

Hauer (1986) addresses the estimation of the expected number of accidents. It is noted that when a certain number of accidents are recorded in a given period, it does not necessarily mean that this will be the average number of accidents in the following period. Therefore, safety estimated based solely on the 2 periods will be inaccurate. To account for this phenomenon, better estimates of the expected number of accidents are required. Using actual accident counts, the author shows how this count gives a very poor estimate of the average number of counts per location, given that the numerical differences between the observed counts and the actual mean are significant. It is then shown how the estimation of the expected number of accidents can be improved using the Bayesian approach. A simplified form of the equation given to estimate the expected number of accidents is represented by Equation 2-2.

$$T = x + \frac{E\{x\}(E\{x\}-x)}{Var\{x\}} \qquad \textbf{Equation 2-2}$$

Where T is the expected number of accidents at a site, based on xx, the observed number of accidents, $E\{x\}$ the overall mean for similar sites, and $Var\{x\}$, the variance of observed accidents across these similar sites. Using this method, the author showed that the estimated expected number of collisions was close to the number observed in a second period for sites with high crashes in the first period. Based on this, and other factors discussed by the author, it is concluded that this approach would give a good estimate of the number of collisions expected without treatment in a before-after study (Hauer, 1986).

Wright, Abbess & Jarrett (1988) in their paper on estimating the regression to the mean effect associated with road accident blackspot treatment, suggest that using a simple before-after comparison of accidents at blackspots as a method for identifying treatment effects is not a practical one as the data are distorted by the regression to the mean effect. In the paper, the various methods for correcting for regression to the mean are assessed to determine the validity of the assumptions. It is further argued that the gamma distribution used in the EB method does not seem to be affected by the varying assumptions for distributions of different collision types. In conclusion, it is suggested that to improve the results of the assessments it is important to develop a good definition for the population of similar sites to ensure very similar characteristics in the sites such that the mean collision over time is stable. It was suggested by the authors that the accuracy of these estimates will be further improved by accumulating data over a longer period of time (Wright, Abbess, & Jarrett, 1988).

Hauer et al. (2002) estimate safety using the Empirical Bayes method. It is argued that the Empirical Bayes method increases the precision of a safety estimate when only 2 or 3 years of collision history are available and corrects for regression to the mean bias. It is noted that even though the Empirical Bayes method has been widely recognized for some time, papers are still being published based on naïve before-after studies that do not account for regression to the mean. In conclusion, the authors stress that though the EB method may seem a complex process, it really is not so it can be easily incorporated into all before-and-after studies (Hauer, Harwood, & Council, 2002).

Persaud and Lyon (2007) document the lessons learned from two decades of experience using the Empirical Bayes before-and-after studies. It is suggested that this method, if properly applied, can produce results that accurately portray the effects of safety treatments which are significantly different and less biased than those completed through other types of studies. The whole purpose of the EB methodology was to account for the effects of regression to the mean bias that occurs when high short term accident counts trigger safety treatment for sites that will experience a reduction in accidents as the counts return to the true long term average of the site in the following years. It was noted by the authors, however, that there still exists much scepticism as to the need for the EB methodology if sufficient years of pre-treatment data are used to determine the true mean of a site before treatment. It was argued that while it is possible to determine the true mean through this method it is difficult to estimate how many years will be required to conclude that there is no regression to the mean in the estimate. Previous analysis had shown that even with 5 years of before data for 2-lane rural highways, it was still not possible to eliminate the regression to the mean bias. The paper then provides examples for calculating the number of accidents per year using the before and after comparison group method verses the Empirical Bayes method to show the large difference in the results. While it is demonstrated that the EB method can produce more accurate results the authors noted that it is important not to use it blindly in that there are problems that can affect the validity of the EB method if they are not accounted for. The first issue identified was the differential effects for different crash types given that treatments affect different crash types differently; to assess the overall effect it is necessary to determine the effect for each crash type and severity, and to weigh these effects accordingly. The second issue identified is the specification of the reference group necessary to calibrate safety performance functions for each of the before and after periods so as to properly account for regression to the mean and external conditions that change over time. The third issue relates to changes in traffic volumes not being accounted for properly. While it is argued that traffic volumes only increase by 2-4% per year and can therefore be ignored, it was shown that if the changes in traffic volumes are not accounted for, certain situations may actual show an apparent larger accident reduction than the true reduction. The paper concludes that current evidence shows that the EB method will produce better results that are more valid than traditional before-and-after studies if they are completed correctly. As such, it was argued, it is worth the effort to do the additional data collection and analysis rather than conducting a simple before-and-after study that would produce questionable results. The importance of properly apply the EB method by taking into consideration all the factors that could invalidate the results was noted. It is suggested by the authors, that as a further step, it should be determined whether improving the results from EB studies can be accomplished by additional research in the development and calibration of the safety performance functions to produce more sophisticated models that would better predict the number of expected collisions which is an integral part of the EB method. Further, the full Bayes approach is proposed for more complex safety performance functions that cannot be easily handled with the generalized linear modeling traditionally used. (Persaud & Lyon, 2007).

2.4. Correcting CMFs to Account for Regression to the Mean

Bahar (2009) authored a research circular on the methodologies for the development and inclusion of accident modification factors in the Highway Safety Manual (HSM). This paper identifies the very methodology used for the inclusion of Accident modification Factors (AMFs) (same as CMFs) in the Highway Safety Manual. As part of this, the various methods for developing AMFs are described in detail to explain how regression to the mean can affect the estimates provided by simple before-and-after studies. It is noted that there are methods such as the Empirical Bayes method for developing AMFs. The author explains that there are many past studies that did not use such methods in developing the AMF for the treatment studied. Moreover, there are cases where the EB method is not applied correctly and the AMF will still include regression to the mean (RTM) bias. As part of the inclusion process it is proposed by Bahar that a correction can be applied to AMFs that are suspected of having RTM bias so that the corrected AMF can be included in the HSM. This process involves the use of correction factors ranging from 0.05 for small RTM bias to 0.25 for large RTM bias. The formula for this procedure is shown below in Equation 2-3 (Bahar, 2009):

$$\text{AMFbiased} - \text{AMFunbiased} = \frac{A}{B} - \frac{A}{(B-X)}$$ **Equation 2-3**

Where:

$A = After\ Crash\ Frequency$

$B = Before\ Crash\ Frequency$

$A/B = AMF\ biased$

$X = RTM\ bias\ assumed\ by\ the\ NCHRP\ 17-27\ research\ team$

Given that X is small compared to B the equation is simplified to Equation 2-4 (Bahar, 2009):

$$\text{AMFunbiased} = \text{AMFbiased} \times (1 + X/B)$$ **Equation 2-4**

The X/B ratio ranges between 0.05 for a small RTM bias and 0.25 for a large RTM bias. A large RTM bias of 0.25 would be assumed if a few years of data were used and a very small proportion of the highest accident sites was selected for treatment. A small RTM of 0.05 would be assumed if a large proportion of all the sites was treated and many years of data were included in the development of the AMF. The document also identifies methods for adjusting AMFs for traffic volume bias.

The HSM relies on the standard error of AMFs to estimate the reliability of the safety effect expected to be achieved by the AMF. As such, when the AMF is adjusted to account for regression to the mean effect, it is also necessary to calculate the adjusted standard error as well. The author also identifies a method for adjusting the standard error is identified as well. A small standard error would mean that the AMF is very reliable. For AMFs to be included in the HSM, they must pass a rigorous inclusion/exclusion process. The AMFs were filtered for results that have a maximum standard error of 0.1 (Bahar, 2009).

2.5. Summary

Researchers have identified the existence of the regression to the mean phenomenon and established that it can significantly impact the accuracy of simple before and after studies for determining the effect of safety treatments. Based on this knowledge, much research has been placed in the development and improvement of methods for taking into account the effect of regression to the mean for safety treatment studies. From the review of these papers, it has been shown that the Empirical Bayes method has been proven to be the most effective method to date for accounting for regression to the mean. However, it has also been identified that although the Empirical Bayes method has been widely accepted as the preferred method to do this, studies are still being published with results of road safety treatment studies that do not account for regression to the mean. As such, in compiling a database of reliable Collision Modification Factors (CMFs) for the Highway Safety Manual (HSM), researchers have developed a qualitative method for adjusting CMFs that are suspected of having regression to the mean bias. However, a method does not seem to exist for determining the extent of the regression to mean error based on the number of years and proportion of sites selected for treatment in the study.

3. ANALYSIS DATA

For the purpose of this research, collision data were required for the empirical analysis. It was decided that real collision data would be used for the initial analysis. Once the results and observations have been determined using the real data, the procedure would then be generalized using simulated data.

3.1. Collision Data

The collision data for California intersections from the year 2000 to the year 2007 were used as the test data. This encompasses all types of intersections which include signalized, stop controlled, 3 legged and 4 legged, all with various numbers of lane approaches and turning lane configurations. The dataset was also categorized with the following fields as identified in the Guidebook for the California State Data Files (Council & Mohamedshah, 2007):

- **Mainline AADT** – Major Annual Average Daily Traffic
- **Cross Street AADT** – Minor Annual Average Daily Traffic
- **Highway Group** – Right independent alignment, left independent alignment, divided, undivided, or other
- **Traffic Control Type** – Stop signs on cross street, main street or both; signals, pre-timed, semi actuated or fully actuated
- **Intersection Type** – Tee, wye (Y), four legged, more than four legged or other
- **Mainline Number of Lanes** – 2, 3, 4, 5, or 6
- **Cross Street Number of Lanes** – 2 or 4

- **Mainline Traffic Flow** – One way or 2 way with left turns permitted or not

- **Mainline Left & Right Turn Channelization** – Curbed, painted, raised bars or no channelization

- **Cross Street Traffic Flow** – One way or 2 way with left turns permitted or not

The data were filtered to extract a similar intersection type for the analysis. The selected intersection type was 4 legged intersections on undivided highways that had stop control on the minor approach with characteristics shown in Table 3-1 and Figure 3-1.

Table 3-1: Characteristics of sample data

Characteristic:	Major Street	Minor Street
Stop control	None	Stop
Number of lanes	4	2
Left turn channelization	None	None
Left turn permitted	Yes	Yes
Direction of travel	2 Way Street	2 Way Street
Right turn channelization	None	None

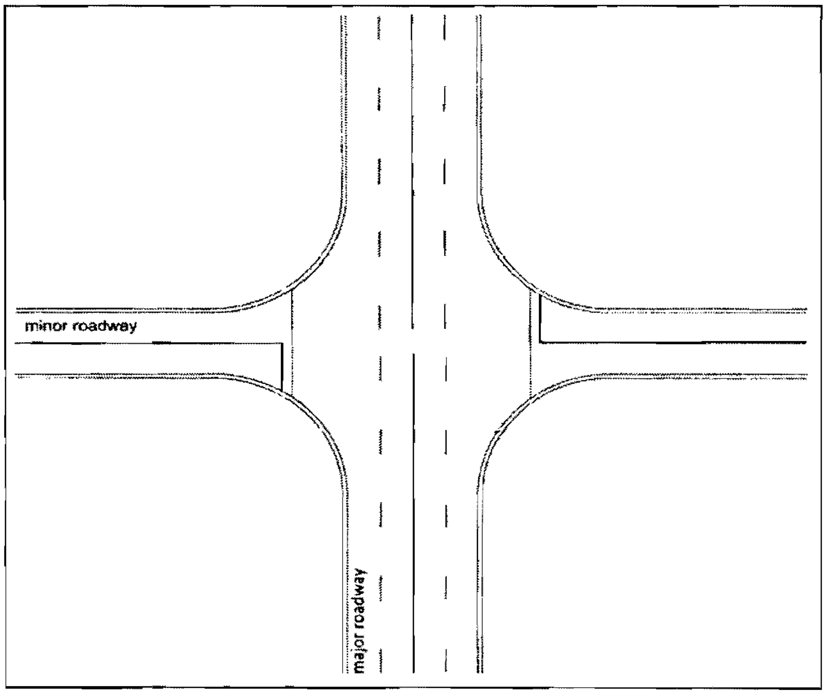

Figure 3-1: Four legged intersection with stop control on the minor approach

The result of this site selection produced a total of 204 sites with collisions statistics shown in Table 3-2. The 8 year collision history for these sites is included in Appendix A.

Table 3-2: Collision Statistics for sample data

Number of sites = 204			
Variable	**Mean**	**Minimum**	**Maximum**
Years	8	8	8
Crashes/site-year	1.70	0	14
Major Road AADT	16,750	2,350	51,750
Minor Road AADT	990	100	9,400

3.2. Safety Performance Function for the Collision Data

The dataset for the California intersections already had Safety Performance Functions (SPFs) developed. For the analysis, it was decided to use the collision data for all types of collisions. As such, the relevant safety performance function for the estimated number of collisions for 4 legged intersections with 4 lanes on the major approach, is given in the form of Equation 3-1 (National Cooperative Highway Research Program, 2008).

$$E\{\kappa\} = \alpha\,(Major\ AADT)^{\beta_1}(Minor\ AADT)^{\beta_2} \qquad \textbf{Equation 3-1}$$

Where:

$\alpha = 6.44E^{-5}$ $\qquad \beta_1 = 0.7693 \qquad \beta_2 = 0.4262$

Resulting in the safety performance function identified in Equation 3-2.

$$E\{\kappa\} = 6.44E^{-5}\,(Major\ AADT)^{0.7693}(Minor\ AADT)^{0.4262} \qquad \textbf{Equation 3-2}$$

3.3. Methodology for Data Simulation

The simulated dataset is defined by a fixed mean (μ) and standard deviation (σ) of a Gamma distribution. Using the acceptance-rejection technique, the fixed variables were used to generate values for the simulated dataset that correspond with that of the Gamma distribution. The acceptance-rejection technique is the method for ensuring that the random numbers generated for the defined dataset are within the parameters of the selected distribution type. It is called the acceptance-rejection technique because the generated number is accepted if it is within the parameters for the distribution type or rejected if it is not, in which case the process is repeated until it is accepted. This simulated set of means would then represent a mean number of collisions for each site in a hypothetical population. The mean for each site was then considered to be the long term average of the number of collisions at that site. To generate the integer values for the number of collisions for each year of that site, a Poisson distribution was used for the data simulation which would be produced based on the mean number of collisions for that site. This was done using the acceptance-rejection technique as well, to ensure each of the values generated corresponded to that of the Poisson distribution for that site. Using a different mean for each site, a Poisson distribution would be simulated to generate the individual occurrences of crashes occurring for each year of that site. The method used for generating the random variates for both the Gamma and Poisson distributions using the acceptance-rejection technique for generating simulated data is explained in the following sections (Banks, Carson, Nelson, & Nicol, 2005).

Gamma Distribution Data Simulation:

Gamma Distribution Function:

$$f(x) = \frac{\beta\theta}{\Gamma(\beta)} \cdot (\beta\theta)^{\beta-1} . e^{-\beta\theta x}$$

Gamma Distribution Constants:

$$\theta = \frac{1}{\mu} \qquad \beta = \frac{1}{\sigma^2\theta^2}$$

Gamma Distribution Data Simulation Process:

Step 1: $a = \frac{1}{(2\beta-1)^{1/2}}$ $b = \beta - \ln(4)$

Step 2: Generate Random numbers R_1 & R_2 and set: $V = \frac{R_1}{1-R_1}$

Step 3: $X = \beta V^a$

Step 4a: If X $>$ $[b + (\beta a + 1)\ln(V) - \ln(R_1{}^2R_2)]$ then reject X and repeat Step 2

Step 4b: If X $>$ $[b + (\beta a + 1)\ln(V) - \ln(R_1{}^2R_2)]$ then use X

Step 5: The mean for each site is: $X = X/(\beta\theta)$

(This process is repeated for the **number of sites** in the population of interest to generate the mean number of collisions for each of the sites.)

Poisson Distribution Data Simulation:

Poisson Distribution Function:

$$p(n) = \frac{e^{-\alpha}\alpha^{n}}{n!}$$

Poisson Distribution Constant:

$$\alpha = X\ (mean\ number\ of\ crashes\ per\ site\ defined\ by\ gamma\ distribution)$$

Poisson Distribution Data Simulation Process:

Step 1: Set N = 0 and P = 1

Step 2: R_1 = Random Number

$P = P * R_1$

Step 3a: if $P > e^{-\alpha}$ then Reject N and make N = N+1 and go back to step 2.

Step 3b: if $P < e^{-\alpha}$ then Accept number of collisions as N

(This process is repeated for the **number of years** of collision data required for each site, which is then repeated for the **number of sites** of data required.)

3.4. Simulated Data

It was decided that the simulated data should be composed of the same number of years as that of the real data. It was, therefore, decided to produce simulated data for 8 years of collisions for 100 sites. The simulated dataset was defined by a fixed mean (μ) of 4 collisions per year and standard deviation (σ) of 1.6 for the Gamma distribution which can be portrayed by the probability density function shown in Figure 3-2. The simulated data for the 8 years of data for 100 sites is included in Appendix B.

$$\text{Mean:} \quad \mu = k\theta \qquad\qquad \text{Variance:} \quad \sigma^2 = k\theta^2$$

Where:

$$\text{Shape factor:} \quad \theta = \frac{\sigma^2}{\mu} \qquad\qquad \theta = \frac{1.6^2}{4} = 0.64$$

$$\text{Scale factor:} \quad k = \frac{\mu}{\theta} \qquad\qquad k = \frac{4}{0.64} = 6.25$$

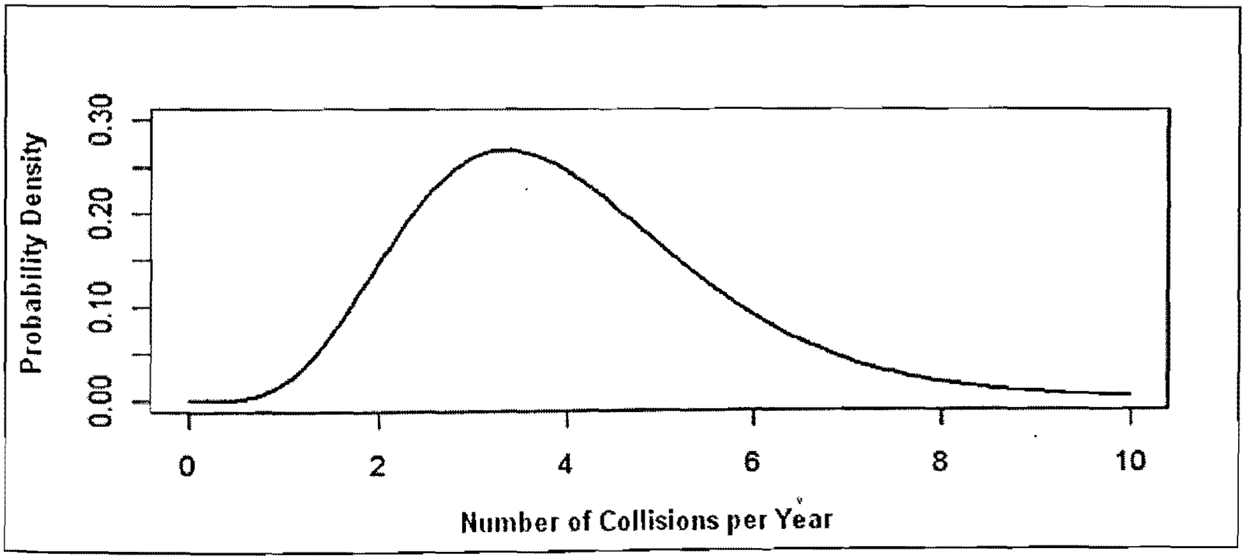

Figure 3-2: Probability Distribution Function for Gamma Distribution of simulated dataset

4. COMPARISON OF METHODS FOR ESTIMATING EXPECTED COLLISIONS

These expected numbers of collisions are calculated to determine the true average number of collisions at the target locations, rather than using the observed number of collisions, which as noted, could be randomly high or low. The following comparison of the results from the different methods will demonstrate these differences and provide some insights into which method is best.

4.1. Methods for Estimating Expected Number of Collisions

From the literature review, it has been established that there are several methods that can be used to estimate the expected number of collisions for a specific location. For the purpose of this comparison, the dataset for the California intersections with stop control on the minor road identified in section 3.1 was used for the comparison of the various methods. The main methods for estimating the expected number of collisions for comparison are:

1. The Naïve Method based on the observed (k) number of collisions
2. The Empirical Bayes Method of Moments Method (EB-MoM)
3. The Full Empirical Bayes Approach based on the Safety Performance Function (EB-SPF)

The mean of other years of data is also included in the comparison. It is not actually a method that can be used in a before-and-after study for estimating the expected number of collisions given that the site would be altered by the treatment in the after period. However, for the purpose of this analysis, it is used to determine a value for empirical testing of the other methods given that no treatment was applied, and it would therefore represent the value in the after period that the other methods are trying to predict. The methodology and variations of these methods used are explained in the following subsections.

4.1.1. The Observed Method Number of Collisions (Naïve Method)

For a given time period, based on the naïve approach, the observed (k) number of collisions during that period is assumed to be the normal number of collisions that will occur per year. Thus, regardless of whether the number of collisions for that year may seem to be abnormally high it is assumed to be the expected number of collisions that will occur in future years. As confirmed by the literature review, this assumption has been strongly challenged by many studies that demonstrate that this method is very prone to the regression to the mean bias. As such, this method is only included in the comparison as it is used as the starting point for the estimates for the other methods, and also as the baseline for the regression to the mean estimates. To determine these values, a year or group of years is selected as the target period, and the sites are then ranked from highest number of collisions to lowest based on the observed number of collisions in the target period. Based on this method it is assumed that for these sites, the expected number of collisions in the following years would stay the same for each site. Using the year 2000 data as the target year, the results of this method are shown in Table 4-1. These values will also be used as the starting point for each of the other methods of estimating the expected number of collisions, and will be compared against the other methods to determine the extent of the difference.

