American Journal of Applied Sciences

Original Research Paper

Evaluation of Subset Matching Methods: Evidence from a Monte Carlo Simulation Study

Lateef Amusa, Temesgen Zewotir and Delia North

Department of Statistics, School of Mathematics, Statistics and Computer Science, University of Kwazulu-Natal, Durban, South Africa

Article history
Received: 20-01-2019
Revised: 13-02-2019
Accepted: 11-04-2019

Corresponding Author:
Lateef Amusa
Department of Statistics, School of Mathematics, Statistics and Computer Science, University of Kwazulu-Natal, Durban, South Africa
Email: amusasuxes@gmail.com

Abstract: In the absence or infeasibility of experiments, matching methods have increasingly been used in making causal claims using observational data. This paper conducts a Monte Carlo simulation study, based on a household panel survey, to compare the performance of some widely used subset matching methods. The methods include the propensity score caliper matching, Mahalanobis distance matching, and coarsened exact matching. Comparisons were made in terms of the ability to reduce covariate imbalances, as well as effective recovery of the real treatment effect. Numerical results from our simulations provided evidence of coarsened exact matching outperforming the other methods. Our results also showed that, except for the Mahalanobis distance matching method, the efficiency of treatment effect estimates decreases with an increasing proportion of treated units.

Keywords: Matching, Balance, Monte Carlo Simulation, Observational Studies, Propensity Score

Introduction

Randomized experiments are the gold standard for estimating causal effects: They guarantee that the treated and control groups are only randomly different from one another with respect to the background covariates. Many matching methods have been proposed for replicating this scenario as much as possible for observed covariates with observational data.

Several methods serve as alternatives to matching, including adjusting for background variables in a regression model, instrumental variables, structural equation models and regression discontinuity designs. However, matching methods have been paid more attention and widely used because of its intuitiveness and more importantly, straightforward diagnostics, by which the performance is evaluated.

Matching is a nonparametric method for taking control of the confounding influence of background covariates or pretreatment control variables in observational or non-experimental data. The main aim of matching is to selectively prune observations from the data so that a better balance between the treated and control groups is achieved with the remaining data, which in other words means that the empirical distributions of the covariates in the two groups are then more similar. Statistical modelling assumptions handle any residual imbalance. The primary merit of matching is that it significantly reduces model dependence (King *et al*., 2011).

There are several matching methods existing in the literature, and they employ different distance measures, algorithms and rules for selecting control group members. Each technique could potentially choose different control group members from the overall control pool to create the matched group. The matched control group composition could, therefore, vary considerably depending on the particular matching algorithm used (Jacovidis, 2017). Matching techniques have been applied either using covariate (Miksch *et al*., 2010) or propensity score matching (Stock *et al*., 2010; Windt and Glaeske, 2010; Drabik *et al*., 2012) with some authors providing evidence for the superiority of propensity score matching (Drabik *et al*., 2012). The literature has shown that propensity score matching is not necessarily the gold standard (Fullerton *et al*., 2016). Depending on the scenario, other matching techniques can induce a better balance on the covariates and furthermore, the performance of propensity score matching highly depends on the correct specification of the propensity score model, choice of covariates and the matching algorithm used (Dehejia and Wahba, 2002; King *et al*.,

2011; Rosenbaum and Rubin, 1984). While many simulation studies have compared the performance of different matching methods, it cannot be taken for granted that their findings are transferrable to another data situation (Franklin *et al*., 2014). Consequently, there is a need for extensive research on identifying which matching methods perform best in several scenarios. Even though there have been a few notable studies that have examined the performance of matching techniques in terms of how well they balance the groups on the covariates, only a few of them have extended the evaluation of the matching techniques to the outcome analyses (Jacovidis, 2017; Austin, 2014; Stone and Tang, 2013).

Accordingly, this study aims to compare the performance of three (3) matching methods that are widely used in applied studies, under systematically manipulated conditions. The performance of each matching method was evaluated in terms of the ability to balance covariates between treated and control groups and efficient recovery of the real treatment effect. The abundance of subset matching methods and their variations is too large to be all compared in one study; without loss of generality, we studied the Propensity Score Caliper Matching (PSCM), Mahalanobis Distance Matching (MDM) and Coarsened Exact Matching (CEM).

Materials and Methods

Matching Methods

In this section, we briefly describe the matching methods we focused on in this study, each of which is commonly used in the applied literature. For the Mahalanobis distance and propensity score matching methods, we assumed a 1-1 matching without replacement, with the greedy matching algorithm being used to define the matched pairs. In matching without replacement, an already matched control unit is no longer available as a potential match for other treated units. In the case of greedy matching, a treated unit is chosen randomly, and the nearest control unit is then selected for matching to this treated unit (Austin, 2009).

