# Uncertain Neighbors: Bayesian Propensity Score Matching for Causal Inference

R. Michael Alvarez* Ines Levin†

April 18, 2014

**Abstract**

In this paper we compare the performance of standard nearest-neighbor propensity score matching with that of an analogous Bayesian propensity score matching procedure. We show that the Bayesian approach has several advantages, including that it makes better use of available information, since it makes less arbitrary decisions about which observations to drop and which ones to keep in the matched sample. Additionally, the Bayesian method produces parameter samples that can be used to easily compute summary measures of the distribution of treatment effects; allows evaluating the sensitivity of treatment effects to alternative priors; and can be used to produce interesting visualizations of the distribution of quantities of interest. We conduct a simulation study to evaluate the performance of standard and Bayesian nearest-neighbor matching when the propensity score model is correctly specified, as well as under different misspecifications of the propensity score model. Lastly, we use both methods to replicate a recent study about the impact of land reform on guerrilla activity in Colombia.

*Professor of Political Science, Division of the Humanities and Social Sciences, California Institute of Technology, rma@hss.caltech.edu.

†Assistant Professor of Political Science, Department of Political Science, University of Georgia, ilevin@uga.edu.

# 1 Introduction

Matching methods are commonly used by social scientists to measure causal effects based on observational data—that is, in situations where the researcher has no control over the assignment of observations to causal states—or to correct randomization failures in the context of experimental or quasi-experimental research (Cochran and Rubin, 1973; Rubin, 1979). These methods allow comparing values of the outcome variable across observations that are similar in every (observed) relevant way except for differences in exposure to a presumed cause. A variety of procedures have been developed to determine the similarity of observations and to match observations conditional on similarity—or more specifically, conditional on a summary measure of the distance between observations. One such distance measure is the propensity score, commonly defined as the probability of being exposed to alternative causal states or *treatments*, conditional on determinants of treatment assignment (Heckman, Ichimura, and Todd, 1997; Rosenbaum and Rubin, 1983). The true propensity score is an unobserved quantity, and is typically estimated using a regression approach. In this paper, we propose and evaluate the performance of a simple method for incorporating information about estimation uncertainty in the propensity score.

Even though the idea of using estimated propensity scores for matching observations emerged decades ago (Rosenbaum and Rubin, 1983, 1985), and though alternative matching procedures have been developed in recent years (see for instance, Diamond and Sekhon 2013; Iacus, King, and Porro 2012), propensity score matching methods are still widely applied and researchers continue proposing extensions and generalizations of propensity score-based procedures (Imai and Dyk, 2004; Imai and Ratkovic, 2012; Zigler and Dominici, 2014). The propensity score approach was developed as a way to solve the dimensionality problems that would ensue if researchers tried to control for observed differences between treatment and control observations by stratifying on multiple covariates. Estimated propensity scores can be used to match observations treated with exposure to alternative causal states, using non-parametric matching procedures, in order to construct matched samples that are well-

balanced on relevant covariates (Dehejia and Wahba, 2002); where *well-balanced* means that covariates are similarly distributed between treatment groups. If the matching method succeeded in balancing all relevant confounders, then differences in outcomes across treatments can be attributed to differences in exposure to the presumed cause.

[Figure 1 about here.]

Despite the ease of use and popularity of the propensity score, it has a number of limitations. Since we do not know the true mechanism underlying treatment assignment, estimated propensity scores may deviate in unknown ways from true assignment probabilities, producing biased measures of causal effects (Rosenbaum, 1999). Furthermore, propensity score matching procedures are usually conducted in two stages: one involving the estimation of propensity scores, and another one in which treated and control observations are matched based on point estimates of the distance measure (Stuart, 2010). But this ignores the fact that propensity scores are themselves estimated quantities, and as such there is always some degree of measurement uncertainty associated with those estimates (Tu and Zhou, 2002). This idea is illustrated in Figure 1 (constructed with synthetic data), where the top-most plot depicts the distribution of point estimates of propensity scores among treatment and control units (information typically used as input to propensity score matching algorithms) and the bottom-most plot depicts 95% credible intervals for the propensity score corresponding to each observation in the sample (information typically disregarded by propensity score matching algorithms). Even though analysts always have some amount of uncertainty about estimated propensity scores, standard matching algorithms treat the distance measure as a fixed quantity that is known with certainty (McCandless et al., 2009). The primary purpose of our paper is to propose and evaluate the performance of a simple method for accounting for estimation uncertainty in the propensity score.

A number of scholars have already made progress in investigating the extent to which incorporating uncertainty in the propensity score would have an impact on standard errors associated with measurements of causal effects. Tu and Zhou (2002) sought to incorporate

uncertainty in the propensity score using a bootstrap method, and found that doing so leads to larger standard errors associated with estimates of treatment effects. However, the reliability of the bootstrapping approach has been called into question in recent years (Abadie and Imbens, 2008). McCandless et al. (2009) were the first to propose using Markov Chain Monte Carlo (MCMC) methods to account for uncertainty in the propensity score. Consistently with the previous two studies, they found that incorporating uncertainty in the propensity score leads to wider Bayesian credible intervals in the context of propensity-score-based stratification and regression. An (2010) followed a similar approach, using Bayesian methods to account for uncertainty in the propensity score in the context of propensity-score-based regression and matching, but in contrast to McCandless et al. (2009), found that doing so leads to lower standard errors.

Both McCandless et al. (2009) and An (2010)'s Bayesian inference procedures involve the simultaneous estimation of propensity score and outcome models.[1] Kaplan and Chen (2012) criticize the simultaneous-estimation approach on the basis that, by estimating both models simultaneously, outcome data is allowed to inform the estimation of the propensity score—a potentially problematic attribute of the procedure, since it may introduce selection bias. While the propensity score should incorporate information about the treatment-assignment mechanism, it should not incorporate information about the outcome or treatment effect. To address the previous issue, Kaplan and Chen (2012) proposed first estimating the propensity score model using MCMC methods; then repeatedly estimating treatment effects (using regression or non-parametric approaches), each time considering a different sample from the posterior distribution of the propensity score; and finally computing the mean and variance of estimated treatment effects across samples.

In this paper, we evaluate the performance of Bayesian propensity score matching (BPSM) using a sequential estimation procedure along the lines of the method proposed by Kaplan

[1]See also Zigler and Dominici 2014, who develop a Bayesian method for jointly estimating propensity score and outcome models that allows accounting for uncertainty in the specification of the propensity score model.

and Chen (2012), which avoids the problems discussed in the previous paragraph while still allowing the incorporation of information about the uncertainty in the propensity score. We find that BPSM has important advantages relative to standard propensity score matching (PSM), including that—as is generally the case with Bayesian estimation procedures—it produces samples of model parameters that can be used to summarize results and compute quantities of interest such as measures of centrality and dispersion (Jackman, 2000), allowing the analyst to get a more accurate sense of the estimation uncertainty in treatment affects.

An important difference between this paper and previous studies of the performance of Bayesian propensity score regression and matching (such as McCandless et al. 2009, An 2010, and Kaplan and Chen 2012), is that we focus on procedures where control units that are not comparable to treatment units are dropped from the analysis and are not taken into account for the computation of treatment effects. Dropping control observations with no close matches in the treatment group is a standard practice that allows achieving better balance in the matched sample (Ho et al., 2007). We argue that accounting for estimation uncertainty is particularly important in the case of matching procedures that entail dropping observations, since the decision of whether to keep or drop a unit is based on the estimated distance measure. Indeed, we find that another reason why BPSM improves upon standard PSM is that the decision of whether to keep or drop observations from the matched sample is done in a less arbitrary manner.

[Figure 2 about here.]

The arbitrariness of discarding units based on point estimates of propensity scores is illustrated in Figure 2. This figure depicts the relationship between estimated propensity scores and the decision of whether to drop observations from the matched sample, when performing nearest-neighbor matching within caliper. In the case of control observations in the synthetic dataset used for constructing this example, dropping decisions are relatively rare for point estimates of propensity scores above 0.2. This suggests that small changes in the distance measure (i.e. point estimates of propensity scores) may alter the composition of

control units kept in the matched sample, especially for observations with point estimates of the propensity score close to the 0.2 cutoff. We argue that incorporating uncertainty in the propensity score by repeatedly matching treatment and control observations based on draws from the posterior distribution of propensity scores (instead of matching only once based on point estimates of the propensity score) can satisfactorily address the arbitrariness problem, and (by making more efficient use of the available information) can lead to estimates of treatment effects that are more representative of those prevailing in the target population. Moreover, using more data should lead to higher precision, reflected in lower standard error and narrower confidence intervals.

The rest of the paper is organized as follows. First, we describe the standard and Bayesian propensity score matching procedures evaluated in this paper. Second, we report the results of simulation studies conducted to compare the performance of both methods. We consider hypothetical situations where the propensity score model was correctly specified, as well as situations where a relevant covariate was omitted from the propensity score model. After that, we replicate a recent study about the impact of land reform on the frequency of guerrilla insurgency in Colombia, using both standard and Bayesian propensity score matching procedures. This application allows us to evaluate the impact of incorporating uncertainty in the propensity score into the computation of average treatment effects in the context of a real-world application. We conclude with a brief discussion of our results and a summary of the benefits of the Bayesian approach.

# 2 Methodology

In this section we describe the propensity score matching procedures implemented in this paper. Let $Z_i$ denote an indicator of treatment assignment, which takes values one for individuals exposed or assigned to the treatment, and zero for individuals in the control group. Further, let $Y_{zi}$ indicate the potential outcome for an individual $i$, which takes values

$Y_{1i}$ if the individual is exposed to the treatment, and $Y_{0i}$ otherwise. The individual treatment effect can be defined as the difference between the potential outcome under the treatment, and the potential outcome under the control. Since each individual is exposed to a single causal state (meaning that for each $i$, either $Z_i = 1$ or $Z_i = 0$), it is not possible to measure individual treatment effects (Holland, 1986). However, numerous causal inference techniques have been developed that allow researchers to estimate average treatment effects.