Table 4-1: Observed number of collisions

Year 2000: Top 10 Sites	Observed: (K)
1st	18
2nd	16
3rd	11
4th	11
5th	10
6th	9
7th	9
8th	8
9th	8
10th	8

4.1.2. Empirical Bayes Method of Moments (EB-MoM) Method

The Empirical Bayes Method of Moments (EB-MoM) method accounts for regression to the mean by using a comparison group to estimate the mean number of accidents and the variance observed from similar sites from a large population. For this purpose, the similar sites would be all the intersections with similar characteristics, regardless of the traffic volumes. The expected number of collisions for this method is calculated by Equation 4-1 (Hauer, 1997).

$$E_A = \frac{\bar{x}^2}{s^2} + \left(\frac{s^2 - \bar{x}}{s^2}\right) K \qquad \textbf{Equation 4-1}$$

Where:

K – Observed number of accidents in the analyzed site, in the selected time period;

E_A – Expected number of accidents in the analyzed site, in the selected time period;

$\bar{x}$ – Average value of observed accident frequencies on entities similar to the study site in parallel time periods;

s^2 – Variance of observed accident frequencies on entities similar to the study site in parallel time periods.

The value for the mean and variance of all observed accident frequencies for all 8 years collision data of the 204 sites in the dataset is found to be:

$$\bar{\text{x}} = 1.705 \text{ (per year)}$$

$$\text{s}^2 = 6.629$$

Using these values in Equation 4-1 and the year 2000 data as the target year as identified in Table 4-1, the results of the EB-MoM method are shown in Table 4-2.

Table 4-2: Expected number of collisions based on Empirical Bayes - Method of Moments method

Year 2000 - Top 10 Sites	(K)	EB-MoM
1st	18	13.81
2nd	16	12.32
3rd	11	8.61
4th	11	8.61
5th	10	7.87
6th	9	7.12
7th	9	7.12
8th	8	6.38
9th	8	6.38
10th	8	6.38

4.1.3. Empirical Bayes Method of Moments (EB-MoM) Adjusted Method

As a further refinement to the Empirical Bayes Method of Moments (EB-MoM) method, an adjusted method was used. The mean and variance required for the method was not only taken from similar sites, but also for sites with similar traffic volumes. The range of the AADTs for the target sites was determined, and the mean and variance was calculated from similar sites that have AADTs within the 85th percentile volume range of the target sites range.

For example, the top 10 sites based on Year 1, have Major AADTs ranging from 9,393 to 45,000 and Minor AADTs ranging from 501 to 2,010. The 85th Percentile volumes would therefore occur between 14,750 and 45,000 for the Major and 730 and 2,010 for the Minor. The resulting mean and variance calculated from these similar sites is given as:

$$\bar{x} = 3.298 \text{ (per year)}$$

$$s^2 = 12.656$$

Using these values in Equation 4-1 and the year 2000 data as the target year, the results of the EB-MoM Adjusted method are shown in Table 4-3.

Table 4-3: Expected number of collisions based on Empirical Bayes - Method of Moments Adjusted method

Year 2000 - Top 10 Sites	(K)	EB-MoM Adjusted
1st	18	14.15
2nd	16	12.68
3rd	11	9.01
4th	11	9.01
5th	10	8.27
6th	9	7.54
7th	9	7.54
8th	8	6.81
9th	8	6.81
10th	8	6.81

To determine the expected numbers of collisions for other proportions of groups of sites from the entire dataset based on the EB Method of Moments Adjusted method, the mean and variance of the similar sites was recalculated for the 85th percentile volume range of the target sites.

4.1.4. Empirical Bayes Safety Performance Function (EB-SPF) Method

The Empirical Bayes Safety Performance Function (EB-SPF) method produces the estimate of expected collisions by combining the history of collisions with the knowledge of similar sites represented in the form of a Safety Performance Function (SPF) estimate. Through the combination of these pieces of information, regression to the mean is accounted for in the resulting expected number of collisions. The expected number of collisions for this method is calculated using Equation 4-2 (Hauer, 1997).

$$E_A = \alpha\, E\{\kappa\} + \ (1-\alpha)K \qquad \textbf{Equation 4-2}$$

Where:

K - Observed number of accidents in the analyzed site, in the selected time period;

E_A - Expected number of accidents in the analyzed site, in the selected time period;

$\alpha -$ *The weight factor expressed as* Equation 4-3 (Hauer, 1997):

$$\alpha = \frac{1}{1+\frac{Var\{\kappa\}}{E\{\kappa\}}} = \frac{E\{\kappa\}}{E\{\kappa\} + \ Var\{\kappa\}} \qquad \textbf{Equation 4-3}$$

Therefore the expected number of collisions is expressed as Equation 4-4.

$$E_A = \frac{E\{\kappa\}^2}{E\{\kappa\} + \ Var\{\kappa\}} + \left(\frac{Var\{\kappa\}}{E\{\kappa\} + Var\{\kappa\}}\right)K = \frac{E\{\kappa\}^2 + \ K\ Var\{\kappa\}}{E\{\kappa\} + \ Var\{\kappa\}} \qquad \textbf{Equation 4-4}$$

Where E{κ}is the expected number of crashes at a site expressed by the SPF in Equation 4-5 for the intersection database used in this study.

$$E\{\kappa\} = \alpha\ (Major\ AADT)^{\beta_1}(Minor\ AADT)^{\beta_2} \qquad \textbf{Equation 4-5}$$

And where Var{κ} is the variance of the expected number of crashes expressed by Equation 4-6 (Hauer, 1997):

$$Var\{\kappa\} = \frac{(E\{\kappa\})^2}{b} \qquad \textbf{Equation 4-6}$$

Where:

$\alpha = 6.44E^{-5}$

$\beta_1 = 0.7693$

$\beta_2 = 0.4262$

$b = 1.5503\ (inverse\ of\ the\ 0.645\ dispersion\ parameter)$

Using these values in Equation 4-4 and the year 2000 data as the target year, the results of the EB-SPF method are shown in Table 4-4.

Table 4-4: Expected number of collisions based on Empirical Bayes - Safety Performance Function method

Year 2000 - Top 10 Sites	Major AADT	Minor AADT	(K)	$E\{\kappa\}$	$Var\{\kappa\}$	EB-SPF
1st	13115	801	18	1.636781	1.727989	10.04
2nd	45000	1501	16	5.522721	19.67279	13.70
3rd	45000	2001	11	6.242681	25.13634	10.05
4th	20986	501	11	1.923954	2.38753	6.95
5th	45000	1501	10	5.522721	19.67279	9.02
6th	15538	700	9	1.760696	1.999532	5.61
7th	9393	2010	9	1.873975	2.2651	5.77
8th	45000	1501	8	5.522721	19.67279	7.46
9th	25201	901	8	2.844315	5.218132	6.18
10th	23475	860	8	2.640301	4.496416	6.02

4.1.5. Empirical Bayes Safety Performance Function (EB-SPF) Recalibrated Method

The Empirical Bayes Safety Performance Function (EB-SPF) Recalibrated method is identical to the EB-SPF method with the one exception that the α value is recalibrated for each year of data so that the safety performance function can better fit that year of data. The formula for this recalibration is identified as Equation 4-7:

$$\alpha_{year} = \alpha \times \frac{\sum M_{\text{other}}}{\sum E\{\kappa\}} \qquad \textbf{Equation 4-7}$$

Where:

$\alpha_{year} = Recalibrated\ \alpha\ for\ the\ target\ year$

$\alpha = Original\ SPF\ value\ 6.44E^{-5}$

$M_{other} = Mean\ number\ of\ collisions\ per\ site\ for\ the\ other\ years$

$E\{\kappa\} = the\ expected\ number\ of\ crashes\ at\ a\ site\ expressed\ by\ the\ SPF$

Based on this formula, new α values were calculated for all 8 years of data shown in Table 4-5.

Table 4-5: Recalibrated α values for each year of data

Year	α_{year}
General	6.44E-05
2000	5.73E-05
2001	5.76E-05
2002	5.71E-05
2003	5.84E-05
2004	5.64E-05
2005	5.75E-05
2006	5.79E-05
2007	5.68E-05

Using the new α values, the expected number of collisions was calculated based on the same methodology for the previously explained Empirical Bayes Safety Performance Function (EB-SPF) method, with the one exception that the Safety Performance function was adjusted as per Equation 4-8.

$$E\{\kappa\} = \alpha_{year}\ (Major\ AADT)^{\beta_1}(Minor\ AADT)^{\beta_2} \qquad \text{Equation 4-8}$$

Where α_{year} was taken from the newly calibrated α for the specified target year from Table 4-5.

Using these values in Equation 4-8 and the year 2000 data as the target year, the results of the EB-SPF method are shown in Table 4-6.

Table 4-6: Expected number of collisions based on Empirical Bayes - Safety Performance Function Recalibrated method

Year 2000 - Top 10 Sites	Major AADT	Minor AADT	(K)	$E\{\kappa\}$	$Var\{\kappa\}$	EB-SPF Recalibrated
1st	13115	801	18	1.636781	1.727989	9.47
2nd	45000	1501	16	5.522721	19.67279	13.34
3rd	45000	2001	11	6.242681	25.13634	9.81
4th	20986	501	11	1.923954	2.38753	6.59
5th	45000	1501	10	5.522721	19.67279	8.78
6th	15538	700	9	1.760696	1.999532	5.30
7th	9393	2010	9	1.873975	2.2651	5.47
8th	45000	1501	8	5.522721	19.67279	7.26
9th	25201	901	8	2.844315	5.218132	5.92
10th	23475	860	8	2.640301	4.496416	5.75

4.1.6. Mean of Other Years

This method is based on the definition of regression to the mean, which is the tendency for the number of collisions at a site to return to the long term average. As such it is assumed that for the high collision sites of a given period, the true mean of each site is the mean number of collisions occurring in the other years. Therefore, the difference between this mean and that of the high collision year is assumed to be the regression to the mean. This is a reasonably good estimate, given that this "mean of other years" is taken from actual collision data for a long period, and so should represent the unbiased long term average of the site.

This estimate of the true mean is simply the average number of collisions from the other years. For example: If the year 2000 data are used for ranking the worst sites based on collision counts, the true mean of the site would be estimated as the average number of collisions per year from years 2001-2007 as shown in Table 4-7.

Table 4-7: Calculation of mean of other sites based on target period of 1 year

2000 Top 10 Ranking Sites	Target Year: 2000	Other Years:							Mean of other years (2001-2007)
		2001	2002	2003	2004	2005	2006	2007	
1	18	18	26	17	13	7	5	8	13.43
2	16	14	9	6	10	13	12	9	10.43
3	11	11	7	8	11	4	9	10	8.57
4	11	6	2	4	5	3	1	5	3.71
5	10	19	13	17	9	9	9	8	12.00
6	9	6	4	2	5	1	3	4	3.57
7	9	2	6	2	4	2	5	1	3.14
8	8	10	10	4	12	10	11	9	9.43
9	8	7	6	4	7	4	10	6	6.29
10	8	7	6	5	7	4	3	5	5.29

4.2. Comparison of Methods

Taking the results from each of the methods for estimating the expected number of collisions outlined in the previous section, the comparison of the results for the top 10 collision sites from the dataset using the year 2000 as the target year for ranking sites based on collision counts is compiled in Table 4-8. The results for these sites are shown graphically in Figure 4-1.

Table 4-8: Comparison of methods for estimating the expected number of collisions for top 10 sites

2000 Top 10 Sites	Major AADT	Minor AADT	Observed: (K)	EB-MoM	EB-MoM Adjusted	EB-SPF	EB-SPF (Recalibrated)	Mean of other years
1st	13115	801	18	13.81	14.15	10.04	9.47	13.43
2nd	45000	1501	16	12.32	12.68	13.70	13.34	10.43
3rd	45000	2001	11	8.61	9.01	10.05	9.81	8.57
4th	20986	501	11	8.61	9.01	6.95	6.59	3.71
5th	45000	1501	10	7.87	8.27	9.02	8.78	12.00
6th	15538	700	9	7.12	7.54	5.61	5.30	3.57
7th	9393	2010	9	7.12	7.54	5.77	5.47	3.14
8th	45000	1501	8	6.38	6.81	7.46	7.26	9.43
9th	25201	901	8	6.38	6.81	6.18	5.92	6.29
10th	23475	860	8	6.38	6.81	6.02	5.75	5.29
Total			**108**	**84.61**	**88.62**	**80.80**	**77.69**	**75.86**

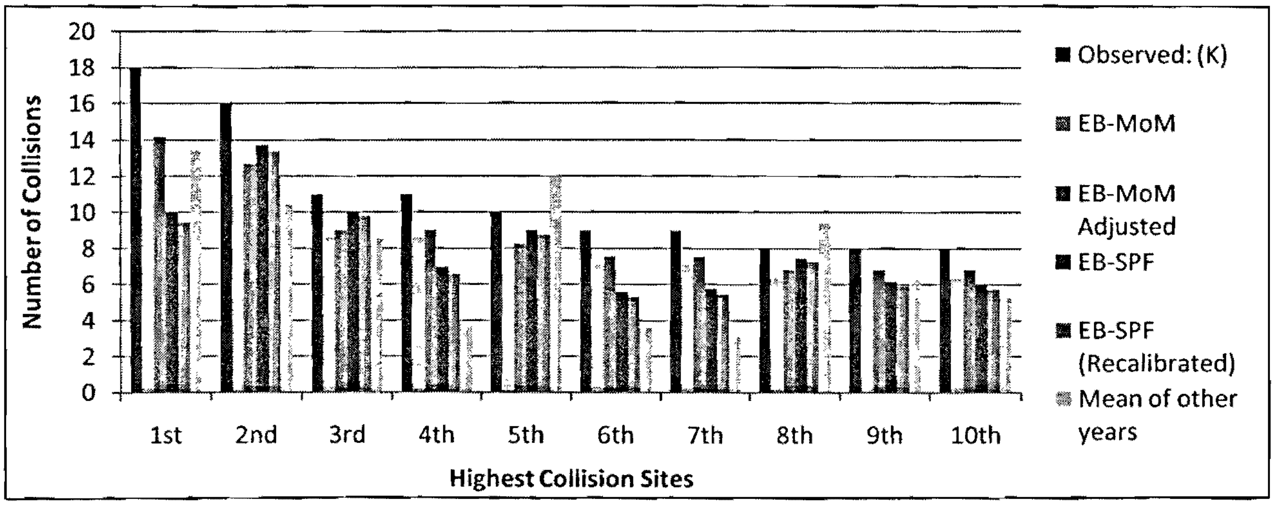

Figure 4-1: Bar chart showing comparison of methods for estimating the expected number of collisions for top 10 sites

From the graphical representation shown in Figure 4-1, with the exception of the 5th and 8th highest ranking sites, the observed number of accidents is higher than all the other methods for estimating the expected total number of collisions of each method. With respect to trends, it is not possible to identify any based on the random variation that is produced by comparing individual sites. Similar to the procedure that would be used for estimating treatment effects in a before-after study, the sites are grouped. For this purpose the summation of the expected number of collisions for all the top 10 sites was done in the last row of Table 4-8. These totals for each method are represented graphically in Figure 4-2.

The graphical representation clearly shows that the number of collisions obtained from the observed count method is much greater than the expected number of collisions from all the other methods as it is subject to regression to the mean bias. Following this, it is observed that the closest estimate to the "mean of other years" estimate is the EB-SPF Recalibrated method. This is expected given that the SPF method can attain a closer estimate by taking into consideration the traffic volumes, and that the recalibration of the alpha parameter makes the SPF an even better fit.

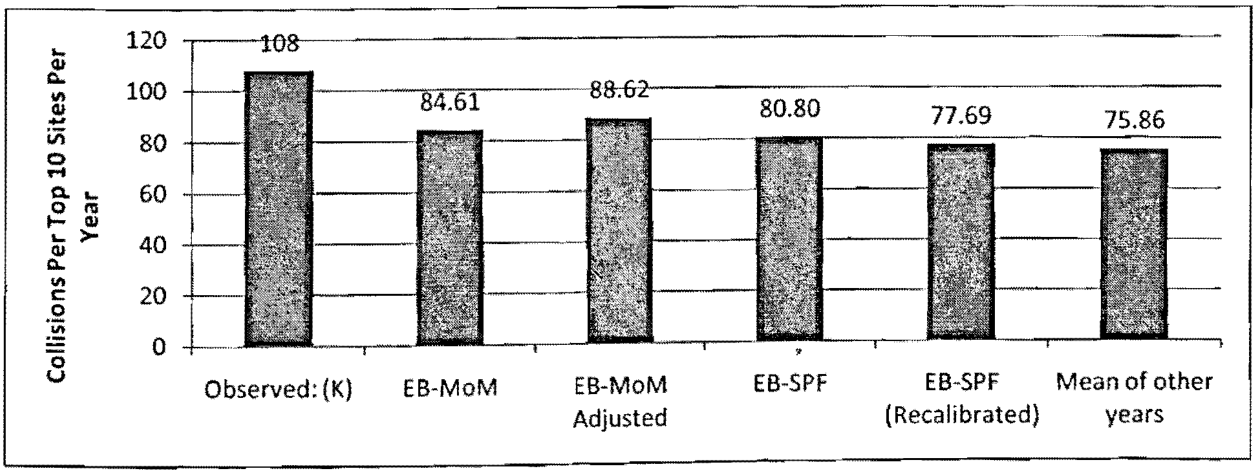

Figure 4-2: Bar chart showing the total of the expected number of collisions for top 10 sites for the different methods

The process was repeated using the year 2000 as the target year for selecting other proportions of highest ranked sites based on collision counts. The results from the varying methods are shown in Table 4-9 and represented graphically in Figure 4-3.

Table 4-9: Expected number of collisions for varying groups using the various methods based on year 2000 data

Year 2000 Top Site Groups	Observed: (K)	EB-MoM	EB-MoM Adjusted	EB-SPF	EB-SPF (Recalibrated)	Mean of other years
10	108	84.61	88.62	80.80	77.69	75.86
20	175	138.76	146.97	131.04	125.71	110.00
30	223	178.80	191.11	168.30	161.24	141.86
40	260	210.67	226.42	196.59	188.21	165.43
50	290	237.34	256.69	223.03	213.43	196.29
60	317	261.78	284.97	242.49	231.82	212.71
70	337	281.02	307.91	262.21	250.60	239.14
80	357	300.26	330.85	278.41	265.81	250.86
90	377	319.50	353.78	292.34	278.78	261.29
100	388	332.06	370.65	308.33	294.17	275.00
153	417	376.84	438.09	358.35	341.33	319.57
204	417	399.19	465.39	386.27	367.36	337.86

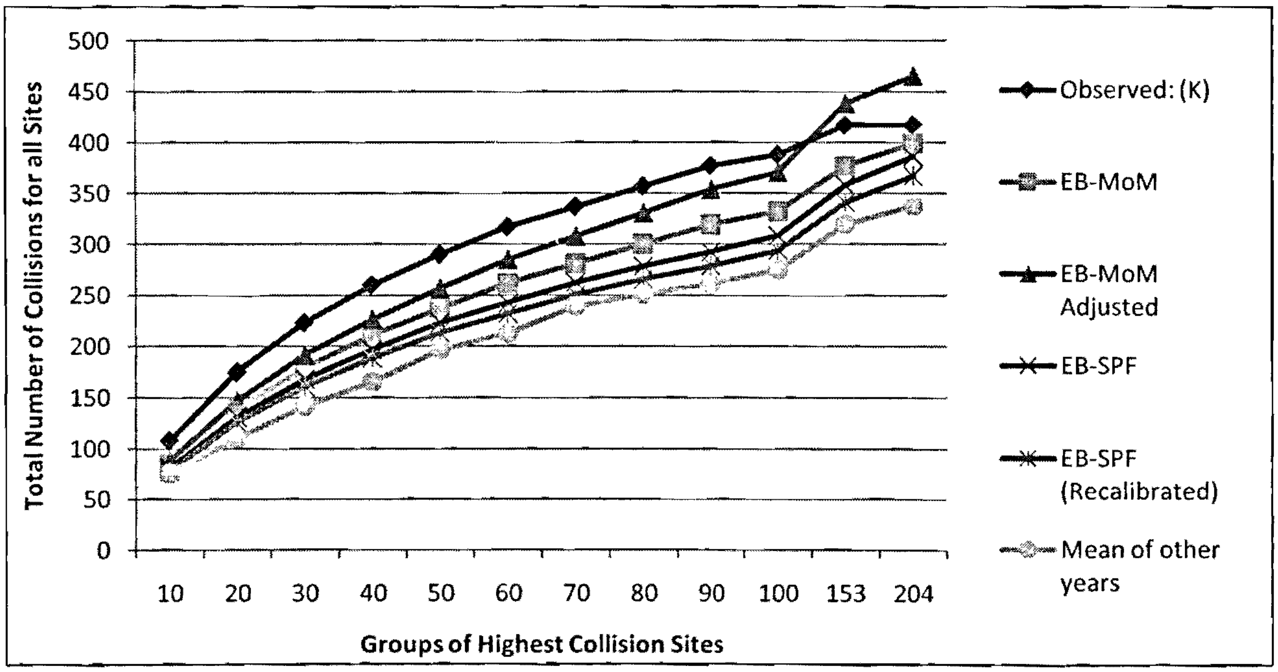

Figure 4-3: Expected number of collisions for varying top groups using the various methods based on year 2000 data

For verification of the trends these calculations were repeated using the additional years of data:

Table 4-10: Expected number of collisions for varying groups using the various methods based on year 2000-2001 data

Year 2000-2001 Top Site Groups	Observed: (K)	EB-MoM	EB-MoM Adjusted	EB-SPF	EB-SPF (Recalibrated)	Mean of other years
10	106	77.18	82.49	82.40	79.49	77.67
20	166	133.56	142.22	123.49	118.53	115.50
30	209.5	171.38	184.31	159.47	152.91	148.00
40	243.5	201.76	218.77	187.42	179.55	174.33
50	272.5	229.17	250.05	209.30	200.36	190.33
60	296.5	251.38	275.92	231.97	222.02	210.17
70	316.5	272.11	300.25	251.68	240.81	227.50
80	333.5	292.84	324.65	266.94	255.20	244.00
90	348.5	307.62	343.76	283.37	270.88	259.17
100	361	325.37	365.87	293.79	280.66	269.33
153	399	376.09	439.27	348.79	332.63	313.17
204	401	399.19	467.31	380.92	362.85	330.00

