COMPARISON OF CRASH MODIFICATION FACTORS FOR ENGINEERING TREATMENTS ESTIMATED BY BEFORE-AFTER EMPIRACLE BAYES AND PROPENSITY SCORE METHODS

DRAFT FINAL REPORT

SOUTHEASTERN TRANSPORTATION CENTER

BO LAN AND RAGHAVAN SRINIVASAN

DECEMBER 2017

US DEPARTMENT OF TRANSPORTATION GRANT

DISCLAIMER

The contents of this report reflect the views of the authors, who are responsible for the facts and the accuracy of the information presented herein. This document is disseminated under the sponsorship of the Department of Transportation, University Transportation Centers Program, in the interest of information exchange. The U.S. Government assumes no liability for the contents or use thereof.

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.

4. Title and Subtitle
Comparison of Crash Modification Factors for Engineering Treatments Estimated by Before-After Empirical Bayes Methods and Propensity Score Methods

5. Report Date
December 2017

6. Source Organization Code
$ 50,000

7. Author(s)
Bo Lan; Raghavan Srinivasan

8. Source Organization Report No.
STC-2016-##-XX

9. Performing Organization Name and Address

Southeastern Transportation Center
UT Center for Transportation Research
309 Conference Center Building
Knoxville TN 37996-4133

10. Work Unit No. (TRAIS)

11. Contract or Grant No.
DTRT13-G-UTC34

12. Sponsoring Agency Name and Address

US Department of Transportation
Office of the Secretary of Transportation–Research
1200 New Jersey Avenue, SE
Washington, DC 20590

13. Type of Report and Period Covered
Final Report: March 2016 – December 2017

14. Sponsoring Agency Code
USDOT/OST-R/STC

15. Supplementary Notes:

16. Abstract

Cross-sectional and the empirical Bayes (EB) before-after are the two most common methods for estimating crash modification factors (CMFs). The EB before-after method has now been accepted as one way of addressing the potential bias due to RTM. However, the EB requires the before and after periods data and they may not be available. In those cases, researchers rely on cross-sectional studies to develop CMFs. One of the primary challenges of cross-sectional studies is it cannot address confounding issue, thus the estimated CMFs may be biased and unreliable.

The propensity score (PS) methods along with cross-sectional regression models is one of the methods that can be used to remove the confounding effects of such factors if they are measured in the data. Though the propensity score methods are widely used in epidemiology and other studies, there are only a few studies using the propensity score methods in CMF derivations in transportation safety.

The intent of this study is to evaluate and compare the performance of cross-sectional regression models that make use of propensity scores with the results from the EB and traditional cross-sectional methods. The cross-sectional method that make use of various propensity score methods were explored in this study. These methods were evaluated and compared with the traditional cross-sectional and the EB methods using two carefully selected simulated datasets. It was found the optimal propensity score distance (PSD) matching with maximum variable ratio of 5 performs best using the two datasets. It correctly identifies the true CMFs in the two datasets while the EB and the traditional cross-sectional methods failed.

17. Key Words
Negative binomial regression, Empirical Bayes Logistic regression, propensity score, logit of Propensity Score, Inverse Probability of Treatment Weight, Stabilized Inverse Probability of Treatment Weight, Standardized Mortality Ratio Weight, Nearest Neighbor Matching, Caliper width, Optimal Matching, Mahalanobis distance matching, Mixed-effects model

18. Distribution Statement

Unrestricted; Document is available to the public through the National Technical Information Service; Springfield, VT.

19. Security Classif. (of this report) Unclassified	20. Security Classif. (of this page) Unclassified	21. No. of Pages 22	22. Price …

Form DOT F 1700.7 (8-72) Reproduction of completed page authorized

TABLE OF CONTENTS

EXECUTIVE SUMMARY

A crash modification factor (CMF) is an estimate of the change in crashes expected after implementation of a countermeasure. Practitioners can use the CMF in quantifying safety in many ways including as part of the roadway management process, roadway safety audits, alternatives development and analysis, and design decisions and exceptions (FHWA, 2014).

The two common methods for estimating CMFs are: cross-sectional and the empirical Bayes (EB) before-after. The EB before-after method has now been accepted as one way of addressing the potential bias due to RTM. However, there are some treatments for which before-after studies may not be possible due to unavailability of data from the before period. In those cases, researchers rely on cross-sectional studies to develop CMFs. One of the primary challenges of cross-sectional studies is confounding which is sometimes due to systematic differences between the reference and treatment groups. In the presence of uncontrolled confounding, any obtained CMFs from the treatment group cannot be attributed solely to a causal effect of the countermeasures, and thus the estimated CMFs may be biased and unreliable.

Many statistical approaches can be used to remove the confounding effects of such factors if they are measured in the data. One such method is propensity score (PS) methods along with cross-sectional regression models. The propensity score is the probability of being assigned to the treatment group given the observed covariates. Though the propensity score methods are widely used in epidemiology and other studies, there are only a few studies using the propensity score methods in transportation safety.

The intent of this study is to evaluate and compare the performance of cross-sectional regression models that make use of propensity scores with the results from the EB and traditional cross-sectional methods. The cross-sectional method with various propensity score methods were explored in this study. These methods were evaluated and compared with the traditional cross-sectional and the EB methods using two carefully selected simulated datasets. The simulated data sets were designed such that the characteristics of the treated and reference sites were quite different, and the CMF estimates from the EB method were statistically different from the true CMF. The intent was to determine if the PS methods would outperform the EB method under these specific circumstances.

The explored propensity score methods including weighting, covariate adjustment, and the match methods. The weighting option incudes weighting by Inverse Probability of Treatment Weighting (IPTW), Stabilized Inverse Probability of Treatment Weighting (SIPTW), and Standardized Mortality Ratio Weighting (SMRW). The covariate adjustment options are propensity score, logit of propensity score (LPS), IPTW, SIPTW, and SMRW. Note that the last four covariate adjustment options have not been found previous studies. Furthermore, neither the weighting method nor the covariate adjustment method has been applied, evaluated, or explored in road safety studies.

It was found the optimal propensity score distance (PSD) matching with max. variable ratio of 5 and the nearest neighbor (NN) Mahalanobis distance (MD) matching with 1 replacement correctly identified the true effects, but the former has most number of the matched control sites and provides much better results. The important findings from this study are as following:

1. The NN MD matching with 1 replacement and the optimal matching by propensity score correctly identify the true effects.
2. The optimal PSD matching with max variable ratio of 5 has the most number of matched control sites and provide the best CMF estimates
3. The NN PSD matching with 5 replacements has the least number of matched control sites and is the worst method for CMF estimates
4. The optimal MD matching with max variable ratio of 5 does not perform as well as the NN MD matching with 5 replacements.
5. The mixed-effects has the better results than the NB models in terms of better estimate for the mean values and smaller stand errors.
6. Weighting by IPTW and SMRW as well as covariate adjustment by LPS, IPTW, and SIPTW generated similar or better CMFs than the EB method.

Based on the findings, we recommend the optimal PSD matching for the CMFs evaluation. However, we cannot conclude that this method will always outer perform the EB method. The weighting by IPTW and SMRW as well as the covariate adjustment by LPS, IPTW, and SIPTW are also suggested for further exploration using different simulated datasets.

1 BACKGROUND

A crash modification factor (CMF) is an estimate of the change in crashes expected after implementation of a countermeasure. Practitioners can use the CMF in quantifying safety in many ways including as part of the roadway management process, roadway safety audits, alternatives development and analysis, and design decisions and exceptions (FHWA, 2014). There are many ways to estimate the CMF associated with an engineering improvement. The methods for estimating CMFs can be divided into two broad categories: cross-sectional and before-after. Before-after studies include "all techniques by which one may study the safety effect of some change that has been implemented on a group of entities (road sections, intersections, drivers, vehicles, neighborhoods, etc.)" (Hauer, 1997, p. 2). On the other hand, cross-sectional studies include those where "one is comparing the safety of one group of entities having some common feature (say, STOP controlled intersections) to the safety of a different group of entities not having that feature (say, YIELD controlled intersections), in order to assess the safety effect of that feature (STOP versus YIELD signs)" (Hauer, 1997, p. 2, 3).
Many safety researchers feel that CMFs developed using cross-sectional studies may not always be reliable because cross-sectional models rarely represent causal relationships. The issues associated with the CMFs derived from cross-sectional models are discussed in some detail in Gross et al., (2010) and Carter et al., (2012).

There is some consensus in the safety research community that properly designed before-after studies provide more reliable estimates of CMFs. In before-after studies, the CMF is estimated based on two parameters: (1) crashes that occurred at the treated sites after the treatment is implemented, and (2) an estimate of the crashes that would have occurred during the same 'after' period had the treatment not been implemented, and the variance of this estimate. Often, sites are not selected for treatment at random; practitioners usually select high crash locations for treatment. This non-random selection can potentially lead to bias due to regression to the mean (RTM). Using the empirical Bayes before-after method has now been accepted as one way of addressing the potential bias due to RTM. However, there are some treatments for which before-after studies may not be possible due to unavailability of data from the before or after period. In those cases, researchers rely on cross-sectional studies to develop CMFs (Miaou and Lum, 1993; Persaud et al., 2009; Donnell et al., 2010; Donnell and Gross, 2011).

One of the primary challenges of cross-sectional studies is confounding which is sometimes due to systematic differences between the reference and treatment groups. In the presence of uncontrolled confounding, any obtained CMFs from the treatment group cannot be attributed solely to a causal effect of the countermeasures, and thus the estimated CMFs may be biased and unreliable. Confounding in road safety studies can arise from a variety of different reasons. The most common form of confounding arises from treatments based on some risk factors. The distributions of the risk factors in the treatment group may be completely different from those in the reference group. The observed difference will then be the result of both confounding and treatment choice, making it difficult to delineate the true effect of the treatment. The CMFs estimated by traditional cross-sectional methods could then be biased and unreliable.

Many statistical approaches can be used to remove the confounding effects of such factors if they are measured in the data. One such method is propensity score (PS) methods along with cross-sectional regression models (Austin, P. C., 2011). The propensity score is the probability of being assigned to the treatment group given the observed covariates. A propensity score can be easily estimated by logistic regression. The propensity score allows one to design and analyze an observational (nonrandomized) study so that it mimics some of the particular characteristics of a randomized controlled trial. It should be noted that the propensity score is a balancing score: conditional on the propensity score, the distribution of observed covariates will generally be similar between treated and untreated subjects. In other words, propensity score can be used to balance the measured covariates between the treatment and reference groups and thus overcome the confounding issue associated with traditional cross-sectional method. In fact, propensity score methods are the most popular causal inference methods for observational studies in epidemiology and etiology fields (Rosenbaum and Rubin, 1983).

There are four different propensity score methods used for removing the effects of confounding when estimating the CMFs: matching on the propensity score, stratification on the propensity score, inverse probability of treatment weighting using the propensity score, and covariate adjustment using the propensity score. Furthermore, there are a few options on how to use each of the above four propensity score methods.

Given the advantages of the propensity score methods, recently, researchers have started applying the propensity score approaches in connection with cross-sectional regression models for improving the reliability of CMFs that are estimated from cross-sectional studies. The next section will list a few studies using the propensity score methods in road safety study.

2 SUMMARY OF RECENT STUDIES USING THE PROPENSITY SCORE METHODS

Though the propensity score methods are widely used in epidemiology and other studies, there are only a few studies using the propensity score methods in CMF derivations in transportation safety. Below are some of the studies found in previous transportation safety studies.

