Accounting for Site-Selection Bias in Before-After Studies for Continuous Distributions: Characteristics and Application Using Speed Data

Pei-Fen Kuo  
Research Assistant  
Zachry Department of Civil Engineering  
Texas A&M University  
3136 TAMU  
College Station, TX 77843-3136

Dominique Lord*  
Associate Professor and Zachry Development Professor I  
Zachry Department of Civil Engineering  
Texas A&M University  
3136 TAMU  
College Station, TX 77843-3136  
Tel. (979) 458-3949  
Fax. (979) 845-6481  
Email: d-lord@tamu.edu

Submitted for publication. Revised Version.

December 27, 2012

*Corresponding Author
ABSTRACT

The before-after study is still the most popular method used by traffic engineers and transportation safety analysts for evaluating the effects of an intervention. Compared to the cross-sectional study, the before-after study has lower within-subject variability since it directly accounts for changes that have occurred at the study sites. However, although this kind of study may offer superior performance, it can still be plagued by important methodological limitations, which could significantly alter the study outcome. They include the regression-to-the-mean (RTM) and site-selection effects. The primary objective of this study consists of presenting a method that can reduce the selection effects when an entry criterion is used in before-after studies for continuous data (e.g. speed, reaction times, etc.), without relying on the use of a control group. The distribution of the data could follow a normal or lognormal distribution. The study objective was accomplished using simulated and observed speed data collected in Florida. The proposed method documented in this paper was compared to the Naïve, Control Group (CG) and the Analysis of Covariance (ANCOVA) methods. The simulation results show that the proposed method provides a more precise estimate than the Naïve method, as expected. In addition, the method performs better than the CG and the ANCOVA methods when similar control group data are not available. The results also show that higher entry criteria, lower between-subject variances, and higher within-subject variances cause higher selection biases. When traffic engineers and urban planners evaluate or compare different strategies, the proposed method can be applied to adjust naïve estimators of treatment effectiveness documented in previous studies without similar control group data.
1. Introduction

The before-after study is still the most popular method used by traffic engineers and transportation safety analysts for evaluating the effects of an intervention. Different methods or techniques exist for conducting before-after studies. They include the Naïve method, the Control Group (CG) method, the Empirical Bayes (EB) method, and the full Bayes method (the latter two have so far mainly been used for analyzing crash data). These methods all share the primary goal of comparing observed data collected in the before period with data assembled in the after period for the same entities. Compared to the cross-sectional study, the before-after study has lower within-subject variability (i.e., the variation associated with multiple measurements observed over time for one subject) since it directly accounts for changes that have occurred at the study sites (Hauer, 1997) (see Appendix A key definitions used in this study). This explains why the before-after study is the most popular approach for evaluating the effects of an intervention. However, although this kind of study offers a superior performance, it can still be plagued by important methodological limitations, which could significantly alter the study outcome.

The limitations that significantly influence the effects of a treatment include the regression-to-the-mean (RTM) and site-selection effects. RTM is a statistical phenomenon that occurs when observations characterized by very high (or low) values in a given time period, and for a specific site, are anticipated to regress towards the long-term mean of a site in a subsequent time period. Site-selection effects, on the other hand, occur when the observations must meet a minimum criterion in order to be included in a study. Setting minimum entry criteria is actually not uncommon when transportation agencies select sites for treatments. Their motivations can take various forms, such as following warrants based on the Manual on Uniform Traffic Control Devices (MUTCD) (FHWA, 2010) (see, e.g., Obeng and Burkey, 2008), availability of data, or specific study-design characteristics. So far, most of the research on these biases has focused on discrete counts, since the before-after study has been applied extensively for crash data analyses (see Hauer, 1997; Persaud et al., 2001). Not only have a significant amount of studies focused on discrete counts, but the majority of the work has also been directed mainly at the RTM. Only a handful of studies have specifically discussed or addressed site-selection effects (Davis, 2000; Lord and Kuo, 2012).

Hence, the primary objective of this study consists of presenting a method that can remove or reduce site-selection effects when an entry criterion is used in before-after studies for
continuous data, without relying on the use of a control group; continuous data that have been used in transportation research where site selection effects could play a role in the study outcome include observed speed, vehicle delay, running time, reaction time and wait time among others (Johnson and Ceerla, 1996; Chen et al., 2002; Owen and O’Mahony, 2003; Dziekan and Kottenhoff, 2007; Tétreault and El-Geneidy; 2010; Watkinsa et al., 2011; Diab and El-Geneidy, 2012; Hou et al., 2012). This study expands on the work performed by Park and Lord (2010), who examined the effects of the RTM for this kind of data. Since the RTM and site-selection effects are distinct biases, there is a need to analyze them separately (discussed in greater details in the next section). The study objective was accomplished using simulated and observed speed data collected in Florida. The simulation analysis was carried out to examine the accuracy of the adjusted method, while the observed data were used to evaluate the effectiveness of the adjusted method in practice. To better illustrate the differences between the RTM and site-selection effects, the same observed dataset used by Park and Lord (2010) was utilized in this study.

2. Background

This section describes the characteristics associated with the RTM, site-selection effects, and the Analysis of Covariance (ANCOVA).