For each individual, let $Y_i$ denote the observed outcome and let $X_i$ denote observed covariates thought to affect both the outcome and the probability of exposure to the treatment, $P(Z_i = 1|X_i)$. If the outcome variable is a binary indicator that takes value one for any individual with probability $P(Y_i = 1|Z_i, X_i)$, the Average Treatment Effect (ATE) in the population can be defined as the change in $P(Y_i = 1|Z_i, X_i)$ caused by a change in $Z_i$:[2]

$$ATE = P(Y_i = 1|Z_i = 1, X_i = x) - P(Y_i = 1|Z_i = 0, X_i = x) \quad (1)$$

In this paper, we consider two alternative approaches to estimating treatment effects: standard nearest-neighbor propensity score matching (PSM) and Bayesian nearest-neighbor propensity score matching (BPSM). In the latter case, we use MCMC methods to estimate the propensity score model, and then estimate the treatment effects using a non-parametric propensity-score matching procedure. Next, we describe the main characteristics of the procedures that we use to measure treatment effects (PSM, BPSM, and post-matching model-based adjustments).

[2]If the outcome variable is continuous, then the ATE in the population can be defined as the change in the conditional expectation of $Y_i$ caused by a change in $Z_i$. That is, $ATE = E[Y_i|Z_i = 1, X_i = x] - E[Y_i|Z_i = 0, X_i = x]$.

## 2.1 Propensity Score Matching (PSM)

In the case of standard nearest-neighbor propensity score matching (PSM), we estimate the propensity score model using a logistic regression approach, such that:

$$logit[P(Z_i = 1|X_i)] = \Gamma X_i \tag{2}$$

where $\Gamma$ is a vector of coefficients and $X_i$ is a vector of observed individual attributes.

Subsequently, we match treatment and control observations on the basis of estimated propensity score, $\hat{PS}_i = \hat{P}(Z_i = 1|X_i)$. We search for matches using a non-parametric procedure, whereby treatment units are sorted in terms of the estimated distance measure from largest to smallest, and then each unit is matched one at a time to the nearest control unit(s) along the distance measure.

Most frequently, we use a slightly modified matching procedure where, for each treatment unit, we only consider control units lying within a maximum distance on the distance measure. The maximum distance used to determine eligible matches is termed a *caliper* and is specified in terms of number of standard deviations of the distance measure (Rosenbaum and Rubin, 1985). For each treatment unit, we select a single match at random from those control units lying within the chosen caliper.

### 2.1.1 Average Treatment Effect

When the outcome variable is a binary indicator, $\hat{P}_0$ denotes the estimated probability of success under the control; and $\hat{P}_1$ denote the estimated probability of success under the treatment. For units within the matched sample, we estimate $\hat{P}_0$ as the average observed outcome taken over all matched control units:

$$\hat{P}_0 = \frac{1}{\tilde{N}_0} \sum_{i=1, i \in \tilde{C}}^{\tilde{N}_0} Y_{0i} \tag{3}$$

where $\tilde{C}$ denotes the matched control group, $\tilde{N}_0$ denotes the number of matched control

units, and $Y_{0i}$ denotes the observed outcome for individual $i$ in the matched control group. Similarly, we estimate $\hat{P}_1$ as the average observed outcome taken over all matched treatment units:

$$\hat{P}_1 = \frac{1}{\tilde{N}_1} \sum_{i=1, i \in \tilde{T}}^{\tilde{N}_1} Y_{1i} \tag{4}$$

where $\tilde{T}$ denotes the matched treatment group, $\tilde{N}_1$ denotes the number of matched treatment units, and $Y_{1i}$ denotes the observed outcome for individual $i$ in the matched treatment group.

We estimate the ATE as the difference between the estimated probability of success under the treatment ($\hat{P}_1$) and the estimated probability of success under the control ($\hat{P}_0$):[3]

$$\hat{ATE} = \hat{P}_1 - \hat{P}_0 \tag{5}$$

Note that while the PSM procedure produces a point estimate of the ATE, it does not produce indicators of dispersion. Therefore, unless additional analyses are conducted after matching, it is not possible to evaluate the uncertainty about treatment effects.

## 2.2 Bayesian Propensity Score Matching (BPSM)

First, we estimate a propensity score model similar to that given in equation (2), but using using MCMC methods. The Bayesian estimation procedure produces samples of the $\Gamma$ vector of parameters of the propensity score model, $[\hat{\Gamma}^{(1)} \dots \hat{\Gamma}^{(K)}]$, where $K$ denotes the total number of saved iterations. These samples can, in turn, be used to calculate samples of estimated linear predictors, $[\hat{\Gamma}^{(1)} X \dots \hat{\Gamma}^{(K)} X]$, and samples of estimated propensity scores, $[\hat{P}^{(1)} \dots \hat{PS}^{(K)}]$, either of which can later be used as distance measures for the matching procedure.

For each $k$ sample, we matched treatment and control units on the basis of the estimated

[3]When the outcome variable is continuous, the average treatment effect is computed as the difference between the average value of the outcome variable in the treatment group, and the average value of the outcome variable in the control group.

propensity score, using a nearest-neighbor propensity score matching procedure similar to that described before. Subsequently, we estimate the probability of success under the control for matched sample $k$, $\hat{P}_0^{(k)}$, as the average observed outcome taken over all matched control units:

$$\hat{P}_0^{(k)} = \frac{1}{\tilde{N}_0^{(k)}} \sum_{i=1, i \in \tilde{C}^{(k)}}^{\tilde{N}_0^{(k)}} Y_{0i} \tag{6}$$

where $\tilde{C}^{(k)}$ denotes the matched control group in sample $k$ and $\tilde{N}_0^{(k)}$ denotes the number of matched control units in sample $k$. Similarly, we estimate the probability of success under the treatment for matched sample $k$, $\hat{P}_1^{(k)}$, as the average observed outcome taken overall all matched treatment units:

$$\hat{P}_1^{(k)} = \frac{1}{\tilde{N}_1^{(k)}} \sum_{i=1, i \in \tilde{T}^{(k)}}^{\tilde{N}_1^{(k)}} Y_{1i} \tag{7}$$

where $\tilde{T}^{(k)}$ denotes the matched treatment group in sample $k$ and $\tilde{N}_1^{(k)}$ denotes the number of matched treatment units in sample $k$. For each matched sample $k$, we estimate the average treatment effect similar to in equation (5):

$$\hat{ATE}^{(k)} = \hat{P}_1^{(k)} - \hat{P}_0^{(k)} \tag{8}$$

Thus, the BPSM procedure does not produce a single point estimate of the ATE, but a sample of size $K$ of estimated ATEs. This sample can be used to produce summary measures of the posterior distribution of the estimated ATE, including measures of centrality such as the mean, and measures of dispersion such as credible intervals. In contrast to PSM, BPSM can be used to obtain not only point estimates of the ATE, but also associated measures of uncertainty.

## 2.3 Post-Matching Regression Adjustment

Since PSM and BPSM are not exact matching procedures, neither will typically produce matched samples that are perfectly balanced on observables. To control for remaining imbalances in covariates, it is suggested that analysts conduct post-matching regression adjustments in the matched sample (Rubin, 1973, 1979; Rubin and Thomas, 2000). According to Ho et al. (2007), results produced by parametric adjustment on the matched sample should be less model-dependent than similar adjustments on the original dataset, since it is only conducted among comparable treatment and control observation and therefore avoids extrapolation. When the outcome variable is binary indicator, we perform regression adjustment on the matched sample by estimating logistic regressions of the following form:[4]

$$logit[P(Y_i = 1|Z_i, X_i)] = \beta Z_i + \Theta X_i \tag{9}$$

where $\beta$ is the coefficient associated with the binary indicator of treatment assignment $Z_i$, and $\Theta$ is a vector of coefficients capturing the effect of covariates $X_i$ (the same set of covariates included in the matching procedure).

In the case of PSM, regression (9) is typically estimated using maximum likelihood, and then a simulation procedure is used to estimate the effect of switching the treatment assignment variable $Z_i$ from zero to one while holding $X_i$ constant. This procedure produces a series simulated treatment effects that can be used to obtain summary measures of the distribution of the ATE, including point estimates and associated measures of uncertainty.

In the case of BPSM, we repeatedly estimate regression (9) within each $k$ matched sample using MCMC methods. Estimating the outcome model using Bayesian methods produces a series of $S$ draws from the posterior distribution of $\beta$ and $\Theta$ parameters for *each* $k$ matched sample, which can then be used to summarize quantities of interest. Thus, the BPSM procedure with post-matching regression adjustment produces $K \times S$ samples of estimated

[4]When the outcome variable is continuous, we perform regression adjustment by estimating linear regressions on the matched sample, under the assumption that: $E[Y_i|Z_i, X_i] = \beta Z_i + \Theta X_i$.

average treatment effects. When the outcome variable is a binary indicator, each draw of the estimated ATE is computed as:[5]

$$\hat{ATE}^{(k,s)} = \hat{P}(Y = 1|Z = 1, X = \bar{X})^{(k,s)} - \hat{P}(Y = 1|Z = 0, X = \bar{X})^{(k,s)} \tag{10}$$

where $\bar{X}$ is a vector of mean or median levels of covariates included in both the matching procedure and the subsequent model-based adjustment. The sample of adjusted ATE estimates can be used to summarize different aspects of the posterior distribution of the average treatment effect, including measures of centrality and dispersion, controlling for imbalances in observed covariates that might have remained following the application of the matching procedure.[6]

# 3 Simulation Studies

In this section of our paper, we undertake a number of different simulation studies to examine the relative performance of the PSM and BPSM approaches. We begin with the obvious simulation: we know the specification of the propensity score and outcome models, and we estimate the true model using both approaches. These simulations indicate that the BPSM approach is superior to PSM for recovering estimates of the treatment effects. The second set of simulations consider what is probably a more likely scenario for applied researchers: the propensity score model is misspecified due to the absence of a confounder. Here we present simulation results from a variety of simple misspecifications, and in all cases we present evidence that the BPSM approach dominates the PSM approach. Thus, the simulation studies presented in this paper suggest that the BPSM approach is superior to the PSM approach.