Donnell et al. explored the propensity score method using a dataset that was part of a study on the safety evaluation of shoulder rumble strips (SRS) and centerline rumble strips (CLRS) applications (Donnell et al, 2017). The data set included 334 treatment sites as well as 13,286 reference sites. First, the study compared the CMFs for the five crash types including the total, injury, runoff road, head-on, and the sideswipe-opposite direction crashes obtained from the EB and the three propensity score matching methods, where the reference sites were matched with treatment sites by propensity score 1-1 matching, propensity score 5-1 matching, and propensity score 10-1 matching with replacement. For all crash types, the 95% confidence intervals of the CMFs from all the four methods include the value of 1, indicating there is no significant treatment effects for the two treatments. Additionally, it was found the variances from the propensity score methods are much bigger than those from the EB method except for the total crashes. They further compared the EB and the same propensity score matching methods by adding a simulated covariate and a designated CMF to the original data set. They concluded the propensity score matching method is better than the EB method in that the CMFs from the propensity score methods are closer to the predefined true CMF. This study did not evaluate the traditional cross-sectional method.

Sasidharan and Donnell (2013) applied the propensity score method to estimate effectiveness of installation of lighting at the intersections. 6464 intersections were available for analysis and about 42% of the intersections contained some form of roadway lighting. Four years (1999–2002) of crash and corresponding roadway inventory data were used in the analysis. The nearest neighbor (NN) propensity score matching, Mahalanobis distance (MD) matching, and stratification of propensity score with regression were used to estimate the CMFs. The results from the three methods indicate that fixed roadway lighting reduces expected nighttime crashes, but does not significantly reduce daytime crashes.

Wood et al (wood et al, 2015) used the propensity score methods to estimate the safety effects of lane widths on urban streets in Nebraska. Matching was performed using both Nearest Neighbor and Mahalanobis matching techniques. The database consisted of ten years' of crash data (2003– 2012) at mid-block segments on urban streets in four Nebraska cities, totaling 18,227 observations (segment-years). CMFs for the target crash types (sideswipe same direction and sideswipe opposite-direction) were estimated using mixed-effects negative binomial or poisson regression with the matched data. They concluded that CMFs for target crash types (sideswipe same direction and sideswipe opposite-direction) were consistent with the values currently used in the Highway Safety Manual (HSM).

Wood and Donnell (Wood and Donnell, 2016) evaluated the CMFs of the continuous green T (CGT) intersections using both the nearest neighbor and genetic matching methods. The matching criteria was Mahalanobis distance. The crash data from 2008 to 2012 in Florida (FL) and from 2009 to 2013 in South Carolina (SC) were used for the analysis. The data included 30 CGT and 38 comparison sites from Florida and 16 CGT and 21 reference sites from South Carolina. The studied crash types include total, fatal and injury, and target crash (rear-end, angle, and sideswipe). No significant treatment effects were identified for all the studied crash types.

A comparison CMFs study of EB, propensity score, and traditional cross-sectional methods were conducted for total and run-off-road crashes by Wood et al (Wood et al, 2015). The sample was small in that there were only 57 treatment sites and 147 reference sites. Data from 1999 to 2008 obtained from Georgia were used to conduct the study. The data set consists of 57 rural road segments that received a Safety Edge treatment and 147 untreated segments. Years 2006-2008 were the after periods for the treatment group and they were used for the cross-sectional studies with or without propensity score matching while the whole period (1999-2008) were used for the EB comparison study. The caliper-based 1:1 nearest neighbor matching without replacement method (Austin 2011) were used. Their results show that the cross-sectional and propensity score based methods yielded similar CMFs when compared to the EB methods. However, the EB estimates had smaller standard errors than either of the traditional or propensity score based cross-sectional methods. Additionally, since the estimated CMFs were very close to 1, it is unclear if these findings are transferable to situations where the CMFs are different from 1.0. It is likely that depending on a particular dataset, EB before-after methods may underperform or perform better than propensity score methods.

2.1 Study Purpose

The intent of this study is to evaluate and compare the performance of cross-sectional regression models that make use of propensity scores with the results from the EB and traditional cross-sectional methods. To this end, a careful study design was performed, and various propensity score related methods were explored. The authors explored the possibility of using real data to compare the different methods. However, as evident from Wood et al (2015), the results may depend on a particular data set. Hence, the decision was made to use simulated date. The details are described in the later sections. Note that the terms, reference sites, comparison sites, or control sites are used interchangeably throughout the report.

The report is organized into eight sections. The first part describes the background of CMFs studies in road safety and the need for a modified cross-sectional study. The second section lists the recent studies using the propensity score methods in CMFs evaluations and provides the purpose of this study. The third section introduces the different statistical methods that were evaluated in this study. Section 4 describes the study design. The next section provides information on the approach used in simulation. Section 6 provides the results from the study. Section 7 provides the discussion and conclusions.

3 METHODS EXPLORED

The two most popular methods for the CMFs evaluations, which are traditional cross-sectional (CS) models and the EB method, were used to compare the performance of the cross-sectional models that utilize the propensity score methods. The three methods are described below:

3.1 The EB method

In the EB approach (Hauer, 1997, Hauer et al., 2002), the change in safety for a given crash type at a location is given by:

$$\lambda - \pi \qquad (1)$$

where λ is the expected number of crashes that would have occurred in the after period without treatment and π is the number of reported crashes in the after period.

In estimating λ, the effects of regression to the mean and changes in traffic volume are explicitly accounted for using safety performance functions (SPFs) relating crashes to traffic flow and other relevant factors (SPFs are typically estimated using data from a reference group of untreated sites).

Annual SPF multipliers were calibrated to account for the temporal effects due to change in weather, demography, crash reporting and so on.

In the EB procedure, the SPF is used to first estimate the number of crashes that would be expected in each year of the before period at locations with traffic volumes and other characteristics similar to the one being analyzed. The sum of these annual SPF estimates (P) is then combined with the count of crashes (x) in the before period at a treatment site to obtain an estimate of the expected number of crashes (m) before treatment. This estimate of m is:

$$m = w_1(x) + w_2(P), \tag{2}$$

where the weights w_1 and w_2 are estimated from the mean and variance of the SPF estimate as:

$$w_1 = P/(P + 1/k) \tag{3}$$

$$w_2 = 1/k(P + 1/k), \tag{4}$$

where k is a constant for a given model and is estimated from the SPF calibration process with the use of a maximum likelihood procedure. (In that process, a negative binomial distributed error structure is assumed with k being the dispersion parameter of this distribution.)

A factor is then applied to m to account for the length of the after period and differences in traffic volumes between the before and after periods. This factor is the sum of the annual SPF predictions for the after period divided by P, the sum of these predictions for the before period. The result, after applying this factor, is an estimate of λ. The procedure also produces an estimate of the variance of λ, the expected number of crashes that would have occurred in the after period without treatment.

The estimate of λ is then summed over all sites in a treatment group of interest (to obtain λ_{sum}) and compared with the count of crashes during the after period in that group (π_{sum}). The variance of λ is also summed over all sections in the treatment group.

The CMF (θ) is estimated as:

$$\theta = (\pi_{sum}/\lambda_{sum}) \ / \ \{1 + [Var(\lambda_{sum})/\lambda_{sum}^2]\}. \tag{5}$$

The standard deviation of θ is given by:

$$Stddev(\theta) = [\theta^2\{[Var(\pi_{sum})/\pi_{sum}^2] + [Var(\lambda_{sum})/\lambda_{sum}^2]\} \ / \ [1 + Var(\lambda_{sum})/\lambda_{sum}^2]^2]^{0.5} \tag{6}$$

Generalized linear modeling is typically used to estimate the required reference group SPF using negative binomial regression. The negative binomial dispersion parameter, k, is also estimated in this process.

3.2 Cross-Sectional Modeling

Similar to the SPF developments in the EB method, crashes are normally assumed to follow a Negative Binominal (NB) distribution to address the overdispersion nature in the crash counts. In this approach the treatment group and the untreated group are pooled and a dummy variable indicating reference/treatment group is included in the models. The coefficient for this dummy variable is an indicator the effectiveness of a treatment.

Let's use intersection signalization treatment as an example to describe how to derive CMF from the crash models. Suppose a prediction model for intersection crashes is as below:

$$\lambda_{i,t} = \alpha \bullet ma\_aadt_{i,t}{}^{\beta_1} mi\_aadt_{i,t}{}^{\beta_2} \exp(\beta_3 Z + \varepsilon_i) \qquad (7)$$

and,

$Y_{i,t}$ = observed number of crashes at site i in year t

$\lambda_{i,t}$ = expected number of crashes at intersection i in year t

$ma\_aadt_{i,t}$ = AADT on major road at site t in year t

$mi\_aadt_{i,t}$ = AADT on minor road at site i in year t

$\alpha = Intercept$

Z=Dummy variable, Z=1 for treatment site, and Z=0 for control site

$\beta_1, \beta_2, \beta_3$ =Coefficients for $ma\_aadt_{i,t}$, $mi\_aadt_{i,t}$, and Z, respectively

The CMF for signalization treatment is exp(β_3). Since the cross-sectional method is not able to address the confounding issue, it may provide biased estimates. Interested readers can see Gross et al., (2010) and Carter et al., (2012) for more details on this issue.

3.3 Cross-sectional with the Propensity Score Methods

Treatment selection is often influenced by subject characteristics, i.e., signalization treatment is normally implemented at stop controlled intersections with high traffic volumes. As a result, characteristics of treated sites often differ systematically from those of untreated sites. Therefore, one must account for systematic differences in characteristics between treated and untreated sites when estimating the effect of the treatment on expected crashes. Additionally, high traffic volume also causes high crashes and it is a confounder which affects the selection of the treatment and the outcome (crashes). A traditional cross-sectional method is not able to address this confounding issue and a modified cross-sectional method is needed. The propensity score method is one of the methods to reduce or eliminate the effects of confounding when using observational data.

A few popular propensity score methods including propensity score weighting, propensity score covariate, and propensity score matching were explored in this study. For the weighting or covariate methods, propensity score as well as a few variables derived from the propensity score can be used as a weight or covariate when developing the NB models. Similar to the traditional cross-sectional (CS) method, all the sites are used to develop the models by these two methods. For the matching method, there are a few ways to match a control site to a treated site in terms of how and which measure(s) should be used. Unlike the above weighting and covariate adjustment propensity score methods, the CMFs are derived from the models developed from the matched data - the unmatched treatment or control sites are excluded in the modeling development. Therefore, the data from the propensity score matching methods are usually smaller than the original dataset for the CMFs development.

3.3.1 Matching

Matching is the most common propensity score method, which involves assembling the treatment sites and the control sites with similar or identical propensity scores or covariates. CMFs can then be estimated from the matched sample. The analysis of the matched samples can then approximate that of a randomized trial by directly comparing outcomes between treatment sites and control sites which did not receive any treatment, using methods that account for the paired nature of the data such as mixed-effect NB or Poisson models based on the matched ID. Many statistical software packages provide options for estimating mixed effects models, e.g., SAS Glimmix procedure can be used to estimate mixed-effects models.

3.3.1.1 Matching distances

The first step for matching is to select the matching distance, in other words, which measure(s) should be used to match a control to a treated site. A few matching distance measures such as Mahalanobis Distance (MD) as well as differences in propensity score or Logit of propensity score (LPS) between a treatment and a control sites can be used. The propensity score distance (PSD) and Mahalanobis Distance (MD) were used in this study.