Consider the unit i (i = 1,..., n), where T_i denotes a treatment variable coded 1 and 0 for the treated and control groups respectively. Let $\{Y_i(t)$: $t \in (0,1)\}$ be the potential outcome variable value, also known as a counterfactual outcome (Rubin, 1974). This implies that $Y_i = T_iY_i(1) + (1\text{-}T_i)Y_i(0)$ is observed. Let X_i be a vector of pretreatment covariates; while, let m_T and m_C be the number of matched treated and control units respectively, for the methods. In estimating the average treatment effects, the Sample Average Treatment effect on the Treated units (SATT) was utilized. SATT = $\frac{1}{n_t}\sum_{i\in T} TE_i$, where $TE_i = Y_i(1)\text{-}Y_i(0)$.

Propensity Score Caliper Matching

Propensity score caliper matching is by far the most widely used matching method in the applied literature (Amusa, 2018). As the name of this method implies, it matches treated and control groups, based on the corresponding propensity scores, which weight covariates by how well they predict group membership. The propensity score was defined by Rosenbaum (1983) as the probability of treatment assignment, given the observed baseline covariates, stated mathematically as:

$$e_i = P\left(T_i = 1 \mid X_i\right), \tag{1}$$

where, it is assumed that, given the X's, the T_i's are independent:

$$P\left(T_1 = t_1 ... T_N = t_n\right) = \prod_{i=1}^{N} e_i^{t_i}\left\{1 - e_i\right\}^{1-t_i} \tag{2}$$

Let π_t and π_c and be the propensity scores for the treated and control group respectively, I_1 be the set of units in the treated group and I_0 be the set of units in the control group. A neighbourhood C (π_c) is defined to contain the c units control group ($c \in I_o$) as a match for the treated group t ($t \in I_1$), where the absolute difference of propensity scores is the smallest among all possible pairs of propensity scores between t and c, i.e.:

$$C\left(\pi_t\right) = \min \| \pi_t - \pi_c \|, c \in I_o \tag{3}$$

Once a particular value for c is found to match t, c is removed from I_o, without replacement. There is a further restriction imposed on the distance between π_t and π_c, and as such, c is selected as a match for t, only if the absolute difference of propensity scores between the two groups meets the following condition:

$$\| \pi_t - \pi_c \| < \xi, k \in I_o, \tag{4}$$

where, ξ is a caliper or a pre-specified tolerance for matching.

This procedure is known as propensity score caliper matching. A caliper size of a quarter of the estimated propensity scores' standard deviation has been suggested in the literature (Rosenbaum and Rubin, 1985).

Mahalanobis Distance Matching

Similar to PSCM, the Mahalanobis distance matching method is built on specific notions of between observations of pretreatment covariates. MDM is unlike PSM which matches are made based on a scalar "Propensity Score", known as a balancing score; MDM matches on covariates by a specified distance,

which consequently ensures that covariates have equal weights. MDM measures the distance between two units, X_t and X_c as:

$$M\left(X_t, X_c\right) = \sqrt{\left(X_t - X_c\right)' S^{-1}\left(X_t - X_c\right)} \quad (5)$$

Where X_t, X_c denote the treated group and control group covariates respectively; S is the sample covariance matrix of X. Once the distance metric d is selected, a matching algorithm can then be applied. The procedure is known as the Mahalanobis distance matching.

Coarsened Exact Matching

The earlier mentioned methods are known as Equal Percent Bias Reduction (EBPR) methods, where improvements in the bound of balance for one covariate will affect each of the other covariates. To avoid this and other shortcomings of the EPBR methods, a new generalized class of matching methods known as Monotonic Imbalance Bounding (MIB), which has Coarsened Exact Matching (CEM) as a particular case, was introduced (Iacus *et al*., 2011; 2012). The strength of this method lies in the fact that, unlike other matching methods where balance is being continually checked until it is improved, CEM inverts the process and thus guarantees that the covariate imbalances between the matched treated and control groups will not be more than the user's pre-chosen level. MIB methods, therefore, improve bounds in the balance of one covariate in isolation as it will not affect the maximum imbalance of each of the other covariates (Iacus *et al*., 2012).

The essential thought of CEM is to coarsen each variable as reasonably as possible temporarily, through automated choices of coarsening using the Sturges rule (Scott, 2009), or any user-defined coarsening could be used. The automated approach was adopted for this study because of its ease and intuition. The exact matching algorithm is then applied to the coarsened data to determine the matches and to prune unmatched units. Finally, the coarsened data are left out, and the original values of the matched data are retained. In other words, after coarsening, the CEM algorithm creates a set of strata, say sЄS, each with same coarsened values of X. Units in strata containing at least one treated and one control unit are retained, while units in the remaining strata are then removed from this sample.