[5]When the outcome variable is continuous, the adjusted $\hat{ATE}$ equals the estimate of the coefficient associated with the treatment variable in the linear outcome model. That is, $\hat{ATE}^{(k,s)} = \hat{\beta}^{(k,s)}$.

[6]The BPSM procedure can be easily implemented with the aid of existing R packages. In this paper, Bayesian estimation of propensity score models was performed using *MCMCpack* (Martin et al., 2011), and nearest-neighbor matching steps were conducted using *MatchIt* (Ho et al., 2011).

## 3.1 Correctly Specified Propensity Score Model

We used Monte Carlo simulation to evaluate the performance of PSM and BPSM under a correctly specified propensity score and outcome model, for different sample sizes (500 and 1,500) and magnitudes of the treatment effect (determined by different values of the true $\beta$ coefficient). For each combination of sample size and magnitude of treatment effects, we generated $J = 1,000$ synthetic data sets using the following procedure:

1. Generate J covariate matrices, $[X^{(1)} \dots X^{(J)}]$, where each $X^{(j)}$ includes a vector of ones, a binary indicator $x_1^{(j)}$, and a continuous variable $x_2^{(j)}$. At each step $j$ of the simulation procedure, draw variables $u^{(j)}$ and $x_2^{(j)}$ from a multivariate normal distribution. Variable $x_1^{(j)}$ is a binary indicator that equals one when a variable $u^{(j)}$ takes positive values and otherwise equals zero. For the baseline simulation study, we assume $x_1^{(j)}$ and $x_2^{(j)}$ are independent by setting the covariance between $u^{(j)}$ and $x_2^{(j)}$ to zero.

2. Conditional on simulated $X^{(j)}$'s and a fixed vector of parameters $\Gamma = [-1, 2, 2]$ (with the first element being an intercept, and the second and third elements capturing the impact of $x_1$ and $x_2$, respectively, on treatment assignment) generate J probabilities of treatment assignment (i.e. true propensity scores), $[PS^{(1)} \dots PS^{(J)}]$, based on a logistic regression model similar to that given in expression (2). Then, generate J indicators of treatment assignment, $[Z^{(1)} \dots Z^{(J)}]$, by repeatedly drawing $Z^{(j)}$'s from a binomial distribution with probabilities given by true propensity scores $PS^{(j)}$'s.

3. Conditional on simulated $x_1^{(j)}$'s, $x_2^{(j)}$'s, $Z^{(j)}$'s, a fixed coefficient $\beta$ capturing the effect of the treatment, and fixed parameters $\theta_1 = 1$ and $\theta_2 = -1$ capturing the impact of $x_1$ and $x_2$ on the outcome, respectively, generate J probabilities of success, $[P_y^{(1)} \dots P_y^{(J)}]$, based on a logistic regression model similar to that given in expression (9). Then, generate J binary outcome indicators, $[Y^{(1)} \dots Y^{(J)}]$, by repeatedly drawing $Y^{(j)}$'s from a binomial distribution with probabilities given by simulated $P_y^{(j)}$'s.

After generating $J = 1,000$ synthetic data sets, we applied PSM and BPSM procedures similar to those described in sections 2.1 and 2.2. For each $j$ data set, we discarded control units outside the support of the propensity score in the treatment group, and then performed one-to-one matching without replacement, searching for matches within a caliper of 0.5 standard deviations of the estimated propensity score (in the case of BPSM, this was repeated at each iteration). For both procedures (PSM and BPSM), estimates of treatment effects were computed after performing post-matching regression adjustment. The simulation procedure allowed us to evaluate both the bias and error associated with each matching procedure by comparing ATEs estimated using PSM and BPSM with true ATEs. True average treatment effects were computed by calculating individual treatment effects (which vary as a function of $x_{1i}$ and $x_{2i}$ for each individual, since the outcome model is non-linear) and then taking the average over the population, for each $j$ step of the simulation procedure.

[Table 1 about here.]

The results of the initial stage of our simulation study are reported in Table 1. The first column of Table 1 gives the percentage of observations kept in the matched sample (PSM rows) or retained *at least once* during the matching procedure (BPSM rows). With the configuration of parameters described above, between 68% (for N = 500) and 70% (for N = 1,500) of observations are matched in the case of PSM, and similar percentages are matched on average at each iteration of BPSM. However, between 94% (for N = 500) and 97% (for N = 1,500) of observations are matched at least once (that is, are kept in the matched sample for at least one iteration of the matching procedure) for BPSM. These results suggest that treatment effects computed using BPSM incorporate information from a larger number of observations, compared to PSM.

The next six columns of Table 1 provide the following information: the mean value and 95% credible interval for the estimated ATE; the bias of the estimated ATE (average of the absolute value of the difference between the true ATE and the estimated ATE), the mean squared error of the estimated ATE (average of the squared difference between the true ATE

and the estimated ATE); and the coverage probability (proportion of simulations where the true ATE fell inside the 95% credible interval for the estimated ATE).

While both PSM and BPSM lead to measures of the estimated ATE that are relatively close to the true ATE, the magnitude of the bias is considerably lower for BPSM for each of the four combinations of sample sizes and magnitude of treatment effects. In order to interpret this last result it is important to take into account all potential sources of bias. On the one hand, neither matching procedure produces perfectly balanced samples, so part of the bias can be explained by remaining imbalances in covariates that might lead to adjusted estimates of treatment effects that are slightly model-dependent. On the other hand, within-caliper nearest-neighbor matching requires dropping control units located far from the treatment group along the distance measure, as well as treatment observations that remain unmatched after all potential control matches have been used (this is the case because matching is conducted without replacement), so part of the bias can also be explained by the fact that the matched sample is not perfectly representative of the original sample but biased toward specific values of observed covariates. We argue that one of the reasons why BPSM usually produces lower bias relative to PSM is that, by making less arbitrary decisions about which observations to keep or drop during the matching procedure, it allows incorporating information from a larger number of observations, and as a results leads to estimated ATEs that are more representative of true ATEs in the target population.

Additionally, the spread and MSE of estimated ATEs (both of which decrease markedly when the sample size increases from 500 to 1,500) are always lower in the case of BPSM; and the probability that the true ATE falls within the 95% credible interval (i.e. the coverage probability) is always higher under BPSM than under PSM. Since BPSM leads to lower bias, spread, and MSE, as well as higher coverage probability compared to PSM for all configuration of parameters and sample sizes, we conclude that BPSM performs considerably better than PSM for recovering treatment effects when the model is correctly specified.

## 3.2 Misspecified Propensity Score Model

In Tables 2-4 we present the results of similar simulation studies with the alteration that the propensity score model (as well as the outcome model used for the post-matching regression adjustment) is misspecified due to the omission of a confounder. Table 2 corresponds to a situation where the missing confounder is an interaction between variables $x_1$ and $x_2$ that has a positive effect on both $PS$ (the true propensity score) and $P_y$ (the probability of a successful outcome). In the case of Table 3, the missing confounder is a third variable $x_3$ drawn from the same multivariate distribution as $u$ (the continuous variable used to determine $x_1$) and $x_2$ that has low covariance with included covariates ($\sigma_{u,x_3} = 0.1$ and $\sigma_{x_2,x_3} = 0.1$). Lastly, Table 4 corresponds to a situation where the missing confounder is again a third variable $x_3$ drawn from the same multivariate distribution as $u$ and $x_2$, but with high covariance with included covariates ($\sigma_{u,x_3} = 0.7$ and $\sigma_{x_2,x_3} = 0.7$). In the last two cases (i.e. Tables 3 and 4), the omitted confounder $x_3$ has a positive effect on both $PS$ and $P_y$.

[Tables 2-4 about here.]

The omission of a confounder leads to biased treatment effects in the case of both nearest-neighbor propensity score matching procedures (PSM *and* BPSM). The magnitude of the bias depends on the impact of the confounder on $PS$ and $P_y$, as well as on the correlation between the excluded confounder and included covariates. Moreover, since the propensity score and outcome models are non-linear, the magnitude of the bias should also depend on the impact of the treatment and included covariates on $PS$ and $P_y$ (in the case of logistic link functions, the further from 0.5 the true values of $PS$ and $P_y$, holding constant the level of the omitted confounder, the lower the magnitude of the bias produced by the omission of a confounder).

We found that the bias is considerably larger for PSM than for BPSM, for the three types of misspecifications and for all combinations of treatment effect size and sample size. Therefore, our results suggest that the magnitude of the bias also depends on the extent

to which the matching procedure incorporates uncertainty about the propensity score. In the cases shown in Tables 3 and 4, we held constant the impact of the omitted confounder on $PS$ and $P_y$, and varied its correlation with the two included covariates. For both PSM and BPSM, the largest bias occurred when the true ATE was small and when the omitted confounder was a third variable with low correlation with included covariates (see top section of Table 3). In that case, either matching procedure would lead the researcher to conclude that the ATE is at least 14 percentage points larger than the true ATE. However, the bias is about 3 percentage points larger for PSM compared to BPSM.