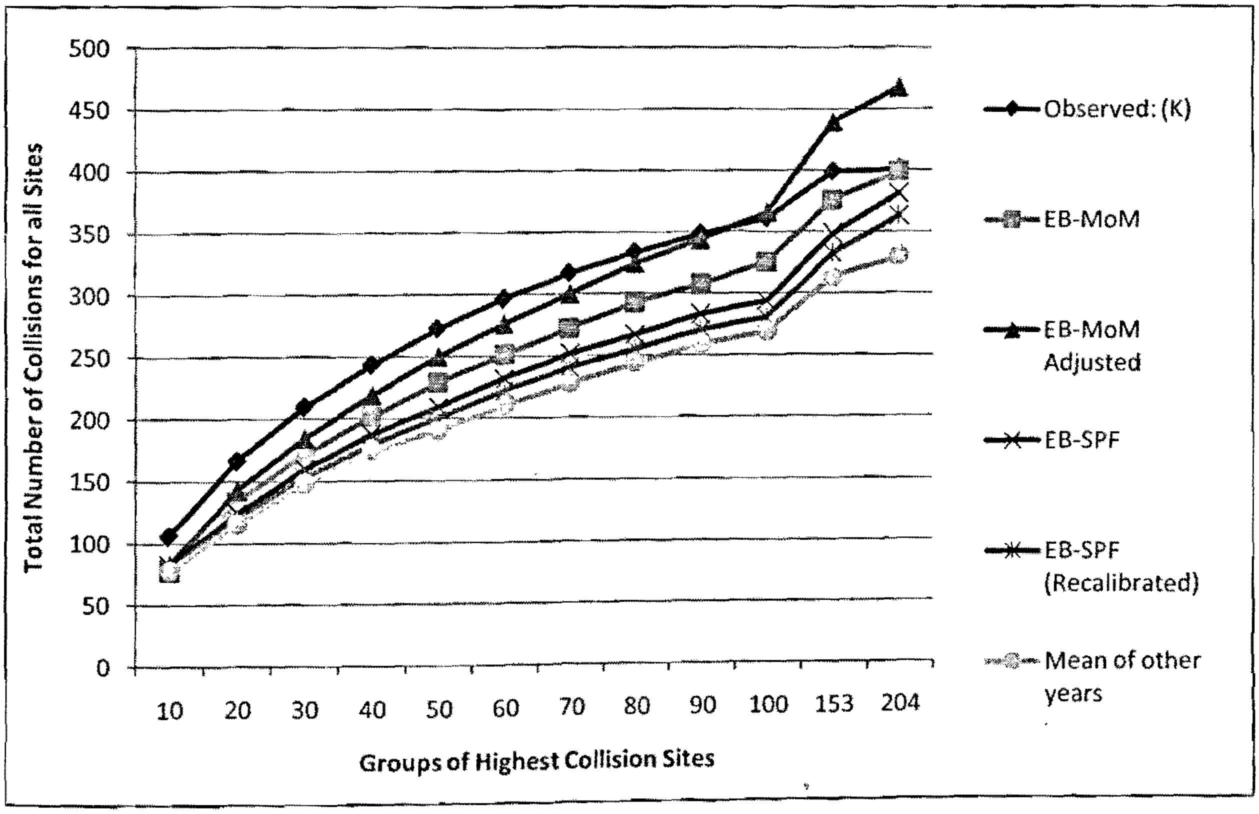

Figure 4-4: Expected number of collisions for varying top groups using the various methods based on year 2000-2001 data

Table 4-11: Expected number of collisions for varying groups using the various methods based on year 2000-2002 data

Year 2000-2002 Top Site Groups	Observed: (K)	EB-MoM	EB-MoM Adjusted	EB-SPF	EB-SPF (Recalibrated)	Mean of other years
10	102.6667	76.44	81.11	80.01	77.14	74.20
20	158	131.34	139.38	118.94	114.10	108.80
30	202.3333	169.89	181.37	152.99	146.57	145.20
40	236	196.56	211.58	183.88	176.19	174.40
50	263.6667	226.20	244.75	208.16	199.32	192.80
60	288.6667	247.67	269.78	229.93	220.09	215.60
70	308.6667	270.62	296.48	249.89	239.14	228.80
80	325.6667	289.12	318.57	265.53	253.93	242.40
90	339.6667	305.39	338.96	280.10	267.70	254.40
100	352.3333	317.20	354.76	293.06	279.95	264.00
153	388.6667	373.12	433.47	346.06	329.93	302.40
204	392.6667	399.19	464.46	378.06	359.96	320.80

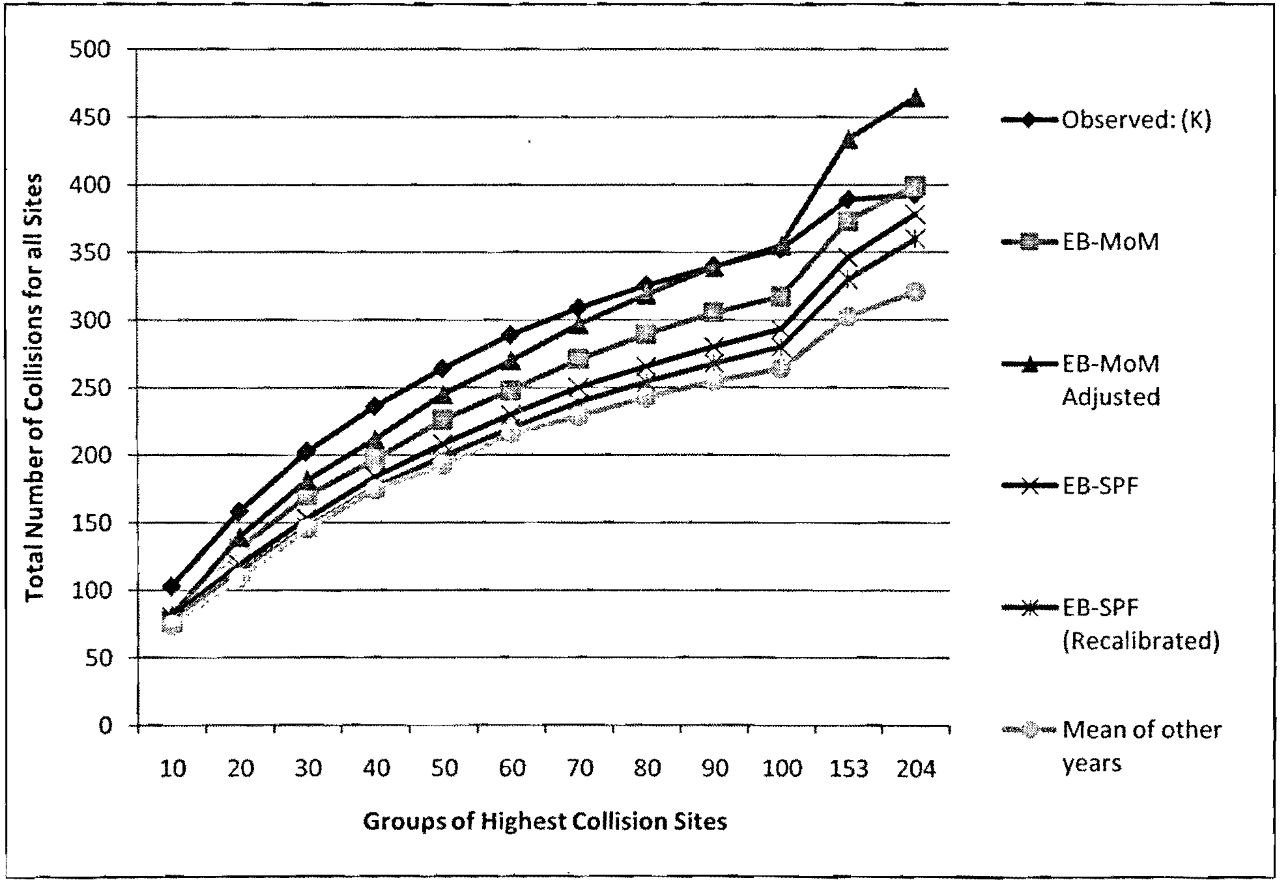

Figure 4-5: Expected number of collisions for varying top groups using the various methods based on year 2000-2002 data

Table 4-12: Expected number of collisions for varying groups using the various methods based on year 2000-2003 data

Year 2000-2003 Top Site Groups	Observed: (K)	EB-MoM	EB-MoM Adjusted	EB-SPF	EB-SPF (Recalibrated)	Mean of other years
10	98.5	76.44	81.11	77.05	74.37	71.25
20	152.75	130.59	138.65	115.60	111.05	109.50
30	195.5	167.66	180.13	152.50	146.46	141.25
40	230.5	195.07	211.08	180.47	173.24	174.50
50	258.75	220.26	239.79	204.67	196.34	198.00
60	282.5	244.70	268.24	224.69	215.40	211.25
70	301.5	267.65	295.17	243.81	233.62	226.50
80	318.25	290.61	321.87	261.60	250.60	235.50
90	331.25	301.68	336.28	275.43	263.70	252.75
100	343.25	316.46	356.01	287.23	274.82	265.00
153	381.75	374.61	443.22	343.20	327.82	298.50
204	386.75	399.19	472.74	374.49	357.23	308.75

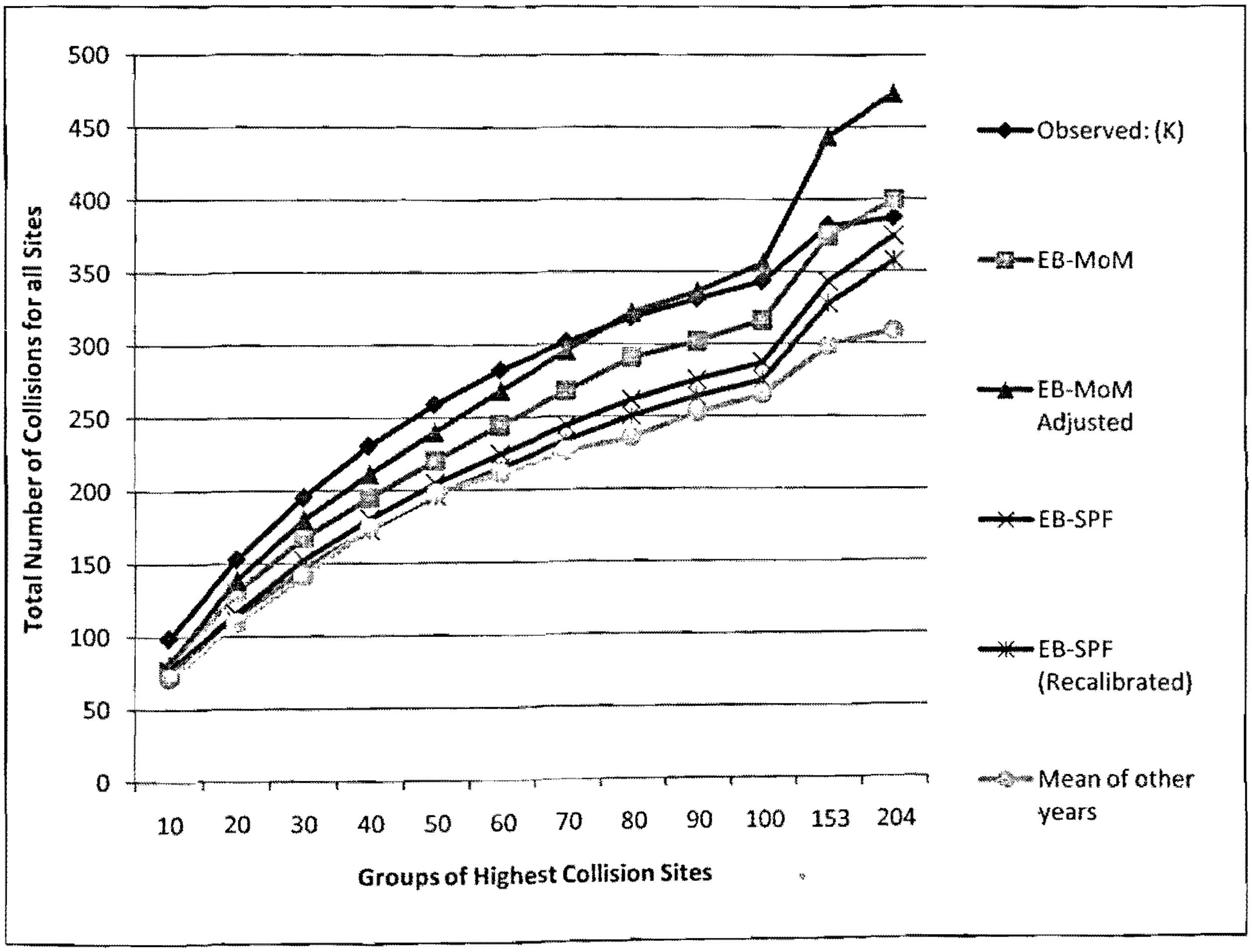

Figure 4-6: Expected number of collisions for varying top groups using the various methods based on year 2000-2003 data

4.3. Comparison Results

Having completed the comparison of the 6 methods for estimating the expected number of collisions to occur in the future years, there are several conclusions that can be drawn regarding the various methodologies:

i) **The observed method (naïve method)**, which is simply based on the assumption that the number of collisions that will occur in the following years is the same as the current has been clearly shown to be incorrect. This was confirmed by past studies in the literature review, but also shown here, where the averages of the other years of actual data from the high collision sites are significantly lower. There are situations where the number of collisions for individual sites may have a higher number of collisions in the following years. However, when selecting a group of sites, the numbers from these sites get averaged and the average of the mean of the other years would still be significantly lower in the following years. This confirms the presence of the regression to the mean phenomenon where the number of collisions in the future years will tend to return to the long term average. In fact the observed number of collisions for the target year could be used to estimate the extent of the regression to the mean for validating the other methods.

ii) **The Empirical Bayes Method of Moments (EB-MoM) method**, which uses the mean and variance of a reference group (of similar sites) to account for regression to the mean, does provide a reasonable expected number of collisions. However, it is noted that the method uses the same mean and variance to correct the entire dataset, resulting in a fixed translation for each number of observed collisions into a corrected number without taking into account traffic volumes or any other considerations. Given that this method produces a fixed correction for each number of collisions, it is not a likely expectation for real world data.

iii) **The Empirical Bayes Method of Moments (EB-MoM) Adjusted method,** which is based on the same methodology as the normal EB-MoM method, seeks to make adjustments to better fit the dataset. Instead of using the same mean and variance to correct all values in the dataset, it uses the mean and variance from sites with similar volume ranges for each group selected in an attempt to take into consideration traffic volumes. This attempt was unsuccessful in trying to get a better estimate for the expected number of collisions. This occurred because the new mean and variance determined for each group was larger than that of the entire dataset. This effectively increased the expected number of collisions bringing them closer to that of the observed counts instead of bringing them closer to the true mean of the sites.

iv) **The Empirical Bayes Safety Performance Function (EB-SPF) method,** which produces the estimate of the expected number of collisions by using the history of collisions combined with the SPF, was able to produce a much closer estimate to the long term average than the previous methods discussed. This was possible because of the use of the safety performance function that effectively considers the traffic volume in defining similarity of a reference site. However, it is noted that it also produces a higher estimate than that of the long term average. It is suspected that this occurs given that the SPF is not specifically calibrated to suit these 204 sites, but also includes other sites with similar features that were excluded from this dataset. This adjustment of the SPF was undertaken in the "EB-SPF Recalibrated" method which is therefore expected to produce a better estimate based on using an SPF that is a better fit.

v) **The Empirical Bayes Safety Performance Function (EB-SPF) Recalibrated method**, builds upon the existing EB-SPF method used previously by accounting for the noted limitation. The alpha (α) value is recalibrated (as shown in Section 4.1.5) to ensure that the Safety Performance Function used is a better fit and is further refined by calculating a separate value for each year of data being analyzed. The recalibration was done by adjusting the alpha (α) value by the ratio of the sum of the SPF estimates for all sites in the target year to the sum of the expected number of collisions for all sites determined by the mean of other years. Based on this it is able to provide the closest estimate to the long term average. While the values are still higher than the long term mean, it is as close as we can get without recalibrating all the parameters of the SPF. It should be noted that this estimate is not in fact being compared to the long term average (the true mean) of the site but rather to an estimate based on averaging 8 years of information, so, a priori, should not necessarily be expected to be a close match.

vi) **The "mean of other years" method**, uses the other unselected years to calculate the long term mean of collisions for the site based on the actual occurrence of these collisions. As such, it is the best estimate that we can deduce for the expected number of collisions in the future years as it is the count of what actually happened. As such, it is assumed to be an unbiased estimate of the true long term

> mean of each site in the absence of a better procedure. In practice, it is not possible to use this as the expected number of collisions as the sites would have been modified based on the treatment selected. However, for this research it can be used to represent the unbiased mean of the expected number of collisions.

For all of the methods used, it is observed from the various graphs plotted that the trends remained consistent for each trial. The various EB methods produced usable estimates of the expected number of collisions; however, it is noted that the Empirical Bayes Safety Performance Function methods produced estimates that were much closer to the actual number of collisions than that of the Empirical Bayes method of Moments Methods. Thus, when it is possible to incorporate the safety performance function into the estimate this should always be done, as the method of moments method is limited with respect to how well it can produce the expected number of collisions at similar sites due to its exclusion of traffic volumes in defining similarity. However, given that the "mean of other years" estimate is based on the actual observations of the following years of data we can assume that it is an unbiased estimate of the true mean of the sites.

5. EMPIRICAL ESTIMATION OF REGRESSION TO THE MEAN EFFECT

The purpose of the empirical exploration is to observe how regression to the mean depends on the number of years of data selected and the percentage of high collision sites selected. In order to undertake this task, the various methods for estimating the regression to the mean will be compared so as to identify a method for the in-depth analysis of the relationship between the regression to the mean effect and the number of years, percentage of high accident sites selected, as well as other variables such as the mean and standard deviation of the dataset that may affect the extent of the regression to the mean. This investigation was conducted using both real and simulated collision data.

5.1. Comparison of Methods for Estimating the RTM Effect

The results of the expected number of collisions produced previously can then be used to calculate how much regression to the mean is accounted for by each method. It is assumed that the best estimate (True RTM) would be the one calculated from the "mean of other years" estimate, which is the number of collisions that actually occurred and represents the unbiased mean that the other methods are seeking to estimate. Using the expected number of collisions produced from the various methods for the top 10 sites for the year 2000 data shown in Figure 4-2, the regression to the mean estimate can be depicted on a similar graph as shown in Figure 5-1. Similar to the results from the methods of estimating the expected number of collisions, the RTM estimate for the "EB-SPF Recalibrated" method is very similar to that of the "mean of other years" estimate.

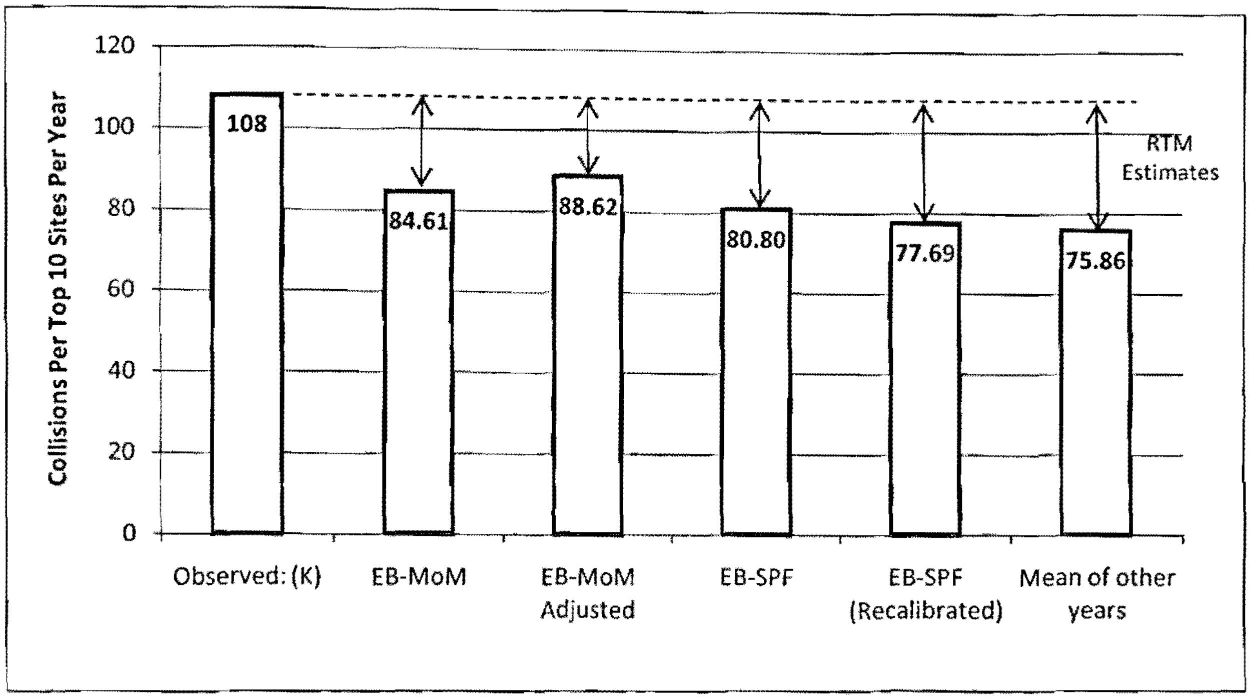

Figure 5-1: RTM representation for Top 10 Sites based on the Year 2000 Collision Data

Using the same methodology for the top 10 sites in order to generalise the results in groups rather than individual sites, the totals can then be expressed as regression to the mean percentages of the observed number of collisions. This RTM Percentage that is calculated using Equation 5-1 is represented as a percentage of the observed collision, which would make it simple to reverse calculate the actual number of collisions that should be expected if regression to the mean was not taken into consideration in an already published study.

$$RTM\ \% = \frac{C - \bar{x}}{C} \times 100\%$$

Equation 5-1

Where:

$C = Average\ Number\ of\ observed\ crashes\ per\ year\ from\ the\ selected\ years$

$\bar{x} = Expected\ Number\ of\ crashes\ per\ year\ estimated\ from\ the\ various\ methods$

Using Equation 5-1 and the values from the expected number of collisions for the methods in Figure 5-1 the RTM Percentages for the top 10 sites were calculated and shown in Table 5-1 and Figure 5-2.