2.1 RTM and Site-Selection Effects

As discussed in Lord and Kuo (2012) and the references herein, RTM and site-selection biases are two different biases. The RTM refers to the concept that observations characterized by very high (or low) values in a given time period, and for a specific site, are anticipated to regress towards the long-term mean of a site in a subsequent time period (Hauer and Persaud, 1983). The RTM is not new and was first observed more than a century ago by Francis Galton (Stigler, 1997). The RTM can be conceptualized mathematically using random variables in two time periods, labeled as 1 and 2, respectively (Copas, 1997; Lord and Kuo, 2012). Let us assume that $Y_1$ and $Y_2$ are two random variables with almost exactly the same distribution (e.g. speed, reaction time, wait time, etc.), but where the conditional expectation $E[Y_2 | Y_1]$ is not equal to
Y₁. It can be shown that the conditional expectation can be defined as a jointly normal distribution:

\[
E[Y_2 | Y_1] = \rho Y_1 + (1 - \rho) \mu
\]

(1)

Where \( \rho \) is the correlation between \( Y_1 \) and \( Y_2 \), and \( \mu \) is the common mean. When the correlation coefficient is equal to 1, no RTM exist since \( E[Y_2 | Y_1] = Y_1 \). On the other hand, when the correlation coefficient is not equal to 1, RTM is observed in the data. Smaller values of \( \rho \) are associated with larger RTM effects because \( E[Y_2 | Y_1] \) is closer to \( \mu \) and farther away from \( Y_1 \). Equation (1) also shows that the magnitude of the RTM can be computed by taking the difference of \( E[Y_2 | Y_1] \) and \( Y_1 \) (Figure 1).

![Figure 1. Graphical representation of the regression-to-the-mean phenomenon (Lord and Kuo, 2012)](image)

Site-selection effects occur when an entry criterion is used for selecting observations that will be included in the before-after study. For continuous data, this gives rise to a truncated
normal distribution; see Cook and Wei (2002) and Lord and Kuo (2012) for discrete counts. This distribution should not be confused with the conditional normal distribution described above. Figure 2 illustrates the characteristics associated with the site-selection effects for continuous data.

![Figure 2. Site-selection bias for different correlation coefficient values.](image)

In Figure 2, the left-hand side is the probability related to the normally distributed data in the before and after periods (without selection). The difference between the mean of these two curves is defined as $\delta_{\text{true}}$. After setting the minimum entry criteria (the dash vertical line), the data distribution in the before period is left-truncated, as indicated by the point curved line on the right-hand side. It should be noted that setting the same entry criteria might cause different effects according to different correlation coefficient ($\rho$) values with the before and after data. If $\rho$ is equal to 1, the estimator of the difference, $\delta_{\rho=1}$, is unbiased because the mean value
increases in the same manner in the before and after periods. If \( \rho \) is equal to 0, the naïve estimator of difference is less than its true value. The mean in the before period increases by \( \varepsilon \) (i.e., \( E(Y_{i1})-E(Y_{i1} \mid Y_{i1} > C) \)), but the mean in the after period remains the same because the before-after data are independent. If \( \rho \) is negative, the estimator of the difference might become lower than the above values. Removing the data with low values in the before period may also remove data with high values in the after period because of the negative correlation. Hence, the difference becomes much smaller because of the higher mean in the before period and a lower mean in the after period.

It should be noted that if sites are randomly selected from the whole population to evaluate the installation of a new treatment (say a new speed reduction system on urban freeways), then there is no site-selection bias when the treatment effectiveness is estimated because the sample should represent the whole population, as long as no entry criterion is used. Conversely, if a minimum entry criterion for choosing the study sites is used, then the random sample represents the characteristics associated with a sub-population, but does not represent those linked to the whole population. In other words, even if a sample is randomly chosen from a sub-population, there is no reason to believe that the estimate of the treatment is actually the same estimate that would be estimated if the sample was taken from the entire population.

To reiterate the differences between site-selection effects and RTM, Table 1 summarizes the mathematical equations used for quantifying the RTM and site-selection effects (i.e., data are left-truncated). It should be noted that with site-selection effects the entry criteria (C) can take any value and is not limited only to one criterion. On the other hand, with the RTM, the observations in the before period are selected based only on the mean. In other words, the site-selection bias may still exist in some conditions where the RTM is ignored or not applicable, such as when the entry criterion is relative low (i.e., close to zero) or are used to remove extreme values by setting a maximum and minimum threshold (multiple criteria) for including observations in the dataset.

<table>
<thead>
<tr>
<th>Effects</th>
<th>Before</th>
<th>After</th>
</tr>
</thead>
<tbody>
<tr>
<td>Site selection</td>
<td>( P(Y_{i1} \mid Y_{i1} &gt; C), i : site )</td>
<td>( P(Y_{i2} \mid Y_{i1} &gt; C) = P(Y_{i2} \mid Y_{i1})P(Y_{i1} &gt; C) )</td>
</tr>
<tr>
<td>Regression-to-the-Mean</td>
<td>( P(Y_{i1} \mid Y_{i1} &gt; \mu_i) )</td>
<td>( P(Y_{i2} \mid Y_{i1} &gt; \mu_i) )</td>
</tr>
</tbody>
</table>
2.2 Analysis of Covariance (ANCOVA)

The Analysis of Covariance (ANCOVA) has been another method proposed for eliminating the selection bias (Chuang-Stein and Tong, 1997; Park and Lord, 2010, Chen et al., 2013). This method has been utilized extensively in epidemiology research (Barnett et al., 2005; Laird, 1983; Twisk, 2003). Basically, the ANCOVA provides a more precise estimate of treatment effect by adjusting each subject’s follow-up measurement according to its baseline measurement. The characteristics of this method are described in the methodology section further below. Although the ANCOVA has successfully been used in the past in the context of a before-after study with entry criteria (i.e., its estimator has a narrower confidence interval than the traditional paired t-test), the method has the same disadvantage as the one used for the CG method: both methods require collecting additional data that will be part of the control group. In many instances, collecting additional data may be prohibitive depending on the study design. On other occasions, it may be infeasible to collect at control locations (i.e., control sites do not have the same characteristics as the treatment sites). Additional information can be found in (Barnett et al., 2005), who applied the ANCOVA for analyzing the effects of skin cancer prevention treatments in Nambour, Australia.