The omission of a confounder does not only lead to biased estimates of treatment effects, but also to higher MSE, and more so for PSM than BPSM. Although the MSE decreases considerably for larger sample sizes (i.e. for $N = 1,500$ compared to $N = 500$), the probability that the true ATE falls within the 95% credible interval for the estimated ATE (i.e. the coverage probability) also decreases markedly with the sample size. The explanation for the last result is that while the omission of a relevant confounder leads to biased credible intervals, regardless of the sample size, the interval is narrower (and therefore less likely to contain the true treatment effect) when the sample size is larger. Although the reduction in coverage probability caused by the model misspecification is considerable for both matching procedures (especially when the omitted confounder has low correlation with included covariates, as in Table 3), it is always more pronounced for PSM and for BPSM.

[Figures 3 and 4 about here.]

Overall, the results of our simulation studies indicate that incorporating information about uncertainty in the propensity score leads to: (1) lower bias of estimates of treatment effects, and (2) lower dispersion among estimates of treatment effects. These two results hold regardless of whether the propensity score model is correctly specified or misspecified due to the omission of a relevant confounder. Figure 3 illustrates these findings for a simulation study conducted under the assumption of small sample size ($N = 500$) and small treatment effect ($\beta = .25$). For the four situations depicted in the figure, the distribution of bias (i.e.

deviations from true effects) has wider spread in the case of PSM, meaning that the absolute value of the bias tends to be larger for this procedure compared to BPSM. Additionally, while both methods lead to biased estimates when the propensity score model is misspecified (panels b, c, and d), the bias is always more pronounced in the case of PSM. Similar results are illustrated in Figure 4, which shows the relationship between bias of PSM and bias of BPSM estimates across simulated datasets. In the case of misspecifications (c) and (d), most points are located in the first quadrant of each plot (indicating that both methods produce upwardly biased estimates of treatment effects), but the magnitude of the bias is systematically higher for PSM compared to BPSM. We found comparable results for different combinations of sample sizes and magnitude of treatment effects.

# 4 Application: Land Reform and Insurgency in Colombia

In order to demonstrate the practical utility of BPSM, we turn in this section to an application: a replication of a recent study by Albertus and Kaplan (2013) about the impact of land reform on guerrilla activity in Colombia. The central question investigated by Albertus and Kaplan (2013) is whether land reform could serve as a tool for reducing guerrilla warfare, by addressing income inequality and improving the living conditions of peasants who might otherwise support insurgency. The authors investigate this question using municipality and municipal-year data covering a 12-year period ranging from 1988 to 2000. One of the methods that they employ to assess the impact of land reform is propensity score matching, where the treatment variable is a binary indicator of "at least 300 plots reformed from 1988 to 2000, and at least three years with fifty or more plots reformed" (Albertus and Kaplan, 2013, p. 215). The outcome variable is a count of the number of guerrilla attacks recorded over the period.[7] Contrary to expectations, the authors find that land reform is usually followed by

[7]Covariates included in the matching procedure include: prior plots reformed, paramilitary attacks, government attacks, poverty, population density, other tenancy, coca region, new colonized region, altitude,

an increase in the number of guerrilla attacks.

In this paper, we do not take issue with the selection and measurement of variables entering the matching procedure, nor with the particular matching algorithm used by the authors (one-to-one propensity score matching with replacement, discarding units outside the common support of the propensity score). We merely observe that the use of one-to-one matching in a context where the number of control units greatly exceeds the number of treatment units, leads analysts to discard a large proportion of control units on the basis of the estimated propensity score.[8] While most of the discarded control units are not comparable to units in the treatment group, some of them are borderline cases that are not significantly further away from the treatment than some of the matched control units. We argue that in situations like this, ignoring the uncertainty the propensity score might lead to arbitrariness in the selection of control units to be kept in the matched sample. In the rest of this section we show how results change when the Bayesian approach is used in order to account for estimation uncertainty in the propensity score.[9]

[Table 5 about here.]

Table 5 provides a comparison of the results found using a PSM procedure identical to the one used by Albertus and Kaplan (2013), and a BPSM procedure that is similar in every way. The only difference between the latter and former procedures is that the matching algorithm is not implemented only once on the basis of point estimates of propensity scores, but repeatedly for numerous draws from the posterior distribution of propensity score, with a measure of the ATE being computed at each iteration. Panel (a) of the table gives results found using data at the municipality level, and panel (b) gives results found using data

and percent minorities.

[8] In the case of the municipal-level data, the ratio of control to treatment units in the unmatched data set is larger than 13, and the use of one-to-one matching leads authors to drop 68% of control units.

[9] In order to ensure that we used exactly the same data as (Albertus and Kaplan, 2013), we downloaded their replication package from `http://esoc.princeton.edu/files/land-reform-counterinsurgency-policy-case-colombia`. We first replicated their analysis using the Stata code included in the replication package, as well as in R. Subsequently, we re-analyzed the data using BPSM.

at the municipality-year level. Consistently with what we found during simulation studies, the proportion of units kept in the matched sample after implementing PSM (11.38% for the municipality data and 10.38% for the municipality-year data) is considerably smaller than the proportion of units matched at least one during the implementation of the BPSM procedure (85.91% for the municipality data and 49.76% for the municipality-year data).

However, the fact that some control units (which would otherwise be dropped under PSM) are used *at least once* during the BPSM procedure, does not necessarily mean that they are used frequently. This feature is illustrated in Figure 5, which gives information about the proportion of the time that treatment units (dark grey bars) and control units (light gray bars) are kept in the matched sample during the BPSM procedure. While treatment units (a total of 65 in municipality dataset, and 754 in in the municipality-year dataset) are used almost all of time, most control units (a total of 893 in municipality dataset, and 10,887 in in the municipality-year dataset) are used less than 20% of the time, with a majority being used rarely or never. Figure 5 suggests that PSM and BPSM are actually not so different in terms of the amount of information that they incorporate; PSM also kept most treatment units and dropped most control units. Is the fact that some additional control units are matched a small proportion of the time under BPSM enough to affect estimates of treatment effects?

[Figure 5 about here.]

Table 5 also gives information about the ATE estimated under each procedure, without and with post-matching regression adjustment. In their paper, Albertus and Kaplan (2013, Table 3, Panel A) report results corresponding to PSM without regression adjustment, together with bootstrap estimates of standard errors. For the municipal-level data, they find that land reform leads to 34.57 more attacks, with a bootstrap standard error of 11.68. When we replicate their analysis, we find similar results, although bootstrap standard errors are slightly higher (12.43), which could happen for random reasons. When we re-analyze the data using BPSM, however, we find that land reform leads to 24.22 more attacks, a considerably smaller amount. Furthermore, when we perform post-matching regression adjustment

at each iteration of the BPSM procedure, we find that land reform leads to a positive but non-significant increase in the number of attacks (the estimated ATE drops to 13.57, with standard error equal to 9.45, and 95% credible interval ranging between -6.71 and 30.64). These findings can be visualized in Figure 6. This figure gives histograms and density curves for the ATE estimated using BPSM, without (top-most panel) and with (bottom-most panel) regression adjustment. In each plot, the smooth black line indicates the average estimated ATE under BPSM, and the dashed grey line indicates point estimates of treatment effects under PSM. Even though results obtained using BPSM are substantively similar to those found by Albertus and Kaplan (2013) using PSM, the magnitude of the effects (i.e. increase in number of attacks after land reform) is considerably smaller under BPSM, to the point that effects are no longer statistically significant after post-matching regression adjustment.

[Figures 6 and 7 about here.]

We found comparable results for the municipality-year data. When matching is done based on point estimates of propensity scores, estimated ATEs are considerably larger than when matching is done repeatedly on the basis of draws from the posterior distribution of propensity scores. In their paper, Albertus and Kaplan (2013, Table 3, Panel A) report that land reform leads to 1.40 more attacks, with a bootstrap standard error of 0.61. When we replicate their analysis using PSM without post-matching regression adjustment, we find similar results. However, when we re-analyze the data using BPSM without post-matching regression adjustment, we find that the average ATE falls to 1.07 (with standard error 0.50 and 95% credible interval ranging between -0.07 and 1.69). This result can be visualized in the top-most plot Figure 7, which shows that the distribution of ATEs is markedly asymmetrical in the case of BPSM, with a thick left tail, and an average value (smooth black line) located to the left of the point estimated of the ATE found using PSM (dashed grey line). The distribution of estimated ATEs found using BPSM becomes more symmetric when post-matching regression adjustment is performed at each step of the BPSM procedure (bottom-most plot). In contrast to the municipal level data, we find that the adjusted

ATE estimated using BPSM with post-matching regression adjustment is still positive and statistically significant (1.03, with standard error 0.28 and 95% credible interval ranging between 0.47 and 1.56), although the magnitude of the effect is still considerably smaller than the one reported by Albertus and Kaplan (2013).

# 5 Conclusion

Point estimates of propensity scores are commonly used for preprocessing the data by matching treatment and control units with similar predicted probabilities of assignment to treatment, in order to construct matched samples where covariates are well-balanced across treatment and control groups. The process typically involves dropping units outside the support of the propensity score (in the treatment, control, or both groups); control units with very low ex-ante probability of being assigned to treatment; treatment units for which no matches are available in the control group; or excess control units (as in cases where the number of control units greatly exceeds the number of treatment units, and each treatment unit is matched to a single or few control units). While propensity score matching allows analyst to overcome the dimensionality problem that would ensue if they tried to match units on the basis of multiple individual covariates; existing procedures also have a number of disadvantages. In particular, analysts tend to disregard the fact that estimated propensity scores have associated measures of uncertainty. We argued that standard approaches that take the propensity score as given can be problematic since they may cause the analyst to make arbitrary decisions regarding whether to keep or drop units from the matched sample.