Table 5-1: RTM Estimate for Top 10 Sites based on the Year 2000 Collision Data

Method	Observed	Estimated Number	RTM %
EB-MoM	108	84.61064	21.66
EB-MoM Adjusted	108	88.6179	17.95
EB-SPF	108	80.80488	25.18
EB-SPF (Recalibrated)	108	77.69275	28.06
Mean of other years	108	75.85714	29.76

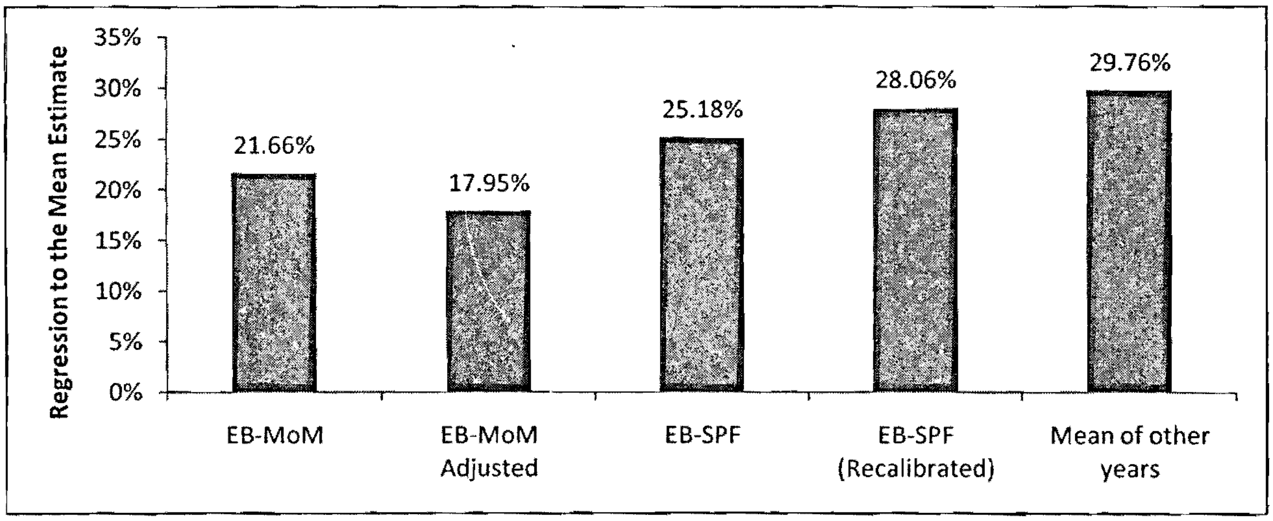

Figure 5-2: RTM Estimate for Top 10 Sites based on the Year 2000 Collision Data

Similar to the expected number of collisions comparison in Section 4, in order to do the comparison for the RTM percentage, Equation 5-1 was used to calculate the RTM estimate using groups of different proportions of high collision sites. The results from these additional groups taken from the total 204 sites are shown with the corresponding RTM percentage estimate in Table 5-2, and graphically in Figure 5-3.

Table 5-2: RTM Estimate for Top Site Groups based on the Year 2000 Collision Data

Site Groups	Observed: (K)	EB-MoM		EB-MoM Adjusted		EB-SPF		EB-SPF Recalibrated		Mean of other years	
	Total	Total	RTM %	Total	RTM %	Total	RTM %	Total	RTM %	Total	RTM %
Top 10	108	84.61	21.66	88.62	17.95	80.80	25.18	77.69	28.06	75.86	29.76
Top 20	175	138.76	20.71	146.97	16.02	131.04	25.12	125.71	28.17	110.00	37.14
Top 30	223	178.80	19.82	191.11	14.30	168.30	24.53	161.24	27.69	141.86	36.39
Top 40	260	210.67	18.97	226.42	12.91	196.59	24.39	188.21	27.61	165.43	36.37
Top 50	290	237.34	18.16	256.69	11.49	223.03	23.09	213.43	26.40	196.29	32.32
Top 60	317	261.78	17.42	284.97	10.10	242.49	23.50	231.82	26.87	212.71	32.90
Top 70	337	281.02	16.61	307.91	8.63	262.21	22.19	250.60	25.64	239.14	29.04
Top 80	357	300.26	15.89	330.85	7.33	278.41	22.01	265.81	25.54	250.86	29.73
Top 90	377	319.50	15.25	353.78	6.16	292.34	22.46	278.78	26.05	261.29	30.69
Top 100	388	332.06	14.42	370.65	4.47	308.33	20.53	294.17	24.18	275.00	29.12

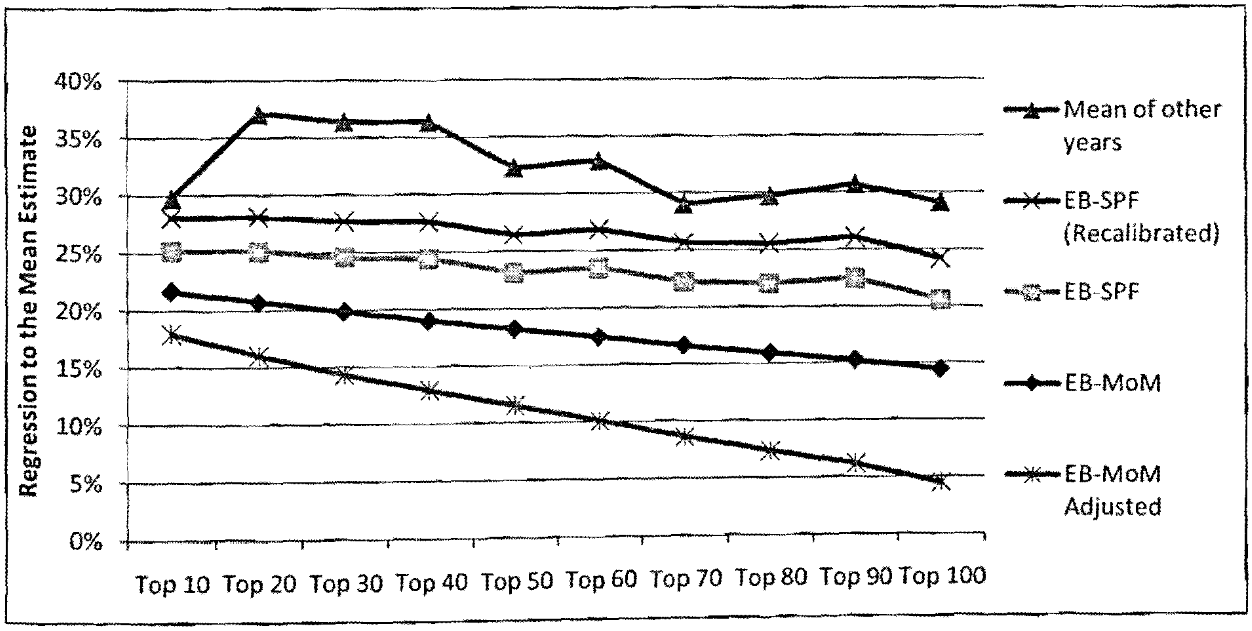

Figure 5-3: Method comparison for RTM estimates for Top Site Groups based on the Year 2000 Collision Data

As expected, as the group of top sites gets larger, the overall regression to the mean reduces, as shown from the general downward trend for each of the methods. The mean of other years is assumed to be our true unbiased mean as this is based on actual collision history. The EB-SPF methods consistently give a closer estimate of the expected number of collisions to the "mean of other years" estimate than the other methods, and also follows a similar trend. This result is because the EB method takes into consideration the traffic volumes, which does influence the total number of collisions. Furthermore, the EB-SPF Recalibrated method attains an even closer estimate as the recalibration of the SPF produces a better fit for each year of data.

On the other hand, the EB-MoM methods did not produce very close results nor did they display similar trends as that of the "mean of other years" estimate. Instead they produce a linear relationship to that of the observed collisions, where the regression to the mean percentage increases proportionally to the proportion of sites selected. This occurs given that this method has a fixed correction for each number of accidents and therefore does not properly account for all the regression to the mean.

To further check the results from this comparison test, the trial was repeated using more groups of 2, 3 and 4 years of data; according to the literature review conducted, it is expected that additional years of data used in the target period would reduce the magnitude of the regression to the mean effect. However, the estimates from the different methods should theoretically still produce the same results relative to each other. Given that the year 2000 was used as the target year, the groups of increased years of data used the year 2000 data as the first year, and included data from subsequent years as required. The results of these calculations are summarized graphically in Figure 5-4, Figure 5-5and Figure 5-6.

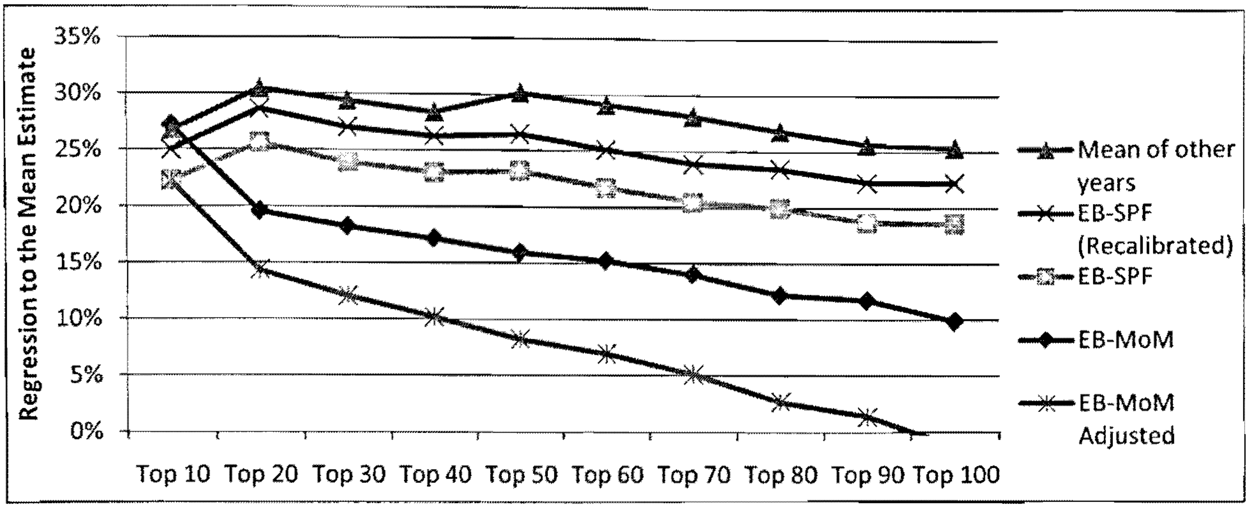

Figure 5-4: Method comparison for RTM estimates for Top Site Groups based on 2 Years (2000-2001) of Collision Data

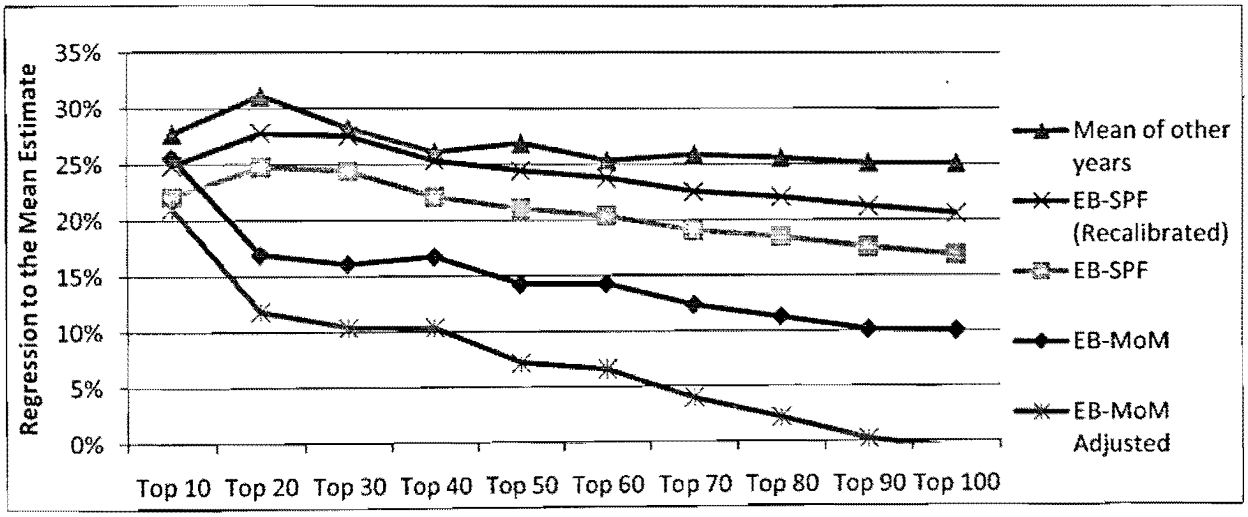

Figure 5-5: Method comparison for RTM estimates for Top Site Groups based on 3 Years (2000-2002) of Collision Data

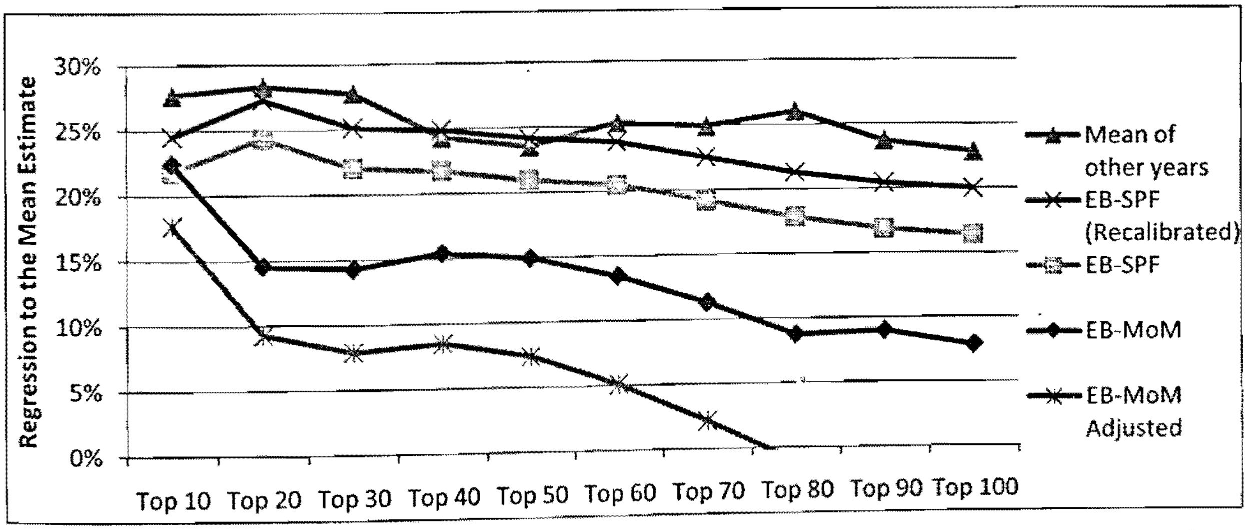

Figure 5-6: Method comparison for RTM estimates for Top Site Groups based on 4 Years (2000-2003) of Collision Data

Using the increased sample years, the same trends were observed as those found using the highest sites and 1 year of data. It is important to note that although the EB-SPF methods are unbiased and give a closer estimate to the actual number of collisions than the EB-MoM methods, they still do not represent the regression to the mean error as accurately as the "mean of other years" estimate, which would be considered to be the best estimate of the true unbiased mean.

Therefore, in applying the procedure for estimating the effect of regression to the mean based on number of years and percentage of high accident sites selected, the sample data with the actual number of collisions from the other years should be used rather than one of the EB methods for estimating the expected future collision frequency.

5.2. Regression to the Mean Analysis using Mean of Other Years Method

Based on the results from of the comparison in Section 5.1, in proceeding with the in-depth analysis of how the regression to the mean estimate is affected by the number of years and percentage of high collision sites selected, the "mean of other years" estimate was used. This method does not rely on any assumptions about the distribution of the dataset and is a straight comparison of the number of collisions in the target period against the mean number of collisions that is determined by a long term average that is assumed to be unbiased. For this purpose, the California intersections dataset for four legged stop controlled intersections as identified in Section 3.1 was used to complete the empirical exploration of how the regression to the mean varies based on number of years and percentage of high collision sites selected.

5.2.1. Methodology

The methodology involves calculating the regression to the mean using the other years for the various percentages of high crash sites (top 5, 10, 20, 30, 40, etc.) for each of the different combinations of years. The methodology in this section will identify the process for **1 Year of selected sites and the top 10 group**. However, for the complete results, this was done for groups of years from 1 to 7 and for groups of sites ranging from 2.5% to 100% of the top sites.

Step 1: Determining the long term average of each site

Starting with the Year 2000, the sites were sorted from highest collision site to the lowest one and the top 10 sites were selected. The long term average for each of these sites was determined by the average for the other 7 years of data as shown in Table 5-3. In addition, the average of the selected years and the other years for the top 10 sites was calculated for use in further steps.

Table 5-3: Mean of other years for Top 10 sites for selecting year 2000 data

	1 Year	**Average of**	**Other Years**							**Average of**
ID	**2000**	**Selected years**	**2001**	**2002**	**2003**	**2004**	**2005**	**2006**	**2007**	**Other years**
11683	18	**18**	18	26	17	13	7	5	8	**13.43**
17332	16	**16**	14	9	6	10	13	12	9	**10.43**
17330	11	**11**	11	7	8	11	4	9	10	**8.57**
7302	11	**11**	6	2	4	5	3	1	5	**3.71**
17333	10	**10**	19	13	17	9	9	9	8	**12.00**
16550	9	**9**	6	4	2	5	1	3	4	**3.57**
9660	9	**9**	2	6	2	4	2	5	1	**3.14**
17334	8	**8**	10	10	4	12	10	11	9	**9.43**
5582	8	**8**	7	6	4	7	4	10	6	**6.29**
15723	8	**8**	7	6	5	7	4	3	5	**5.29**
Average	10.8	**10.8**	10	8.9	6.9	8.3	5.7	6.8	6.5	**7.586**

Largest to smallest

Step 2: Determining the long term average of each site for the other 7 years

This process in step 1 was repeated for each of the other 7 years of data from 2001-2007 to produce 7 similar tables.

Step 3: Determining the average for the top 10 sites for each of the 8 groups

The averages of the selected years and the comparison years were calculated for each of the 8 tables produced in step 1 and are displayed in Table 5-4. The average of the 8 was used for the RTM percentage calculation to take into account the fact that the results can vary significantly from year to year.

Table 5-4: Average values for the top 10 sites of all 8 groups of 1 year of target data

Selected Year:	2000	2001	2002	2003	2004	2005	2006	2007	Average
Target Period Average	10.8	11.4	10.5	10.4	9.8	8.3	9.8	8.9	9.99
True Average	7.59	8.06	7.70	6.54	8.27	7.76	7.54	7.71	7.65

Step 4: Determining the Regression to the Mean Percentage

The results of the 8 trials were then averaged to calculate the regression to the mean percent for the top 10 sites for a 1 year period of selected data using the RTM % Equation 5-1:

$$RTM\ \% = \frac{9.99 - 7.65}{9.99} \times 100\% = 23.44\%$$

5.2.2. Results

The methodology was repeated for each of the other combinations of selected years, (2, 3, 4, 5, 6 and 7) as well as for each of the other groups of top sites (Top 2.5% to all 100%) to determine the RTM percentage for each of these. The resulting RTM percentages are shown in Table 5-5 and graphically in Figure 5-7.

Table 5-5: Regression to the Mean Estimate based on number of years and percentage of high accident sites selected

High Crash Sites		Regression to the Mean Estimate Percentage (%)							
Number of sites	Percent of total	1 Year	2 Years	3 Years	4 Years	5 Years	6 Years	7 Years	8 Years
5	2.45 %	27.49	15.00	7.76	7.74	4.78	3.54	4.14	0.00
10	4.9 %	23.44	13.93	11.42	7.30	6.13	5.58	2.10	0.00
20	9.8 %	21.68	14.59	11.84	10.53	9.39	9.50	7.22	0.00
31	15.2 %	21.70	12.94	10.18	8.13	7.12	5.58	5.21	0.00
41	20.1 %	21.18	12.08	7.77	5.38	4.68	3.34	3.34	0.00
51	25 %	19.43	11.07	7.89	5.61	4.22	3.55	3.06	0.00
102	50 %	13.32	8.48	6.24	4.59	3.56	2.96	1.72	0.00
153	75 %	3.88	3.95	3.39	2.09	2.26	1.47	1.15	0.00
204	100 %	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

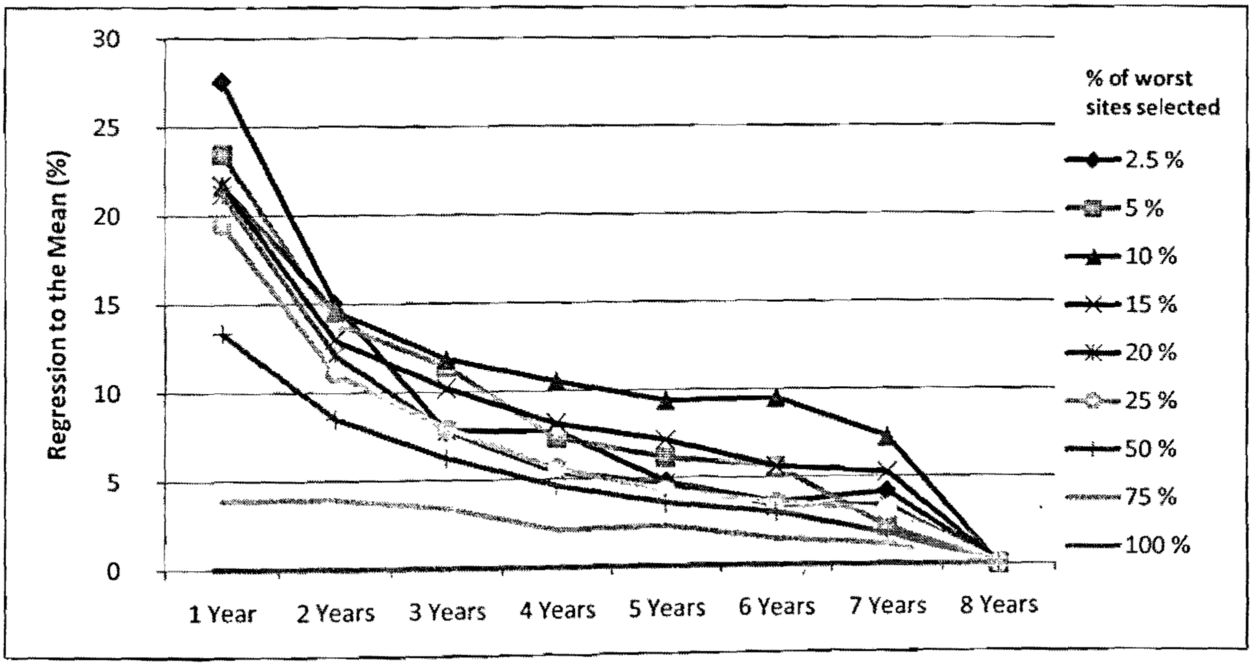

Figure 5-7: Graph of Regression to the Mean Estimate Vs Number of Years

As expected there is no regression to the mean calculated from the 100% group, given that if all sites are used, there is no bias represented in the method in which they are selected, and the long term mean for the group will be the same as the mean for the selected year. Furthermore, it is important to note that the 8 year selection shows no regression to the mean, since the dataset has only 8 years of data available; it could be possible that there is a small amount of regression to the mean still occurring if more years of data were available to assess this effect.