We denote by T_s and C_s, the treated and control units, respectively in stratum s; m_T^s as the number of matched units in T_s; m_C^s is the number of matched units in C_s. The number of matched units are, respectively, $m_T = U_{s\epsilon S} m_T^s$ and $m_C = U_{s\epsilon S} m_C^s$, for the treated and control units. Unmatched units receive zero weight, while to each matched unit i in stratum s, CEM assigns the weights:

$$W_i = \begin{cases} 1, & i \epsilon T^s \\ \dfrac{m_C m_T^s}{m_T m_C^s}, & i \epsilon C^s \end{cases} \quad (6)$$

Simulation Scheme

In this section, we describe the design of the Monte Carlo simulations which were used for data generation and to compare the performance of the considered matching methods. The performance was assessed using the following criteria: (a) Quality of matches: The ability to induce balance on measured background covariates; (b) Absolute bias of estimated treatment effects; (c) Root Mean Squared Error (RMSE) of estimated treatment effects.

The data-generating process and analyses were conducted with R packages, "MatchIt" (Ho *et al*., 2011) and "Matching" (Sekhon, 2011), in the environment of R version 3.4.1 (R Core Team, 2016).

We replicate previous simulation designs that had been used to evaluate matching methods (Iacus *et al*., 2012; Jacovidis, 2017; Austin, 2011), with slight modifications – the proportion of treated units where varied. Data were generated to mimic the Lalonde non-experimental data described in the next section.

Data Generation – Covariates Balance

Data were generated to mimic the structure and properties of the famous non-experimental Lalonde-PSID data. A small portion of the data is a U.S. job training program provided to participants for 12-18 months to help them find a job (Lalonde, 1986). The dataset comprises the original Lalonde's experimental treated units and non-experimental control units from the Panel Study of Income Dynamics (PSID), which includes 185 treated and 2490 control units. The choice of this dataset is driven by its importance in the evaluation literature since there has been considerable knowledge accumulated on evaluating non-experimental estimators, using this data.

The dataset comprises ten covariates: Four continuous covariates including age (age), years of education (education), real earnings in 1974 (re74) and 1975 (re75); as well as six binary covariates including marital status (married), black race (black), Hispanic race (Hispanic), lack of a high school diploma (nodegree) and indicator variables for unemployment in 1974 (u74) and 1975 (u75).

Using the idea of Austin (2011), we related the ten covariates with the probability of treatment selection via the following logistic regression model:

$$\begin{aligned} Logit\left(\pi_{i,t}\right) &= \alpha_{0,t} + \alpha_1 age + \alpha_2 education + \alpha_3 re74 \\ &+ \alpha_4 re75 + \alpha_5 married + \alpha_6 black + \alpha_7 hispanic \\ &\alpha_8 nodegree + \alpha_9 u74 + \alpha_{10} u75 \end{aligned} \quad (7)$$

The treatment group membership was regressed on the covariates for the study data and was used as coefficients (α_1, α_2,…, α_{10}) above. The intercept $\alpha_{0,t}$ was modified such that the proportion of treated units is varied at four different levels: 0.17, 0.20, 0.25, 0.33. For each unit *i*, in each of 1000 replications from this process, treatment status (denoted by *T*) was generated from a Bernoulli distribution with parameter $\pi_{i,t}$, i.e., $T_i \sim$ Ber ($\pi_{i,t}$), so that the number of pre-match treated and control units in the sample varies over replications.

Data Generation – Recovery of the True Treatment Effect

Next, outcome scores (*Y*) were generated as follows:

$$\begin{aligned} Y &= 1000T + \beta_1 age + \beta_2 education + \beta_3 re74 \\ &+ \beta_4 re75 + \beta_5 married + \beta_6 black + \beta_7 hispanic \\ &+ \beta_8 nodegree + \beta_9 u74 + \beta_{10} u75 + \varepsilon \end{aligned} \tag{8}$$

ATT was fixed at 1000 and $\varepsilon \sim N(0,10)$ as assumed by Iacus *et al*. (2012). Also, like Jacovidis (2017), the covariances between the covariates and outcome variable were obtained for the study data and were used to calculate the coefficients (β_1, β_1,…, β_{10}) above. A total of 1000 replications of each dataset were generated and matched with each method.

Performance Assessment

As stated in Section 1, the performance of the matching methods were evaluated relative to the unmatched data, under two criteria: (i) quality of matches and (ii) recovery of the true treatment effect. For each criterion, we varied the proportion of units who received the treatment (subsequently referred to as proportion of treated) at 17%, 20%, 25% and 33% levels, which corresponds to treatment-to-control ratios of 1:5, 1:4, 1:3 and 1:2, respectively.