3.3.1.1a Propensity Score Distance

The propensity score was defined by Rosenbaum and Rubin (1983) to be the probability of treatment assignment ($Z_i = 1$) at site i conditional on observed covariates: $\Pr(Z_i = 1)\,|\boldsymbol{X_i}$. Where $\boldsymbol{X_i}$ is a set of covariates for site i. It can be seen that the propensity score is a balancing score: conditional on the propensity score, the distribution of measured baseline covariates is similar between treated and untreated entities. Thus, in a set of entities all of whom have the same propensity score, the distribution of observed baseline covariates will be the same between the treated and untreated entities (Austin, 2011).

In this study, the propensity score for a treatment was estimated using a binary logistic regression shown below:

$$\Pr(Z_i = 1)\,|\boldsymbol{X_i} = \frac{exp(\boldsymbol{\beta X_i})}{\boldsymbol{1}+exp(\boldsymbol{\beta X_i})} \qquad (8)$$

Where,
$\Pr(Z_i = 1)\,|\boldsymbol{X_i}$ is the propensity score for a treatment at site i;
$\boldsymbol{X_i}$ is a set of covariates for site i.
$\boldsymbol{\beta}$ is the vector of parameters associated with covariates $\boldsymbol{X_i}$;

The absolute value of difference in propensity score is defined as propensity score distance (PSD) and were used to match treatment site i and control j in this study.

$$PSD_{ij} = \left|\Pr(Z_i = 1) - \Pr(Z_j = 0)\right| \qquad (9)$$

3.3.1.1b Mahalanobis Distance

Mahalanobis Distance (MD) measures the distance between the two observations Xi in a treatment group and Xj in a control group. The MD is calculated using the following equation.

$$MD_{ij} = (\boldsymbol{X_i} - \boldsymbol{X_j})' \textstyle\sum^{-1} (\boldsymbol{X_i} - \boldsymbol{X_j}) \qquad (10)$$

Where Xi−Xj is the matrix of the differences in values between treatment site i and control site j for the variables included in the MD calculation, and Σ is the covariance matrix of **X** between the two groups. It is intuitive to include all the significant variables from the logistic regression for the propensity score calculation for the MD calculation. In addition to all the significant covariates, propensity score were also included in the MD calculation in this study. Note that the MD matching measure works best with continuous variables.

3.3.1.2 Matching Methods

Matching is the most popular method that makes use of the propensity score. NN match with caliper option and the optimal matching are the two most popular and useful matching methods, and most statistics software have procedures to implement these two matching methods. Additionally NN match also has options for matching with and without replacement. PSMATCH procedure in SAS 14.2 was used to explore these two matching methods in this study.

3.3.1.2a Nearest Neighbor Matching

NN matching (Rubin, 1973) method is also called Greedy NN matching as it uses a "greedy" algorithm, which cycles through each treated unit one at a time, selecting the available control unit with the smallest distance to the treated site in terms of PSD or MD respectively in this study. The algorithm makes "best" matches first and "next-best" matches next, in a hierarchical sequence until no more matches can be made.

1:1 vs k:1 matching

In its simplest form, 1:1 nearest neighbor matching selects the control site j with the smallest distance for each treated site i. The advantage for 1:1 matching is that the matched sets are more similar compared to k:1 NN matching. k:1 NN matching allows k controls to be matched to the same treated sites. One issue regarding 1:1 matching is that it can discard a large number of observations and thus may lead to reduced precision for the CMF estimates. On the other hand, while k:1 NN matching increases the matched sample set, it reduces the similarity between the matched treated and control groups. In other words, selecting the number of matches involves a bias-variance trade-off. Selecting multiple controls for each treated site will generally increase bias since the 2nd, 3rd, and 4th closest matches are further away from the treated site compared to the 1st closest match. On the other hand, utilizing multiple matches can decrease variance due to the larger matched sample size. An additional concern with k:1 NN matching is that, without any restrictions, it can lead to some poor matches, if for example, there are no control sites with propensity scores similar to a given treated site.

Caliper width

One strategy to avoid poor matches is to impose a caliper and only select a match if it is within the caliper width. Best matches are those with the lowest matching distances with the range of the caliper width. This can lead to difficulties in interpreting effects if many treated sites do not receive a match, but can help avoid poor matches. For more details, see Rosenbaum and Rubin (1985).

With or without replacement

One way to resolve the issue when many treated sites do not receive a match is matching with replacements, indicating one where control sites are matched to multiple treated sites based on the aforementioned criteria. Matching with replacement can often decrease bias because controls that look similar to many treated individuals can be used multiple times. This is particularly helpful in settings where there are few control individuals comparable to the treated individuals (e.g., Dehejia and Wahba, 1999). However, inference becomes more complex when matching with replacement, because the matched controls are no longer independent. It is also possible that the estimate of the treatment effect will be based on just a small number of controls when matching with replacement.

3.3.1.2b Optimal matching

One issue of the NN matching is that the order in which the treated subjects are matched may change the quality of the matches. Optimal matching avoids this issue by taking into account the overall set of matches when choosing individual matches, minimizing a global distance measure (Rosenbaum, 2002). Generally, NN matching performs poorly when there is intense competition for controls, and performs well when there is little competition (Gu and Rosenbaum, 1993).

In SAS PSMATCH procedure, all matches are selected simultaneously and without replacement to minimize the total absolute difference in propensity score or MD across all matches. Maximum variable ratio allows a specified maximum number of control sites to be matched to each treated site to achieve a minimal global PSD or MD. Note that this option uses much more computer memory compared to the NN matching.

For this study, both PSD and MD were used as matching distance in both NN matching and optimal matching. Additionally, in NN matching, 1:1, 5:1, and 10:1 with replacement(s) were also explored, and PS caliper width was set to be 0.25. In optimal matching, the maximum variable ratio was set to be 5 to ensure larger matched samples.

3.3.1.3 CMFs Development using the Propensity Score Matching Methods

For the cross-sectional with the propensity score matching methods, the CMFs can be derived in two ways:

1. The CMFs can be developed in the same way as the traditional CS method using the NB model. Instead of the whole dataset, the matched datasets are used
2. Since the matched dataset includes matched treated and control sites. it may be more appropriate to use a method specifically for the matched data, where treatment effects are estimated after controlling the effects in each matched control sites. Mixed-effects model is one of the options for this purpose, where matched ID is used for the random effects. The Equation for the mixed-effects model is:

$$\lambda_{i,t} = \alpha \bullet ma\_aadt_{i,t}{}^{\beta_1} mi\_aadt_{i,t}{}^{\beta_2} \exp(\beta_3 Z + \varepsilon_{matched\_ID}) \qquad (11)$$

It can be seen Equation (11) is similar to Equation (7) for the traditional CS. In mixed-effects model, instead of a site level error ε_i, a matched pair error $\varepsilon_{matched\_ID}$ is used indicating each matched pair share the same error.

Both methods were used for the CMFs development. Poisson models were also used if the NB models could not be estimated due to possible convergence issues.

3.3.2 Weighting Using the Propensity Score

Unlike the matching methods, this method uses all of the sites to derive CMFs. In this method, propensity scores are used to calculate statistical weights for each site. The following three measures derived from propensity score can be used as a weight to balance the distributions of measured covariates between the treatment and reference groups so that the treatment assignment is independent of the measured covariates.

Inverse Probability of Treatment Weighting (IPTW)

Inverse probability of treatment weighting was first proposed by Rosenbaum (1987) as a form of model-based direct standardization. In this method, a site's weight is equal to the inverse of the probability of receiving the treatment that the site actually received. IPTW can be calculated as following:

For treatment sites, IPTW=1/PS; and
For reference sites, IPTW=1/(1-PS)

Stabilized Inverse Probability of Treatment Weighting (SIPTW)

The above IPTW may be inaccurate or unstable for sites with a very low probability of receiving the treatment received, as a small propensity score will receive a huge weight and vice versa. To address this issue, the use of stabilizing weights, which is called Stabilized Inverse Probability of Treatment Weight (SIPTW), has been proposed (Robins, Hernan, & Brumback, 2000). The equations to calculate SIPTW is as below:

For treatment sites, IPTW=PSm/PS; and
For reference sites, IPTW=(1-PSm)/(1-PS).

Where PSm is the mean value of the propensity score in the dataset.
From the above equations, SIPTW can be seen as a smoothed IPTW by the mean value of propensity score such that there are no extremely big or small values of the weight.

Standardized Mortality Ratio Weighting (SMRW)
Standardization is a way to validly summarize treatment effects in the presence of treatment effect heterogeneity. Note that matching implicitly standardizes the estimate to the treated population. "Standardizing" to the treated population can also be achieved using standardized mortality/morbidity ratio weights (SMRW) (Sato and Matsuyama, 2003, Stürmer et al, 2014). These weights create a pseudo-population of the untreated, which has the same covariate distribution as the treated. The SMRW can be defined as:

For treatment sites, SMRW=1; and
For untreated reference sites, SMRW= PS/(1-PS).

For the cross-sectional model with the propensity score weighting methods, CMFs can be derived in the same way as in the traditional CS method, after including a weight option during the estimation.

3.3.3 Covariate Adjustment Using the Propensity Score

For this approach, the propensity score is included as a covariate in the model development. Thus CMFs associated with the treatment can be estimated while adjusting for the probability of receiving that treatment. This is one of the commonly used propensity score method in the medical literature (Weitzen et al, 2004, Shah et al., 2005, Stürmer et al., 2006). Implicit in the use of this method is that effect of treatment is being compared between treated and untreated subjects with the same propensity score. This method assumes that treated and untreated sites with the same propensity score have the same distribution of measured variables (Austin, 2009). Some researchers believe this method allows the investigator to estimate the outcome associated with the treatment while adjusting for the probability of receiving that treatment, thus reducing confounding (Haukoos and Lewis, 2015), while other researchers argue that covariate adjustment does not allow for balancing of covariates across treated and control groups as achieved with other propensity score methods. They believe this method cannot properly account for confounding issue and thus may provide biased results (Garrido, 2016, Austin, 2009, and Hadea and Lu, 2014). Austin (Austin, 2009) conducted an evaluation of the propensity score methods using empirical and simulated clinical data. He found this method provided biased results associated with treatment effect estimation when the propensity score is used as a covariate in nonlinear regression models, such as logistic regression and Cox proportional hazards models. Hadea and Lu (2014) found this method provided biased results even using a linear model.

Though only PS is used as a covariate in the literature, we think that those three weighting options in the above PS weighting method, if used as covariates, could also be used to balance the distributions of the measured covariates between the treated and control groups and thus improve the CMF evaluation. Additionally, the logit of Propensity Score (LPS) which is defined as log(PS/(1-PS) could also be used for this purpose. Thus, five measures including PS, LPS, IPTW, SIPTW, and SMRW were explored and included as a predictor in Equation (7), respectively, and a log linear relationship to the expected crashes was assumed in this study.

To the authors' knowledge, this propensity score weighting and covariate adjustment methods have not been evaluated in crash studies. Additionally, there is no literature on the covariate method using the above four measures (LPS, IPTW, SIPTW, SMRW). For this reason, this method was evaluated by using propensity score, IPTW, SIPTW, SMRW, and LPS as a predictor respectively.

Similarly, CMFs can be calculated in the same way as in the traditional CS method. Instead of using the weighting option in the PS weighting method, PS is used as a covariate when developing the NB models.

4. STUDY DESIGN

As aforementioned, the objective of this study is to evaluate the above propensity score methods and compared to the CMFs from the EB and traditional CS methods. To this end, an appropriate study design is needed.