It is also noted that for the top 2.5% and 5% worst sites (5 sites and 10 sites, respectively) there seems to be inconsistencies in the trends as highlighted by those lines on the graph which are crossing below the others when they should be at the top, according to expectation. To better observe this result, the graph of the RTM estimate is plotted against the percentage of high accident sites as shown in Figure 5-8.

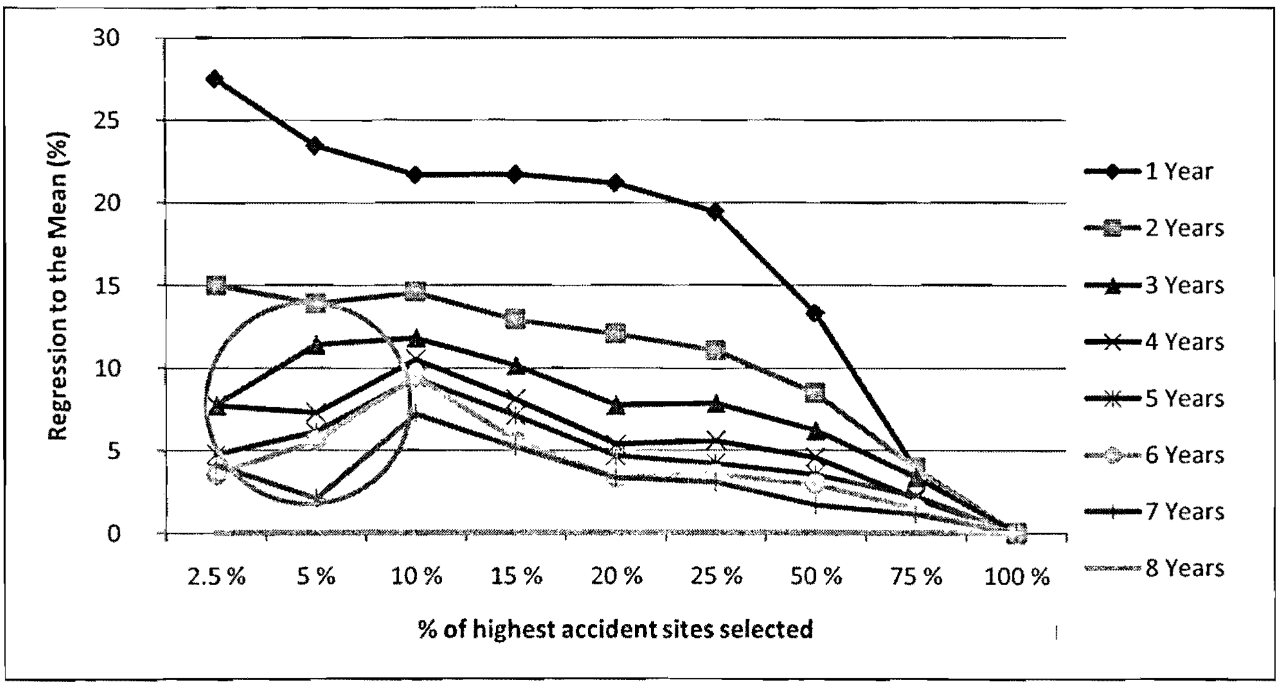

Figure 5-8: Graph of Regression to the Mean Estimate Vs Percentage of High Collision Sites

It is now clearly shown in Figure 5-8 that the data for these groups are not following the expected trend. This can be attributed to the fact that these groups have very few sites, 5 & 10 respectively. With such small numbers of sites, any small inconsistencies in the collision trends for a few of the sites in those groups would have an exaggerated impact on the results. Regardless of this inconsistency, given that this is an estimate, it is possible to use interpolation to correct those values that are inconsistent with the trends to produce a table showing the estimate of the regression to the mean as it relates to number of years and percentage of high accident sites selected. These interpolated values are included in Table 5-6.

Table 5-6: Regression to the Mean Estimate based on number of years and percentage of high accident sites selected

Percent of total high accident sites	**Regression to the Mean Estimate Percentage (%)**							
	1 Year	2 Years	3 Years	4 Years	5 Years	6 Years	7 Years	8 Years Or more
2.5%	27.49	21.77*	16.04*	15.08*	14.12*	15.38*	11.09*	0.00
5%	23.44	18.72*	14.00*	12.89*	11.77*	12.30*	9.14*	0.00
10 %	21.68	14.59	11.84	10.53	9.39	9.50	7.22	0.00
15 %	21.70	12.94	10.18	8.13	7.12	5.58	5.21	0.00
20 %	21.18	12.08	7.77	5.38	4.68	3.34	3.34	0.00
25 %	19.43	11.07	7.89	5.61	4.22	3.55	3.06	0.00
50 %	13.32	8.48	6.24	4.59	3.56	2.96	1.72	0.00
75 %	3.88	3.95	3.39	2.09	2.26	1.47	1.15	0.00
100 %	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

* *Interpolated values*

5.3. Regression to the Mean Analysis using the Simulated Data

It is expected that using simulated data would remove the inconsistencies in the trends that were identified in Section 5.2. To observe the RTM effects based on number of years selected and percentage of high accident sites for the simulated data, the same methodology as identified in Section 5.2 is repeated. The difference is that it is done using instead the simulated dataset identified in Section 3.4 . This dataset was produced for 8 years of data as well and is based on a gamma distribution with a mean of 4 collisions per year and a standard deviation of 1.6 collisions per year which is different from that used in Section 5.2.

The resulting regression to the mean estimate percentages from the repeat in the methodology are shown in Table 5-7 and graphically in Figure 5-9. Similarly to Section 5.2, the graph of the RTM estimate is plotted against the % of high collision sites, as shown in Figure 5-10.

Table 5-7: RTM Estimate based on number of years and percentage of high accident sites selected using simulated dataset

High Crash Sites		Regression to the Mean Estimate Percentage (%)							
Number of sites	Percent of total	1 Year	2 Years	3 Years	4 Years	5 Years	6 Years	7 Years	8 Years
5	5 %	39.40	27.45	22.13	16.31	15.50	16.10	13.32	0.00
10	10 %	34.61	26.17	19.30	17.37	13.23	12.69	10.20	0.00
15	15 %	31.78	24.67	18.36	16.61	12.89	12.16	8.53	0.00
20	20 %	30.49	21.65	17.36	14.28	12.23	9.10	6.81	0.00
25	25 %	28.73	18.90	14.10	12.72	8.74	7.57	5.31	0.00
30	30 %	27.69	17.30	12.27	10.51	9.51	8.12	5.88	0.00
50	50 %	16.35	11.58	9.28	7.60	6.28	5.29	4.75	0.00
75	75 %	4.45	2.61	2.20	1.98	1.67	1.44	1.47	0.00
100	100 %	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

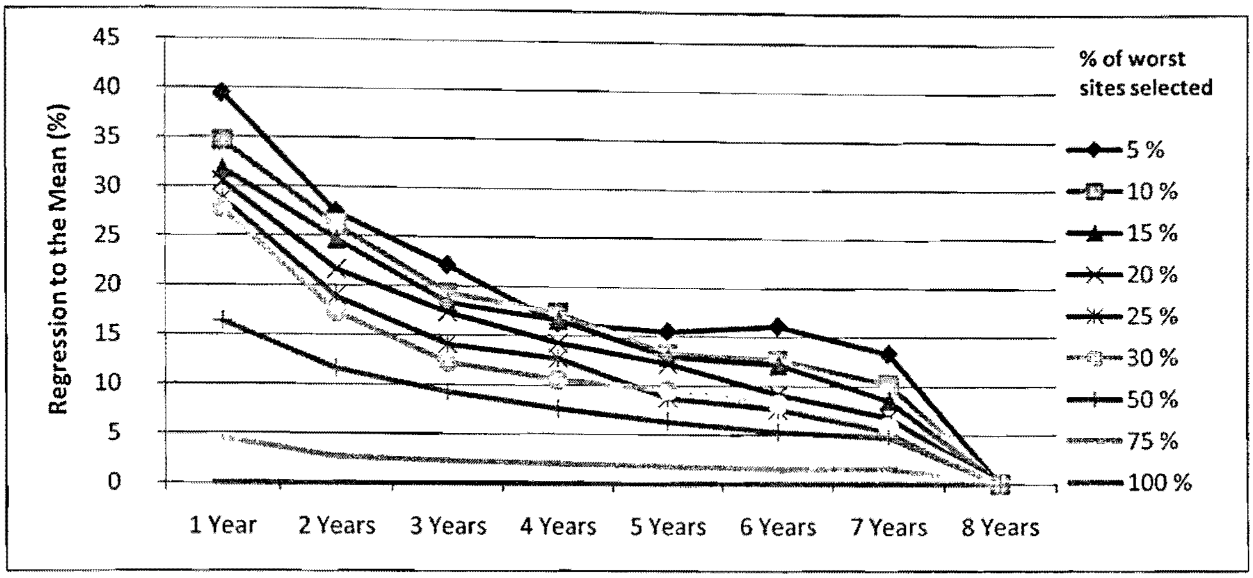

Figure 5-9: Graph of RTM Estimate Vs Number of Years Selected for simulated dataset

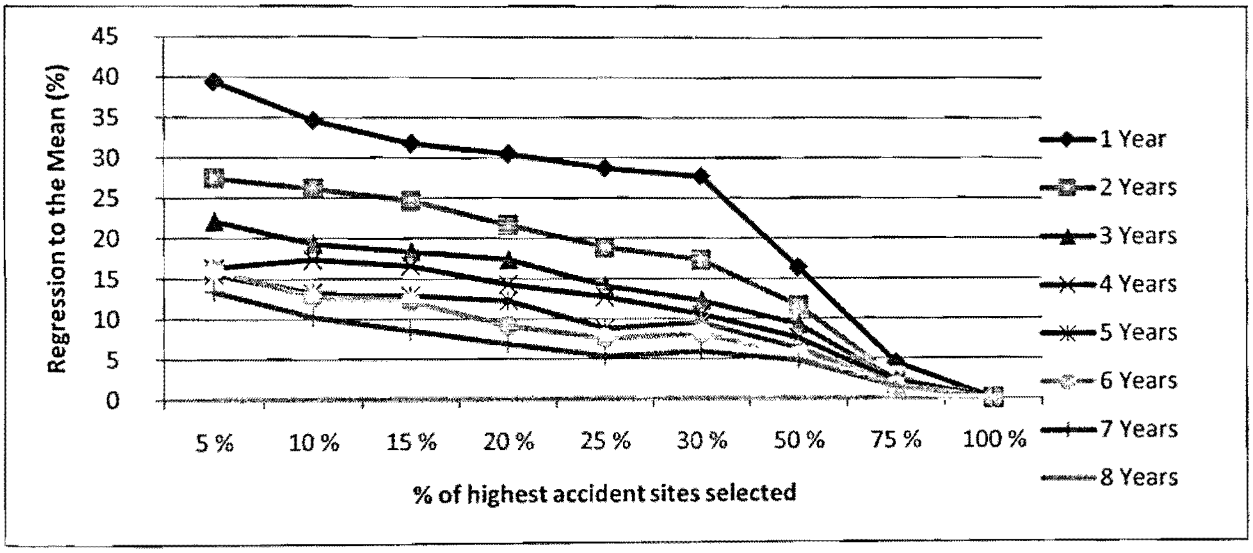

Figure 5-10: Graph of RTM Estimate Vs Percentage of High Collision Sites for simulated dataset

As expected, these graphs of the RTM Estimate Percentage for the simulated dataset followed the same expected trends as those shown in Section 5.2, but do not display any of the inconsistencies observed in the real dataset.

5.4. Comparison of the Regression to the Mean Estimate of Real vs. Simulated Data

For both analyses, the regression to the mean in general was larger using the smaller groups of high accident sites and larger for fewer years of selected data. These trends correspond with the expected results. Based on this empirical analysis, it could be concluded that the regression to the mean is affected by the number of years and percentage of high accident sites used.

For the regression to the mean estimate calculated using the real dataset, inconsistencies were found with the top 5 and top 10 groups of sites as their graphs did not follow the expected general trends. This is not a flaw in the methodology and instead is due to the small number of sites in these groups which would be largely affected by any discrepancies in the trends for any one of the sites. This would occur if that site had a high number of accidents occurring in the other comparison years as well. However, it is expected that with a better distributed dataset or a simulated dataset this should not occur. When the analysis was completed using the simulated dataset, it was found that no such discrepancies resulted from these two top groups. Based on this, it is justifiable to have corrected the inconsistent results from the RTM results for the real dataset using interpolation from the other results.

Although the trends for both analyses were similar, it is noted that the magnitudes of the RTM Percentages are different for the same number of years and groups. This would be largely due to the differences in the dataset parameters as defined by the mean and standard deviation. For the next section, variations in the mean and standard deviation are used to observe their effects on the regression to the mean estimates.

5.5. Generalization of the Regression to the Mean Estimate Results

Based on the results from Sections 5.2, 5.3 and 5.4 it is clear that the function for estimating the regression to the mean is not simply based on the number of years and percentage of high accident sites selected. Without being able to generalise the results, the only conclusions that can be made about the regression to the mean magnitude are that:

i) it increases as less years of data are selected and vice versa, and

ii) it increases as smaller proportions of the high collision sites from a population are used, and vice versa.

Generalizing the results would provide a quantifiable measure for how much the regression to the mean estimate depends on these factors. It was noted in the previous section that the results could be different due to the datasets having different mean and standard deviation values. As such, it is important for the effects of these variables to be investigated. This was accomplished by calculating the regression to the mean estimate for datasets with varying means and standard deviations to observe the relationship.

Using the same methodology for creating simulated datasets based on a specified mean and standard deviation detailed in Section 3.3, combinations of datasets were produced using the following parameters:

- **Mean**: Ranging from 1 to 12 collisions per year with intervals of 1 collision per year.
- **Standard Deviation**: Ranging from 0.25 times the mean to 1.25 times the mean at intervals of 0.1.

The effect of the number of years and percentage of high sites selected on the regression to the mean estimate would still be observed since the number of years is, in effect, considered automatically in varying the mean. To determine the RTM estimate for each of these combinations, 132 datasets were simulated (12 means x 11 standard deviations). The standard deviations are used as multiples of the mean rather than fixed values so as to normalize the results for comparison. These combinations are shown in Table 5-8.

Table 5-8: Combinations of Mean and Standard Deviation used for observing the RTM effect

Mean	Standard Deviations (Calculated by the mean x the factors below)										
	x 0.25	x 0.35	x 0.45	x 0.55	x 0.65	x 0.75	x 0.85	x 0.95	x 1.05	x 1.15	x 1.25
1	0.25	0.35	0.45	0.55	0.65	0.75	0.85	0.95	1.05	1.15	1.25
2	0.50	0.70	0.90	1.10	1.30	1.50	1.70	1.90	2.10	2.30	2.50
3	0.75	1.05	1.35	1.65	1.95	2.25	2.55	2.85	3.15	3.45	3.75
4	1.00	1.40	1.80	2.20	2.60	3.00	3.40	3.80	4.20	4.60	5.00
5	1.25	1.75	2.25	2.75	3.25	3.75	4.25	4.75	5.25	5.75	6.25
6	1.50	2.10	2.70	3.30	3.90	4.50	5.10	5.70	6.30	6.90	7.50
7	1.75	2.45	3.15	3.85	4.55	5.25	5.95	6.65	7.35	8.05	8.75
8	2.00	2.80	3.60	4.40	5.20	6.00	6.80	7.60	8.40	9.20	10.00
9	2.25	3.15	4.05	4.95	5.85	6.75	7.65	8.55	9.45	10.35	11.25
10	2.50	3.50	4.50	5.50	6.50	7.50	8.50	9.50	10.50	11.50	12.50
11	2.75	3.85	4.95	6.05	7.15	8.25	9.35	10.45	11.55	12.65	13.75
12	3.00	4.20	5.40	6.60	7.80	9.00	10.20	11.40	12.60	13.80	15.00

5.5.1. Methodology

Given that the data simulation is based on random sampling, the simulated dataset for each of the 132 combinations was reproduced 10 times. The RTM estimate was then taken as the average of the 10 RTM estimates that were produced for each of the 132 combinations in order to produce more stable results.

To simplify the process, it was determined that it was only necessary to generate 1 year of data for each of the 100 sites, given that the mean of each site is already known. Recall from Chapter 3 that the mean of each of the 100 sites is generated from the data simulation of the Gamma distribution where the mean and standard deviation of the dataset is defined. This mean is then used to generate each discrete number of collisions occurring each year using the Poisson distribution. It was necessary to generate many years of data in Section 5.3 given that the mean was determined from the average number of collisions in the other years. However, for this process, given that the mean is known, it would have been unnecessary to use this method to try determining the true long term average for each site.

Further to this, as noted earlier, the RTM estimate does not need to be recalculated each time for the number of years of data selected as was done in the previous sections. Since we are introducing the mean as a variable, having the number of years as a variable as well would simply be a redundant as the number of years is related to the mean and is automatically considered. For the purpose of the RTM estimate, doubling the number of years would have the same effect as if the mean were doubled.

The process to generate the RTM for the 132 combinations of mean and standard deviations is outlined by an example of each stage in the process.

Step 1: Calculate the Regression to the Mean for individual combination

The parameters were set for the Gamma distribution data simulation as the first mean and standard deviation from Table 5-8. The dataset was sorted from highest accident site to lowest accident site and the averages of the number of collisions for each of the 10 sets were calculated. The average of the corresponding true mean values for each of the groups was also calculated, and the RTM percentage was calculated for each group based on Equation 5-2.

$$RTM\,\% = \frac{C - M}{C} \times 100\% \qquad \textbf{Equation 5-2}$$

Where:

$C = Average\ Number\ of\ generated\ crashes\ per\ group\ of\ sites$

$M = Known\ mean\ of\ generated\ crashes\ per\ group\ of\ sites$

The results of this calculation were recorded as shown in Table 5-9. This trial was repeated 10 times and the average of the RTM % for each of the 10 trials was recorded.

Table 5-9: RTM Calculation from Data Simulation

Mean:	**1.0**	**Std. Dev:**	**0.25**
Group	**Ave. Selected Yr**	**Ave. True Mean**	**RTM (%)**
Top 10	3.1	1.05	66.27
Top 20	2.55	0.957367	62.46
Top 30	2.066667	0.955705	53.76
Top 40	1.8	0.973751	45.90
Top 50	1.64	0.955831	41.72
Top 60	1.533333	0.948823	38.12
Top 70	1.342857	0.946831	29.49
Top 80	1.175	0.946525	19.44
Top 90	1.044444	0.9476	9.27
Top 100	0.94	0.944055	-0.43

Step 2: Calculate the Regression to the Mean for all Std. Dev. of each mean

The process in step 1 was then repeated for each of other standard deviations and recorded in Table 5-10, and the averages for each group were calculated.

Table 5-10: RTM percentage for simulated data for combinations with mean value of 1

Mean: 1	**Regression to the Mean Percentage for Each Standard Deviation Combo (%)**											
Std Dev	**0.25**	**0.35**	**0.45**	**0.55**	**0.65**	**0.75**	**0.85**	**0.95**	**1.05**	**1.15**	**1.25**	**Ave.**
Top 10	63.65	62.11	61.79	59.17	54.07	51.59	51.44	48.45	47.88	48.37	29.05	52.51
Top 20	58.28	56.63	54.34	52.98	50.11	45.72	49.02	44.42	43.32	41.48	27.27	47.60
Top 30	51.80	49.50	48.20	47.57	45.98	40.33	43.66	39.79	39.62	37.30	26.00	42.70
Top 40	45.30	43.26	42.26	42.82	40.19	36.77	39.41	35.88	36.34	31.63	20.75	37.69
Top 50	41.07	39.04	39.44	39.58	36.58	33.93	35.37	31.11	29.04	26.40	16.22	33.43
Top 60	35.10	33.89	33.71	33.73	30.20	27.39	28.95	25.32	22.43	21.22	11.47	27.58
Top 70	25.79	24.75	24.49	25.60	22.19	20.18	22.28	19.31	16.04	16.23	7.74	20.42
Top 80	15.95	14.77	15.45	17.17	14.71	12.77	16.92	13.67	10.43	8.99	2.63	13.04
Top 90	6.11	5.79	6.77	9.32	6.76	5.68	10.58	7.25	4.68	2.77	0.00	5.97
Top 100	0.00	0.00	0.00	0.70	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Step 3: Calculate the Regression to the Mean for all combinations

The process in step 1 & 2 was repeated for all the remaining means, and the tables were populated as per step 2, and all the averages summarized into Table 5-11.