In sum, it is necessary to develop a new method that can easily adjust or reduce naïve estimators in before-after studies that employ continuous data and are characterized by an entry criterion, without relying on data collected for the control group. This will be performed by comparing the new method (subsequently called the adjusted method in the text) with the other commonly known before-after methods: the Naïve, CG, and ANCOVA.

3. Calculating the Site-Selection Bias

The estimator of the difference ($\delta$) is equal to the mean value in the after period ($\Lambda_2$) minus the mean value in the before period ($\Lambda_1$). After setting an entry criterion in the before period, the continuous data should follow a truncated normal distribution (more common) or a truncated lognormal distribution. This study focuses on the truncated normal distribution, but the bias adjusted method for the lognormal normal distribution can be found in Appendix B. There are two reasons for focusing on the normal distribution. First, parameters and site-selection bias for the lognormal distribution are more complex to estimate than for the normal distribution. Second, the above problem can be easily solved by transforming the data to a normal distribution.
According to Cook and Wei (2002), the unbiased estimator of the difference is equal to the naïve estimator minus site-selection bias:

\[
\hat{\delta}_{\text{adjusted}} = \left[ \hat{\Lambda}_2 - \hat{\Lambda}_1 \right] - \text{site-selection bias} \\
= \left[ \bar{Y}_2 - \bar{Y}_1 \right] - \left( \rho - 1 \right) \frac{\sqrt{(\phi + \sigma^2)} f(d)}{1 - F(d)}
\]  

(2)

Where,
\( \delta \): The difference;
\( \rho \): The correlation coefficient of the observed data in the before period (Y_1) and after period (Y_2);
\( \phi \): Between-subject variance;
\( \sigma^2 \): Within-subject variance; and, for notation convenience
\[
d = \frac{C - \Lambda_1}{\sqrt{\phi + \sigma^2}}, \text{ C after Normalization}
\]

It should be noted that C and \( \mu \) are the entry criteria and mean respectively (same as Table 1).

Because the true values for the above parameters, \( \rho, \sigma^2, \phi, \Lambda_1 \), are seldom known, the following equations are used here as their estimators (the \( \hat{\text{•}} \) represents the parameter estimate):

\[
\hat{\rho} = \text{Corr}(Y_{i1}, Y_{i2} \mid Y_{i1} > C)
\]  

(3)

\[
\hat{\sigma}^2 + \hat{\phi} = \text{Var}(Y_{i2} \mid Y_{i1} > C)
\]  

(4)

\[
\hat{\Lambda}_1 = E(Y_{i1} \mid Y_{i1} > C)
\]  

(5)

The above estimators are unbiased when C is relatively low. For more precise estimators, several methods have previously been proposed (Barr and Sherrill, 1999; Cohen Jr, 1950; Formann, 2008). Equations (6) and (7) show the latest estimators (Formann, 2008):

\[
\hat{\Lambda}_1 = E(Y_{i1} \mid Y_{i1} > C) - \sqrt{\sigma^2 + \phi} \frac{f(d)}{1 - F(d)}
\]  

(6)
\[
\hat{\delta}^2 + \hat{\phi} = \frac{\text{Var}(Y_i \mid Y_i > C)}{1 - \left( \frac{f(d)}{1 - F(d)} \right)^2 + d \frac{f(d)}{1 - F(d)}} \quad (7)
\]

However, \( d \), which is a function of \( \hat{\Delta} \) and \( \hat{\delta}^2 + \hat{\phi} \), means that Equations (6) and (7) have to be solved iteratively. Hence, this study still used Equations (3), (4) and (5) for the simulation analysis. For more details about how to solve Equations (6) and (7), the reader is referred to Formann (2008).

4. Methodology and Simulation Protocol

This section describes the methodology and simulation protocol used for estimating the bias for the four previously discussed methods as well as evaluating the adjusted method effectiveness for different conditions. More specifically, a simulation analysis was carried out for the following reasons:

1) The true values of the treatment effectiveness for the whole population are seldom known in practice. Usually, the subjects or observations are not selected randomly and are truncated using an entry criterion because of the limited funds or selection guidelines (i.e., high risk sites, minimum speed, etc.), as discussed above. Furthermore, using observed datasets that contain a limited number of observations makes the evaluation of the accuracy of the bias-adjusted method difficult. This difficulty is caused by the fact that the real population mean, between-subject variance, and site-selection bias are usually unknown. With the help of simulation, the analyst has full control over the input variables.

2) Although theoretical equations to estimate the site-selection bias are available, these equations assume the sample to be infinite (or extremely large), which does not reflect how these equations would be utilized with observed data or in practice. In certain fields, datasets frequently have a limited number of observations that are caused by various factors, such as the costs associated with the data collection process.
3) Simulation was utilized to examine the effects when dissimilar control groups are used. Furthermore, the simulation results can be used to show the marginal changes of the site-selection bias when the main characteristics of the treatment group data vary (such as between-subject variance and within-subject variance). Section 4.2 provides more information about items (2) and (3).

This study only assumed one year for the before and after periods, respectively, in order to simplify the calculations.