Simulated datasets have the advantage since the true CMFs are known, and the different methods can be evaluated based on difference between the estimated and the true CMFs. The best method can be identified based on the difference between estimated CMF and the true CMF. In order to consider a realistic scenario, the study design should allow for distribution of the major covariates to be significantly different between the treated and control groups, and in this case, the traditional cross-sectional method may provide biased CMF estimates. In addition, it is possible significant differences between reference and treatment groups may provide a biased estimates from the EB before-after method. On the other hand, the reference group should have enough sites that can be matched to the treated sites so that it is possible to implement and evaluate the matching methods.

We also wanted to use different true CMF values and dispersion parameters reported in previous studies when generating the crash counts at each site. Additionally, the simulated annual average daily traffic volumes (AADTs) and crash counts should be realistic. We finally decided to simulate urban stopped intersections as control sites and signalization as a countermeasure. In order to evaluate the EB before after method, before-after data are needed. We decided to generate 11 years of data and assumed that a signalization treatment occurred in year 6 at some of the intersections with high traffic volumes on the major road (ma_aadt), such that there were 5 years before and after periods, respectively, for estimated the CMFs using the EB before-after method. The 5 years of after period data was selected for evaluating the traditional cross-sectional as well as the PS methods.

As mentioned earlier, we need to have some control sites with similar distributions of the major covariates as the treated group so that the matching methods can be applied. Thus, part of the top ranked intersections were set to be part of the control sites. The majority of the control sites, however, were from the intersections with lower ma_aadts. In this way, we hope there are some control sites which can be used for matching and there is a significant difference between the treated and control sites since the majority of the control sites were from the intersections with much lower ma_aadts.

After assigning treatment and control sites and following Donnell et al. (2017), a confounder was then assigned to each treated and control site with different distributions in the two groups. In order to simulate realistic traffic volumes on major and minor roads as well as crash counts, a literature review was conducted for recent studies on SPFs developments for urban stop-controlled intersections. The crash counts were then generated for each site assuming a negative binomial distribution using a SPF similar to those found in previous studies.

After the datasets were generated, the summary statistics of the variables in the SPFs and crash counts by treatment/ control groups were compared to ensure that was a difference in the two groups. The generated datasets were then used for the EB and the traditional CS analysis. The final datasets were then used to conduct the propensity score methods evaluation. The below has the details on how to generate the simulated datasets.

5. SIMULATED DATA

In deriving the simulated data, it was assumed, as is common, that the crash count over "similar" sites follows a negative binomial distribution (NBD). The NBD may be derived by "heterogenous Poisson sampling" which assumes that the crash count Y_i at a site over time is Poisson distributed with unknown mean λ_i per unit of time at site i and that these means λ_i follow a Gamma distribution over similar sites, such that

$$E(Y) = E(\lambda) \text{ and}$$

$$Var(Y) = E(\lambda) + E^2(\lambda)/\varphi \tag{12}$$

Where, φ is the dispersion parameter of the NBD.

The data used to examine the propensity score methods were generated from a Poisson-Gamma distribution (Lord, 2006; Lan et al, 2009). The simulation framework for the stop-controlled intersection dataset used for the study is as follows:

Step 1: For year 1, randomly generate entering traffic volumes on the major road (5000 ~ 50,000 ma_aadt) which follows a truncated normal distribution with mean of 20,000 and standard deviation of 6000. Similarly generate aadt on minor road (500 ~ 5,000 mi_aadt) with mean of 2,000 and standard deviation of 600. The number of sites was set to be 100,000 to ensure that control sites are significantly different from the treated sites in terms of ma_aadts as well as there are enough sites for the evaluation after identifying subsets of the data.

Step 2: ma_aadts for the remanig 10 years were generated with random variation (within 5%), such that most of the traffic volumes would be around the mean value 20,000 AADT.

Similarly, traffic volumes were generated randomly on the minor road in the range of 500 ~ 5000 mi_aadt across 11 years with random variation (within 5%).

Step 3: Sort the ma_aadt from high to low. The top 3000, 5000, and 10000, sites were selected. Then we randomly assigned top 20% -5% of the above high ma_aadt sites to be treated sites. The treated sites are 470, 492, 1026, and 2017 respectively. The remaining sites with high ma_aadts were partially assigned to the control group. The number of the assigned control sites with high ma_aadts was about 2-6 times the treated sites.

Step 4: Since in practice the treatment group is quite different from a reference group, we wanted to identify control sites that were significantly different from the treated sites in terms of the average ma_aadt. The 20,000 – 30,000 sites with lowest ma_aadts in the initially generated 100,000 sites were thus selected to be in the control group.

Step 5: The treated and control sites with 11 years' ma_aadts and mi_aadts data were generated after Step 4. We then generated a confounding variable v_w with different distribution between the treated and control groups.

$$v\_w|Z \sim N(50*(1+0.5Z),(10*(1+0.5Z))^2) \quad (13)$$

Where,
Z=Dummy variable, Z=1 for treatment site, and Z=0 for control site

Step 6: It is assumed the treatment was implemented in year 6. Year 1 to year 5 is before period and years 7- 11 is after period. CMFs were applied to treated sites in the after period when generating the crash counts from the NB models from the following model.

$$\lambda_{i,t} = \alpha \times ma\_aadt_{i,t}{}^{\beta_1} mi\_aadt_{i,t}{}^{\beta_2} \exp(\beta_4 v\_w + \varepsilon_i) \times (1 + CMF \times Z_{t>T_0}) \quad (14)$$

Where,

β_4 =parameter associated with the confounding variable V,

$Z_{t>T_0}$= dummy variable, $Z_{t>T_0}$=1 for treated sites when year > treatment year $T_{0;}$

Otherwise $Z_{t>T_0}$=0.

Safety performance function (SPF) parameters α, β_1, β_2 were developed from the CMFs in California state *(Bhim, 2006)* and Virginia states (Garber and Rivera, 2010). The SPF used was:

$$\lambda_{i,t} = 0.00004 \times ma\_aadt_{i,t}{}^{0.6191} mi\_aadt_{i,t}{}^{0.4813} \exp(0.015\text{v\_w} + \varepsilon_i) \times (1 + CMF \times Z_{t>T_0}) \quad (15)$$

Where CMF=1.3, 1.0, 1.15, 0.85 respectively for the four datasets.

Step 7: Calculate the expected number of crashes $\lambda_{i,t}$ for intersection i across 11 years from the above SPF (Equation 15).

Step 8: Generate a scale factor δ_i from a Gamma distribution with the mean equal to 1 and the dispersion parameter φ: $\delta_i \sim gamma(\varphi, \frac{1}{\varphi})$. It is necessary to use the parameterization of the gamma distribution $\delta \sim gamma(a, b)$ when its mean and variance are defined as $E(\delta) = ab$ and $\mathrm{Var}(\delta) = ab^2$, respectively. It can be shown that when $E(\delta) = 1$ and $\mathrm{Var}(\delta) = 1/\varphi$ (where $a = \varphi$ and $b=1/\varphi$), the Poisson-gamma function gives rise to a NB distribution with $\mathrm{Var}(Y) =\mu+\mu^2/\varphi$ (Lord 2006, Cameron and Trivedi 1998).

Step 9: Calculate the modified mean $\mu_{i,t} = \lambda_{i,t} \times \delta_i$

Step 10: Generate a discrete value $Y_{i,t}$ for the observed count at intersection i in year t from a Poisson distribution with mean $\mu_{i,t}$.

It is worth mentioning that the simulation was performed to generate the 4 datasets, with dispersion parameters φ of 0.5, 1.0, 1.5, and 2.0, respectively, to reflect the range of typical values reported on relevant studies. CMFs of 0.85, 1.0, 1.15, and 1.30 were also applied so that each simulated dataset has a different CMF value.

For the generated datasets, the EB evaluation and traditional CS were conducted first. The results showed that both methods almost perfectly estimated the CMFs from the simulated datasets, probably due to there are enough sites that are similar to the treated sites in the control group. From these data sets, treated sites were further subset based on the site-level naïve CMFs while the control sites stayed the same. The final simulated datasets were selected such that the characteristics of the treated and control sets were significantly different which led to the estimated CMFs from the EB method to be significantly different from the true CMF – the intent was to determine whether the PS method would provide an estimated CMF that was closer to the true CMF compared to the EB method. Since the two datasets with the larger number of treated sites were not possible to implement for the optimal matching method due to the computer memory limits, the CMFs from the other methods from these two large datasets are not included in the report. The final two simulated datasets and their summary statistics are described in Tables 1 and 2.

Table 1 summary statistics for simulated dataset 1

(True CMF=1.30 and dispersion parameter=1/1.5)

Control/Treated	Variable	Min	Max	Mean	Std Dev
Control group 21000 sites (1000 sites with high ma_aadt)	confounding variable V	8.25	90.58	50.06	10
	ma_aadt	11219	47528	15712	4807
	ma_aadt	485	5273	1963	597
	Crashes/site.year	0	12.91	1.29	3.99
	ma_aadt - before	10122	47818	15790	4659
	mi_aadt - before	491	4956	1963	591
	ma_aadt - after	12018	51087	15636	5002

	mi_aadt - after	462	5560	1963	608
	Crashes/site.year - before	0	12.20	1.30	2.83
	Crashes/site.year - after	0	14.60	1.28	2.81
Treated group 386 sites	confounding variable V	26.05	117.53	76.39	14.61
	ma_aadt - before	26782	43712	31854	2671
	mi_aadt - before	589	4507	2126	684
	ma_aadt - after	30469	46245	33138	2423
	mi_aadt - after	556	4341	2124	696
	Crashes/site.year - before	0	16	3.19	6.14
	Crashes/site.year - after	1	19	4.65	8.02

Table 2 summary statistics for simulated dataset 2

(True CMF=1.00 and dispersion parameter=2)

Control/Treated	Variable	Minimum	Maximum	Mean	Std Dev
Control group 23000 sites (3000 sites with high ma_aadt)	confounding variable V	11.3	89.74	49.97	9.94
	ma_aadt	8998	47528	16049	7115
	ma_aadt	487	5273	1967	618
	Crashes/site.year	0	31.73	1.28	6.61
	ma_aadt - before	8185	47818.00	16063.40	6891.79
	mi_aadt - before	496	4956.00	1967.03	612.05
	ma_aadt - after	9727	51087.00	16035.29	7369.87
	mi_aadt - after	462	5560.00	1966.55	628.78
	Crashes/site.year - before	0	31.60	1.29	4.53
	Crashes/site.year - after	0	34.00	1.28	4.55
Treated group 343 sites	confounding variable V	31.63	115.94	74.91	14.07
	ma_aadt - before	24439	40561	30202	2733
	mi_aadt - before	681	3860	2104	623
	ma_aadt - after	28057	41851	31136	2601
	mi_aadt - after	654	3887	2109	641
	Crashes/site.year - before	0	33.60	3.35	11.06
	Crashes/site.year - after	0	36.80	3.00	10.94

It can be seen the ma_aadt and average crashes per site per year in the treated group are higher than those in the reference group as expected. This phenomenon is normal in real data.

6. EVALUATION RESULTS

The two simulated datasets were explored using the EB, traditional CS, and the various propensity score methods as aforementioned. As mentioned earlier, the after period (year 7 to year 11) was used for the traditional CS and the propensity score methods while the period from year 1 to year 11 was used for the EB evaluation. It is worth mentioning that the caliper width of propensity score for the NN matching by propensity score or MD was set to 0.25 in this study. The NN matching was explored using propensity score with 5 replacements and MD match using 1, 5, and 10 replacements respectively. The optimal matching by propensity score and MD was also investigated, respectively, and the maximum variable ratio was set to be 5.