Quality of Matches

In terms of the quality of matches, the methods were compared in terms of their ability to induce covariates balance between treated and control groups. This was achieved using the absolute standardized mean difference and percent bias reduction for all the covariates. The Absolute Standardized Mean Difference (ASMD), according to Rosenbaum and Rubin (1985), is defined as:

$$ASMD_k = \begin{cases} \left| \dfrac{\bar{x}_t - \bar{x}_c}{\sqrt{\dfrac{s^2t + s^2c}{2}}} \right| * 100\%, \text{ for continuous covariates} \\ \left| \dfrac{\hat{p}_t - \hat{p}_c}{\sqrt{\dfrac{\hat{p}_t(1-\hat{p}_c) + \hat{p}_c(1-\hat{p}_c)}{2}}} \right| * 100\%, \text{ for dichotonous covariates}, \end{cases} \tag{9}$$

where, $\bar{x}_t$ and $\bar{x}_c$ denote the sample mean of the treated and control units, respectively for the *kth* covariate; s^2t and s^2c denote the sample variance of the treated and control units, respectively for the *kth* covariate; $\hat{p}_t$ and $\hat{p}_c$ denote the mean of the *kth* dichotomous variable in the treated and control units, respectively.

It has been suggested that a standardized mean difference of at most 10% is quite sufficient at balancing a given covariate between the treatment groups (Austin, 2007; Normand *et al*., 2001).

Following the convention, the Percent Bias Reduction (*PBR*) for each covariate was also utilized. A threshold value of at least 80% is acceptable for judging the effectiveness of a matching method in reducing covariate imbalances (Cochran and Rubin, 1973; Pan and Bai, 2015). The percent bias reduction is defined as follows:

$$PBR_k = \frac{|B_{k,before\ matching}| - |B_{k,after\ matching}|}{|B_{k,before\ matching}|} * 100\%, \tag{10}$$

where, $B_{k,befor\ matching}$ and $B_{k,after\ matching}$ denote the mean difference in the *kth* covariate between the treated and control units, before and matching respectively.

For each of the ten covariates, the absolute standardized mean difference and percent bias reduction values were averaged across the 1000 simulated datasets.

Recovery of the True Treatment Effect

In each of the matched sets, we estimated the SATT estimators based on the difference in means between the observed outcome in the treated units and the control units. The performance of estimated treatment effects was assessed by its absolute bias, calculated as $|\bar{\hat{\gamma}} - \gamma|$ and root mean square error (RMSE), calculated as $\sqrt{\left(\bar{\hat{\gamma}} - \gamma\right)^2 + \text{var}(\hat{\gamma})}$, where $\bar{\hat{\gamma}}$ is the mean of the 1000 estimated treatment effects.

Results

In this section, we present results from the simulation study. We compared the matching methods in terms of covariates balance and the performance of treatment effect estimates.

Covariates Balance

The covariates balance assessment was varied at 17%, 20%, 25% and 33% proportions of treated units, as shown respectively in Table 1 to 4. As confirmed by the balance metrics, the raw data which we simulated from, is highly imbalanced - all the covariates have high standardized mean difference values - more substantial than the recommended 10% threshold value (Austin, 2007; Normand *et al*., 2001; Stuart, 2010).

Table 1: Balance assessment of covariates for 33% proportion of treated units (treatment-control ratio of 1:2)

	ASMD				PBR (%)		
Covariates	RAW	PSCM	MAH	CEM	PSCM	MAH	CEM
Age	0.56	0.06	0.45	0.01	89.84	18.91	97.81
Education	0.57	0.04	0.24	0.01	92.42	57.78	98.22
re74	2.01	0.07	1.40	0.06	96.57	30.63	97.17
re75	2.86	0.09	2.00	0.34	97.00	30.05	88.14
Black	0.86	0.05	0.55	0.00	94.61	36.49	100.00
Hispanic	0.14	0.04	0.03	0.00	73.62	76.90	100.00
Married	0.64	0.04	0.49	0.00	93.16	24.50	100.00
Nodegree	0.53	0.04	0.25	0.00	92.02	53.35	100.00
u74	0.66	0.03	0.61	0.00	95.07	8.22	100.00
u75	0.67	0.03	0.60	0.00	95.56	10.16	100.00

Note: The presented values are averages from each of the 1000 replications

Table 2: Balance assessment of covariates for 25% proportion of treated units (treatment-control ratio of 1:3)