We proposed a simple modification of standard propensity score matching procedures that can be easily implemented using Bayesian estimation. The Bayesian approach has several advantages, including that it can be use to calculate point estimates of treatment effects, as well as associated measures of uncertainty, without the need of resorting to bootstrapping or post-matching simulation procedures. Since matching under BPSM is done

probabilistically (on the basis of information about estimation uncertainty in the propensity score) instead of deterministically, it leads to less arbitrary decisions about whether to keep or drop observations from the matched sample—an attribute which we hypothesized would lead to estimates of treatment effects exhibiting higher precision and lower bias. The results of our simulation study were in line with our expectations; they indicated that incorporating information about uncertainty in the propensity score leads to lower bias and dispersion of estimates of treatment effects, compared to standard propensity score matching. Furthermore, we replicated a published study that employed PSM and showed the utility of the BPSM approach in that application. All of the evidence presented in this paper documents the utility of the BPSM approach.

There are many future directions in which the BPSM approach can be extended. One important extension is the incorporation of prior information into propensity score and outcome models. For example, we plan to explore the use of beliefs held by biased observers regarding the direction and magnitude of treatment effects, as well as other sources for prior information. Furthermore, the Bayesian approach used here can be extended to other causal modeling applications.

# References


Abadie, Alberto and Guido W. Imbens. 2008. On the Failure of the Bootstrap for Matching Estimators. *Econometrics* 76(6): 1537-57.

Albertus, Michael and Oliver Kaplan. 2013. Land Reform as a Counterinsurgency Policy : Evidence from Colombia. *Journal of Conflict Resolution* 57(2): 198-231.

Albertus, Michael and Oliver Kaplan. 2013. Replication Data for "Land Reform as a Counterinsurgency Policy: The Case of Colombia." Empirical Studies of Conflict Project (ESOC).

An, Weihua. 2010. Bayesian Propensity Score Estimators: Incorporating Uncertainties in Propensity Scores into Causal Inference. *Sociological Methodology* 40(1): 151-189.

Cochran, William G. and Donald B. Rubin. 1973. Controlling for Bias in Observational Studies: A Review. *Sankhya: The Indian Jounal of Statistics: Series A* 35(4): 417-66.

Dehejia, Rajeev H. and Sadek Wahba. 2002. Propensity Score-Matching Methods for Nonexperimental Causal Studies. *Review of Economics and Statistics* 84(1): 151161.

Diamond, Alexis and Jasjeet S. Sekhon. 2013. Genetic Matching for Estimating Causal Effects: A General Multivariate Matching Method for Achieving Balance in Observational Studies. *Review of Economics and Statistics* 95(3): 932-45.

Gelman, Andrew and Jennifer Hill. 2007. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. New York: Cambridge University Press.

Heckman, James J., Hidehiko Ichimura, and Petra E. Todd. 1997. Matching As An Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme. *Review of Economic Studies* 64 (4): 605-54.

Ho, Daniel E., Kosuke Imai, Gary King, and Elizabeth A. Stuart. 2007. Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis* 15 199-236.

Ho, Daniel E., Kosuke Imai, Gary King, and Elizabeth A. Stuart. 2011. MatchIt: Nonparametric Preprocessing for Parametric Causal Inference. *Journal of Statistical Software* 42(8).

Holland, Paul W. 1986. Statistics and Causal Inference. Journal of the American Statistical Association 81(396): 945-60.

Iacus, Stefano M., Gary King, and Giuseppe Porro. 2012. Causal Inference without Balance Checking: Coarsened Exact Matching. *Political Analysis* 20(1): 1-24.

Imai, Kosuke and David A van Dyk. 2004. Causal Inference With General Treatment Regimes: Generalizing the Propensity Score. *Journal of the American Statistical Association* 99(467): 854-66.

Imai, Kosuke and Marc Ratkovic. 2012. Covariate Balancing Propensity Score. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 76(1) 24363.

Jackman, Simon. 2000. Estimation and Inference Are Missing Data Problems: Unifying Social Science Statistics via Bayesian Simulation. *Political Analysis* 8(4):307-32.

Kaplan, David and Jianshen Chen. 2012. A Two-Step Bayesian Approach for Propensity Score Analysis: Simulations and a Case Study. *Psychometrica* 77(3): 581-609.

Martin, Andrew D., Kevin M. Quinn, and Jong Hee Park. 2011. MCMCpack: Markov Chain Monte Carlo in R. *Journal of Statistical Software* 42(9).

McCandless, Lawrence C., Paul Gustafson, and Peter C. Austin. 2009. Bayesian Propensity Score Analysis for Observational Data. *Statistics in Medicine* 28(1): 94-112.

Rosenbaum, Paul R. 1999. Propensity Score. In Armitage P, Colton T, eds., *Encyclopedia of Biostatistics.* New York, NY: John Wiley, 3551-5.

Rosenbaum, Paul R. and Donald B. Rubin. 1983. The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika* 70:41-55.

Rosenbaum, Paul R. and Donald B. Rubin. Constructing a Control Group Using Multivariate Matched Sampling Methods that Incorporate the Propensity Score. *American Statistician* 39:33-8.

Rubin, Donald B. 1973. The Use of Matched Sampling and Regression Adjustment to Remove Bias in Observational Studies. *Biometrics* 29(1): 185-203.

Rubin, Donald B. 1979. Using Multivariate Matched Sampling and Regression Adjustment to Control Bias in Observational Studies. *Journal of the American Statistical Association* 74(366): 318-28.

Rubin, Donald B. and Neal Thomas. 2000. Combining Propensity Score Matching with Additional Adjustments for Prognostic Covariates. *Journal of the American Statistical Association* 95(450): 573-85.

Stuart, Elizabeth A. 2010. Matching Methods for Causal Inference: A Review and a Look Forward. *Statistical Science* 25(1):121.

Tu, Wanzhu and Xiao-Hua Zhou. 2002. A Bootstrap Confidence Interval Procedure for the Treatment Effect Using Propensity Score Subclassification. *Health Services & Outcomes Research Methodology* 3: 135-147.

Zigler, Corwin Matthew and Francesca Dominici. 2014. Uncertainty in Propensity Score Estimation: Bayesian Methods for Variable Selection and Model-Averaged Causal Effects. *Journal of the American Statistical Association* 109(505): 95-107.

# A Tables and Figures

Table 1: Simulation Study, Correctly Specified Model

a. Small Effect ($\beta$=0.25)

| | Units matched at least once (%) | ATE | 95% Interval 2.5% | 97.5% | Bias | MSE | Coverage Probability |
|---|---|---|---|---|---|---|---|
| *N = 500* | | | | | | | |
| PSM | 67.9 | 5.62 | -5.65 | 16.61 | 0.48 | 37.00 | 93.9 |
| BPSM | 94.3 | 5.16 | -4.86 | 15.37 | 0.02 | 25.92 | 95.6 |
| *N = 1,500* | | | | | | | |
| PSM | 70.3 | 5.82 | -0.62 | 12.16 | 0.68 | 10.9 | 94.2 |
| BPSM | 96.6 | 5.22 | -0.49 | 11.09 | 0.09 | 7.70 | 96.3 |

b. Large Effect ($\beta$=1.5)

| | Units matched at least once (%) | ATE | 95% interval 2.5% | 97.5% | Bias | MSE | Coverage Probability |
|---|---|---|---|---|---|---|---|
| *N = 500* | | | | | | | |
| PSM | 67.9 | 25.7 | 16.26 | 35.28 | 0.77 | 25.12 | 94.3 |
| BPSM | 94.3 | 24.9 | 15.75 | 33.80 | -0.04 | 19.94 | 96.7 |
| *N = 1,500* | | | | | | | |
| PSM | 70.3 | 25.7 | 20.29 | 31.19 | 0.72 | 7.48 | 95.6 |
| BPSM | 96.6 | 24.7 | 19.48 | 29.90 | -0.24 | 5.90 | 97.3 |

Number of simulations: 1,000 for each combination of treatment effect size and sample size.

Table 2: Simulation Study, Missing Interaction Term

a. Small Effect ($\beta$=0.25)

| | Units matched at least once (%) | ATE | 95% Interval 2.5% | 97.5% | Bias | MSE | Coverage Probability |
|---|---|---|---|---|---|---|---|
| *N = 500* | | | | | | | |
| PSM | 62.31 | 7.40 | -4.23 | 18.85 | 2.21 | 43.11 | 92.10 |
| BPSM | 92.39 | 6.95 | -3.92 | 18.05 | 1.76 | 32.17 | 94.70 |
| *N = 1,500* | | | | | | | |
| PSM | 64.19 | 7.27 | 0.64 | 13.88 | 2.08 | 15.43 | 90.9 |
| BPSM | 94.61 | 6.82 | 0.57 | 13.18 | 1.63 | 11.26 | 93.5 |

b. Large Effect ($\beta$=1.5)

| | Units matched at least once (%) | ATE | 95% interval 2.5% | 97.5% | Bias | MSE | Coverage Probability |
|---|---|---|---|---|---|---|---|
| *N = 500* | | | | | | | |
| PSM | 62.31 | 28.14 | 17.97 | 38.08 | 3.38 | 38.52 | 89.9 |
| BPSM | 92.39 | 27.36 | 17.42 | 37.06 | 2.61 | 28.85 | 93.8 |
| *N = 1,500* | | | | | | | |
| PSM | 64.19 | 27.96 | 22.15 | 33.74 | 3.19 | 17.97 | 82.9 |
| BPSM | 94.61 | 27.07 | 21.37 | 32.72 | 2.31 | 11.92 | 90.4 |

Number of simulations: 1,000 for each combination of treatment effect size and sample size.