4.1 Notation

Before describing the methodology, it is important to define the notations used in the study:

- \( C, C' \): The entry criterion (minimum and maximum);
- \( n \): The sample size;
- \( \lambda_{i1}^T, \lambda_{i1}^{CG} \): The mean response rate for site i (T: treatment group, CG: control group) in the before period (k=1);
- \( \lambda_{i2}^T, \lambda_{i2}^{CG} \): The mean response rate for site i (T: treatment group, CG: control group) in the after period (k=2);
- \( Y_{i1}^T, Y_{i1}^{CG} \): The observed response for site i (T: treatment group, CG: control group) in the before period for continuous data, \( Y_{i1}^T, Y_{i1}^{CG} > C \);
- \( Y_{i2}^T, Y_{i2}^{CG} \): The observed response for site i (T: treatment group, CG: control group) in the after period for continuous data;
- \( \hat{\delta}_{\text{ANCOVA}} \): The estimator of difference by using the ANCOVA; this is the coefficient of the “group” variable in the regression model; and,

\[
Y_{i2} = \text{constant} + a \times (Y_{i1} - \bar{Y}_{i1}) + b \times \text{group} + \text{error},
\]

Where group=1, treatment group; group=0, control group."
Given the notation above, it is now possible to define the equations related to the observed data $Y_{i1}$ using the two-sided truncated normal distribution:

$$f(Y_{i1}; \Lambda_1, \phi, \sigma^2, C, C') = \frac{1}{\sqrt{\phi + \sigma^2}} \Phi \left( \frac{Y_{i1} - \Lambda_1}{\sqrt{\phi + \sigma^2}} \right) - \Phi \left( \frac{C - \Lambda_1}{\sqrt{\phi + \sigma^2}} \right)$$

$$E(Y_{i1} | C' > Y_{i1} > C) = \Lambda_1 + \frac{f(C - \Lambda_1; \phi, \sigma^2) - f(C' - \Lambda_1; \phi, \sigma^2)}{F(C' - \Lambda_1; \phi, \sigma^2) - F(C - \Lambda_1; \phi, \sigma^2)} \times \left( \sqrt{\phi + \sigma^2} \right)$$

where $Y_{i1} = \Lambda_1 + u_i + e_{ik}, u_i \sim N(0, \phi), e_{ik} \sim N(0, \sigma^2)$

In Equation (9), when $C'$ is close to $\infty$, the equation is defined as a left-truncated normal distribution, which is used when the selected study subjects are larger than the threshold (i.e., entry criterion), and vice versa.

4.2 Scenario Analysis

Based on the site-selection bias estimation equation and its theoretical derivations, four scenarios were used to examine possible factors related to site-selection bias. They include a direct comparison of the methods, between-subject variance, within-subject variance, and sample size. Scenario 1 compared the four most common types of before-after studies: the Naïve, using a control group (CG), ANCOVA, and the adjusted method. Recall that Equations (10), (11), (12) and (13) are used for estimating the difference. Scenario 2 assumed the between-subject variance varied from 9 ($=3^2$, small heterogeneity) to 225 ($=15^2$, very large heterogeneity). There are seven levels of variance: $3^2, 4^2, 5^2, 7^2, 9^2, 11^2,$ and $15^2$. It should be pointed out that the between-subject variances have been observed using real speed data; they ranged from $4^2$ to $7^2$ (Muchuruza and Mussa, 2004). For Scenario 3, the bias when the within-subject variance varied from 9 (small heterogeneity), 16 (medium heterogeneity), and 25 (large heterogeneity) are examined. Finally, Scenario 4 investigated the effects of the sample size on the selection bias: 10 (small), 30 (medium), and 100 (large).
For all scenarios, the entry criteria was assumed to vary from $C=58$ (i.e., $Y_{ij} > 58$) to $C=73$. The data generation procedure followed a normal distribution to generate the mean response rate. $m$ ($i = 1$ to $m$) observations were selected randomly for the treatment group only when the response was larger than the entry criteria ($\sum_{j=1}^{r} Y_{ij} > C$). These sites are labeled as $Y_{i1}^T$ and $Y_{i2}^T$, respectively. The sample size was equal to 100, while the difference was set to 10 mph. Note that scenarios 1 and 2 were analyzed simultaneously.

The equations for the four methods are as follows:

(1). Naïve method: 
$$\hat{\delta}_{\text{naive}} = \hat{\Lambda}_2 - \hat{\Lambda}_1 = \frac{\sum_{i=1}^{m} Y_{i2}^T - \sum_{i=1}^{m} Y_{i1}^T}{m}$$  \hspace{1cm} (10)

(2). CG method: 
$$\hat{\delta}_{\text{CG}} = (\hat{\Lambda}_2^T - \hat{\Lambda}_1^T) - (\hat{\Lambda}_2^{CG} - \hat{\Lambda}_1^{CG})$$  \hspace{1cm} (11)

(3). ANCOVA method: 
$$\hat{\delta}_{\text{ANCOVA}} = \hat{b}$$  \hspace{1cm} (12)

(note: $b$ was described above Equation (8))

(4). Adjusted method:
$$\hat{\delta}_{\text{adjusted}} = \hat{\delta}_{\text{naive}} - (\hat{\rho} - 1)\sqrt{\frac{\hat{\sigma}^2}{\hat{\phi} + \hat{\sigma}^2}} f(\hat{d}) \left/ (1 - F(\hat{d})) \right.$$  \hspace{1cm} (13)

$$= \hat{\delta}_{\text{naive}} - (\text{Corr}(Y_{ij1}, Y_{ij2}) - 1)\sqrt{\text{Var}(Y_{ij2})} \left[ f\left(\frac{C - \bar{Y}_i}{\sqrt{\text{Var}(Y_{ij2})}}\right) \right] \left/ \left(1 - F\left(\frac{C - \bar{Y}_i}{\sqrt{\text{Var}(Y_{ij2})}}\right)\right) \right.$$  \hspace{1cm} (13)

As discussed above, the CG and ANCOVA methods both adjust the naïve estimator by removing the bias using information from the control group data, while the adjusted method can remove site-selection bias using information from the treatment group (truncated normal distribution).
4.3 Simulation protocol

The simulated data were generated using the software R (R Development Core Team, 2006). The steps were as follows:

1) For each between-subject variance ($\phi$), generate the subject-specific random effect ($u_i$), which follows a normal distribution $N(0, \phi)$. The sample mean in the before period is 70, and error term of each site also follows a normal distribution $N(0, \sigma^2)$. $\sigma$ is equal to 5 for all scenarios except for Scenario 3, which examines the effects for different $\sigma$ values. Then, observed data ($Y_i$) in the before period were generated by combining the above terms: mean, random effect, and error for each site $i$ ($Y_{ik} = \Lambda_k + u_i + e_{ik}$).