Table 5-11: Average of RTM percentages for all standard deviations for each mean

	Regression to the Mean Percentage from all Averaged results of Varying Std Deviations (%)											
Means	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**Ave.**
Top 10	52.51	39.20	32.26	27.91	24.99	21.60	19.71	17.19	16.37	15.51	13.46	14.12	**24.57**
Top 20	47.60	35.65	29.78	25.58	22.53	19.93	18.10	16.49	15.37	14.72	12.86	13.10	**22.64**
Top 30	42.70	31.88	26.60	22.84	19.95	18.06	16.06	14.85	13.55	12.56	11.42	11.46	**20.16**
Top 40	37.69	27.81	23.28	20.22	17.27	15.58	13.79	13.00	11.61	10.70	9.64	9.92	**17.54**
Top 50	33.43	24.62	20.13	17.43	14.71	13.24	11.50	11.04	9.84	8.84	8.22	8.20	**15.10**
Top 60	27.58	20.35	16.83	14.51	12.31	10.92	9.31	8.89	7.99	7.11	6.69	6.84	**12.44**
Top 70	20.42	15.87	13.52	11.46	9.58	8.43	7.09	6.85	6.38	5.35	4.86	5.51	**9.61**
Top 80	13.04	11.30	9.74	8.26	6.70	6.08	4.59	4.86	4.30	3.54	3.31	4.04	**6.65**
Top 90	5.97	5.72	5.67	4.75	3.73	3.64	2.33	2.85	2.29	1.75	1.71	2.40	**3.57**
Top 100	0.34	0.61	1.27	0.73	0.39	0.84	0.25	0.75	0.43	0.18	0.21	0.80	**0.57**

5.5.2. Relationship between RTM percent, Mean and Number of High Accident Sites

The relationship between the RTM percent, mean and groups of high collision sites can be shown by holding the standard deviation constant and plotting the resulting graphs. This was done using a low (0.35), middle (0.75) and high (1.15) values of the range the Standard Deviations represented as multipliers of the mean in Figure 5-11, Figure 5-12 and Figure 5-13.

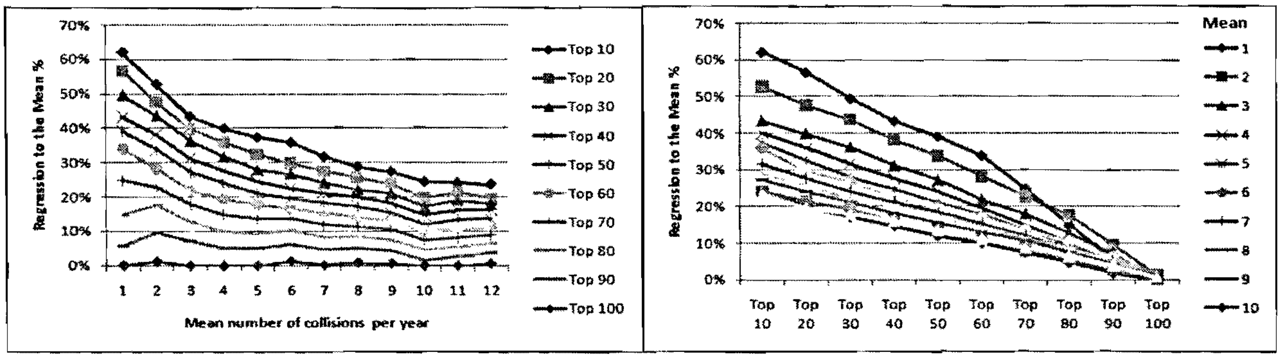

Figure 5-11: Graphs of Regression to the Mean % Vs Percentage of High Sites and Mean for Std. Dev. of 0.35xMean

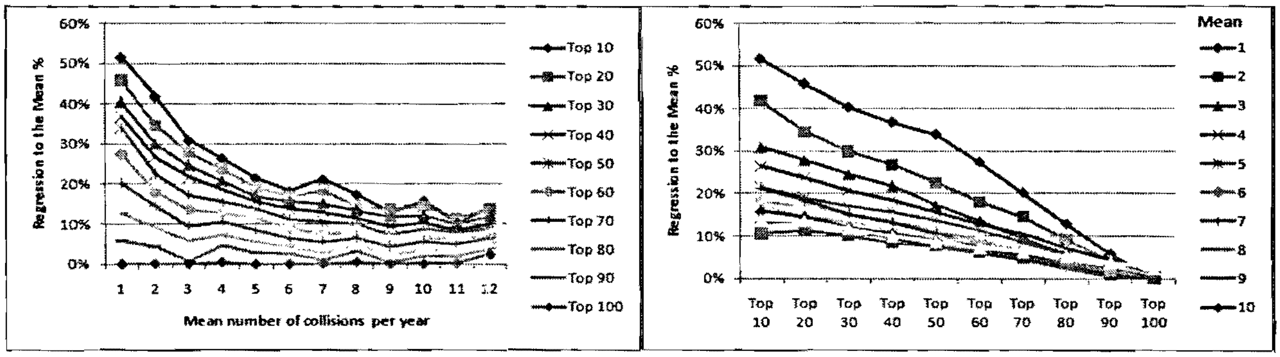

Figure 5-12: Graphs of Regression to the Mean % Vs Percentage of High Sites and Mean for Std. Dev. of 0.75xMean

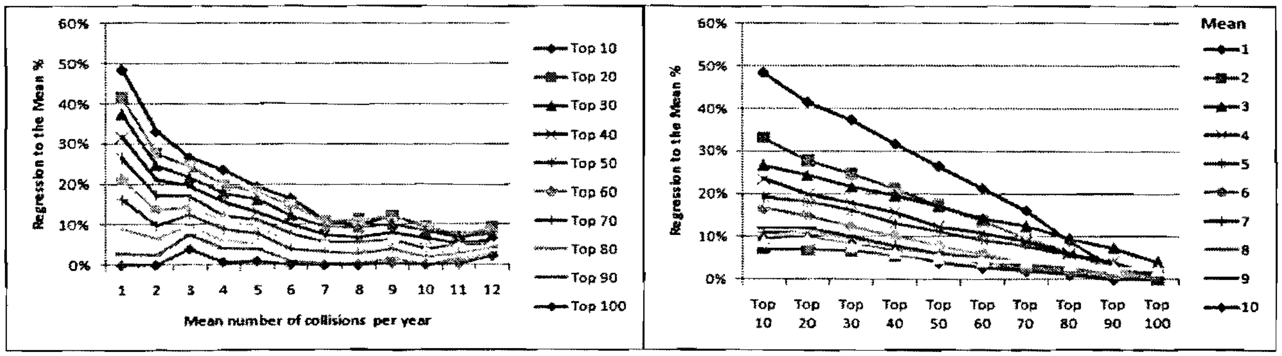

Figure 5-13: Graphs of Regression to the Mean % Vs Percentage of High Sites and Mean for Std. Dev. of 1.15xMean

As a further step the average results from all the standard deviations was plotted against the mean number of collisions per year and groups of sites to eliminate the standard deviation as a variable. This is shown in Figure 5-14 and Figure 5-15. The results of these graphs show the same trends as outlined in the previous sections where the RTM percentage is higher for lower means and smaller percentages of the high collision sites.

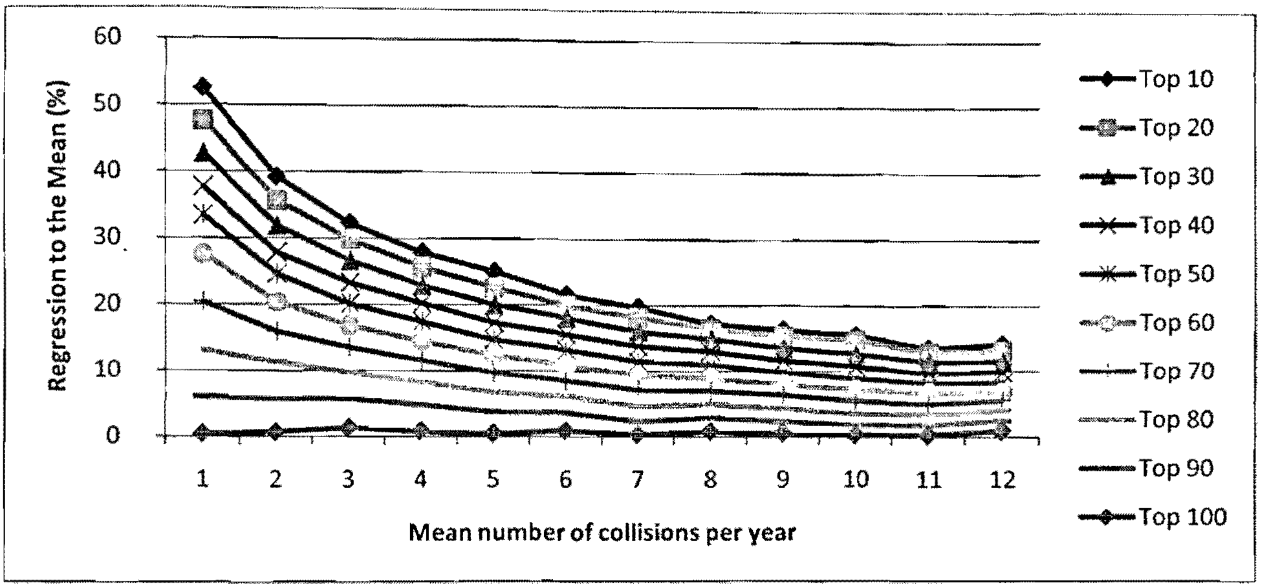

Figure 5-14: Graph of Regression to the Mean % Vs Mean for average of all standard deviations

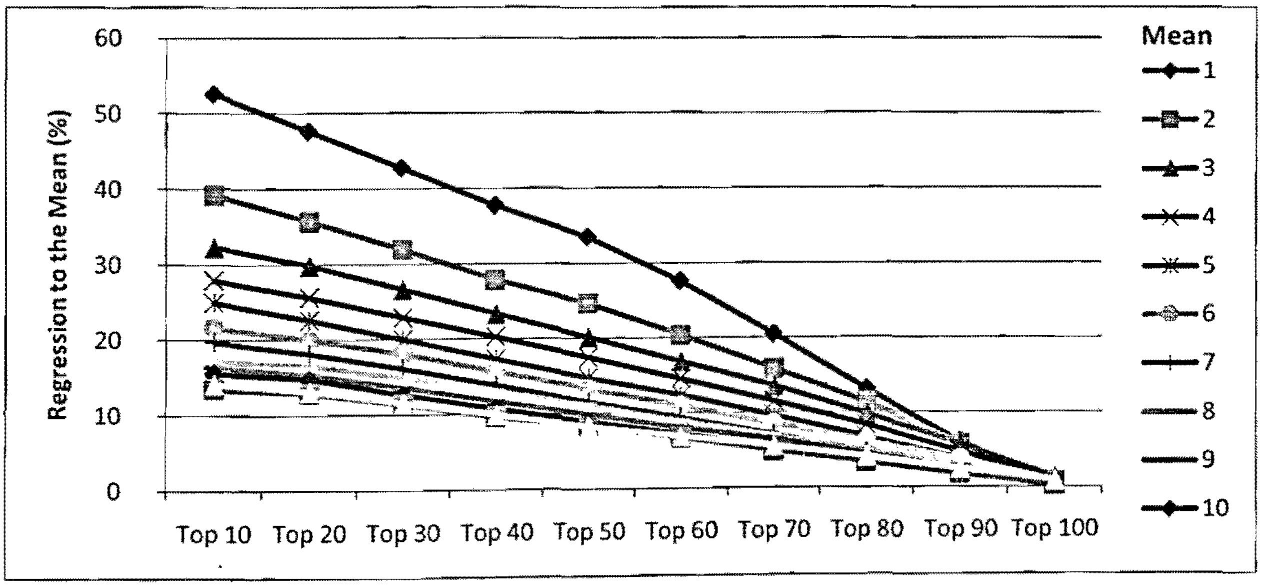

Figure 5-15: Graph of Regression to the Mean % Vs Group of top sites for average of all standard deviations

5.5.3. Relationship between RTM Percentage, Number of High Accident Sites, and Standard Deviation

The relationship between the RTM %, size of the group of high collision sites and standard deviation can be shown by holding the means constant and plotting the resulting graphs. This was done using a low (2), middle (7) and high (11) values of the range means shown in Figures 5-16, 5-17 and 5-18. Here we observe that as the standard deviation increases, the RTM % decreases, and as smaller percentages of the high sites are used the RTM % becomes larger.

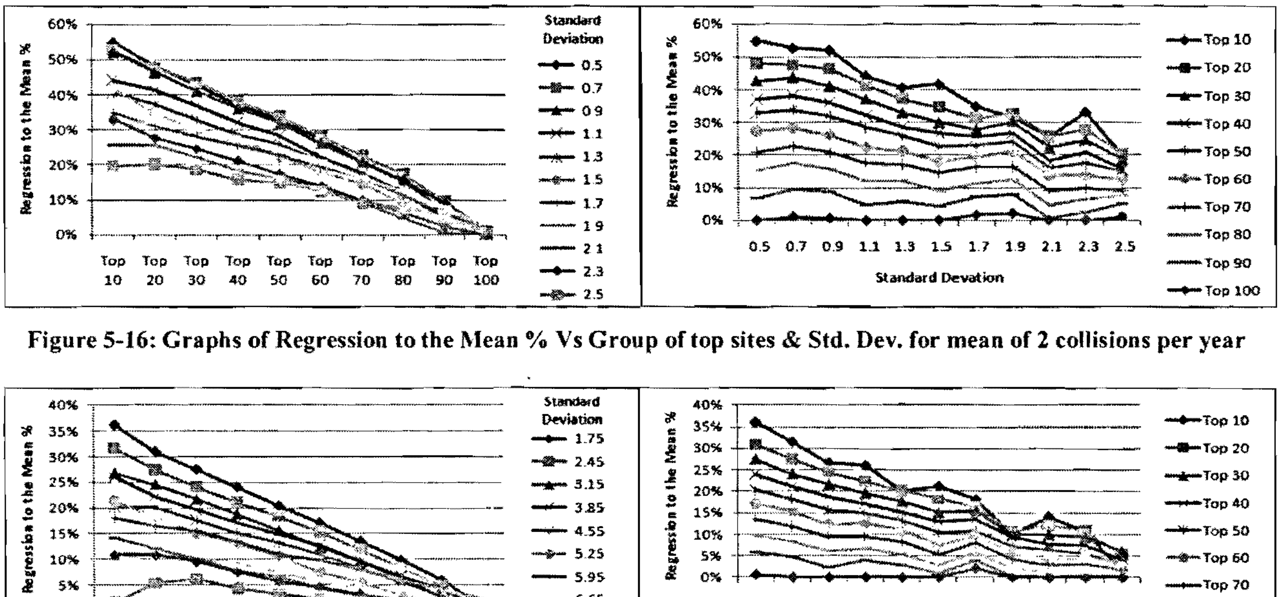

Figure 5-16: Graphs of Regression to the Mean % Vs Group of top sites & Std. Dev. for mean of 2 collisions per year

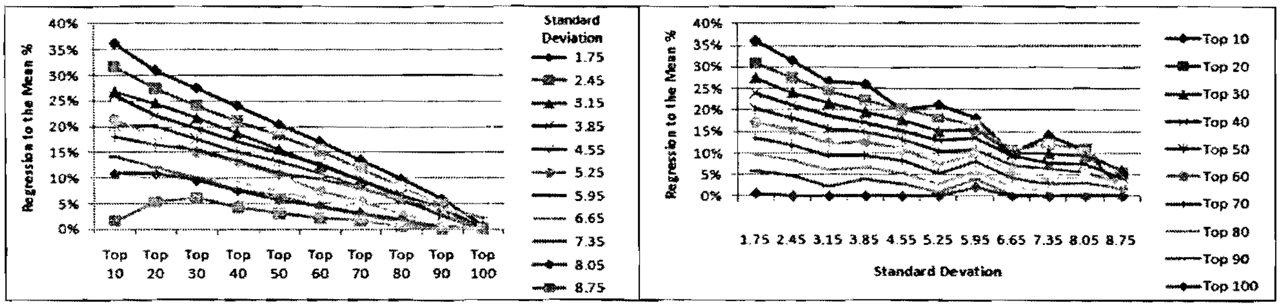

Figure 5-17: Graphs of Regression to the Mean % Vs Group of top sites & Std. Dev. for mean of 7 collisions per year

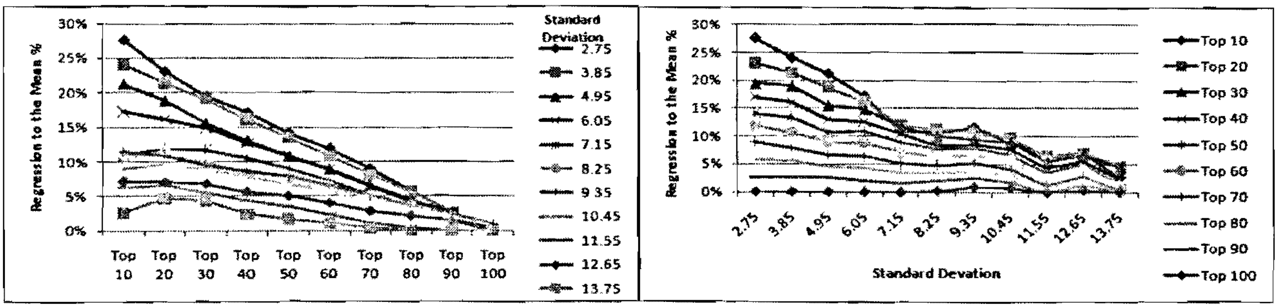

Figure 5-18: Graphs of Regression to the Mean % Vs Group of top sites & Std. Dev. for mean of 11 collisions per year

5.5.4. Regression to the Mean Estimate Generalization Summary

Based on the resulting data, it is shown that the regression to the mean percentage is distinctively dependent on three factors, the mean, standard deviation, and proportion of high sites used. These relationships are simply summarized as:

i) The regression to the mean percentage estimate decreases as the mean of the dataset increases. Further, given that the mean is related to the number of years of data collected/selected, we can also conclude from this relationship that as the number of years of collision data increases, the regression to the mean percentage estimate decreases. This is an expected result, as higher collision numbers per year tend to display a lower percentage difference rate for fluctuations.

ii) The regression to the mean percentage estimate decreases as the standard deviation of the dataset increases. This is an interesting result as one might expect there to be more regression to the mean with a larger variance. However, given that the sites are ranked from highest to lowest, a larger variance would create sites with much more collisions per year in the top groups, which is almost similar to having a higher mean.

iii) The regression to the mean percentage estimate decreases as larger groups of the top sites from the dataset are selected. This is the expected result, given that a smaller group of the top sites would be more likely to be abnormally higher than the true mean of that group of sites. Similarly, when the entire population is used, there would be no regression to the mean occurring.

The three main relationships identified can be illustrated on a scatter plot matrix shown in Figure 5-19. The relationships identified between the RTM percentage and each of the variables is confirmed and this accumulation of data can be used to formulate a function for the relationships.

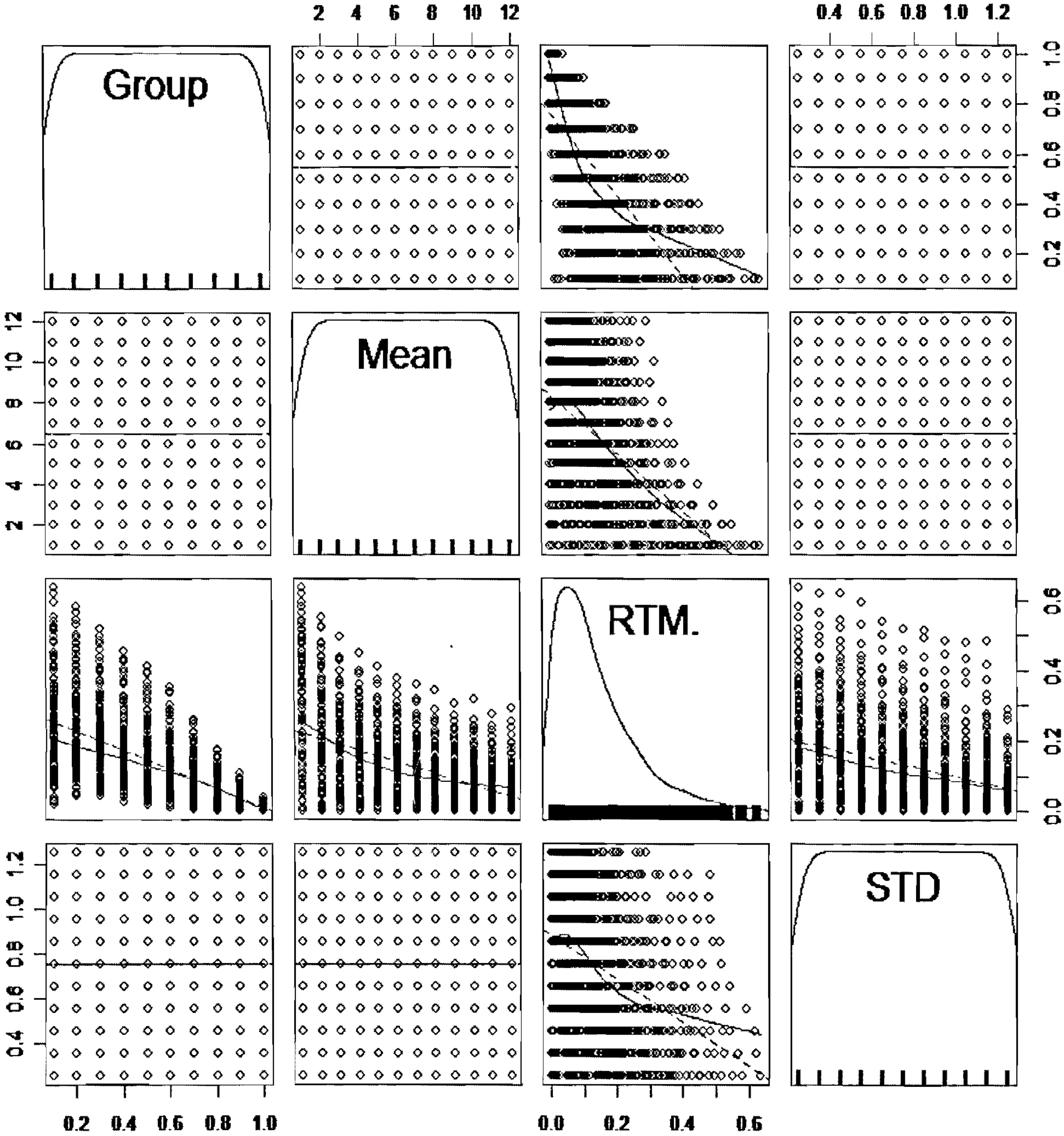

Figure 5-19: Scatter plot matrix of the RTM against the 3 dependent variables, the mean, standard deviation, and proportion of high sites used (The R Foundation for Statistical Computing software, 2009)

5.6. Function to Determine the Regression to the Mean Estimate

Having shown and defined the trends and relationships for the three main variables that affect the regression to the mean estimate, the next step in the process was to derive a function based on the accumulation of simulated data and results produced. From the scatter plot matrix shown in Figure 5-19, the relationships between the variables and the regression to the mean estimate are all fairly linear. Thus, for the purpose of this research model it was decided to use a linear relationship to derive the function. The function would make it possible to quickly and easily estimate the regression to the mean percentage that should be accounted for when undertaking a treatment study.