6.1 CMFs from Dataset 1

There are 386 treated sites and 21,000 reference sites dataset 1 and the true CMF=1.30. Among the 21000 reference sites, there are 1000 sites with ma_aadts similar to those in the treated sites and 20,000 sites with much lower ma_aadts. The CMFs in terms of mean and 95% confidence limits by various methods are shown in Table 3.

If the true CMF 1.30 is located between the 95% confidence limits of the estimated CMFs, we concluded that the method correctly identifies the true effects and gave a "Yes" in Table 3. However, if 1.3 is near the upper of lower 95% limits, to be conservative, we still give a "No".

For the propensity score covariate adjustment method, each of the propensity score covariates is insignificant and less significant than the dummy variable Z, and hence, each of them was excluded from the SPFs, and the CMFs from this method is identical to the CMF from the traditional CS.

For the methods where the CMFs were calculated from the SPFs developed using all the sites, only the weighting method using SIPTW correctly identifies the true CMF, all other methods including the EB and traditional CS methods fail.

Table 3 Comparison of CMFs for Dataset 1 (True CMF=1.30)

mean	95% confidence limits		method	model	Control Sites	Treated sites	Treatment Effects Identified?
1.427	1.382	1.473	EB	NB	21000	386	No
1.399	1.328	1.473	Traditional Cross-Sectional	NB	21000	386	No
1.385	**1.328**	**1.443**	weight=IPTW	NB	21000	386	No
1.258	**1.107**	**1.429**	Weight=SIPTW	NB	21000	386	Yes
1.415	1.372	1.460	weight=SMRW	NB	21000	386	No
1.399	*1.328*	*1.473*	covariate adjustment (propensity score)	NB	21000	386	No
1.399	*1.328*	*1.473*	covariate adjustment (LPS)	NB	21000	386	No
1.399	*1.328*	*1.473*	covariate adjustment (IPTW)	NB	21000	386	No
1.399	*1.328*	*1.473*	covariate adjustment (SIPTW)	NB	21000	386	No
1.399	*1.328*	*1.473*	covariate adjustment (SMRW)	NB	21000	386	No
0.674	0.521	0.872	NN matching by PSD (5 replacements)	NB	10	371	No
0.602	0.515	0.703	NN matching by PSD (5 replacements)	mixed-effects Poisson	10	371	No
1.391	1.232	1.571	NN matching by MD (1 replacement)	NB	58	362	Yes
1.289	**1.170**	**1.420**	NN matching by MD (5 replacements)	NB	198	362	Yes
1.262	**1.158**	**1.375**	NN matching by MD (5 replacements)	mixed-effects NB	198	362	Yes
1.328	**1.212**	**1.454**	NN matching by MD (10 replacements)	NB	313	362	Yes
1.302	**1.233**	**1.375**	NN matching by MD (10 replacements)	mixed-effects Poisson	313	362	Yes
1.331	**1.240**	**1.428**	optimal matching by PSD (max variable ratio=5)	NB	1158	386	Yes
1.358	**1.296**	**1.423**	optimal matching by PSD (max variable ratio=5)	mixed-effects Poisson	1158	386	Yes
1.364	**1.276**	**1.459**	optimal matching by MD (max variable ratio=5)	NB	1158	386	Yes
1.463	1.391	1.538	optimal matching by MD (max variable ratio=5)	mixed-effects Poisson	1158	386	No

For the matching methods, the NN matching using propensity score with 5 replacements has the worst results because only 10 control sites were repeatedly matched to the 371 treated sites. All other matching methods, except optimal matching by MD using mixed effects model, correctly identify the true effects. It can be seen the CMFs from the mixed-effects models are generally consistent with from the NB models, but with a lower standard errors. The lower standard error means a more stable estimates. Thus the mixed-effects model is deemed to be a better option for the CMF estimation for these matching methods. The CMFs from the mixed-effects Poisson models are listed in the Table where mixed-effects NB models are not available due to convergence issue.

Figures 1 and 2 illustrate the pattern of the standardized mean differences (SMD) for the propensity score, LPS, confounding variable v_w, and ma_aadt matched by the NN propensity score and MD with 5 replacements, respectively, where the SMD is computed by dividing the difference (Treated - Control) in the means of the variable in the two groups by an estimate of the standard deviation. SMD can be used to show the quality of the matching - the closer to 0 the absolute value of SMD, the better balanced in the treated and control groups. Note that mi_aadt is not significant in the logistic regression, and is not shown here.

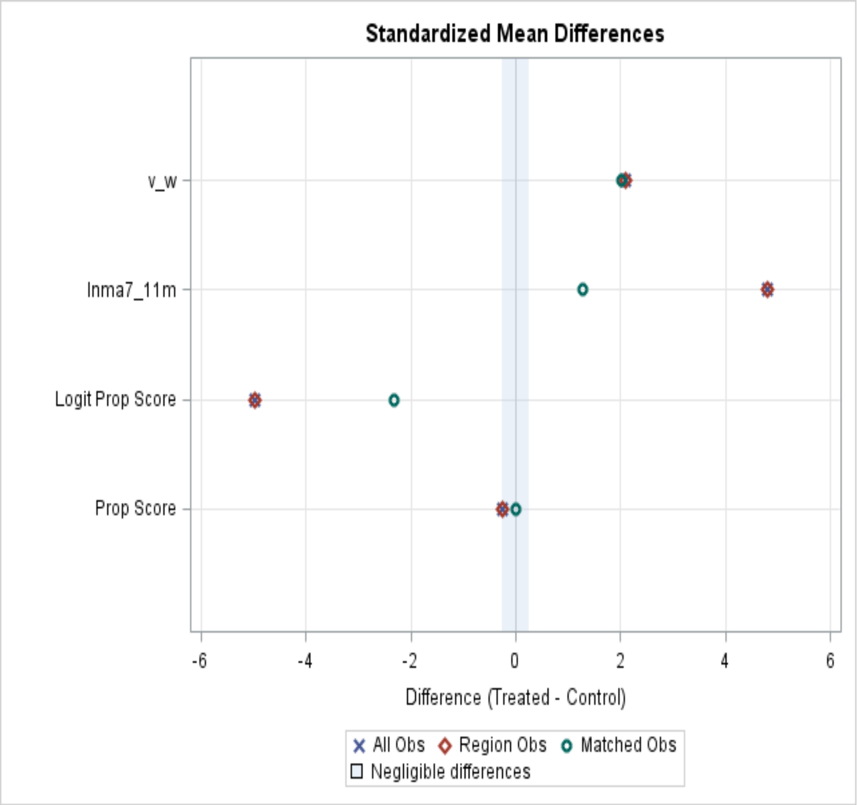

Figure 1 Standardized Mean Differences from NN matching by PSD
(5 replacements, dataset 1 with true CMF=1.30)

"All Obs" in the legends in Figure 1 stands for all the data, "Region Obs" is the common region data indicating only those observations whose propensity score line in the common support region be used for calculation. The common support region is also the largest interval that contains propensity scores for subjects in both groups. "Matched Obs" refers to the matched data.

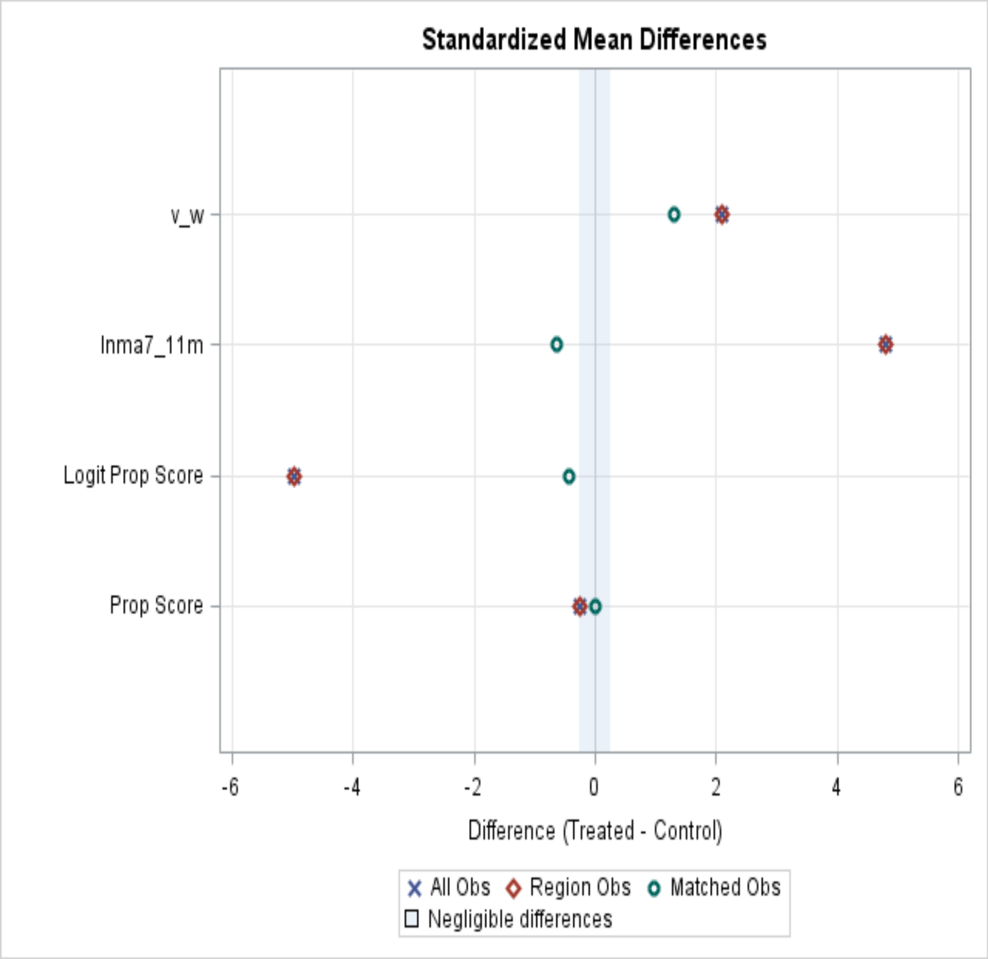

Figure 2 Standardized Mean Differences from NN matching by MD
(5 replacements, dataset 1 with true CMF=1.30)

For the NN matching by PSD in Figure 1, the SMD of zero for the propensity score in the matched data indicates propensity score is perfectly matched. SMD for the LPS and lnma7_11m which is the logarithm value of average ma_aadt in the after period, are much smaller than those in the original or the common region data indicating these variables are better balanced in the matched data. The SMD for the confounding variable v_w is almost the same as before matching, this is expected as we purposely implemented the different means for v_w into the two groups.

In Figure 2, for the NN matching by MD with 5 replacements, the absolute value of the SMDs are much smaller than those in Figure 1, meaning the variables in the treated and control groups are much better balanced compared to the NN matching by PSD.

Note that zero of SMD in Figures 1 and 2 cannot be interpreted as a perfect match as it is only about the mean difference. To better understand the balance of the variables in the matched data, we can take a look at the boxplots of the variables from the matched, common region, and the whole dataset. The boxplots of v_w and lnma7_11m are shown in Figures 3 to 6. It can be seen the balance of these two variables is much more improved by NN matching using the MD. Additionally, lnma7_11m in Figure 5 by the NN using PSD has a much wider range in the matched control group and a narrower range in the treated group indicating a poor match. Since ma_aadt is a major factor associated with the crashes, this method has the worst results because of the poor match in ma_aadt and only 10 control sites in the matched data.