2) Observed data in the after period were generated using a similar procedure, and the only difference was the mean in the after period which was made equal to 80.

3) Generate the data for 5,000 sites, but randomly select 100, 30 and 10 sites ($m$) depending on the scenario.

4) Then, $m$ sites are selected as the sample, whose observed (speed) values are larger than the entry criteria (58, 61,…, 73). The effectiveness can be estimated using Equations (10), (11), (12), and (13).

5) When the control group is used, $\delta_{control}$ is equal to 0. In other words, there is no difference in the mean rate between the before and after periods for the control group.

6) Repeat steps 2 to 5 for a total of 1,000 times, and estimate the biases of various estimators ($\hat{\delta}_{1000} - \delta$).
5. Simulation Results

This section describes the results based on the simulation. The results are presented for each scenario.

5.1 Scenario 1 - results

Figure 3 shows the site-selection bias for the Naïve, CG, ANCOVA, and the adjusted methods. Overall, this figure shows that the bias is reduced as the between-subject variance increases, except when C is particularly small (e.g., C=58). This was expected given the characteristics of Equations (2) and (13). The greater the entry criteria, the more biased the estimate will be. Among the four methods, the Naïve method (Figure 3a) has highest site selection bias; δ can be underestimated by as much as 49%. As discussed by Cook and Wei (2002), unless the correlation is close to one (φ ≫ σ²), ̂δ will be biased if an entry criterion is used (e.g., the bias never equals zero even when φ = 225).

When the CG or ANCOVA method is used, the bias can be theoretically eliminated. For the CG method (Figure 3b), the control group needs to have the same characteristics (i.e., the same sample mean, variance, and entry criteria) as those of the treatment group used for the Naïve method. As explained above, it may be difficult to find in practice datasets with the exact same characteristics as for the treatment group.

The application of the control group is further explored in Figure 4. Figure 4a is the same as Figure 3b, and is used to compare the results with the other figure below. Figure 4b shows that when the control group has a slightly lower mean (0.95 × mean), the site-selection bias is far from zero but smaller than the bias calculated from the Naïve method. Furthermore, the bias is in the opposite direction (CG with a lower mean will over-estimate δ). Conversely, using a control group having a higher mean value (1.2 × mean) causes a negative site-selection bias, which is even higher than with the Naïve method (Figure 4c). In other words, the site-selection bias caused by using a dissimilar control group might be even higher than just using the Naïve method. Figures 4 (d), (e), and (f) show the site-selection bias for the ANCOVA for the same mean values as in Figures 4 (a), (b) and (c). Similar to the CG method, the ANCOVA still provides a biased estimate when the control group is dissimilar (i.e., Figures (e) and (f)). It should be noted that the site-selection biases using the ANCOVA method were same for different entry criteria. Finally, using the adjusted method reduced site-selection bias by about
50%, even when the biased estimators of $\rho, \sigma^2, \phi$, and $\Lambda$ are used (Figure 3d), or are not known with certainty. Compared to all methods, the adjusted method provides a more precise estimate than the Naïve method and performs better than the CG and ANCOVA methods when similar control group data are not available. It should be pointed out that the adjusted method will not completely eliminate the site-selection effects unless $\rho, \sigma^2$, and $\phi$ are fully known. Obviously, these values are rarely known in practice.
Figure 3. Site-selection bias for the Naïve, CG, ANCOVA, and our adjusted methods.
Figure 4. Site selection bias for the CG and ANCOVA method for the following characteristics: (a) same mean, (b) and (e) 0.95 × mean, (c) and (f) 1.2 × mean.
5.2 Scenario 2 – results

As discussed in the first scenario, the site-selection bias decreases when the between-subject variance increases, but the decreasing rate becomes almost flat for values above $11^2$ (Figure 3). With the Naïve method, the bias is never eliminated, as compared to the CG and ANCOVA methods. It should also be noted that when the between-subject variance tends towards zero (subjects have a low heterogeneity), the site-selection bias is the largest. This finding is consistent with Cook and Wei (2002), because the correlation coefficient is close to zero when between-subject variance is zero.

5.3 Scenario 3 – results

Figure 5 shows that the value of within-subject variance changes the value of the bias. This figure illustrates that the bias increases when the error term becomes larger, which is consistent with the characteristics of Equation (2). Moreover, the adjusted method reduces the bias by 50% for all within-subject variance. The results from scenarios 2 and 3 indicate that the lower between-subject variance and higher within-subject variance cause larger selection bias.

5.4 Scenario 4 – results

Figure 6 shows that the sample size of the treatment group does not affect the bias considerably. For all sample sizes, the biases estimated from the adjusted method still reduces the bias by 50%. Hence, Equation (13) which was derived by assuming that the sample size is close to infinity ($\infty$) may be used for estimating site-selection biases when the sample size is small.
Figure 5. Site selection bias for the naïve and adjusted method when the within-subject variance is equal to $3^2$, $4^2$, and $5^2$. 
Figure 6. Site selection bias for the adjusted method when the sample size is equal to (a) 100, (b) 30, (c) 10, and (d) Estimated.
6. Examining RTM And Site-Selection Biases Using Observed Data

In order to better illustrate the simulation results, the equations described in the first part of the paper were applied to a dataset that was used to assess the effects of increasing the speed limit on multilane highways in Florida. For this dataset, the speed distribution was assumed to be normally distributed, which is not uncommon as discussed by Park and Lord (2010). The normal distribution is one of most common distributions used to characterize speed data (Park et al., 2010). The dataset was collected and first explored by Muchuruza and Mussa (2004) which did not completely eliminate the site-selection bias. The same dataset was subsequently analyzed by Park and Lord (2010) to examine the effects of the RTM in the before-after evaluation of the continuous data.

