Research Article

Can Cui, Yunshu Zhang*, Shu Yang, Brian J. Reich, and David A. Gill

Matching estimators of causal effects in clustered observational studies

https://doi.org/10.1515/jci-2024-0061
received October 14, 2024; accepted May 21, 2025

Abstract: Marine conservation preserves fish biodiversity, protects marine and coastal ecosystems, and supports climate resilience and adaptation. Despite the importance of establishing marine protected areas (MPAs), research on the effectiveness of MPAs with different conservation policies is limited due to the lack of quantitative MPA information. In this article, by leveraging a global MPA database, we investigate the causal impact of MPA policies on fish biodiversity. To address challenges posed by this clustered and confounded observational study, we construct a bias-corrected matching estimator of the average treatment effect assuming treatment is assigned at the cluster level and a cluster-weighted bootstrap method for variance estimation. We establish the theoretical guarantees of the matching estimator and its variance estimator. Under our proposed matching framework, we recommend matching on both cluster-level and unit-level covariates to achieve efficiency. The simulation study results demonstrate that our matching strategy minimizes the bias and achieves the nominal confidence interval coverage. Applying our proposed matching method to compare different MPA policies reveals that the no-take policy is more effective than the multi-use policy in preserving fish biodiversity.

Keywords: causal inference, conservation, potential outcomes, weighted bootstrap

MSC 2020: 62D20

1 Introduction

1.1 Causal impact of marine protected areas (MPAs) on biodiversity

Preserving marine biological diversity is an important objective of governments, scientists, local communities, and conservationists. MPAs have been established worldwide to keep sustainable and resilient marine ecosystems by restricting destructive and extractive activities within their boundaries [1,2]. Despite widespread use, the effectiveness of many MPAs and different types of MPA policies in conserving marine biodiversity remains unclear [1]. Very few studies employ rigorous causal inference methods to assess MPA impacts and even less so to investigate the relative effects of different conservation policies [3]. Such studies, however, are

*** Corresponding author: Yunshu Zhang,** Department of Statistics, North Carolina State University, Raleigh, North Carolina, United States of America, e-mail: yzhan234@ncsu.edu
Can Cui: Department of Statistics, North Carolina State University, Raleigh, North Carolina, United States of America, e-mail: ccui5@alumni.ncsu.edu
Shu Yang: Department of Statistics, North Carolina State University, Raleigh, North Carolina, United States of America, e-mail: syang24@ncsu.edu
Brian J. Reich: Department of Statistics, North Carolina State University, Raleigh, North Carolina, United States of America, e-mail: bjreich@ncsu.edu
David A. Gill: Nicholas School of the Environment, Duke University, Durham, North Carolina, United States of America, e-mail: david.gill@duke.edu

Open Access. © 2025 the author(s), published by De Gruyter. This work is licensed under the Creative Commons Attribution 4.0 International License.

important and have significant policy implications, as prohibiting fishing activities that are potentially important for local food and livelihood security can result in significant social costs and harm (e.g., Kamat [4], Bennett and Dearden [5]).

Gill et al. [6] investigated the effectiveness of MPA management and its impacts on fish populations. They developed a database of ecological, management, social, and environmental conditions in and around hundreds of MPAs globally. In their study, management attributes such as available capacity were strongly associated with increases in fish biomass observed in MPAs. Nonetheless, the relative effects of different types of MPAs (referred to as policies or treatments), such as those that restrict fishing (hereafter called multiuse (MU) MPAs) and those that prohibit all fishing (hereafter called no-take (NT) or MPAs) require further investigation.

While the Gill et al. [6] database represents one of the largest global datasets of MPA conditions and ecological outcomes to date, its properties present significant challenges for applying traditional causal inference methods. First, given the intractability of conducting randomized experiments in many conservation settings, the global MPA dataset is observational and thus subject to confounding biases not present when treatment is randomized [7]. MU and NT MPAs are likely to be located in areas with different social, environmental, and regulatory conditions. Direct comparisons of the biodiversity between MU and NT MPAs are fallible. Second, the MPA data are spatially clustered as nearby sites are usually under the same conservation policy, whether it be because they lie within the same MPA, specific management zone within an MPA (e.g., no diving area), or larger-scale management policy area (e.g., regional or national level fishing policies). Individual sites also share similar geographical, environmental, and social features that are possibly dependent on each other. Therefore, estimating the causal impacts of policies such as MPAs requires appropriate methods for clustered and confounded data.

1.2 Previous work: Causal inference in observational studies

Although randomized experiments serve as the gold standard, observational studies can estimate causal effects when all confounding variables are well balanced between treatment groups. To adjust for the imbalance in observed confounding covariates, matching [8] is often applied to isolate causal effects due to its transparency and intuitive appeal.