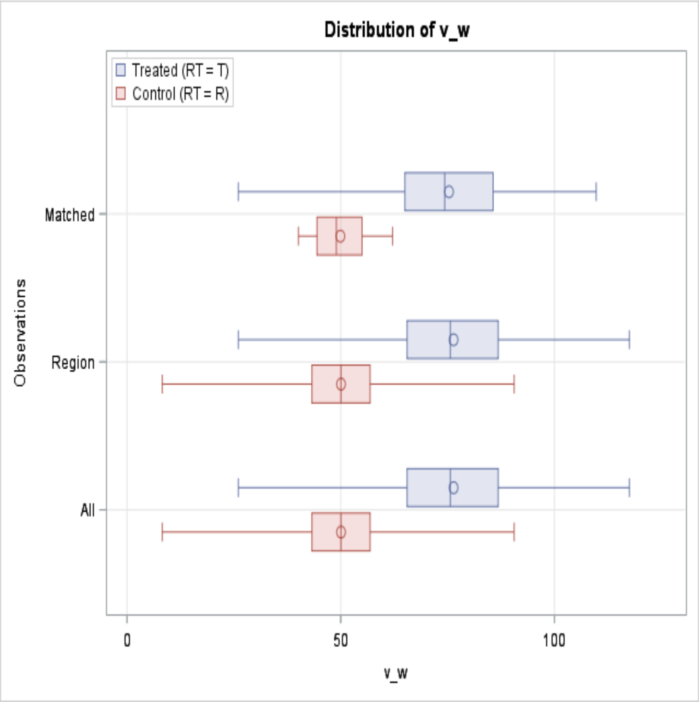

Figure 3 Distribution of confounding variable v_w from NN matching by PSD (5 replacements, dataset 1 with true CMF=1.30)

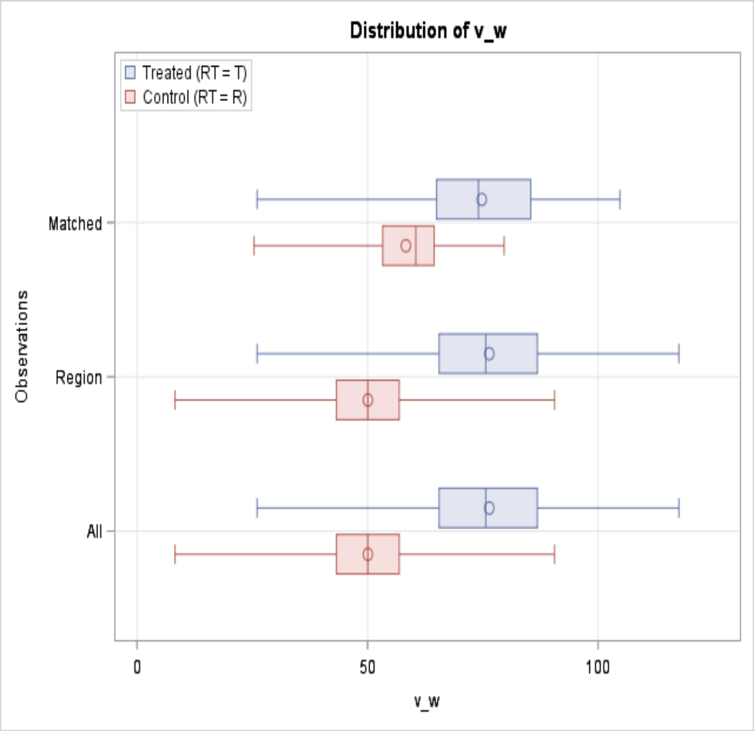

Figure 4 Distribution of confounding variable v_w from NN matching by MD
(5 replacements, dataset 1 with true CMF=1.30)

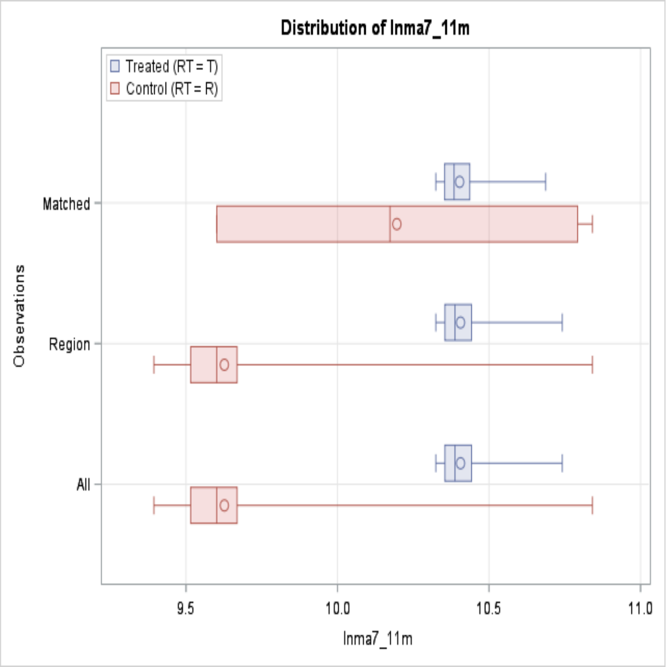

Figure 5 Distribution of lnma7_11m from NN matching by PSD
(5 replacements, dataset 1 with true CMF=1.30)

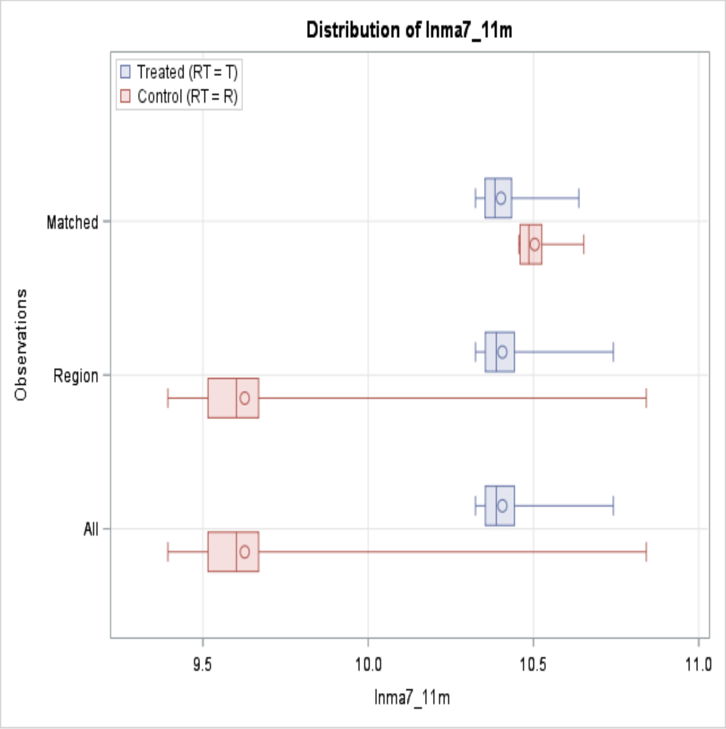

Figure 6 Distribution of lnma7_11m from NN matching by MD

(5 replacements, dataset 1 with true CMF=1.30)

From Table 3, surprisingly, the CMFs by the optimal matching using MD is not as good as the CMFs by the optimal matching using PSD. For the later method, the objective function is to have a minimal absolute difference of propensity score, however, the point of the SMD of propensity score is not located near the zero line as the two NN matching methods. The reason for that is, unlike RMSE, the SMD is just about the mean difference. Although the SMD graph is not as good as the one by the NN MD method, the boxplots in Figures 8 to 9 show the balance for variables v_w and lnma7_11 are as good as those by the NN MD. The number of matched control sites by this method is 1158 while it is just 198 by the NN MD method. A larger sample size of the control group allows a smaller standard error of CMF as shown in Table 3. Given the similar CMF mean vales by the two methods, the optimal matching by PSD is better due to a smaller standard error of the CMF estimates.

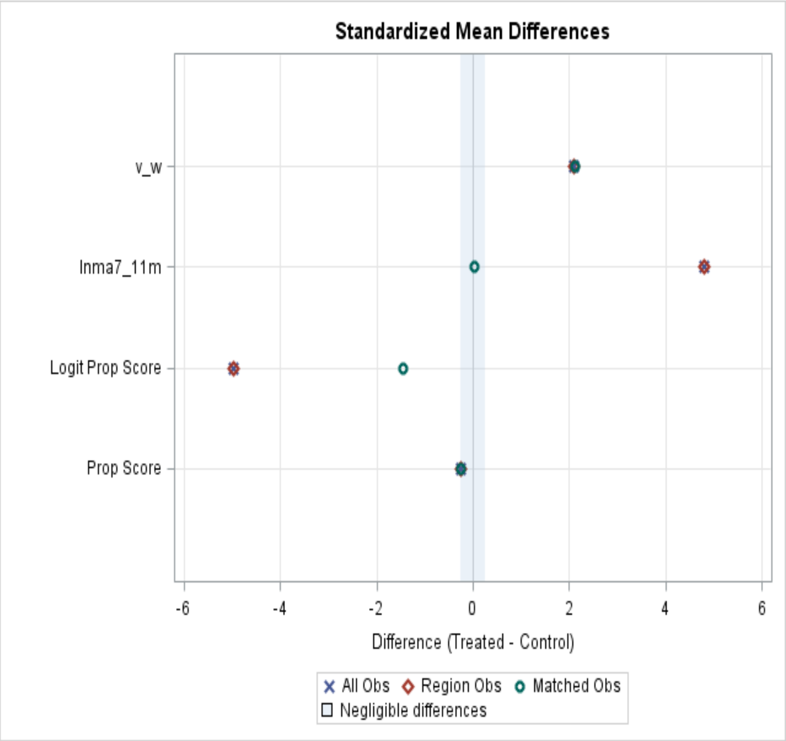

Figure 7 Standardized Mean Differences from Optimal matching by PSD
(max. variable ratio=5, dataset 1 with true CMF=1.30)

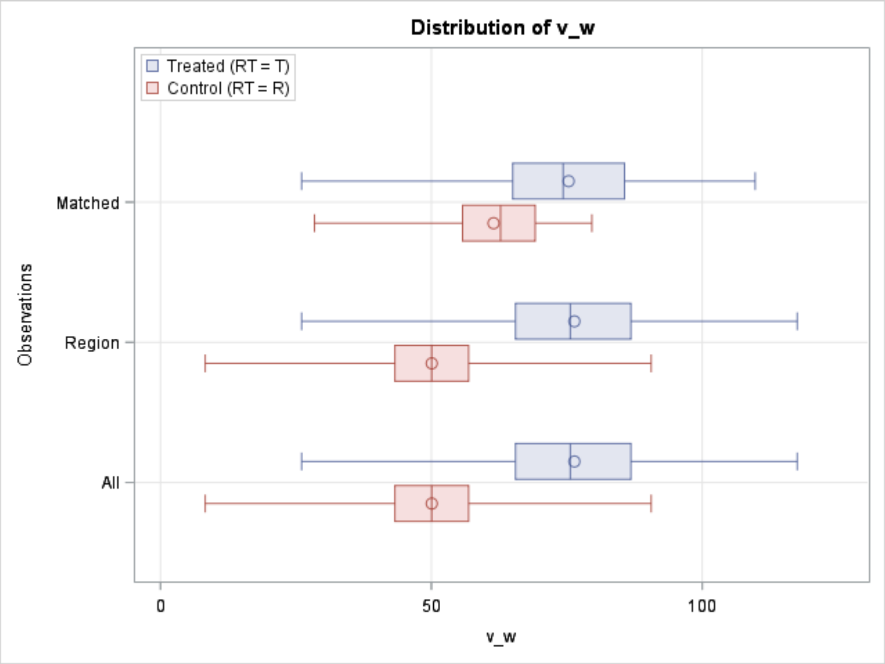

Figure 8 Distribution of confounding variable v_w from Optimal matching by propensity score