Muchuruza and Mussa (2004) examined the impacts on observed driving speeds when the speed limit was increased from 65 mph to 70 mph on four- and six-lane highways. Their analysis was carried out at 18 study sites, 10 for four-lane and eight for six-lane highways, respectively. Table 2 summarizes the important characteristics of the data. Using the Naïve before-after method, they showed that the average speed increased by 5 mph. To examine the effects of the site-selection bias, Equations (13) was used to obtain the adjusted difference. The first entry criterion was assumed to be 62, because the suggested initial assumption for the entry criteria should be equal to the smallest observed data (i.e., \( C = \min Y_{ij} \)) (Johnson et al., 1970). Then, the minimum entry criteria equal to 63, 64, 66, 67, and 68 were subsequently employed. Four-lane and six-lane highways were initially analyzed separately and were then combined to increase the sample size (which may not be technically necessary). By grouping the data, the adjusted estimators for six-lane highways can therefore be compared with the results described in (Park and Lord, 2010).
Table 2. Before-After speed data (Muchuruza and Mussa, 2004)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Four Lane Highway</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I-75</td>
<td>At mile marker 89, WB, Site 351, WB</td>
<td>66</td>
<td>74</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td></td>
<td>At mile marker 89, EB, Site 351, EB</td>
<td>68</td>
<td>78</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Overpass E. of SR 85, WB, Site 9901, WB</td>
<td>67</td>
<td>74</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Overpass E. of SR 85, EB, Site 9901, EB</td>
<td>69</td>
<td>74</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td></td>
<td>C-280 overpass, WB, Site 9901, WB</td>
<td>68</td>
<td>74</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td></td>
<td>C-280 overpass, EB, Site 9901, EB</td>
<td>67</td>
<td>74</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Between SR257 &amp; US221, WB, Site 9928, WB</td>
<td>67</td>
<td>70</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Between SR257 &amp; US221, EB, Site 9928, EB</td>
<td>69</td>
<td>71</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>East end of Aucilla River, WB, Site 9928, WB</td>
<td>67</td>
<td>70</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>East end of Aucilla River, EB, Site 9928, EB</td>
<td>69</td>
<td>71</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td></td>
<td>67.7</td>
<td>73</td>
<td>5.3</td>
</tr>
<tr>
<td>Six Lane Highway</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I-75</td>
<td>Between I-10 &amp; CR136, NB, Site 320, NB</td>
<td>66</td>
<td>73</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Between I-10 &amp; CR136, SB, Site 320, SB</td>
<td>66</td>
<td>74</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Between CR234 &amp; SR21, NB, Site 9904, NB</td>
<td>68</td>
<td>71</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Between CR234 &amp; SR21, SB, Site 9904, SB</td>
<td>67</td>
<td>71</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Between CR210 and I-295, NB, Site 9905, NB</td>
<td>67</td>
<td>72</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Midpoint CR210 and I-295, SB, Site 9905, SB</td>
<td>63</td>
<td>72</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Near Flagler CL, NB, Site 9905, NB</td>
<td>69</td>
<td>72</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Near Flagler CL, SB, Site 9905, SB</td>
<td>64</td>
<td>72</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td></td>
<td>mean</td>
<td></td>
<td>66.25</td>
<td>72.125</td>
<td>5.88</td>
</tr>
</tbody>
</table>

Figure 7 (a) shows the results for the Naïve and adjusted estimators for four-lane and six-lane highways. This figure clearly illustrates that the estimators for four-lane highways have a higher selection bias than for six-lane highways. For example, when the entry criterion is equal to 62, the adjusted estimator (δ=5.93, removing site-selection bias) for the six-lane group is just slightly higher than the Naïve estimator (δ=5.88). However, when the entry criterion is equal to 65, the adjusted estimator (δ=5.57) from the four-lane roads group is higher than the Naïve estimator (δ=5.30). For the same entry criterion equal to 66, the site selection bias for the four-lane roads group is 0.38 (δ=5.38-5.0) while the site selection bias for the six-lane roads group is 0.05. Although the absolute difference appears to be small, the relative difference in
percentages is actually large enough (e.g., 7.6%) that the bias needs to be accounted for; the curious reader is referred to Lord (2006) for additional details about the impact of a biased estimate even when the absolute difference is small. Furthermore, the true site-selection bias may be 15.2% (=7.6%×2) or more, since the adjusted method only captures 50% site-selection bias while biased estimator of mean and variance were used.