	ASMD				PBR (%)		
Covariates	RAW	PSCM	MAH	CEM	PSCM	MAH	CEM
Age	0.59	0.06	0.45	0.01	90.52	23.75	97.90
Education	0.58	0.04	0.22	0.01	92.36	61.95	98.09
re74	2.16	0.06	1.26	0.05	97.23	41.35	97.57
re75	3.17	0.08	1.91	0.36	97.62	39.56	88.49
Black	0.92	0.05	0.45	0.00	94.90	50.89	100.00
Hispanic	0.14	0.04	0.00	0.00	72.57	99.79	100.00
Married	0.73	0.04	0.46	0.00	94.35	37.44	100.00
Nodegree	0.57	0.04	0.26	0.00	92.82	53.79	100.00
u74	0.73	0.03	0.59	0.00	95.89	19.51	100.00
u75	0.71	0.03	0.55	0.00	95.37	21.83	100.00

Note: The presented values are averages from each of the 1000 replications

Table 3: Balance assessment of covariates for 20% proportion of treated units (treatment-control ratio of 1:4)

	ASMD				PBR (%)		
Covariates	RAW	PSCM	MAH	CEM	PSCM	MAH	CEM
Age	0.63	0.06	0.48	0.01	91.01	24.85	97.96
Education	0.59	0.05	0.22	0.01	91.57	63.41	97.89
re74	2.29	0.06	1.16	0.05	97.19	49.53	97.71
re75	3.43	0.08	1.84	0.40	97.55	46.15	88.30
Black	0.98	0.05	0.48	0.00	94.84	50.69	100.00
Hispanic	0.14	0.04	0.00	0.00	71.86	100.00	100.00
Married	0.81	0.04	0.47	0.00	95.03	42.18	100.00
Nodegree	0.61	0.04	0.27	0.00	92.58	56.00	100.00
u74	0.79	0.04	0.55	0.00	95.49	30.67	100.00
u75	0.74	0.04	0.51	0.00	94.92	30.20	100.00

Note: The presented values are averages from each of the 1000 replications

Table 4: Balance assessment of covariates for 17% proportion of treated units (treatment-control ratio of 1:5)

	ASMD				PBR (%)		
Covariates	RAW	PSCM	MAH	CEM	PSCM	MAH	CEM
Age	0.69	0.06	0.55	0.01	91.51	20.13	98.13
Education	0.61	0.05	0.21	0.01	91.58	65.77	97.67
re74	2.46	0.07	1.01	0.06	97.08	58.82	97.51
re75	3.75	0.10	1.73	0.44	97.45	53.88	88.23
Black	1.06	0.05	0.53	0.00	95.04	50.36	100.00
Hispanic	0.14	0.04	0.00	0.00	66.70	100.00	100.00
Married	0.91	0.04	0.53	0.00	95.25	42.28	100.00
Nodegree	0.65	0.05	0.26	0.00	92.76	60.16	100.00
u74	0.85	0.04	0.46	0.00	95.44	45.43	100.00
u75	0.77	0.04	0.44	0.00	94.77	43.08	100.00

Note: The presented values are averages from each of the 1000 replications

In terms of the absolute standardized mean difference, except for the Mahalanobis distance matching method - which resulted in values extremely above the 10% threshold for almost all covariates, propensity score and coarsened exact matching methods had qualitatively comparable balance in the measured covariates. This pattern was consistent across the proportion of treated units. It is however worthy of note that coarsened exact matching had absolute standardized mean difference values of zero for the six continuous covariates across all treatment-control ratios considered.

In terms of the PBR, the performance of coarsened exact matching was excellent - all ten covariates had the Cochran and Rubin's acceptable threshold value of at least 80% PBR value. Mahalanobis distance matching had the worst performance. Propensity score caliper matching also had close to such an excellent performance, barring one covariate which consistently had PBR values below the 80% threshold. CEM further consistently had quantitatively higher PBR values. This pattern was consistent across the considered proportion of treated units. Overall, coarsened exact matching performed best in balancing covariates between the treated and control groups.