(Max. variable ratio=5, dataset 1 with true CMF=1.30)

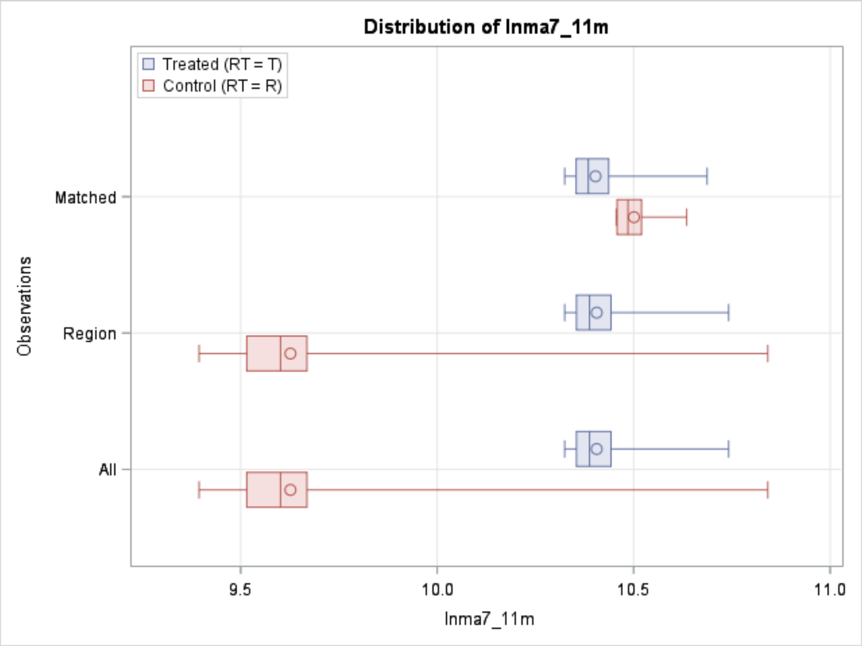

Figure 9 Distribution of lnma7_11m from Optimal matching by PSD

(Max. variable ratio=5, dataset 1 with true CMF=1.30, lnma7_11m=log (average of ma_aadt))

6.2 CMFs from Dataset 2

The dataset 2 has 343 treated sites and 23000 control sites and its true CMF is 1.00. The CMFs results are in Table 4. Unlike dataset 1, each of the covariate adjustment factors is significant and three of them correctly identify the tree effects. The weighting by SIPTW, which has good CMF estimate for dataset 1, does not perform well for dataset 2. For the matching methods, NN matching by PSD continues to have least matched control sites (5 sites only) and provide the poorest CMF estimates. On the other hand, optimal matching by PSD continue to provide the most promising CMFs estimates. Additionally NN MD matching with 1 or 5 replacements correctly estimates the true effects. The matching with 1 replacement has better estimates than the matching with 5 replacements. The traditional CS, weighting by IPTW or SMRW, as well as the covariate adjustment method using LPS, IPTW, or SIPTW all correctly estimate the true CMF. Again, the optimal MD matching doesn't perform well.

Two statistical measures for balance assessment, which are the standardized mean difference between the treatment and control groups and the treated-to-control variance ratio, are listed in Table 5. The variance ratio is the treated-to-control variance ratio for each variable, calculated by dividing the variance in the treated by the variance in the control group. For good variable balance, the absolute standardized mean difference should be less than or equal to 0.25, and the variance ratio should be between 0.5 and 2 (Rubin 2001, p. 174; Stuart 2010, p. 11). The two statistical measures for v_w and lnma7_11m are shown in Table 5.

For the NN PSD matching, the SMD for lnma7_11m is getting larger compared to those from all data or common region data. It indicates the balance of this variable is becoming worse after matching. The SMD for v_w becomes smaller after match, but the variance ratio is getting bigger and far away from 2. It is not clear if the balance for v_w is indeed improved based on the conflicting information in the MSD and the variance ratio.

For the optimal PSD matching, the SMD is significantly reduced from 2.9 to -0.53 and the variance ratio is improved for variable lnma7_11m, indicating the balance of lnma7_11m is greatly improved after matching. Compared to SMD of 3.2 from the matched data by the NN PSD matching, the optimal PSD matching greatly improve the balance of lnma7_11m. Again, distribution of v_w is almost the same as in the original data since we deliberately gave the different means in the two groups when simulating the dataset. The CMFs from this match is almost identical to the true CMF.

Table 4 Comparison of CMFs for Dataset 2 (True CMF=1.00)

mean	95% confidence limits		method	model	Control Sites	Treated sites	Treatment Effects Identified?
0.884	0.851	0.917	EB	NB	23000	343	No
1.016	**0.938**	**1.100**	Traditional Cross-Sectional	NB	23000	343	Yes
1.054	**1.001**	**1.111**	weighting by IPTW	NB	23000	343	Yes
1.390	1.180	1.638	weighting by SIPTW	NB	23000	343	No
1.037	**1.002**	**1.074**	weighting by SMRW	NB	23000	343	Yes
1.154	1.041	1.280	covariate adjustment (PS)	NB	23000	343	No
1.047	**0.956**	**1.148**	covariate adjustment (LPS)	NB	23000	343	Yes
1.008	**0.931**	**1.093**	covariate adjustment (IPTW)	NB	23000	343	Yes
1.017	**0.940**	**1.101**	covariate adjustment (SIPTW)	NB	23000	343	Yes
1.357	1.066	1.726	covariate adjustment (SMRW)	NB	23000	343	No
0.656	*0.235*	*1.831*	NN matching by PSD (5 replacements)	NB	5	341	No
0.654	*0.239*	*1.790*	NN matching by PSD (5 replacements)	mixed-effects NB	5	341	No
1.057	**0.873**	**1.282**	NN matching by MD (1 replacement)	NB	101	343	Yes
1.012	**0.942**	**1.087**	NN matching by MD (1 replacement)	mixed-effects Poisson	101	343	Yes
1.137	0.995	1.301	NN matching by MD (5 replacements)	NB	311	343	No
1.123	0.989	1.276	NN matching by MD (5 replacements)	mixed-effects NB	311	343	No
1.195			NN matching by MD (10 replacements)	NB	509	343	
1.182	1.050	1.331	NN matching by MD (10 replacements)	mixed-effects NB	509	343	No
1.035	**0.910**	**1.177**	optimal matching by PSD (max variable ratio=5)	NB	1029	343	Yes
0.991	**0.934**	**1.051**	optimal matching by PSD (max variable ratio=5)	mixed-effects Poisson	1029	343	Yes
1.149	1.037	1.272	optimal matching by MD (max variable ratio=5)	NB	1029	343	No
1.234	1.188	1.281	optimal matching by MD (max variable ratio=5)	mixed-effects Poisson	1029	343	No

Table 5 Variable balance information in the matched data

Match Method	Variable	Observations	Mean Difference	Standard Deviation	Standardized Mean Difference	Variance Ratio
NN matching by PSD with 5 replacements	Inma7_11m	All	0.7326	0.2511	2.917	0.053
		Region	0.7326		2.917	0.053
		Matched	0.8044		***3.203***	
	v_w	All	24.9439	12.1828	2.047	2.006
		Region	24.9439		2.047	2.006
		Matched	21.5501		1.769	3.251
Optimal matching by PSD with max. variable ratio of 5	Inma7_11m	All	0.7326	0.2511	2.917	0.053
		Region	0.7326		2.917	0.053
		Matched	-0.1343		-0.535	0.125
	v_w	All	24.9439	12.1828	2.047	2.006
		Region	24.9439		2.047	2.006
		Matched	25.5264		2.095	2.027

Note: Standard deviation of All observations used to compute standardized differences

6.3 Comparison of the Methods

From Tables 3 and 4, we can see the CMFs from the two datasets are not always consistent for each method. The comparison of the performance in terms of if each method can correctly identify the true effects for the two datasets is listed in Table 6. If the method can correctly estimate the true effects for the two datasets, we give a "Yes", if CMF is missing for either datasets then it is blank, otherwise it's a "No". There are cases where the true CMF may be located in the 95% confidence limits but very close to the lower or upper limits. To be conservative, we define that method failed to estimate the true effects, i.e. NN matching by MD with 5 replacements for dataset 2 is such a case.

It can be seen, out of 22 methods or options, only three of them can consistently correctly identify the true effects using the two simulated datasets. The three methods include the optimal PSD matching with max variable ratio of 5 with NB or mixed-effects model, and the NN MD matching with 1 replacement using the NB model. It is worth mentioning that the mixed-effects model is not available for the NN MD matched data due to convergence issue. The CMFs estimated using the two datasets are not always consistent indicating different results may be possible using other simulated datasets.

Table 6 Comparison of the Methods in Datasets 1 and 2

Method	Model	Treatment Effects Identified in both Datasets?
EB	NB	No
Traditional Cross-Sectional	NB	No
weight=IPTW	NB	No
Weight=SIPTW	NB	No
weight=SMRW	NB	No
covariate adjustment (propensity score)	NB	No
covariate adjustment (LPS)	NB	No
covariate adjustment (IPTW)	NB	No
covariate adjustment (SIPTW)	NB	No
covariate adjustment (SMRW)	NB	No
NN matching by PSD (5 replacements)	NB	No
NN matching by PSD (5 replacements)	mixed-effects model	No
NN matching by MD (1 replacement)	NB	Yes
NN matching by MD (1 replacement)	mixed-effects model	
NN matching by MD (5 replacements)	NB	No
NN matching by MD (5 replacements)	mixed-effects model	No
NN matching by MD (10 replacements)	NB	
NN matching by MD (10 replacements)	mixed-effects model	No
optimal matching by PSD (max variable ratio=5)	**NB**	**Yes**
optimal matching by PSD (max variable ratio=5)	**mixed-effects model**	**Yes**
optimal matching by MD (max variable ratio=5)	NB	No
optimal matching by MD (max variable ratio=5)	mixed-effects model	No

The CMFs identified from the matched datasets by the optimal PSD matching are much better than those by the NN MD matching with 1 replacement in terms of closer values to the true CMFs and estimated standard errors even though the later method also correctly identifies the CMFs. The important findings from this study are as following:

1. The NN MD matching with 1 replacement and the optimal matching by propensity score correctly identify the true effects.
2. The optimal PSD matching with max variable ratio of 5 has the most number of matched control sites and provide the best CMF estimates
3. The NN PSD matching with 5 replacements has the least number of matched control sites and is the worst method for CMF estimates
4. The optimal MD matching with max variable ratio of 5 does not perform as well as the NN MD matching with 5 replacements.
5. The mixed-effects has the better results than the traditional NB models in terms of better estimate for the mean values and smaller stand errors.

6. Weighting by IPTW and SMRW as well as covariate adjustment by LPS, IPTW, and SIPTW generated similar or better CMFs than the EB method.