When the entire sample is analyzed, the difference between the Naïve estimators and adjusted estimators is more obvious (Figure 7 (b)). The adjusted method increases the Naïve estimator by partially removing the selection bias, while the average adjusted estimator from Park and Lord (2010) remains the same (note that the effects linked to the RTM are removed in their paper from each site’s estimator not from the average estimator). Moreover, Figure 7 shows that higher entry criteria cause higher selection bias, which underestimates the difference even more, and Equation (13) can partially eliminate selection bias but not remove all of it (as discussed above). It should be noted that differences are underestimated when the true difference is positive (|δ_{naive}|<|δ_{adjusted}|, δ > 0 ). On the other hand, the difference is overestimated when the true difference is negative (|δ_{naive}|>|δ_{adjusted}|, δ < 0 ). Generally, Figures 7 (a) and (b) show that higher criteria cause larger (more negative) site-selection bias and the adjusted estimators are closer to the real estimator than the one estimated using the Naïve method. These results are consistent with the (theoretical) Equation (13) and simulation results described above (Scenario 1). However, if the dataset follows another distribution, such as the lognormal distribution, the method described in Appendix B (see Equation B-3) can be used to adjust site-selection biases.
Figure 7. Estimated differences for the Naïve and adjusted methods.
7. Step-by-step procedure for using the adjusted method

This section describes the step-by-step procedure for using the adjusted method in order to better illustrate how it can be used with observed data or to adjust the results documented in previous studies (if enough information is available). The same dataset as the one used in the previous section is utilized here. For this description, the entry criterion was set to 66 mph. Among all sites, 13 were identified as having a speed equal to or above 67 mph in the before period: $Y_1^T = [68, 67, 69, 68, 67, 67, 69, 67, 69, 68, 67, 67, 69]$. The corresponding speeds in the after period are: $Y_2^T = [78, 74, 74, 74, 70, 71, 71, 71, 71, 72, 72, 72, 72]$. The steps are as follows:

**Step 1: Calculate the naïve estimate.**

$$\hat{\delta}_{\text{naive}} = \hat{\Lambda}_2 - \hat{\Lambda}_1 = \frac{\sum_{i=1}^{13} Y_{2i} - \sum_{i=1}^{13} Y_{1i}}{13} = \frac{942 - 882}{13} = 4.62$$

**Step 2: Estimate the value of variables in the estimator of the adjusted method (Equation 13).**

$$\hat{\rho} = \text{Corr}(Y_{1i},Y_{2i}) = \frac{\sum Y_{1i}Y_{2i} - \sum Y_{1i}\sum Y_{2i}}{\sqrt{m\sum Y_{1i}^2 - (\sum Y_{1i})^2}\sqrt{m\sum Y_{2i}^2 - (\sum Y_{2i})^2}}$$

$$= \frac{63913 - 882\times942}{\sqrt{13\times59850 - 882^2}\sqrt{13\times63913 - 942^2}} = 0.08$$

$$\hat{\phi} + \hat{\sigma}^2 = \text{Var}(Y_{2i}) = \frac{1}{m}\sum Y_{2i}^2 - \left(\frac{1}{m}\sum Y_{2i}\right)^2 = 5.1$$

$$\hat{d} = \frac{C - Y_{1i}}{\sqrt{\text{Var}(Y_{2i})}} = \frac{66 - (882/13)}{\sqrt{5.1}} = -0.82$$

**Step 3: Calculate the difference based on the adjusted method (Equation 13).**

$$\hat{\delta}_{\text{adjusted}} = \hat{\delta}_{\text{naive}} - (\hat{\rho} - 1)\sqrt{(\hat{\phi} + \hat{\sigma}^2)} f(\hat{d})/\left(1-F(\hat{d})\right)$$

$$= 4.62 - (0.08 - 1)\sqrt{5.1}\frac{0.13}{(1-0.21)}$$

$$= 4.62 - (-0.33) = 4.95$$
Repeat the above steps to get the naïve and the adjusted estimators for different entry criteria.

8. Summary and Conclusions

This study has examined how setting an entry criterion influences the estimation of interventions for continuous data. The proposed method documented in this research provides a useful approach to adjust the naïve estimator using the sample data without relying on the data collected for the control group. In this paper, the proposed method, a.k.a. the adjusted method, was compared to three commonly used methods in before-after studies: the Naïve, the CG, and ANCOVA methods. Four scenarios were evaluated: a direct comparison of the methods, different between-subject variances, different within-subject variances, and different sample sizes. The analysis was carried out using simulated data and observed speed data.

The study results showed that among all methods evaluated, the Naïve is the most significantly affected by the selection bias, which was expected. Using a control group or the ANCOVA method can eliminate the site-selection bias and RTM, as long as the characteristics of the control group are exactly the same as for the treatment group. However, the control group data that have characteristics based on a truncated distribution or sample may not be available in practice. Based on the simulated scenarios, the study results showed that higher entry criteria, higher within-subject variance, and smaller between-subject variance will cause a higher site selection bias. Equation (13) can partially eliminate site-selection bias even when biased estimators of the mean, variance, and correlation coefficient of a truncated normal distribution are used or are not known with certainty. The analysis performed using observed speed data collected in Florida supports the simulation results.

Based on the above findings, readers should have a better understanding about what constitutes site-selection bias and, consequently, when there may be a need to adjust previously estimated treatment effectiveness documented in published studies. For example, if a loose entry criterion was used in a naïve before-after study (such as using the mean value, 70 mph, in the simulation analysis shown above), the naïve estimator might overestimate the effectiveness when the treatment is in effect not working ($\delta_{true} = 0$, but $\delta_{naive}$ could show a reduction up to 4 mph). Also, the results shown in this paper can help identify potential problems in previous studies that used a CG or an EB approach for cases when the characteristics between the reference and treatment group data are known to be different (e.g., the sample mean for the control group being 1.2 time higher than for the treatment group, as described above). The site-selection bias estimated using these two methods might be even higher than the bias associated with the naïve
estimator. Finally, the analysis carried out in this work show that higher site-selection biases will exist in naïve before-after studies for the following conditions: observations (1) are selected with a higher entry criterion, (2) are more homogeneous (e.g., having a large proportion of drivers traveling with the same driving speed), or (3) have a larger variance for each measurement (e.g., a long time gap between the before and after periods).