Table 3: Simulation Study, Missing Confounder (Low Correlation)

a. Small Effect ($\beta$=0.25)

| | Units matched at least once (%) | ATE | 95% Interval 2.5% | 97.5% | Bias | MSE | Coverage Probability |
|---|---|---|---|---|---|---|---|
| *N = 500* | | | | | | | |
| PSM | 61.19 | 21.60 | 9.25 | 33.47 | 17.40 | 340.31 | 20.10 |
| BPSM | 90.08 | 18.51 | 7.97 | 28.93 | 14.31 | 228.85 | 22.70 |
| *N = 1,500* | | | | | | | |
| PSM | 63.61 | 21.68 | 14.74 | 28.54 | 17.48 | 319.10 | 0.10 |
| BPSM | 92.49 | 18.48 | 12.49 | 24.46 | 14.27 | 212.67 | 0.40 |

b. Large Effect ($\beta$=1.5)

| | Units matched at least once (%) | ATE | 95% interval 2.5% | 97.5% | Bias | MSE | Coverage Probability |
|---|---|---|---|---|---|---|---|
| *N = 500* | | | | | | | |
| PSM | 61.19 | 37.55 | 26.77 | 48.09 | 15.43 | 266.41 | 18.40 |
| BPSM | 90.08 | 34.32 | 24.65 | 43.52 | 12.20 | 168.62 | 28.50 |
| *N = 1,500* | | | | | | | |
| PSM | 63.61 | 37.86 | 31.78 | 43.93 | 15.75 | 258.85 | 0.30 |
| BPSM | 92.49 | 34.42 | 28.96 | 39.72 | 12.31 | 159.06 | 0.80 |

Number of simulations: 1,000 for each combination of treatment effect size and sample size.

Table 4: Simulation Study, Missing Confounder (High Correlation)

a. Small Effect ($\beta$=0.25)

| | Units matched at least once (%) | 95% Interval ATE | 2.5% | 97.5% | Bias | MSE | Coverage Probability |
|---|---|---|---|---|---|---|---|
| *N = 500* | | | | | | | |
| PSM | 55.24 | 17.39 | 3.25 | 30.97 | 13.42 | 231.42 | 54.10 |
| BPSM | 86.14 | 12.83 | 2.34 | 23.52 | 8.85 | 103.32 | 61.70 |
| *N = 1,500* | | | | | | | |
| PSM | 57.39 | 17.63 | 9.56 | 25.46 | 13.65 | 204.26 | 10.60 |
| BPSM | 89.49 | 12.82 | 6.77 | 18.98 | 8.84 | 87.16 | 19.30 |

b. Large Effect ($\beta$=1.5)

| Units matched | | 95% interval | | | | | Coverage |
|---|---|---|---|---|---|---|---|
| | at least once (%) | ATE | 2.5% | 97.5% | Bias | MSE | Probability |
| *N = 500* | | | | | | | |
| PSM | 55.24 | 34.23 | 22.24 | 46.09 | 12.90 | 207.09 | 44.70 |
| BPSM | 86.14 | 28.93 | 18.83 | 38.62 | 7.60 | 80.72 | 69.10 |
| *N = 1,500* | | | | | | | |
| PSM | 57.39 | 34.53 | 27.74 | 41.36 | 13.17 | 185.60 | 3.10 |
| BPSM | 89.49 | 28.82 | 23.03 | 34.46 | 7.45 | 62.75 | 28.20 |

Number of simulations: 1,000 for each combination of treatment effect size and sample size.

Table 5: Replication of Albertus and Kaplan 2012, Table 3, Panel A

a. Unit of analysis: Municipality

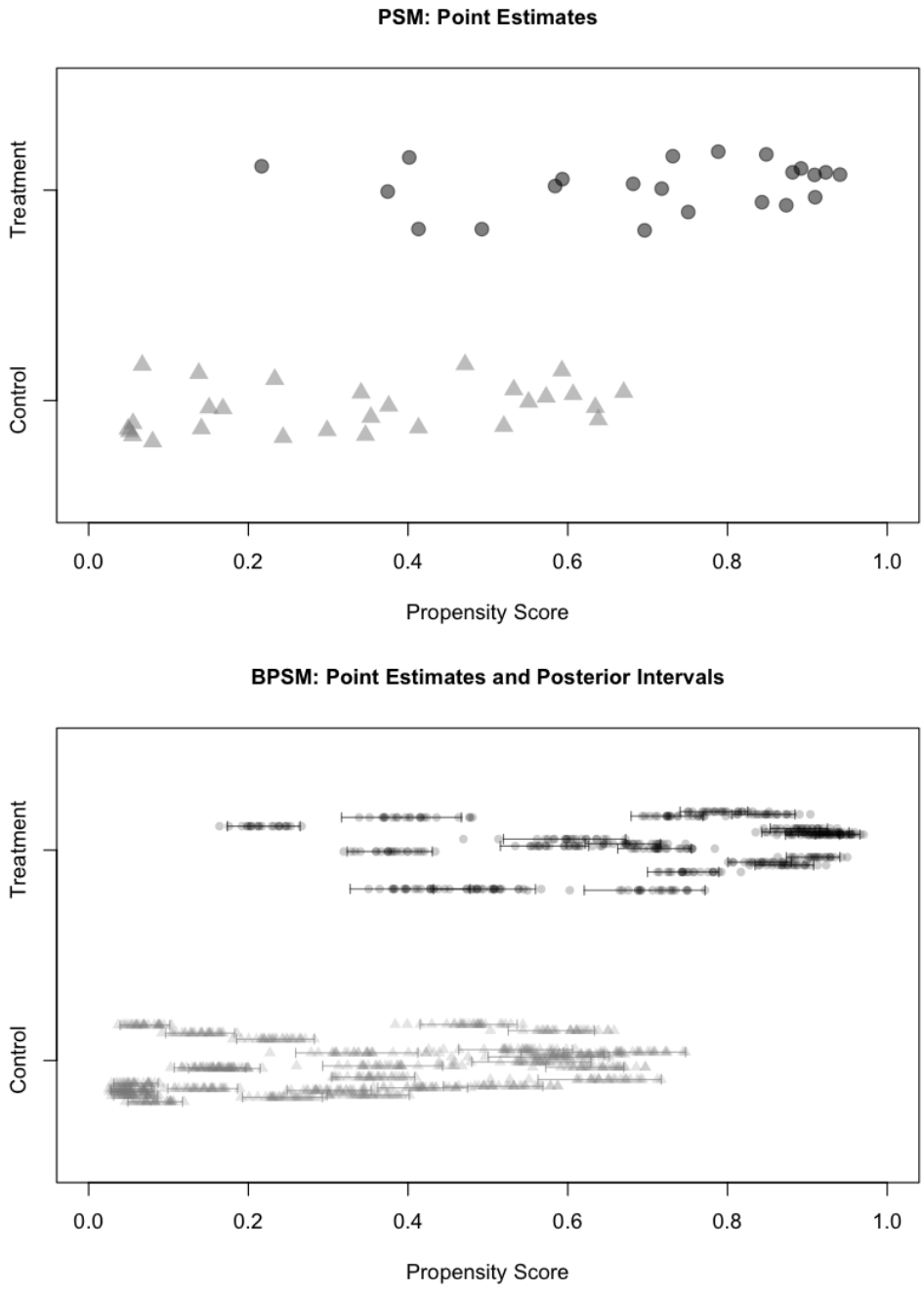
| | Units matched at least once (%) | ATE | SE |
|---|---|---|---|
| PSM | 11.38% | | |
| w/o reg. adj. | | 34.57 | 12.43 |
| with reg. adj. | | 20.48 | 6.55 |
| BPSM | 85.91% | | |
| w/o reg. adj. | | 24.22 | 8.04 |
| with reg. adj. | | 13.57 | 9.45 |

b. Unit of analysis: Municipality-year

| | Units matched at least once (%) | ATE | SE |
|---|---|---|---|
| PSM | 10.38% | | |
| w/o reg. adj. | | 1.40 | 0.54 |
| with reg. adj. | | 1.18 | 0.26 |
| BPSM | 49.76% | | |
| w/o reg. adj. | | 1.07 | 0.50 |
| with reg. adj. | | 1.03 | 0.28 |