Based on the above findings, we recommend the optimal PSD matching for the CMFs evaluation. It can be seen the optimal PSD is even outer performs the EB method in this simulated study in that the EB method fails to identify the true effects. However, it is worth mentioning that the explored EB method for this comparison study used only the reference sites to develop the SPF. There are other variations when using the EB method that have been used when there are significant differences between treated and untreated groups, e.g., the treated group in the before period data as well as a dummy variable indicating treated or reference site is used for the SPF development. Moreover, since the CMFs pattern estimated using the two datasets are not always consistent by each of the methods and the two datasets were purposely selected where the EB method failed to identify the true effects. It may be too early to conclude that the recommended optimal PSD method is better than the EB method for all CMF evaluations. More simulation studies are needed before definitive conclusions can be made. The weighting by IPTW and SMRW as well as the covariate adjustment by LPS, IPTW, and SIPTW are also suggested for further exploration using different simulated datasets as these five methods consistently have similar or better CMFs than the EB method based on the two datasets examined in this effort.

7. CONCLUSIONS

The cross-sectional method that make use of various propensity score methods were explored in this study. These methods were evaluated and compared with the two most common CMF evaluation methods – the traditional cross-sectional and the EB methods using two carefully selected simulated datasets. First, 11 years of traffic volumes on both major and minor roads were generated from truncated normal distribution to ensure realistic traffic volume that found on the roads. Then treated sites were randomly assigned to the sites with high traffic volumes on major roads. The control sites included part of the high ma_aadt sites to ensure a match is possible, but the majority was from the very low ma_aadt sites to simulate some realistic situations. After that, a confounding variable which is significantly different in the treated and control groups was also added to the datasets. A hypothetic treatment was assumed to be implemented in year 6 at the treated sites and the true CMFs was assumed for each of the datasets. Finally, the crash counts were generated through a NB model. The final two datasets were selected where the EB failed to identify the true effects, with the goal of determining whether any of the PS methods would perform better under these conditions.

The explored propensity score methods including weighting, covariate adjustment, and the match methods. The weighting option incudes weighting by IPTW, SIPTW, and SMRW. The covariate adjustment options are propensity score, LPS, IPTW, SIPTW, and SMRW. Note that the last four covariate adjustment options have not been found previous studies. Furthermore, neither the weighting method nor the covariate adjustment method has been applied, evaluated, or explored, for road safety studies.

The optimal PSD matching with maximum variable ratio of 5 and the NN MD matching with 1 replacement correctly identified the true effects, but the former has most number of the matched control sites and provides much better results. The important findings from this study are as following:

1. The NN MD matching with 1 replacement and the optimal matching by propensity score correctly identify the true effects.
2. The optimal PSD matching with max variable ratio of 5 has the most number of matched control sites and provide the best CMF estimates
3. The NN PSD matching with 5 replacements has the least number of matched control sites and is the worst method for CMF estimates
4. The optimal MD matching with max variable ratio of 5 does not perform as well as the NN MD matching with 5 replacements.
5. The mixed-effects has the better results than the NB models in terms of better estimate for the mean values and smaller stand errors.
6. Weighting by IPTW and SMRW as well as covariate adjustment by LPS, IPTW, and SIPTW generated similar or better CMFs than the EB method.

Based on the findings, we recommend the optimal PSD matching for the CMFs evaluation. However, we cannot conclude that this method will always perform better the EB method. It is worth mentioning that there are some approaches that can be applied to the traditional EB method when significant differences exist between the treated before and reference sites, i.e. include the before period data for the SPF developments. Future studies may compare this approach with the PS methods. The weighting by IPTW and SMRW as well as the covariate adjustment by LPS, IPTW, and SIPTW are also suggested for further explore using different simulated datasets as these five methods consistently have similar or better CMFs than the EB method using the two datasets.

There are some options that could be explored in future studies. They are sensitivity analysis on the number of replacements, and the without replacement option for the NN matching method, and various maximum variable ratios for the optimal matching. More simulated datasets are certainly needed to confirm our findings.

Although there were a few CMFs evaluations using the propensity score matching methods for road safety studies, to the authors' knowledge, the recommended optimal PSD matching method has not been evaluated in road safety studies. In addition, very few studies have evaluated the various propensity score methods included in this study. It is our hope that this study provides useful insights that could be used for better CMFs evaluations.

REFERENCES

1. FHWA (2014), *Crash Modification Factors in Practice*, Report FHWA-SA-13-017, http://safety.fhwa.dot.gov/tools/crf/resources/cmfs/docs/product_summary_final.pdf
2. Hauer E. . Observational Before-after Studies in Road Safety: Estimating the Effect of Highway and Traffic Engineering Measures on Road Safety, Pergamon Press, Elsevier Science Ltd, Oxford, U.K. 1997.
1. Gross, F., Persaud, B., and Lyon, C. (2010), A guide to developing quality crash modification factors, Report FHWA-SA-10-032, Federal Highway Administration.
2. Carter, D., Srinivasan, R., Gross, F., and Council, F. (2012), Recommended protocols for developing crash modification factors, NCHRP Project 20-07 (Task 314), National Cooperative Highway Research Program, Washington, D.C.
3. Miaou, S.P., Lum, H., 1993. Modeling vehicle accidents and highway geometric design relationships. Accident Analysis and Prevention 25 (6), 689–709.
4. Persaud, B., Craig, L., Kimberly. E., Nancy, L., Frank, G., 2009. Safety evaluation of offset improvements for left-turn lanes. FHWA-HRT-09-035.
5. Donnell, E. T., R. J. Porter, and V. N. Shankar. Framework for Assessing the Safety Effects of Roadway Lighting. Safety Science, Vol. 48, No. 10, December 2010, pp. 1436-1444.
6. Donnell, E.T., Gross, F., 2011. Case-control and cross-sectional methods for estimating crash modification factors: comparisons from roadway lighting and lane and shoulder width safety effect studies. Journal of Safety Research 42 (2), 117–129.
7. Austin, P. C., An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. Multivariate Behavioral Research, 46:399–424, 2011.
8. Rosenbaum PR, Rubin DB. Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. Journal of the Royal Statistical Society Series B 1983;45(2):212–218.
9. Donnell E., Hanks, E., Porter, R. J., Cook, L., Srinivasan, R., Li F., Nguyen, M., Eccles, K. Project A-6: Highway Safety Statistical Paper Synthesis. FHWA, 2017

10. Sasidharan, L., Donnell, E.T. (2013). Application of propensity scores and potential outcomes to estimate effectiveness of traffic safety countermeasures: Exploratory analysis using intersection lighting data. Accident Analysis and Prevention 50 (2013) 539–553.
11. Wood, J. S., Gooch, J. P., Donnell, E.T. (2015). Estimating the safety effects of lane widths on urban streets in Nebraska using the propensity scores-potential outcomes framework. Accident Analysis and Prevention 82 (2015) 180–191
12. Wood, J. S., Donnell, E.T. (2016). Safety evaluation of continuous green T intersections: A propensity scores-genetic matching-potential outcomes approach. Accident Analysis and Prevention 93 (2016) 1–13.
13. Wood, J. S., Donnell, E.T., Porter, R. J. (2016). Comparison of safety effect estimates obtained from empirical Bayes before–after study, propensity scores-potential outcomes framework, and regression model with cross-sectional data. Accident Analysis and Prevention 75 (2015) 144–154.
14. Hauer, E., Harwood, D W, Council, F M, Griffith, M.. Estimating Safety by the Empirical Bayes Method: A Tutorial. Transportation Research Record, Journal of the Transportation Research Board, 1784, 126-131. 2002.
15. Rubin DB. Matching to remove bias in observational studies. Biometrics 1973;29:159–184.
16. Rosenbaum PR, Rubin DB. The bias due to incomplete matching. Biometrics 1985;41:103–116. [PubMed: 4005368]
17. Dehejia RH, Wahba S. Causal effects in nonexperimental studies: Re-evaluating the evaluation of training programs. Journal of the American Statistical Association
18. Rosenbaum, PR. Observational Studies. 2. Springer Verlag; New York, NY: 2002.
19. Gu X, Rosenbaum PR. Comparison of multivariate matching methods: structures, distances, and algorithms. Journal of Computational and Graphical Statistics 1993;2:405–420.
20. Rosenbaum, P. R. (1987). Model-based direct adjustment. The Journal of the American Statistician, 82, 387–394.
21. Robins JM, Hernan MA, Brumback B. Marginal structural models and causal inference in epidemiology. Epidemiology 2000;11:550–560. [PubMed: 10955408]
22. Sato T, Matsuyama Y. Marginal structural models as a tool for standardization. Epidemiology. 2003 Nov;14(6):680–6. [PubMed]

23. Stürmer, T, Wyss, R., Glynn, R. J., Brookhart, M.A., Propensity scores for confounder adjustment when assessing the effects of medical interventions using non-experimental study designs. Journal of Internal Medicine, Volume 275, Issue 6, 2014. Pages 570–580.
24. Weitzen S, Lapane KL, Toledano AY, Hume AL, Mor V. Principles for modeling propensity scores in medical research: a systematic literature review. Pharmacoepidemiol Drug Saf. 2004;13: 841–53.
25. Shah BR, Laupacis A, Hux JE, Austin PC. Propensity score methods give similar results to traditional regression modeling in observational studies: a systematic review. J Clin Epidemiol. 2005; 58:550–9.
26. Stürmer, T, Joshi M, Glynn RJ, Avorn J, Rothman KJ, Schneeweiss S.2006. A review of the application of propensity score methods yielded increasing use, advantages in specific settings, but not substantially different estimates compared with conventional multivariable methods. J Clin Epidemiol. 2006; 59:437–47.
27. Haukoos JS, Lewis RJ. The Propensity Score. JAMA Guide to Statistics and Methods. 314(15):1637-8. 2015. doi: 10.1001/jama.2015.13480.
28. Melissa M. Garrido, M.M., Propensity Score as a Covariate in Linear Models. JAMA. 315(14):1521-1522. 2016
29. Hadea, E. M. and Lu, B. 2014. Bias associated with using the estimated propensity score as a regression covariate. Stat Med. 33(1): 74–87. doi:10.1002/sim.5884.
30. Austin PC. 2009. The relative ability of different propensity score methods to balance measured covariates between treated and untreated subjects in observational studies. Med Decis Making. 29(6):661-77. doi: 10.1177/0272989X09341755
31. Lord, D. Modeling motor vehicle crashes using Poisson-Gamma models: Examining the effects of low sample mean values and small samples size on the estimation of the fixed dispersion parameter. Accident Analysis and Prevention 38, 751-766. 2006
32. Lan, B., Persaud B., Lyon, C., and Bhim, R. Validation of a Full Bayes methodology for observational before–after road safety studies and application to evaluation of rural signal conversions. Accident Analysis & Prevention. Volume 41, Issue 3, Pages 574-580, 2009.

33. Bhim, R. Observational Before and After Safety Study of Installing Signals at Rural Intersections: Using the Empirical Bayes (EB) and Conventional Methods. Ryerson University. 2006.
34. Garber, N.J., Rivera, G. (2010). Safety Performance Functions For Intersections On Highways Maintained By The Virginia Department Of Transportation, Final Contract Report Vtrc 11-Cr1.
35. Cameron, A.C., Trivedi, P.K. Regression Analysis of Count Data. Cambridge University Press, Cambridge, UK. 1998.
36. Rubin, D. B. (2001). “Using Propensity Scores to Help Design Observational Studies: Application to the Tobacco Litigation.” Health Services and Outcomes Research Methodology 2:169–188.
37. Stuart, E. A. (2010), Matching methods for causal inference: A review and a look forward. Stat Sci. 2010; 25(1): 1–21. doi: 10.1214/09-STS313.