Given the nature of the work documented in this paper, there are many avenues for further work. First, since this research used a biased estimator for the mean, variance, and correlation coefficient to adjust the naïve estimator, it may be beneficial to apply more advanced techniques to estimate the parameters of a truncated normal distribution model in order to obtain more precise estimates. Aside from properly estimating the site-selection bias, an important research question would be to determine the magnitude of the error associated with its estimated value. Based on the simulation results, the errors of the adjusted method are close to the ones for the Naïve method. However, it might not be true for every scenario analyzed, since there is currently no theoretical equation that exists to estimate the actual magnitude of the variance associated with the site-selection bias. This should be further examined. Second, more work needs to be done when multiple entry criteria in before-after studies are used, especially when those come from different types of data. For instance, the implementation of a traffic countermeasure may be based on two warrants: a site having more than five crashes in the past year, and where the observed 85%-percentile driving speed must be over 60 mph. Third, guidelines should be developed to define what the entering criterion should be when it is not known (e.g., minimum value, speed limit, etc.). Finally, a simpler approach for displaying the site-selection bias should be examined. Tables based on the sample mean, entry criteria, and the level of between-subject and within-subject variance could perhaps be provided in design manuals or similar types of documents for continuous variables.
**Appendix A:**

Between-subject variance: The variation of one-time measurements observed from different subjects.
Within-subject variance: The variation of multiple-time measurements observed from one subject.
Entry criteria: The conditions that researchers used to select subjects for treatments. It could be one minimum/maximum value, a range of value, or other formats.

**Appendix B: Applying the adjusted method to data following the lognormal distribution**

This appendix shows how to adjust site-selection bias when the data follow the lognormal distribution. Let \( Z_{ij}^T, Z_{ij}^{CG} \) be the new observation response for site i and time period j, where the logarithm of \( Z \) is \( Y \). The variable \( Y \) still follows the normal distribution (same setting as before). Their distributions are given as follows:

\[
Z_{ij}^{T,CG} \sim LN(\Lambda_j, \sigma^2 + \phi) \\
Y_{ij}^{T,CG} \sim N(\Lambda_j, \sigma^2 + \phi)
\]

Let the safety effectiveness (\( \theta = \frac{\hat{Z}_2^T}{\hat{Z}_1^T} \)) be the target index to estimate the effectiveness of treatments. The naive estimator of \( \theta \) is:

\[
\hat{\theta} = \frac{\hat{Z}_2^T}{\hat{Z}_1^T} = \frac{E(Z_{ij}^T)}{E(Z_{ij}^T)} = \frac{e^{\Lambda_{2} + \frac{1}{2}(\sigma^2 + \phi)}}{e^{\Lambda_{1} + \frac{1}{2}(\sigma^2 + \phi)}} = e^{\Lambda_{2} - \Lambda_{1}}
\]  

(B.1)

In order to simplify the equation, the logarithm of \( \theta \) can be given as:
\[
\log(\hat{\theta}) = \log(e^{\Lambda_2 - \Lambda_1}) = \Lambda_2 - \Lambda_1 = \delta \tag{B.2}
\]

It should be noted that \( \delta \) is the "difference"; the same index that was discussed in the main text of the paper for the normally distributed data. Therefore, Equation (13) can be used to get the adjusted estimator of \( \theta \). The new adjusted estimator is given as follows:

\[
\hat{\theta}_{\text{adjusted}} = e^{\left(\delta_{\text{new}} - (\hat{\rho} - 1)\sqrt{\phi^2 + \sigma^2}f(\hat{\delta}) / (1 - F(\hat{\delta}))\right)}
\tag{B.3}
\]

In addition, the method described in Equation (B.3) is much easier to use than the traditional approach proposed by Cook and Wei (2002) because the conditional lognormal distribution is the lognormal distribution with mean

\[
E(Z_{i2} | Z_{i1}) = \exp\left[\mu_2 + \rho(\ln Z_{i1} - \Lambda_i) + \frac{1}{2}(1 - \rho^2)\sqrt{\sigma^2 + \phi}\right] \text{ which in turn gives}
\]

\[
E(Z_{i2} - Z_{i1} | Z_{i1} > C) = E(Z_{i2} | Z_{i1} > C) - E(Z_{i1} | Z_{i1} > C)
\]

\[
= \left\{\exp\left[\mu_2 + \rho(\ln Z_{i1} - \Lambda_i) + \frac{1}{2}(1 - \rho^2)\sqrt{\sigma^2 + \phi}\right] | Z_{i1} > C\right\} - \exp\left[\Lambda_i + \frac{(\sigma^2 + \phi)}{2}\right] \times \left[1 - F\left(\frac{\log(c) - \Lambda_i - \sigma^2 - \phi}{\sqrt{\sigma^2 + \phi}}\right)\right]
\tag{B.4}
\]

Furthermore, the unbiased estimator of difference, \( \delta \), is given as:

\[
E(Z_{i2} - Z_{i1}) = \delta = \exp\left[\Lambda_i + \frac{(\sigma^2 + \phi)}{2}\right] - \exp\left[\Lambda_2 + \frac{(\sigma^2 + \phi)}{2}\right]
\tag{B.5}
\]

Comparing Equations (B.4) and (B.5), one can see that it is very difficult to estimate the selection bias estimator for the difference \( \delta \). In other words, using Equation (B.3) for estimating
the selection bias for the safety effectiveness, $\theta$, is much easier than estimating bias of the difference, $\delta$, when the data are assumed to follow the lognormal distribution in before-after studies.

**References**


Lord, D., 2006. Modeling motor vehicle crashes using Poisson-gamma models: Examining the effects of low sample mean values and small sample size on the estimation of the fixed dispersion parameter. Accident Analysis & Prevention 38, 751-766.