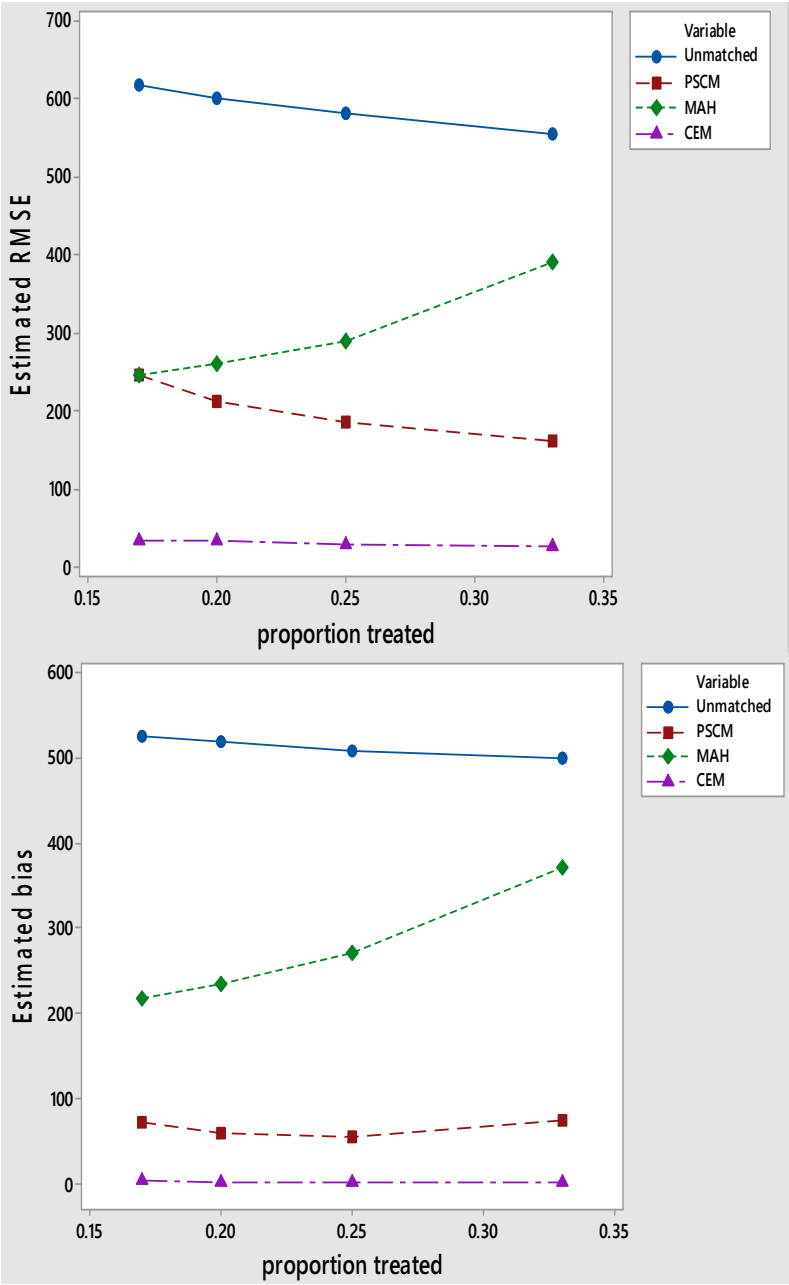

Fig. 1: Top panel: Root mean square error of estimated treatment effects; Bottom panel: Absolute bias of estimated treatment effects

Table 5: Absolute Bias and Root mean square of the matching methods relative to the unmatched data

	Proportion of treated: 17%		Proportion of treated: 20%		Proportion of treated: 25%		Proportion of treated: 33%	
Method	AB	RMSE	AB	RMSE	AB	RMSE	AB	RMSE
Unmatched	524.49	618.07	518.24	600.94	507.63	581.37	498.95	555.44
PSCM	72.60	246.62	59.46	211.75	55.33	186.13	75.98	161.54
MAH	218.59	246.14	235.35	260.59	271.01	290.81	371.49	392.15
CEM	4.33	35.19	2.30	33.16	2.22	29.28	1.80	27.91

AB means absolute bias
Note: values were averaged over 1000 Monte Carlo replications

Performance of Treatment Effect Estimates

The absolute bias (AB) and root mean square error (RMSE) of mean difference in outcomes between treated and control units of the matched data, across the considered proportion of treated units, are reported in Table 5 and Fig. 1.

Relative to the unmatched data, all the three matching methods had lower absolute bias and RMSE values. Regardless of the proportion of treated units, coarsened exact matching (CEM) produced the least absolute bias and RMSE values - absolute bias ranged from 1.80 to 4.33; RMSE ranged from 27.91 to 35.19. Also, the absolute bias and RMSE values of CEM reduced as the proportion of treated units increased from 17% to 33%. The same pattern was, however, not observed for the other two methods, while only the RMSE values for propensity score caliper matching (PSCM) followed the same pattern.

Unlike the other methods, Mahalanobis Distance Matching (MDM) produced absolute bias and RMSE values, which increased as the proportion of treated units increased from 17% to 33%.

Discussion

In this study, we presented a Monte Carlo simulation study of three subset matching methods, namely; propensity score caliper matching, Mahalanobis distance matching and coarsened exact matching. We evaluated the performance of these methods based on the ability to induce balance on measured background covariates, as well as the performance of treatment effect estimates via the assessment of their absolute biases and root mean square errors.