5.6.1. Linear Model for the Regression to the Mean Percentage

To produce the linear model, it was necessary to compile the regression to the mean estimates for all 10 proportion groups for the 132 combinations of the mean and standard deviation in the 4 main categories: group, mean, standard deviation and RTM percentage; the result of this effort was an "RTMDataset". Using the R statistical software (The R Foundation for Statistical Computing software, 2009) available from the website (http://www.r-project.org), the function for the linear model was invoked using the following command:

Call:

```
lm(formula = RTM. ~ STD + Mean + Group, data = RTMDataset)
```

This produced the following results:

```
Residuals:
Min         1Q          Median      3Q          Max
-0.167629   -0.030711   -0.006566   0.021365    0.234725


Coefficients:
            Estimate    Std. Error  t value     Pr(>|t|)
(Intercept) 0.4858226   0.0054974   88.37       <2e-16 ***
STD         -0.1322461  0.0046734   -28.30      <2e-16 ***
Mean        -0.0162959  0.0004281   -38.06      <2e-16 ***
Group       -0.2688363  0.0051452   -52.25      <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.05369 on 1316 degrees of freedom
Multiple R-squared: 0.791,      Adjusted R-squared: 0.7905
F-statistic:  1660 on 3 and 1316 DF, p-value: < 2.2e-16
```

Based on the T test, having T values with magnitudes much larger than 1.96 and corresponding p-values less than 0.001, the results for the model, i.e., the coefficients, are considered to be statically significant. Using these coefficients the function was developed. Based on the assumption that increasing the number of years of before data will increase the mean by the same factor, the number of years was incorporated into the function, which is shown in Equation 5-3.

$$\text{RTM}\,\% = \left(0.486 - 0.132\ \times \frac{\sigma}{\mu} - 0.0163\ \times \mu \times Y_n - 0.269\ \times \frac{p}{100}\right) \times 100 \qquad \textbf{Equation 5-3}$$

Where:

$\mu - Mean\ number\ of\ accidents\ per\ year\ of\ the\ dataset$

$\sigma - Standard\ Deviation\ of\ the\ number\ of\ accidents\ per\ year\ of\ the\ dataset$

$Y_n - \ Number\ of\ years\ of\ target\ data\ selected$

$p - Percentage\ of\ high\ accident\ sites\ selected\ from\ the\ entire\ dataset$

Note: Any negative RTM % is assumed no regression to the mean

5.6.2. Linear Model Verification

The linear model for the regression to the mean estimate was produced using simulated data, and therefore cannot be guaranteed to produce accurate results. To verify the linear model, the results for the RTM estimate determined by Section 5.2 for the California dataset was compared to the results produced by the regression to the mean estimate model (Equation 5-3). The regression to the mean estimates calculated using the actual observed collisions are shown in Table 5-12, while those produced from the predicted values using Equation 5-3 (with mean of 1.7 and standard deviation of 2.57 determined from the dataset) are shown in Table 5-13.

Table 5-12: Regression to the Mean Estimate based on real collision history observations

High Crash Sites		**Regression to the Mean Estimate Percentage (%)**							
Number of sites	Percent of total	1 Year	2 Years	3 Years	4 Years	5 Years	6 Years	7 Years	8 Years
5	2.45 %	27.49	21.77	16.04	15.08	14.12	15.38	11.09	0.00
10	4.9 %	23.44	18.72	14.00	12.89	11.77	12.30	9.14	0.00
20	9.8 %	21.68	14.59	11.84	10.53	9.39	9.50	7.22	0.00
31	15.2 %	21.70	12.94	10.18	8.13	7.12	5.58	5.21	0.00
41	20.1 %	21.18	12.08	7.77	5.38	4.68	3.34	3.34	0.00
51	25 %	19.43	11.07	7.89	5.61	4.22	3.55	3.06	0.00
102	50 %	13.32	8.48	6.24	4.59	3.56	2.96	1.72	0.00
153	75 %	3.88	3.95	3.39	2.09	2.26	1.47	1.15	0.00
204	100 %	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table 5-13: Regression to the Mean Estimate based on predicted values

High Crash Sites		**Regression to the Mean Estimate Percentage (%)**							
Number of sites	Percent of total	1 Year	2 Years	3 Years	4 Years	5 Years	6 Years	7 Years	8 Years
5	2.45 %	25.21	22.43	19.65	16.88	14.10	11.32	8.54	5.76
10	4.9 %	24.54	21.76	18.98	16.20	13.43	10.65	7.87	5.09
20	9.8 %	23.19	20.42	17.64	14.86	12.08	9.30	6.52	3.74
31	15.2 %	21.85	19.07	16.29	13.51	10.74	7.96	5.18	2.40
41	20.1 %	20.50	17.73	14.95	12.17	9.39	6.61	3.83	1.05
51	25 %	19.16	16.38	13.60	10.82	8.05	5.27	2.49	0.00
102	50 %	12.43	9.66	6.88	4.10	1.32	0.00	0.00	0.00
153	75 %	5.71	2.93	0.15	0.00	0.00	0.00	0.00	0.00
204	100 %	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

The difference between the RTM estimate percentages determined from the real collision history and those predicted using the RTM Estimate function are shown in Table 5-14. The patterns are very similar from a visual observation. It is observed that the major difference between the results occurs as a result of the limitation of the real dataset having a maximum of 8 years of data available. Recall that due to this limitation, it was assumed that there is no regression to the mean occurring when 8 years of data were selected. However, according to the function developed from the simulated data, there would still be some RTM present in the small proportions of high collision sites even with 8 years of data. On the whole, the maximum difference between any two regression to the mean percentage values with the same number of years and percentage of sites is only 7.2. Overall for the 72 RTM % values in the table, the average difference between the RTM % values of the two methods is only 1.4, which is fairly small considering the large amount of RTM bias that would be corrected.

Table 5-14: Difference between RTM estimate derived from real collision history and the predicted values

High Crash Sites		Regression to the Mean Estimate Percentage Difference (%)							
Number of sites	Percent of total	1 Year	2 Years	3 Years	4 Years	5 Years	6 Years	7 Years	8 Years
5	2.45 %	2.28	-0.67	-3.62	-1.79	0.03	4.06	2.55	-5.76
10	4.9 %	-1.10	-3.04	-4.98	-3.32	-1.65	1.66	1.28	-5.09
20	9.8 %	-1.52	-5.83	-5.80	-4.33	-2.69	0.20	0.70	-3.74
31	15.2 %	-0.15	-6.13	-6.11	-5.38	-3.62	-2.38	0.03	-2.40
41	20.1 %	0.68	-5.65	-7.18	-6.78	-4.71	-3.27	-0.49	-1.05
51	25 %	0.27	-5.31	-5.72	-5.22	-3.82	-1.72	0.58	0.00
102	50 %	0.89	-1.17	-0.64	0.49	2.24	2.96	1.72	0.00
153	75 %	-1.83	1.02	3.24	2.09	2.26	1.47	1.15	0.00
204	100 %	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

6. COLLISION MODIFICATION FACTOR CORRECTIONS FOR REGRESSION TO THE MEAN

As identified in the literature review, conducting improper before and after studies could result in a large RTM bias in the resulting collision modification factor. Having devised a method for estimating the regression to the mean estimate based on number of years of data selected, proportion of high accident sites, mean and standard deviation, it is possible to reverse calculate the CMF to adjust it according to the RTM estimate that should have been considered during an original, already published study that did not account for RTM. This mainly occurs when few years of data are used, potentially resulting in the expected number of collisions (mean of data) being higher than the true value, which would make lead to an overestimation of the reduction in collisions due to a given treatment. For example, if only 2 years of data are used, it may be concluded that an intersection should normally have 1.9 collisions per year when, in fact, using the long term average the true mean should be 1.7 collisions per year. Should a treatment be used at a site and it is determined that the number of collisions is now 1.5 collisions per year after the treatment, one would conclude that the percent reduction for that treatment is 21% ((1.9 – 1.5)/1.9) which is represented by a collision modification factor of 0.79, when in fact it is just an 11% ((1.7-1.5) /1.7) reduction, which should be represented by a collision modification factor of 0.89.

6.1. Derivation of Method

To correct the collision modification factor, it is first necessary to identify the method for which the CMF is developed. The CMF is derived from the formula in Equation 6-1.

$$\mathrm{CMF} = \frac{C_A}{C_B} \qquad \textbf{Equation 6-1}$$

Where:

$C_A -$ *is the number of collisions per year in the after period*

$C_B -$ *is the number of collisions per year in the before period*

The main concern with the RTM phenomenon is based on having an incorrect assessment of the true mean for the before period of crashes and using that mean as the estimate of crashes expected in the after period without the treatment. The corrected number of before collisions would therefore be calculated using the RTM percentage equation derived in Section 5.

$$\text{Corrected CMF} = \frac{C_A}{C_{BC}} \qquad \textbf{Equation 6-2}$$

Where $C_{BC} -$ *is the corrected number of before collisions per year*:

$$C_{BC} = C_B \times (1 - \mathrm{RTM}) \qquad \textbf{Equation 6-3}$$

Therefore:

$$\mathit{Corrected\ CMF} = \frac{CMF}{(1-RTM)} \qquad \textbf{Equation 6-4}$$

Alternately, correction factors can be derived to modify the CMFs by a simple multiplication.

$$\text{Correction factor} = \frac{1}{(1-\mathrm{RTM})} \qquad \textbf{Equation 6-5}$$

6.2. Correction Factors for the Real Dataset

From the RTM percentage estimates derived for the California intersection sites, a table can be produced to display correction factors using Equation 6-5 as shown in Table 6-1.

Table 6-1: Correction factors for Real Dataset

Percent of total high accident sites	Number of Years of Data Used							
	1 Year	2 Years	3 Years	4 Years	5 Years	6 Years	7 Years	8 Years Or more
2.5%	1.38	1.28	1.19	1.18	1.16	1.18	1.12	1.00
5%	1.31	1.23	1.16	1.15	1.13	1.14	1.10	1.00
10 %	1.28	1.17	1.13	1.12	1.10	1.11	1.08	1.00
15 %	1.28	1.15	1.11	1.09	1.08	1.06	1.05	1.00
20 %	1.27	1.14	1.08	1.06	1.05	1.03	1.03	1.00
25 %	1.24	1.12	1.09	1.06	1.04	1.04	1.03	1.00
50 %	1.15	1.09	1.07	1.05	1.04	1.03	1.02	1.00
75 %	1.04	1.04	1.04	1.02	1.02	1.01	1.01	1.00
100 %	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

These correction factors range from 1.0 for 8 years or more of site data used to a maximum of 1.28 for 2 years of site data collected. For 1 year of site data collected, the correction factor reaches a maximum of 1.38. Based on the fact that using 1 year of data can easily have crash counts for the highest collisions sites more than double the true mean, this high correction factor is not unreasonable. However, it should be noted that safety studies should not be undertaken based on only 1 year of data collection as the estimates of treatment effect would be highly volatile (i.e., they would have a high variance).

The Highway Safety Manual (HSM) process for correcting AMFs (CMFs) for regression to the mean assumes the values of X/B ratios range between 0.05 for a small RTM bias to 0.25 for a large RTM bias (NCHRP 17-27 Project Team, iTrans, 2007). This equates to CMF correction factors of 1.05 to 1.25, values that correspond to the results of the empirical analysis completed in this research, for which has correction factors range up to 1.28, with the exception of the single year correction factors. While these are similar, this only confirms the validity of the HSM correction values. The chart cannot be used for other datasets since the previous results showed that the RTM estimate varies for site populations having different means and variances.

6.3. Generalized Correction Factors Derived from Simulated Data

Based on the function (Equation 5-3) derived for estimating the regression to the mean percentage and the formula (Equation 6-4) for determining the correction factor, it is possible to correct any CMF, providing the mean and standard deviation of the number of collisions in the population from which the treated sites were selected are known.

$$\mathrm{RTM}\,\% = \left(0.486 - 0.132 \times \frac{\sigma}{\mu} - 0.0163 \times \mu \times Y_n - 0.269 \times \frac{p}{100}\right) \times 100$$

$$Corrected\ CMF = \frac{CMF}{(1 - RTM)}$$

For example: Using a before period of 2 years, a CMF based on treatment of the top 20% of sites of a dataset with a mean of 5.2 collisions per year and a standard deviation of 3.4 collisions per year, suppose that a simple before-and-after study resulted in a CMF with a value of 0.75. The RTM percentage and corrected CMF would be calculated as:

$$\mathrm{RTM}\,\% = \left(0.486 - 0.132 \times \frac{\mathbf{3.4}}{\mathbf{5.2}} - 0.0163 \times \mathbf{5.2} \times \mathbf{2} - 0.269 \times \frac{\mathbf{20}}{100}\right) \times 100 = 17.64\%$$

$$\text{Correction factor} = \frac{1}{(1-0.1764)} = 1.214$$

$$Corrected\ CMF = CMF \times Correction\ Factor = 0.75 \times 1.214 = 0.911$$

7. SUMMARY, CONCLUSIONS AND RECOMMENDATIONS

Based on the research presented in this thesis, there are a number of significant findings, conclusions and recommendations that can be drawn.

7.1. Summary

Based on the literature review, it was noted that there have been many methods developed to determine the number of collisions that are expected to occur at a specific location or group of locations. These were narrowed down and split into four main categories that include: a simple observational estimate, an Empirical Bayes Method of Moments approach, an Empirical Bayes Safety Performance Function approach and an average of the long term average of the recorded collision history. These methods were evaluated and compared, and resulted in the Empirical Bayes Safety Performance Function method being identified as best suited to estimate the expected number of collisions that would have occurred in the absence of treatment in a before-after evaluation of treatment effects. These methods were also used to estimate how the regression to the mean effect at high accident sites typically selected for treatment depends on the number of years and percentage of high accident sites selected from a dataset.

Based on the comparison between the regression to the mean results from the real and simulated datasets, it was noted that the regression to the mean also depends on the mean and standard deviation of the crashes in the population of sites in the dataset. To estimate the dependence on these factors, data simulations were performed for various combinations of means and standard deviations so as to eventually formulate a function for calculating the regression to the mean percentage. This could then be used as a correction factor for CMFs that have been developed in published research without accounting for regression to the mean.

7.2. Conclusions

Based on the experimental methodology completed in this thesis, the presence of the regression to the mean phenomenon is clearly confirmed and can significantly affect the results of safety treatment studies. The main concern with regression to the mean is based on having an incorrect assessment of the true mean for the before period of crashes and using that mean as the estimate of crashes expected in the after period without the treatment. As such, to properly account for regression to the mean, it is important to accurately determine the true mean for the before period. It is also concluded that the Empirical Bayes approach with the Safety Performance function is best for estimating the expected number of collisions for a before-and-after study to account for regression to the mean. Through the empirical exploration it was established that the magnitude of the regression to the mean percentage depends on four main factors, which are:

i) The mean number of collisions per year of the dataset that the sites are selected from. As the mean increases, the regression to the mean percentage decreases.

ii) The standard deviation of the number of collisions per year of the dataset that the sites are selected from. As the standard deviation of the dataset increases, the regression to the mean percentage decreases.

iii) The percentage of high collision sites selected from the total sites in the dataset. As the percentage of high collision sites from the total dataset increases, the regression to the mean percentage decreases.

iv) The number of years of collision data used or selected. As the number of years increases, the regression to the mean percentage decreases.

Using collision data simulations, it was possible to generate datasets with various combinations of means and standard deviations in order to determine how the regression to the mean is affected by these. A linear regression model was used to generate the following function for the regression to the mean estimate:

$$\text{RTM}\,\% = \left(0.486 - 0.132 \times \frac{\sigma}{\mu} - 0.0163 \times \mu \times Y_n - 0.269 \times \frac{p}{100}\right) \times 100$$

Where μ is the mean number of collisions in the population of sites from which the treatment sites are drawn, σ is the standard deviation related to this mean, Y_n is the number of years selected and p is the percentage of high accident sites of the site population from which the treated sites are drawn.

If a treatment evaluation study does not account for regression to the mean using the EB Method, it was found that the Highway Safety Manual (HSM) methodology is as good an estimate as any to correct the CMF to account for the RTM bias. The results from this analysis correspond to the range of the correction factors identified in the HSM for correcting CMFs that are subject to RTM bias. Further to this, the equation for correcting CMFs can be used in conjunction with the RTM Percentage function to determine an estimated correction factor to adjust the CMF, providing that the mean and standard deviation of the population of sites from which the treatment sites are drawn are known.

$$Corrected\ CMF = \frac{CMF}{(1 - RTM)}$$

An important conclusion reached is that CMFs developed from studies that ignore RTM will either require a large correction or a small one depending on the percentage of high collision sites used, the number of years used, and the mean and standard deviation of the population of sites from which the treatment sites are drawn. Depending on the magnitude of these corrections, a positive safety effect of a treatment may be completely negated or even turned into a negative effect. As such, it is crucial that all factors that could cause bias are either eliminated or are controlled at the beginning of the safety treatment study to ensure that the results are valid.

7.3. Recommendations

Based on this study, it is recommended that the use of Evidence Based Road Safety (EBRS) be promoted to develop more Collisions Modification Factors (CMFs) that can be used by transportation practitioners to better design and treat roadways. However, when doing this, it is important to always account for phenomenon such as regression to the mean which can severely affect the results produced by before-and-after studies. In order to counteract this, it is strongly recommended that the Empirical Bayes (EB) method with Safety Performance Functions be used to estimate the expected number of collisions that would have occurred in the after period in the absence of treatment. As identified by many papers, it is important that the EB method be applied properly; however there are many issues that can cause the EB method to produce inaccurate results, such as using a safety performance function that is not properly calibrated for the given dataset.

Where the EB method was not used, it is possible to apply the methodology identified in the Highway Safety Manual for correcting collision modification factors suspected of having RTM bias. Further to this, the function identified in this research for estimating the amount of regression to the mean can be used to develop a specific correction for the Collision Modification Factor. While these methods can improve the accuracy of a Collision Modification Factor that did not account for regression to the mean, it is important to note that these should not be used as substitutes for conducting before-and-after studies correctly with the Empirical Bayes method.

8. FURTHER STUDY

From the results attained from this research there are some areas that could benefit from further study. These include:

- Conducting the empirical analysis for the expected number of collisions and regression to the mean estimate using other datasets to verify that the trends do not vary significantly.

- Analyzing datasets for various location types to determine more typical mean and standard deviation values so as to expand the applicability of the function for estimating the amount of regression to the mean at different location types.

- Using additional datasets to calibrate a model for estimating the regression to the mean estimate and to determine whether there is any merit in using a more complex model than a simple linear model.

REFERENCES

AASHTO. (2010). *Highway Safety Manual.* Retrieved 2010, from http://www.highwaysafetymanual.org

Abbess, C., Jarrett, D. F., & Wright, C. C. (1981). Accidents at blackspots: Estimating the effectiveness of remedial treatment, with special reference to the "regression-to-mean" effect. *Traffic Engineering and Control , 22(10),* pp. 535-542.

Bahar, G. (2009). *Methodology for the Development and Inclusion of Accident Modification Factors (AMFs) in the First Edition of the Highway Safety Manual (Draft Version).*

Banks, J., Carson, J. S., Nelson, B. L., & Nicol, D. M. (2005). *Discrete-Event System Simulation* (Fourth Edition ed.). New Jersey: Pearson Prentice Hall.

Belluz, L., & Forbes, G. (2003). Synthesis of Safety For Traffic Operations. *Annual Conference of the Transportation Association of Canada.* St. John's Newfoundland and Labrador: Transportation Association of Canada.

Bonneson, J. A., & Pratt, M. P. (2008). Procedure for Developing Accident Modification Factors from Cross-Sectional Data. *Transportation Research Record No. 2083, Transportation Research Board* , pp. 40-48.

Council, F. M., & Mohamedshah, Y. M. (2007). *Guidebook for the California State Data Files - Highway Safety Information System.* US DOT. Washington, DC: Federal Highway Administration Office of Safety and Traffic Operations Research & Development.

Elvik, R. (2004). To What Extent Is There Bias by Selection? Selection for Road Safety Treatment in Norway. *Transportation Research Record No. 1897, Transportation Research Board* , pp. 200–205.

Forbes, G. (2003). *Synthesis of Safety For Traffic Operations.* Intus Road Safety Engineering Inc. Ottawa: Transport Canada.

Hauer, E. (1980). Bias-By-Selection: Overestimation of the Effectiveness of Safety Countermeasures Caused by the Process of Selection for Treatment. *Accident Analysis and Prevention , Vol 12*, pp. 113-117.

Hauer, E. (1997). *Observational Before/After Studies in Road Safety.* London: Pergamon Press.

Hauer, E. (1986). On The Estimation of the Expected Number of Accidents. *Accident Analysis and Prevention , Vol. 18, No.1*, pp. 1-12.

Hauer, E., Harwood, D. W., & Council, F. M. (2002). Estimating Safety by the Empirical Bayes Method, A Tutorial. *Transportation Research Record No. 1784, Transportation Research Board* , pp. 126-131.

Lord, D., & Bonneson, J. A. (2006). Role and Application of Accident Modification Factors Within Highway Design Process. *Transportation Research Record No. 1961, Transportation Research Board* , pp. 65–73.

Maher, M., & Mountain, L. (2009). The sensitivity of estimates of regression to the mean. *Accident Analysis and Prevention , 2009* (41), pp. 861–868.

National Cooperative Highway Research Program. (2008). *NCHRP Report 617: Accident Modification Factors for Traffic Engineering and ITS Improvements.* Washington, D.C.: Transportation Research Board.

NCHRP 17-27 Project Team, iTrans Consulting Inc. (2009). *Highway Safety Manual Knowledge Base Document.* Washington, D.C.: Transportation Research Board.

NCHRP 17-27 Project Team, iTrans. (2007). *Highway Safety Manual Knowledge Base Companion - Inclusion Process and Literature Review Procedure for Part D.* Washington DC: Transportation Research Board.

Persaud, B. N., Retting, R. A., Garder, P. E., & Lord, D. (2001). Safety Effects of Roundabout Conversions in the United States. *Transportation Research Record No. 1751, Transportation Research Board* , pp. 1-8.