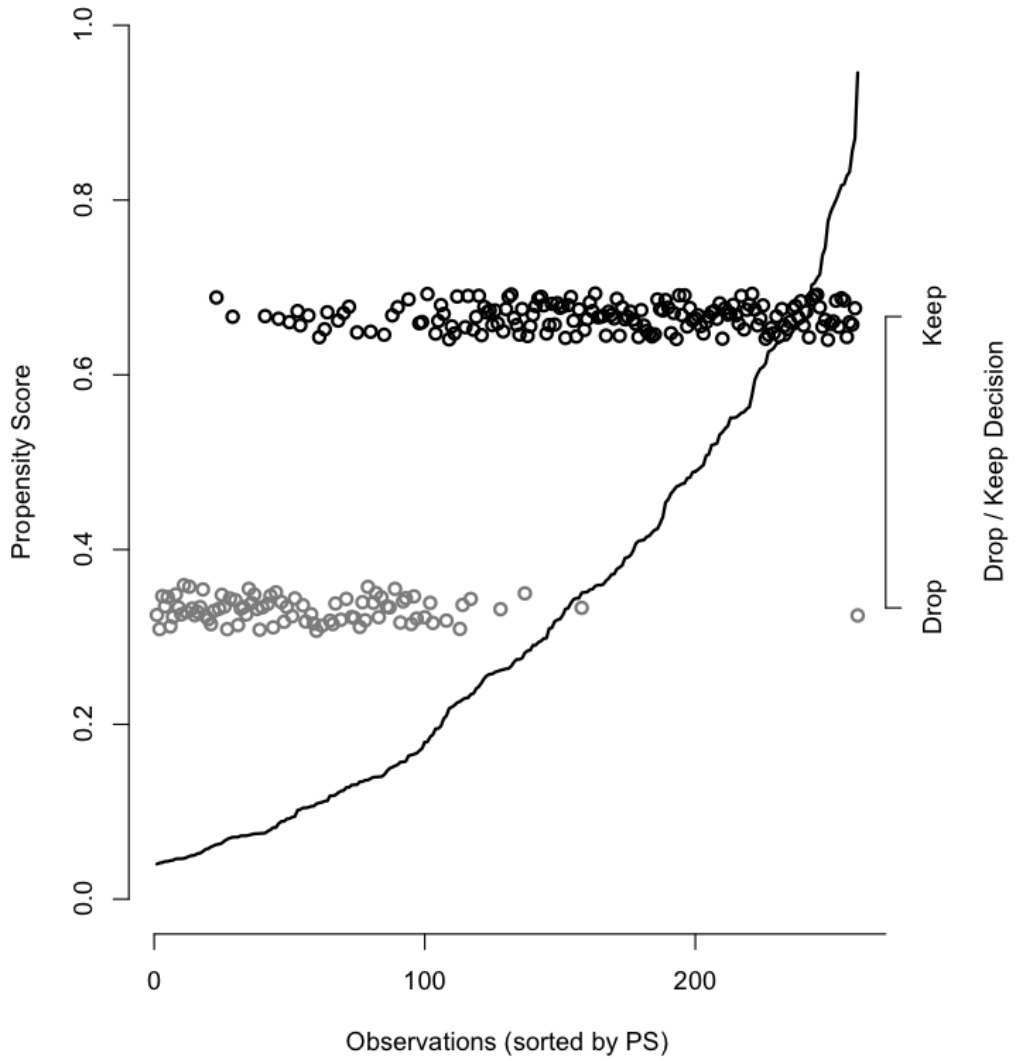
Note: PSM indicates standard propensity score matching and BPSM indicates Bayesian propensity score matching. When PSM was conducted without regression adjustment, standard errors were computed using bootstrapping.

Figure 1: Estimated Propensity Scores



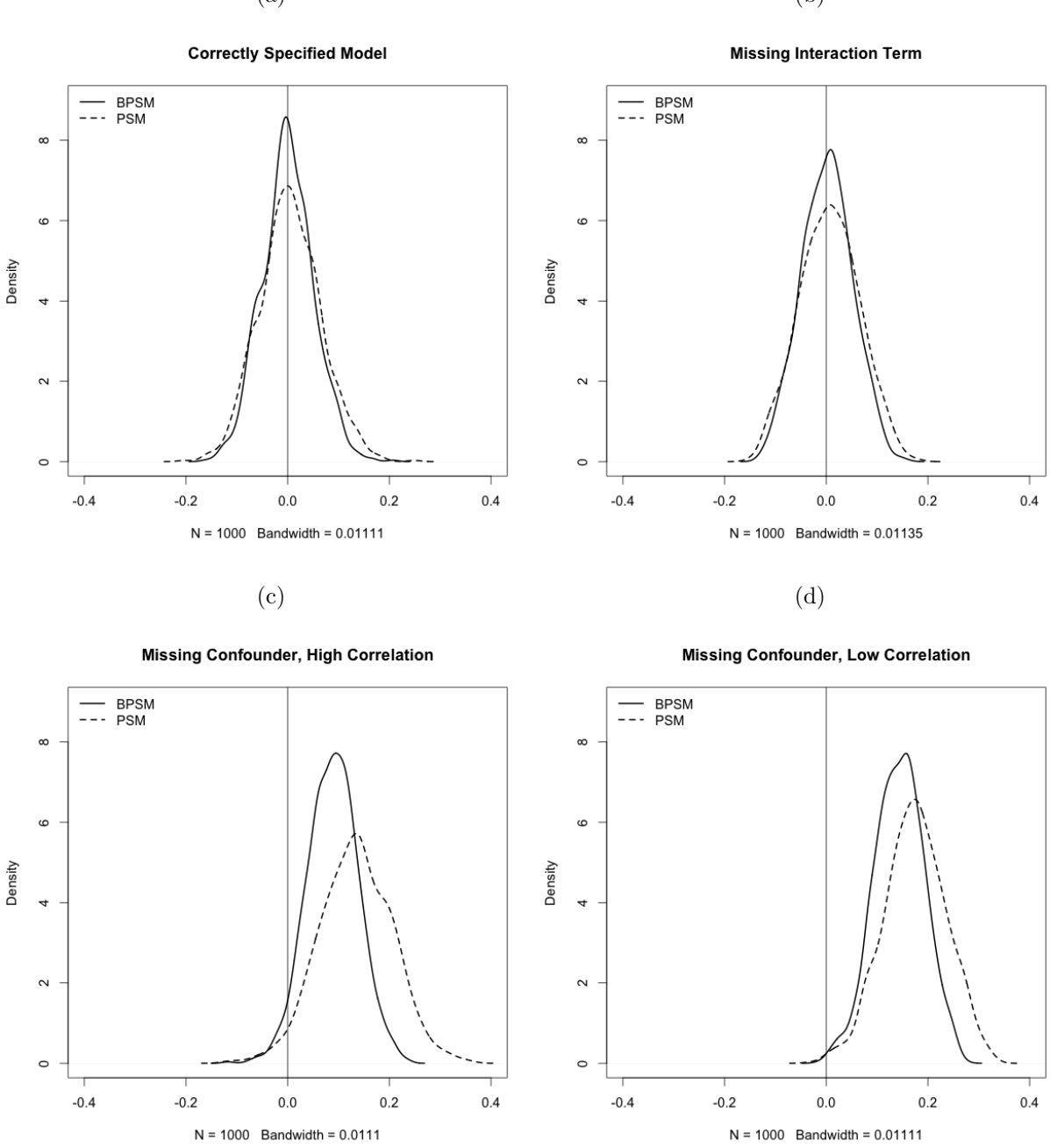

Note: The figure illustrates information used (top-most plot) or ignored (bottom-most plot) by standard propensity score matching procedures, for 50 observations drawn at random from a synthetic dataset. While PSM only considers point estimates of propensity scores, the Bayesian propensity score matching procedure proposed in this paper takes into account information about other aspects of the posterior distribution of estimated propensity scores, including estimation uncertainty.

Figure 2: Example of Drop/Keep Decision for Control Observations



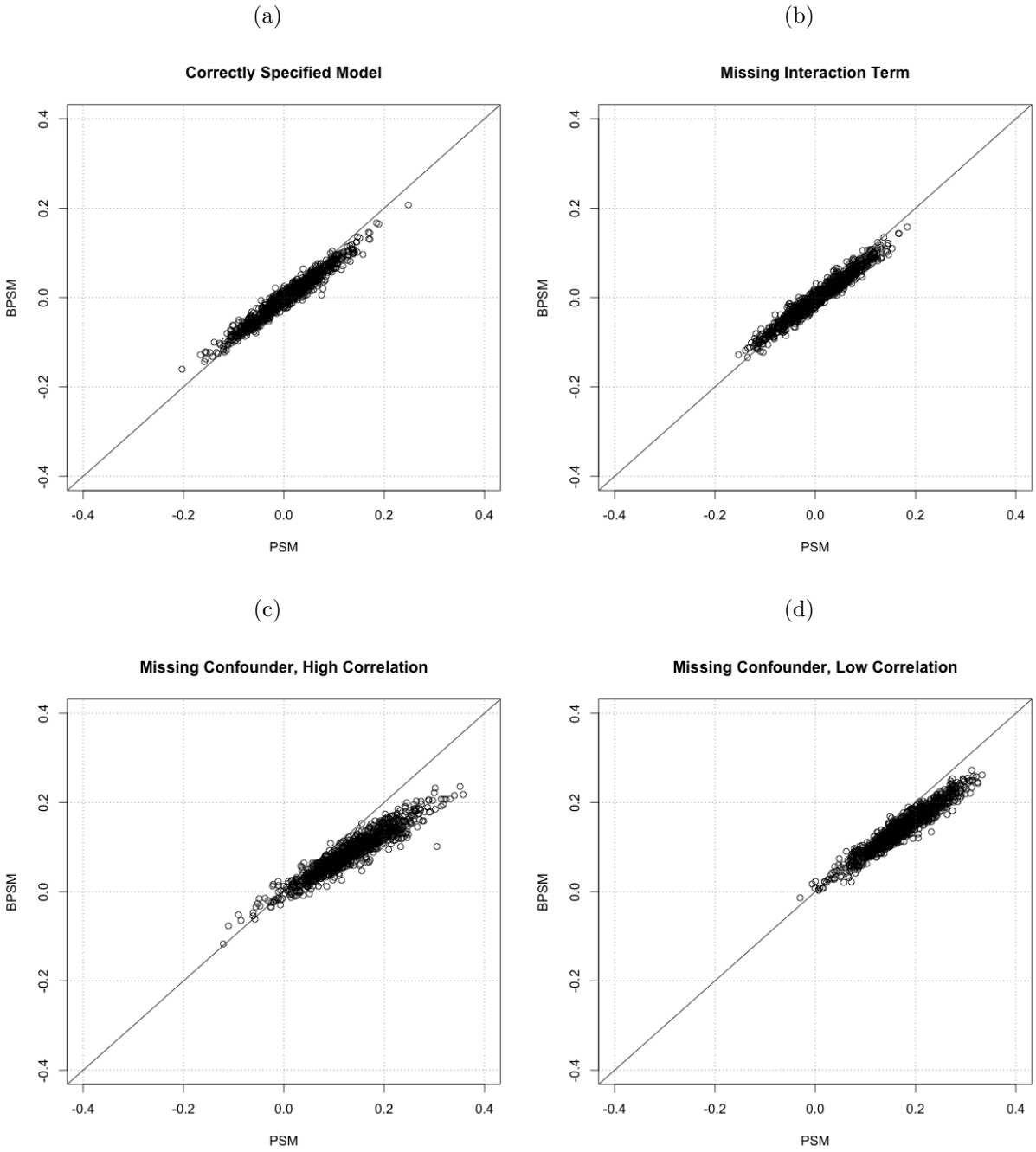

Note: The figure depicts the relationship between point estimates of propensity scores and drop/keep decisions with a one-shot nearest-neighbor propensity score matching procedure, for a synthetic dataset containing 500 observations. Drop decisions are relatively rare for observations with point estimates of propensity score above a certain cutoff (around .2).