This study revealed that coarsened exact matching is the most effective in balancing covariates. As effective as CEM appears to be, the choice of coarsening can make or mar its performance: If the elements of the coarsening values are too small, then too many observations may be discarded. It may then lead to inefficient solutions in the analysis stage: if they are set too high, more observations will be retained, but more covariate imbalances, model dependence and statistical bias, will be introduced (Iacus *et al*., 2012). It is fine if there is a constant treatment effect (discarding units will not change the estimand of interest) but discarding units in the case of heterogeneous treatment effects may dramatically shift the estimand being estimated.

In assessing the recovery of the true treatment effect, Mahalanobis distance matching was the most biased. Mahalanobis distance matching also resulted in the highest RMSE across all considered proportions of treated units. Overall, coarsened exact matching had the least absolute bias and RMSE across all considered proportions of treated units.

Matching based on propensity score methods is by far the most widely used in applied studies to date. Previous research findings reveal that propensity score caliper matching was the best PSM technique (Bai, 2011). However, it is worthy of note that when the sample size is small or violates the statistical assumptions, caliper matching will possibly become problematic, because it usually ignores the cases when they do not have matched pairs or do not meet the caliper's criterion. Thus, it requires larger sample sizes to be very effective. Also worthy of note about matching on propensity scores is the correct specification of the propensity score model. In practice, an excellent alternative to distance driven matching methods may be to estimate the propensity score using a more flexible approach than logistic regression, for example, by using ensemble methods (Lee *et al*., 2010).

A significant strength of this study is the utilization of a real data set that has been used to evaluate the performance of matching methods and to provide a suitable structure for simulating the 1,000 data sets. It has the advantage of simplifying data generation procedures and avoiding making arbitrary choices. This study has a few limitations: Firstly, we have not exhausted all possible matching methods that have been described in the literature. Secondly, we assumed a one-to-one pair matching and therefore did not consider the many-to-one or many-to-many matching methods. Thirdly, we only assumed matching without replacement. Lastly, Optimal matching (Rosenbaum, 1989) - another alternative to the utilized greedy, nearest neighbour matching method, was not considered in this study. The results of our simulation study are limited to scenarios represented by the simulated data, which are typical in the applied social sciences. Parameters of the data generation model were based on model coefficients of a widely used panel study of income dynamics survey.

Conclusion

In comparison to the other subset matching methods, the utilized simulation study has provided sufficient evidence for the outperformance of coarsened exact matching method to the other considered methods, in terms of balancing covariates and efficiency in estimation of treatment effects. Future studies should include more matching methods; simulations should be expanded to consider a broader range of settings, including a non-linear model and heterogeneous treatment effects.

Acknowledgment

We appreciate the anonymous reviewers for their valuable comments which improved this manuscript. The corresponding author would also like to appreciate the University of Ilorin, Nigeria, for granting him study leave to pursue his Ph.D. studies in South Africa.

Author's Contributions

Lateef Amusa designed the study, wrote the simulation codes, and analyzed the data. Temesgen Zewotir and Delia North critically reviewed the manuscript and gave constructive which improved the manuscript.

Ethics

This article is original and contains unpublished material. All authors declare and attest to no conflicts of interest in relation to this study.

References

Amusa, L.B., 2018. Reducing bias in observational studies: An empirical comparison of propensity score matching methods. Turkiye Klinikleri J. Biostatist., 10: 13-26. DOI: 10.5336/biostatic.2017-58633

Austin, P.C., 2007. The performance of different propensity score methods for estimating marginal odd ratios. Stat. Med., 26: 3078-94. DOI: 10.1002/sim.2781

Austin, P.C., 2009. Some methods of propensity-score matching had superior performance to others: Results of an empirical investigation and Monte Carlo simulations. Biometr. J., 51: 171-84. DOI: 10.1002/bimj.200810488

Austin, P.C., 2011. Optimal caliper widths for propensity□score matching when estimating differences in means and differences in proportions in observational studies. Pharmaceutical Stat., 10: 150-61. DOI: 10.1002/pst.433

Austin, P.C., 2014. A comparison of 12 algorithms for matching on the propensity score. Stat. Med., 33: 1057-69. DOI: 10.1002/sim.6004

Bai, H., 2011. A comparison of propensity score matching methods for reducing selection bias. Int. J. Res. Meth. Educ., 34: 81-107. DOI: 10.1080/1743727X.2011.552338

Cochran, W.G. and D.B. Rubin, 1973. Controlling bias in observational studies: A review. Sankhya Serial A, 35: 417-46. DOI: 10.1017/CBO9780511810725.005

Dehejia, R.H. and S. Wahba, 2002. Propensity score-matching methods for nonexperimental causal studies. Rev. Econom. Stat., 84: 151-161. DOI: 10.1162/003465302317331982

Drabik, A., G. Büscher, K. Thomas, C. Graf and D. Müller *et al*., 2012. Patients with type 2 diabetes benefit from primary care-based disease management: A propensity score matched survival time analysis. Populat. Health Manage., 15: 241-247. DOI: 10.1089/pop.2011.0063

Franklin, J.M., J.A. Rassen, D. Ackermann, D.B. Bartels and S. Schneeweiss, 2014. Metrics for covariate balance in cohort studies of causal effects. Stat. Med., 33: 1685-1699.