Persaud, B., & Lyon, C. (2007). Empirical Bayes before–after safety studies: Lessons learned from two decades of experience and future directions. *Accident Analysis and Prevention , Vol 39*, pp. 546–555.

Shen, J., & Gan, A. (2003). Development of Crash Reduction Factors Methods, Problems, and Research Needs. *Transportation Research Record No. 1840, Transportation Research Board* , pp. 50-56.

The R Foundation for Statistical Computing software. (2009). R version 2.10.1. http://www.r-project.org/.

University of North Carolina Highway Safety Research Center. (2010). *Crash Modification Factors Clearinghouse.* (U.S. Department of Transportation Federal Highway Administration) Retrieved 2010, from http://www.cmfclearinghouse.org

Wright, C., Abbess, C., & Jarrett, D. (1988). Estimating the regression-to-mean effect associated with road accident black spot treatment: Towards a more realistic approach. *Accident Analysis & Prevention , Volume 20* (Issue 3), pp. 199-214.

APPENDIX

Appendix A: California Intersection Data - Four Legged Intersections with Stop on Minor

No	SITE	Average AADT		Total Crashes Per Site Per Year								
	ID	Major	Minor	2000	2001	2002	2003	2004	2005	2006	2007	Average
1	1274	3738	21	2	0	0	1	0	0	0	0	0.38
2	1582	14820	5280	0	0	0	4	4	4	1	6	2.38
3	1583	14833	2860	0	2	0	3	2	3	1	1	1.50
4	1584	14845	2750	0	0	0	2	1	0	2	2	0.88
5	1593	20213	625	2	1	1	11	5	3	0	6	3.63
6	1595	20213	1225	2	0	0	2	3	1	2	6	2.00
7	1745	11650	1600	3	0	1	0	0	1	2	0	0.88
8	1746	11650	3675	1	5	5	5	1	4	4	2	3.38
9	1747	11650	2200	4	3	1	1	0	2	4	1	2.00
10	1749	15025	2090	8	4	4	6	4	4	4	3	4.63
11	2956	6278	31	0	0	0	0	0	0	1	0	0.13
12	2958	6216	31	0	0	0	0	1	2	1	0	0.50
13	3091	3512	38	0	1	0	0	0	0	0	1	0.25
14	3208	4677	91	1	0	0	1	0	0	0	1	0.38
15	3209	5422	101	1	0	1	1	0	0	0	2	0.63
16	3210	6167	150	0	0	0	0	0	0	0	1	0.13
17	3211	7013	250	0	0	0	0	0	0	0	0	0.00
18	3212	6978	201	0	0	0	0	0	1	0	2	0.38
19	3213	6926	250	0	0	0	0	1	1	0	0	0.25
20	3214	6880	700	3	2	0	1	1	0	0	4	1.38
21	3215	6736	600	1	0	0	0	0	0	0	4	0.63
22	3216	6667	75	0	0	0	0	0	0	0	1	0.13
23	3690	15000	370	0	1	0	0	0	0	1	0	0.25
24	3691	15000	800	2	2	2	1	0	2	0	0	1.13
25	3733	5550	1750	0	0	0	0	0	0	0	0	0.00
26	3734	5550	300	0	0	0	0	0	0	0	0	0.00
27	4816	8175	751	2	0	0	1	1	0	2	0	0.75
28	4817	8175	371	2	1	2	0	1	1	0	0	0.88
29	4818	8175	2475	1	3	2	0	0	1	0	1	1.00
30	4820	4688	375	1	1	0	1	0	0	0	1	0.50
31	5338	19028	101	0	2	1	0	1	1	2	3	1.25
32	5579	23442	501	6	1	2	7	1	1	0	1	2.38
33	5580	23546	801	3	5	0	5	0	4	4	1	2.75
34	5582	23893	901	8	7	6	4	7	4	10	6	6.50
35	5586	24326	800	0	1	1	1	0	0	1	0	0.50
36	5720	12622	500	6	6	4	4	4	2	1	1	3.50
37	5721	13289	401	2	2	0	1	2	2	0	1	1.25
38	5723	13660	400	2	3	3	1	0	0	0	1	1.25
39	5724	13772	400	2	0	3	2	1	2	0	0	1.25
40	5731	14536	601	0	0	0	0	0	0	0	0	0.00
41	5741	17231	901	1	2	0	0	0	1	0	2	0.75
42	5743	17868	601	0	1	0	0	0	0	1	1	0.38
43	5747	19051	501	1	0	0	0	0	0	0	0	0.13
44	5749	19689	501	1	0	0	1	1	0	1	0	0.50
45	5753	20963	1300	0	0	0	1	2	2	1	0	0.75
46	5757	16711	401	0	1	0	0	0	0	0	1	0.25
47	5759	16971	601	0	1	1	1	0	0	0	0	0.38
48	5761	17231	500	0	0	1	0	0	1	0	0	0.25
49	5884	19888	301	0	1	1	1	0	0	0	0	0.38
50	5885	19888	451	1	2	1	1	2	2	1	0	1.25
51	5886	19888	301	0	0	0	1	1	0	0	0	0.25
52	5888	11306	501	2	2	0	1	0	1	1	0	0.88
53	5889	11849	1101	0	0	2	0	2	0	1	1	0.75
54	5890	12936	401	1	5	3	1	2	2	1	1	2.00
55	5892	14114	1601	1	1	0	1	1	1	2	1	1.00
56	6163	21720	1000	3	5	6	9	3	8	1	2	4.63
57	6369	20465	800	0	0	2	0	0	0	0	0	0.25
58	6376	17246	1800	0	0	0	0	0	0	0	1	0.13
59	6630	27073	700	2	1	2	3	1	0	2	0	1.38
60	7060	31628	1600	0	1	0	0	0	0	0	0	0.13
61	7280	28500	751	5	1	6	5	3	0	1	0	2.63
62	7302	21301	501	11	6	2	4	5	3	1	5	4.63
63	7304	21774	551	4	4	4	3	2	2	2	3	3.00
64	7859	14038	1501	4	5	0	1	0	4	2	1	2.13
65	7860	14038	1501	4	2	0	1	2	2	2	0	1.63
66	7982	6875	1050	3	0	0	2	0	0	1	0	0.75
67	8065	13425	2000	6	1	1	0	1	0	0	0	1.13
68	8066	13425	501	2	0	1	2	1	2	0	1	1.13
69	8068	21616	100	0	2	1	2	1	1	2	0	1.13

No	SITE	Average AADT		Total Crashes Per Site Per Year								
	ID	Major	Minor	2000	2001	2002	2003	2004	2005	2006	2007	Average
70	8221	31688	501	3	4	7	4	1	0	13	6	4.75
71	8222	31688	700	2	2	4	3	8	2	2	2	3.13
72	8223	31688	201	4	2	2	4	2	3	5	1	2.88
73	8227	31688	401	3	5	6	7	6	4	8	4	5.38
74	8229	31688	1800	1	2	2	2	2	4	3	4	2.50
75	8602	11713	401	0	1	1	1	1	1	1	3	1.13
76	8663	23650	601	4	11	4	4	2	5	2	3	4.38
77	8723	9213	1400	3	2	1	1	1	6	3	3	2.50
78	8810	5978	2000	0	3	0	0	0	0	0	0	0.38
79	8812	9025	910	2	0	1	0	0	0	0	0	0.38
80	8813	8903	1610	1	1	0	1	0	0	2	3	1.00
81	8814	8782	430	0	0	1	0	0	0	0	0	0.13
82	8815	8674	1250	0	1	0	0	0	0	0	0	0.13
83	8816	8593	170	0	0	0	0	0	0	0	0	0.00
84	8817	8525	150	0	0	1	0	0	0	1	0	0.25
85	8842	4125	1900	1	2	0	1	0	0	3	2	1.13
86	8896	9098	946	2	2	3	4	1	1	0	2	1.88
87	8897	9238	1375	2	1	0	3	2	2	1	3	1.75
88	8899	9501	814	1	0	2	1	0	0	0	0	0.50
89	8901	11832	929	2	0	0	0	2	2	2	1	1.13
90	8902	11909	528	1	1	2	0	0	2	0	2	1.00
91	9310	5429	680	0	1	0	0	1	1	0	0	0.38
92	9455	13098	2053	3	4	6	5	6	7	5	2	4.75
93	9458	13014	1119	4	6	2	6	12	3	3	5	5.13
94	9482	6888	493	2	3	1	2	1	3	1	4	2.13
95	9660	11647	2934	9	2	6	2	4	2	5	1	3.88
96	9661	11483	969	1	4	0	4	3	0	2	3	2.13
97	9662	11339	1190	2	1	3	1	4	0	0	2	1.63
98	9663	11175	1010	6	6	8	5	4	4	3	1	4.63
99	9664	11031	1249	3	6	5	1	4	3	3	3	3.50
100	9666	14936	3513	6	2	6	4	3	2	2	4	3.63
101	10180	7722	61	0	0	0	1	0	0	1	1	0.38
102	10181	8338	161	0	0	0	1	0	0	0	0	0.13
103	10182	8954	401	0	1	1	1	0	0	0	0	0.38
104	10183	9569	1601	0	2	0	1	0	0	0	0	0.38
105	10184	10184	2551	2	2	2	1	0	0	1	3	1.38
106	10185	10375	1850	2	2	1	2	1	0	0	4	1.50
107	10186	15675	500	1	0	0	2	0	0	1	2	0.75
108	10187	9465	1551	3	2	3	0	0	1	1	1	1.38
109	10188	9074	361	0	0	0	1	0	0	0	0	0.13
110	10189	8684	161	0	1	0	2	0	0	0	1	0.50
111	10282	32388	4500	1	2	1	0	0	0	2	3	1.13
112	10284	31735	1500	2	3	2	5	2	4	4	5	3.38
113	10417	7747	1050	0	0	4	0	0	0	0	0	0.50
114	10418	7768	1350	2	0	0	0	0	0	0	0	0.25
115	11218	28530	1801	2	4	3	2	0	1	0	3	1.88
116	11220	29589	1501	2	2	4	1	2	2	3	0	2.00
117	11221	30154	850	1	5	2	3	2	1	1	2	2.13
118	11223	31143	1101	3	4	0	7	2	3	2	2	2.88
119	11224	31637	1401	6	4	0	2	1	0	1	1	1.88
120	11229	35944	1700	1	2	7	3	3	3	3	8	3.75
121	11237	50196	1001	3	2	3	1	2	1	0	1	1.63
122	11351	45375	1940	5	1	0	0	0	0	3	2	1.38
123	11357	45375	1701	2	0	1	1	0	1	0	1	0.75
124	11614	29644	701	0	1	0	1	0	0	1	0	0.38
125	11615	29885	701	4	1	4	0	2	1	1	1	1.75
126	11617	30326	701	2	1	2	2	0	1	0	0	1.00
127	11618	30486	750	0	0	0	0	0	0	0	0	0.00
128	11619	30767	810	1	1	0	1	1	0	0	0	0.50
129	11620	31007	501	0	0	0	0	1	0	0	0	0.13
130	11621	31248	501	0	0	0	0	0	0	0	1	0.13
131	11622	31448	501	1	0	0	0	0	0	0	0	0.13
132	11623	31689	601	1	1	0	0	0	0	1	0	0.38
133	11624	31930	601	0	0	0	0	0	0	0	1	0.13
134	11625	32170	501	0	1	0	0	1	0	0	1	0.38
135	11683	14242	801	18	18	26	17	13	7	5	8	14.00
136	11830	28488	1201	8	5	8	6	5	3	1	0	4.50
137	11834	44188	1201	3	1	2	1	1	0	0	0	1.00
138	11835	44188	1201	3	1	5	5	3	4	4	1	3.25

No	SITE	Average AADT		Total Crashes Per Site Per Year								
	ID	Major	Minor	2000	2001	2002	2003	2004	2005	2006	2007	Average
139	12642	19485	11	0	0	0	0	1	3	0	0	0.50
140	12979	17393	63	2	2	1	0	2	2	4	1	1.75
141	12980	17328	11	4	2	7	1	3	2	2	3	3.00
142	12982	17198	11	1	1	1	0	1	1	1	0	0.75
143	13011	10943	630	1	0	0	0	0	1	1	1	0.50
144	13012	10852	81	2	4	2	1	3	2	1	0	1.88
145	13013	10754	241	8	5	4	7	2	0	2	3	3.88
146	13014	10660	550	1	0	4	0	1	0	0	0	0.75
147	13015	10563	50	1	2	1	1	2	0	2	0	1.13
148	13016	10471	221	1	2	1	1	1	0	1	3	1.25
149	13127	17795	701	0	0	0	0	0	0	0	0	0.00
150	13138	17088	71	0	0	0	0	0	0	1	0	0.13
151	13305	22813	31	1	1	1	2	7	1	4	10	3.38
152	13310	30531	2101	1	1	2	2	0	2	1	1	1.25
153	13311	30563	1901	0	1	3	0	3	1	2	1	1.38
154	13312	30594	1701	0	2	4	1	3	0	2	1	1.63
155	13743	16276	51	2	0	6	5	5	3	11	7	4.88
156	13744	15999	101	1	1	2	1	2	0	1	1	1.13
157	13793	12563	2000	7	3	5	2	5	2	2	5	3.88
158	14013	5519	501	3	1	2	5	1	0	0	0	1.50
159	14270	6817	530	1	1	1	1	1	1	1	3	1.25
160	14335	11642	101	3	5	1	2	1	1	6	0	2.38
161	14406	6461	154	3	0	0	1	1	3	4	0	1.50
162	14408	6156	619	2	2	2	1	0	1	0	1	1.13
163	14409	6156	401	0	0	1	0	0	1	0	0	0.25
164	14410	6156	350	2	1	0	1	1	1	0	1	0.88
165	14411	6156	836	0	1	0	1	0	1	1	1	0.63
166	14419	6189	82	0	0	0	0	1	0	0	1	0.25
167	14420	6190	164	0	0	0	0	0	0	0	0	0.00
168	14421	6192	726	0	0	1	0	0	0	0	0	0.13
169	14905	3100	150	1	0	0	0	0	1	0	0	0.25
170	14950	6785	501	3	0	2	5	0	1	1	0	1.50
171	15554	21050	330	2	2	6	1	7	6	4	8	4.50
172	15555	20487	1020	6	4	3	1	4	4	4	2	3.50
173	15556	19862	340	0	1	2	2	2	1	1	1	1.25
174	15557	19300	301	0	1	3	1	1	0	2	2	1.25
175	15723	25538	860	8	7	6	5	7	4	3	5	5.63
176	15725	25660	1201	4	11	10	14	8	10	6	6	8.63
177	15727	27961	1201	6	3	5	4	1	3	0	1	2.88
178	15888	17803	1200	6	6	2	7	8	7	9	12	7.13
179	16075	12015	301	1	0	0	2	0	0	0	1	0.50
180	16077	12149	301	0	1	0	1	0	0	0	0	0.25
181	16549	17656	1450	0	1	1	1	1	0	0	3	0.88
182	16550	16502	700	9	6	4	2	5	1	3	4	4.25
183	16551	15972	150	4	2	2	3	4	0	0	0	1.88
184	16763	24205	501	2	1	0	1	0	0	0	0	0.50
185	16764	24196	801	0	0	0	1	1	0	0	0	0.25
186	16765	23113	2301	0	1	1	0	0	1	2	0	0.63
187	17064	2938	570	0	0	0	0	0	0	0	0	0.00
188	17065	4163	150	0	0	0	0	0	0	0	0	0.00
189	17066	4442	340	0	0	0	0	0	0	0	0	0.00
190	17067	4000	590	0	0	0	0	0	0	0	0	0.00
191	17068	3751	250	0	0	0	0	0	0	0	0	0.00
192	17071	5881	1460	0	0	0	0	0	0	0	0	0.00
193	17072	6431	470	0	0	0	0	1	1	0	0	0.25
194	17073	6431	600	0	0	0	0	0	0	0	0	0.00
195	17075	7025	1100	1	1	2	0	1	0	0	0	0.63
196	17076	7025	1230	0	0	0	0	0	0	0	0	0.00
197	17077	7025	590	4	0	0	0	0	2	0	0	0.75
198	17095	4894	1150	1	1	1	1	1	0	0	0	0.63
199	17096	4894	1560	0	0	0	0	0	1	0	0	0.13
200	17330	40688	2001	11	11	7	8	11	4	9	10	8.88
201	17332	40688	1501	16	14	9	6	10	13	12	9	11.13
202	17333	40688	1501	10	19	13	17	9	9	9	8	11.75
203	17334	40688	1501	8	10	10	4	12	10	11	9	9.25
204	17336	40688	1501	3	2	7	3	4	5	1	6	3.88

Appendix B: Simulated Dataset

Site No	Total Crashes Per Site Per Year								
	Year 1	Year 2	Year 3	Year 4	Year 5	Year 6	Year 7	Year 8	Average
1	6	4	11	7	7	15	8	11	8.63
2	8	11	10	4	4	6	12	7	7.75
3	9	4	8	8	4	7	7	8	6 88
4	10	9	9	6	7	3	5	3	6.50
5	4	9	10	2	7	3	9	6	6.25
6	2	11	5	5	7	9	5	5	6.13
7	4	6	3	4	12	7	4	8	6.00
8	8	7	6	1	6	11	6	3	6 00
9	6	4	5	1	7	5	13	7	6.00
10	6	6	9	8	7	2	4	5	5.88
11	7	4	7	3	4	5	13	3	5.75
12	2	8	6	3	9	5	4	7	5.50
13	4	8	9	4	4	8	2	4	5.38
14	6	7	6	4	3	4	7	5	5.25
15	5	7	4	6	3	7	7	3	5.25
16	9	4	5	5	7	5	3	4	5.25
17	6	4	5	7	4	4	7	4	5.13
18	7	5	3	1	8	8	7	2	5.13
19	4	8	5	6	6	2	7	2	5.00
20	7	3	5	8	8	4	2	3	5.00
21	4	6	3	4	8	4	7	3	4.88
22	6	6	4	2	4	6	4	7	4.88
23	4	2	7	3	6	11	3	2	4.75
24	3	5	4	4	4	2	6	9	4.63
25	6	5	7	1	6	3	2	6	4.50
26	6	7	4	6	1	3	3	6	4.50
27	3	1	1	6	3	7	10	4	4.38
28	6	5	2	6	2	3	4	7	4.38
29	7	0	4	2	4	6	4	7	4.25
30	3	6	5	7	4	3	3	3	4.25
31	4	4	4	3	4	5	3	6	4.13
32	5	4	3	3	4	5	3	6	4.13
33	2	3	6	5	2	1	6	7	4.00
34	3	4	4	8	2	4	3	4	4.00
35	1	5	1	5	5	5	6	4	4.00
36	4	3	5	4	4	6	3	3	4.00
37	3	1	3	2	7	2	10	4	4 00
38	1	4	4	3	3	5	6	5	3.88
39	2	4	5	4	7	4	3	2	3.88
40	4	3	4	1	3	6	7	3	3.88
41	4	3	8	5	0	3	2	6	3.88
42	2	8	3	2	3	4	4	4	3.75
43	4	4	3	4	1	6	6	2	3.75
44	6	5	2	5	2	3	3	4	3.75
45	6	5	2	4	0	2	5	5	3.63
46	4	3	6	2	1	4	3	6	3.63
47	4	4	8	1	1	4	1	6	3.63
48	5	5	2	4	5	3	3	2	3.63
49	3	5	4	7	3	2	1	4	3.63
50	3	3	4	2	3	3	6	4	3 50

Site No	Total Crashes Per Site Per Year								
	Year 1	Year 2	Year 3	Year 4	Year 5	Year 6	Year 7	Year 8	Average
51	6	3	5	2	1	5	2	4	3.50
52	3	3	2	2	3	4	5	5	3.38
53	5	1	5	3	2	5	4	2	3.38
54	6	4	5	3	4	1	2	1	3.25
55	4	2	3	1	3	5	4	4	3.25
56	2	5	5	2	4	4	2	2	3.25
57	1	2	3	3	4	5	3	5	3.25
58	3	6	2	3	2	3	4	2	3.13
59	6	3	3	2	3	2	2	4	3.13
60	4	5	1	5	6	1	3	0	3.13
61	3	4	4	4	3	3	1	2	3.00
62	4	5	3	1	4	5	1	1	3.00
63	5	3	6	3	1	2	2	2	3.00
64	1	1	3	5	0	4	1	7	2.75
65	2	2	0	4	3	2	6	3	2.75
66	3	0	2	3	2	4	3	5	2.75
67	6	2	1	1	4	1	5	2	2.75
68	2	2	4	0	6	2	3	2	2.63
69	2	2	5	3	1	3	1	4	2.63
70	3	4	3	1	2	1	4	2	2.50
71	3	2	3	1	2	3	2	4	2.50
72	2	3	2	2	1	0	4	5	2.38
73	3	3	1	1	0	1	4	6	2.38
74	3	2	2	3	3	1	2	3	2.38
75	3	1	3	1	8	2	0	0	2.25
76	3	2	3	1	1	5	2	1	2.25
77	4	0	3	5	1	1	3	0	2.13
78	5	3	0	1	2	2	1	2	2.00
79	1	2	3	5	1	0	3	0	1.88
80	1	3	0	2	1	2	3	2	1.75
81	1	0	3	1	1	1	2	1	1.25
82	0	1	0	0	3	0	1	3	1.00
83	1	3	0	1	0	3	0	0	1.00
84	0	0	0	0	0	0	0	0	0.00
85	0	0	0	0	0	0	0	0	0.00
86	0	0	0	0	0	0	0	0	0.00
87	0	0	0	0	0	0	0	0	0.00
88	0	0	0	0	0	0	0	0	0.00
89	0	0	0	0	0	0	0	0	0.00
90	0	0	0	0	0	0	0	0	0.00
91	0	0	0	0	0	0	0	0	0.00
92	0	0	0	0	0	0	0	0	0.00
93	0	0	0	0	0	0	0	0	0.00
94	0	0	0	0	0	0	0	0	0.00
95	0	0	0	0	0	0	0	0	0.00
96	0	0	0	0	0	0	0	0	0.00
97	0	0	0	0	0	0	0	0	0.00
98	0	0	0	0	0	0	0	0	0.00
99	0	0	0	0	0	0	0	0	0.00
100	0	0	0	0	0	0	0	0	0.00