Figure 3: Simulation Study: Distribution of Bias




Note: Results correspond to a simulation study conducted under the assumption of small sample size ($N = 500$) and small treatment effect ($\beta = .25$). Number of simulations: 1,000.

Figure 4: Simulation Study: Comparison of Bias




Note: Each point corresponds to a different simulated dataset. Results correspond to a simulation study conducted under the assumption of small sample size ($N = 500$) and small treatment effect ($\beta = .25$). Number of simulations: 1,000.

Figure 5: Land Reform and Insurgency in Colombia, Percentage of Units Used During BPSM Procedure

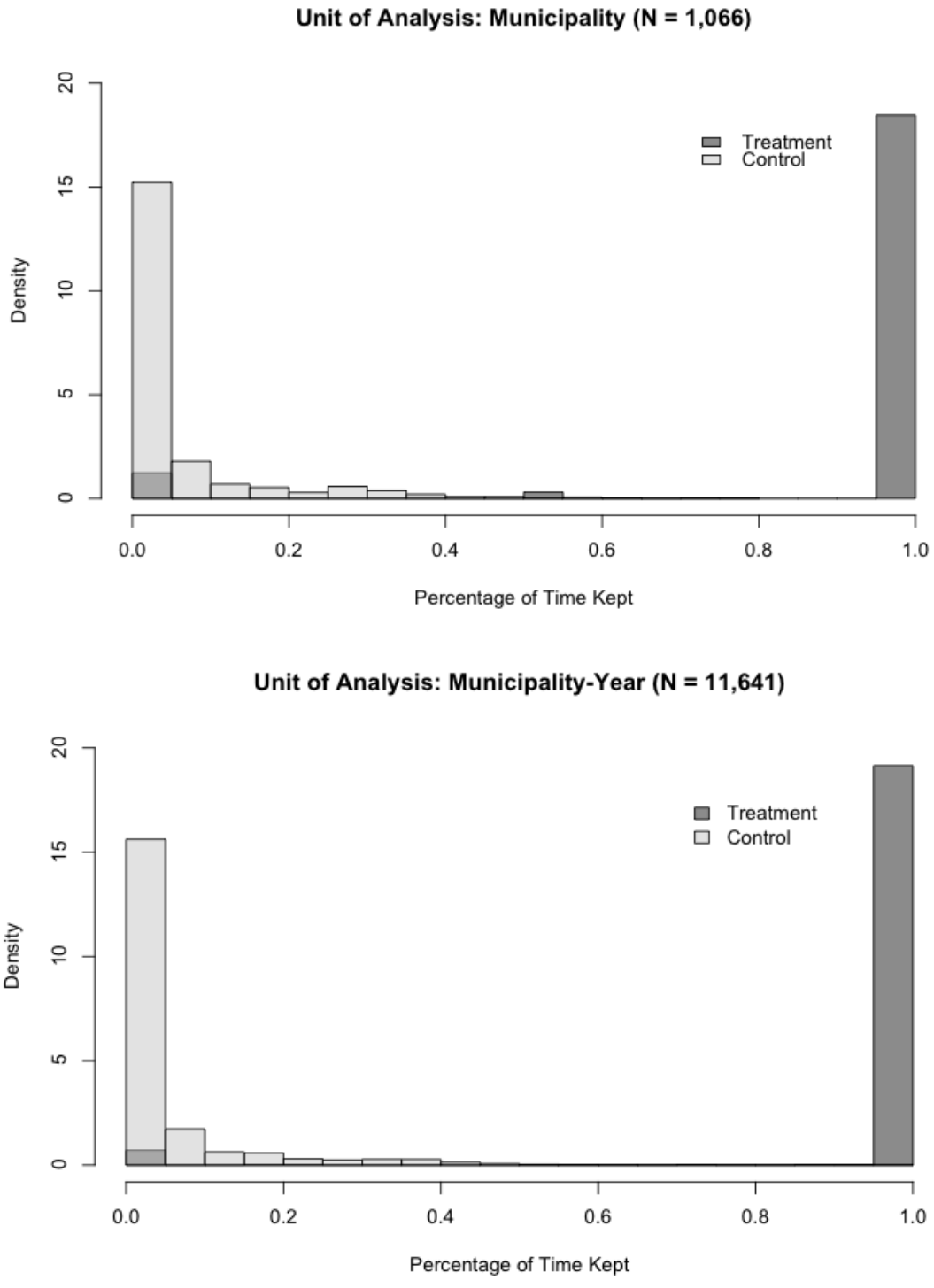


Figure 6: Land Reform and Insurgency in Colombia, Average Treatment Effects (Unit of Analysis: Municipality)

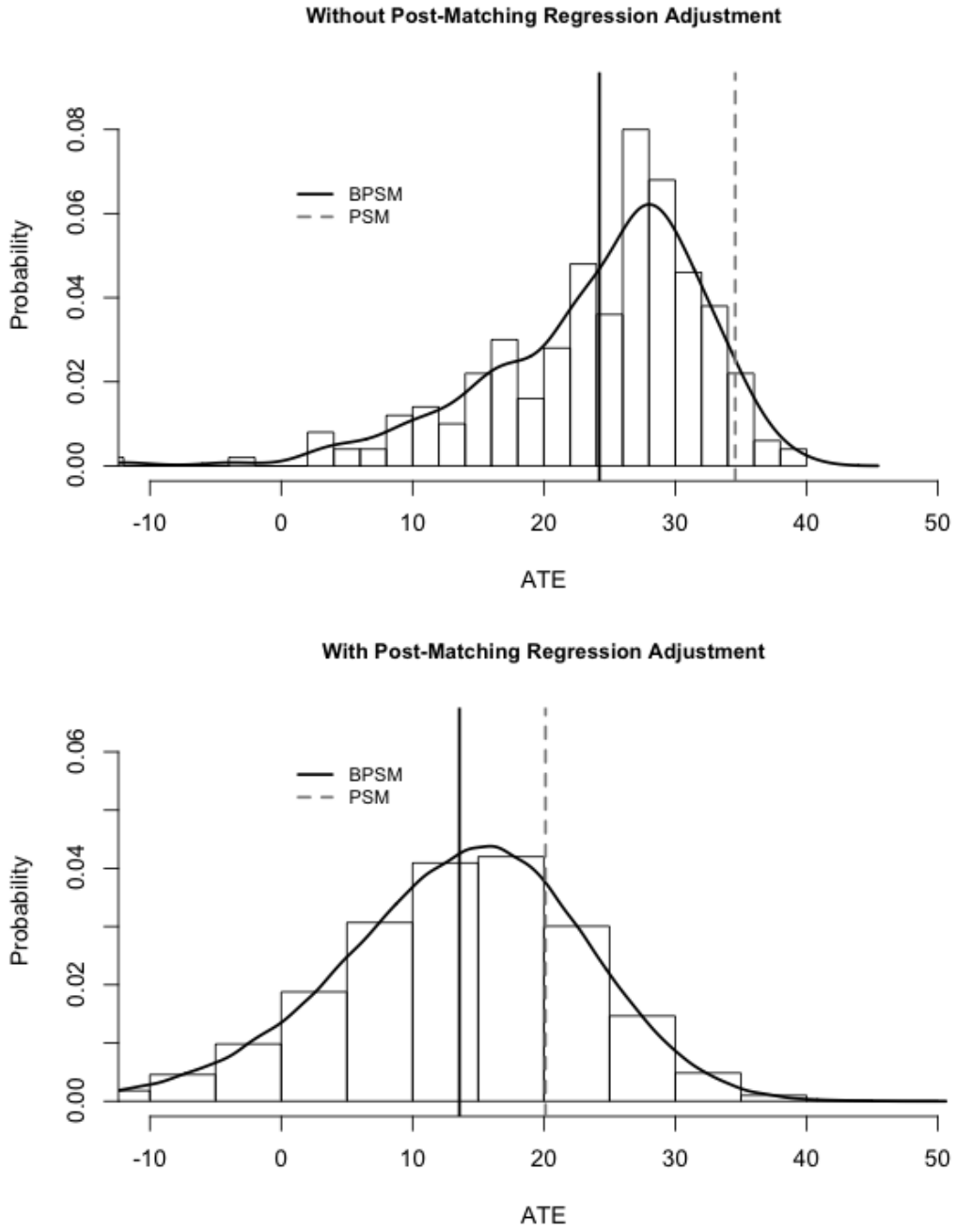


Figure 7: Land Reform and Insurgency in Colombia, Average Treatment Effects (Unit of Analysis: Municipality-Year)