Fullerton, B., B. Pöhlmann, R. Krohn, J.L. Adams and F.M. Gerlach *et al*., 2016. The comparison of matching methods using different measures of balance: Benefits and risks exemplified within a study to evaluate the effects of German disease management programs on long□term outcomes of patients with type 2 diabetes. Health Services Res., 51: 1960-1980. DOI: 10.1111/1475-6773.12452

Ho, D.E., K. Imai, G. King and E.A. Stuart, 2011. MatchIt: Nonparametric preprocessing for parametric causal inference. J. Stat. Software, 42: 1-28.

Iacus, S.M., G. King and G. Porro, 2011. Multivariate matching methods that are monotonic imbalance bounding. J. Am. Stat. Assoc., 106: 345-61. DOI: 10.1198/jasa.2011.tm09599

Iacus, S.M., G. King and G. Porro, 2012. Causal inference without balance checking: Coarsened exact matching. Political Anal., 20: 1-24. DOI: 10.1093/pan/mpr013

Jacovidis, J.N., 2017. Evaluating the performance of propensity score matching methods: A simulation study. James Madison University.

King, G., R. Nielsen, C. Coberley, J.E. Pope and A. Wells, 2011. Comparative effectiveness of matching methods for causal inference. Unpublished Manuscript.

LaLonde, R.J., 1986. Evaluating the econometric evaluations of training programs with experimental data. Am. Econ. Rev. DOI: 10.1257/aer.p20151025

Lee, B.K., J. Lessler and E.A. Stuart, 2010. Improving propensity score weighting using machine learning. Stat. Med., 29: 337-46. DOI: 10.1002/sim.3782

Miksch, A., G. Laux, D. Ose, S. Joos and S. Campbell *et al*., 2010. Is there a survival benefit within a German primary care-based disease management program? Am. J. Managed Care, 16: 49-54. DOI: 10.1186/s13098-015-0065-9

Normand, S.L.T., M.B. Landrum, E. Guadagnoli, J.Z. Ayanian and T.J. Ryan *et al*., 2001. Validating recommendations for coronary angiography following an acute myocardial infarction in the elderly: A matched analysis using propensity scores. J. Clin. Epidemiol., 54: 387-98. DOI: 10.1016/S0895-4356(00)00321-8

Pan, W. and H. Bai, 2015. Propensity score analysis: Concepts and issues. Propensity Score Anal.: Fundamentals Dev.

R Core Team, 2016. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Rosenbaum, P.R., 1983. The central role of the propensity score in observational studies for causal effects. Biometrika, 70: 41-55. DOI: 10.1093/biomet/70.1.41

Rosenbaum, P.R. and D.B. Rubin, 1984. Reducing bias in observational studies using subclassification on the propensity score. J. Am. Stat. Assoc., 79: 516-524. DOI: 10.2307/2288398

Rosenbaum, P.R. and D.P. Rubin, 1985. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. Am. Stat., 39: 33-38. DOI: 10.1017/CBO9780511810725.019

Rosenbaum, P.R., 1989. Optimal matching for observational studies. J. Am. Stat. Assoc., 84: 1024-1032. DOI: 10.1080/01621459.1989.10478868

Rubin, D.B., 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. J. Educ. Psychol., 66: 688-688. DOI: 10.1037/h0037350

Scott, D.W., 2009. Sturges' rule. Wiley Interdisciplinary Reviews: Comput. Stat., 1: 303-306. DOI: 10.1002/wics.35

Sekhon, J.S., 2011. Multivariate and propensity score matching software with automated balance optimization: The matching package for R. J. Stat. Software, 42: 1-52.

Stock, S., A. Drabik, G. Büscher, C. Graf and W. Ullrich *et al*., 2010. German diabetes management programs improve quality of care and curb costs. Health Affairs, 29: 2197-2205. DOI: 10.1377/hlthaff.2009.0799

Stone, C.A. and Y. Tang, 2013. Comparing propensity score methods in balancing covariates and recovering impact in small sample educational program evaluations. Practical Assess. Res. Evaluat.

Stuart, E.A., 2010. Matching methods for causal inference: A review and a look forward. J. Inst. Math. Stat. DOI: 10.1214/09-STS313.

Windt, R. and G. Glaeske, 2010. Effects of a German asthma disease management program using sickness fund claims data. J. Asthma, 47: 674-679. DOI: 10.3109/02770900903556421