While statistical methods to estimate causal effects in observational studies are growing, most methods apply to unstructured data (i.e., without clustering). However, clustering often exists because subjects may be grouped by experimental design, geography, or by sharing higher-level features. Examples include health and educational studies, where patients are nested in hospitals and students are clustered in classrooms or schools. Such clustered data structure poses additional challenges when inferring the causal effect. In our motivating example, the MPA database is naturally clustered, where several sites are nested in the MPA. Capturing MPA level as well as site-level features (e.g., local social or environmental conditions) is important to remove confounding biases when evaluating the effectiveness of environmental policies.

To estimate causal effects in clustered data, Cafri et al. [9] showed that treatment effect estimation is more accurate when accounting for cluster-level confounding variables. Even if sufficient individual-level covariates are included, ignoring cluster-level confounding covariates would leave a bias in estimation. In the matching framework, Zubizarreta and Keele [10] developed an algorithm for optimal cardinality matching, designed to balance covariate distributions between groups. However, this approach does not necessarily yield the most efficient estimator for the average treatment effect (ATE). Several alternative methods incorporating propensity scores have been proposed for clustered data [11–13]. For a comprehensive review of propensity score methods in clustered observational studies, see the study by Chang and Stuart [14]. However, King and Nielsen [15] discussed the inefficiency and failure of balancing covariate distributions between treatment groups using the propensity score. They attribute the inefficiency of matching on propensity scores to its goal of mimicking a completely randomized trial rather than a block-randomized trial as well as error in estimating the propensity score.

1.3 Our contribution: a matching strategy in clustered observational studies

This article focuses on matching as a nonparametric approach and intuitively mimics a cluster-randomized experiment. We aim to estimate the causal effect by matching estimators under the framework in Abadie and Imbens [16]. Following the characteristics in the MPA database, we analyze the clustered data where the treatment is clustered within the MPA. Nearby sites tend to be assigned the same MPA policy, and one MPA usually contains a single policy only. Cluster-level and unit-level covariates are available, and the outcome is collected at the unit level. To account for the conditional bias when matching on multiple covariates, we adopt the bias-corrected matching estimator [17] in clustered data for two common estimands, the ATE and average treatment effect on the treated, and establish the large sample properties. Under this data structure, matching on cluster-level covariates is sufficient to remove the confounding biases. However, we recommend including relevant unit-level covariates in matching to achieve higher efficiency and lower variance. We show reduced variance in theory and simulation to demonstrate the advantages of matching on both cluster-level and unit-level covariates.

To account for clustered dependence, we propose a cluster-weighted bootstrap method for variance estimation, which combines the idea of cluster bootstrap [18] and weighted bootstrap [19]. Based on a linearization of the matching estimator, the weighted bootstrap method creates residuals so that matching estimators can be viewed as the sample averages of residuals. The variance of the matching estimator can then be approximated by bootstrapping the residuals with appropriate weights. This method preserves the distribution of the number of times that each unit is matched in the resampling procedure. Thus, it avoids the failure of the standard bootstrap in this setting, as discussed by Abadie and Imbens [20].

The rest of this article is organized as follows. In Section 2, we introduce the motivating data and describe challenges in establishing causal effects due to the nature of the data structure. Section 3 introduces the notation, assumptions, and estimands of interests. Section 4 explores the large sample properties of matching estimators in clustered data. Section 5 presents the cluster-weighted bootstrap procedure for variance estimation. In Section 6, we apply the proposed matching estimator in the MPA data to investigate the causal effect of different marine protection policies on fish biodiversity. In Section 7, a simulation study is reported to evaluate the performance of the proposed matching estimator in clustered data. Finally, we conclude our findings in Section 8.

2 MPA data and exploratory analysis

The MPA dataset created by Gill et al. [6] includes social, environmental, and ecological information in 9987 sites within 215 MPAs worldwide (Figure 1). The number of sites in each MPA ranges from 1 to 1619, with a mean of 46 and a median of 8. Among 9987 sites, 3988 sites receive the NT policy, whereas 5999 are under the MU policy. The outcome variable is total fish biomass at each site, recorded in underwater visual surveys. There are 13 continuous covariates and 4 categorical covariates that describe the MPA-level and site-level features (Table 1). MPA-level covariates include MPA size and country. The other covariates include site-level social and environmental conditions, as well as sampling protocol, location, and date.

Sites within the same MPA usually receive the same policy (i.e., same fishing regulations), and each site belongs only to one MPA. As a result, the dataset is naturally clustered where observed sites are nested within the MPA, and conservation policies are geographically clustered. The cluster structure brings difficulty in causal inference due to potential confounding. Both cluster-level and site-level covariates could contribute to the confounding bias. Sites in the same MPA share common environmental, MPA-level, and geographical characteristics, affecting both the fish population and MPA policy assignment [3,6,21]. Site-specific covariates, including depth, distance to population centers (also called “markets”), size of neighboring human population, and chlorophyll concentration, are also relevant to the ecological outcome [6,22–24]. Within the same MPA, implementing either the MU or NT policy could be heavily influenced by preexisting ecological conditions, local tourism, fishing, or politics [25,26], which are ideally captured as site-specific covariates.

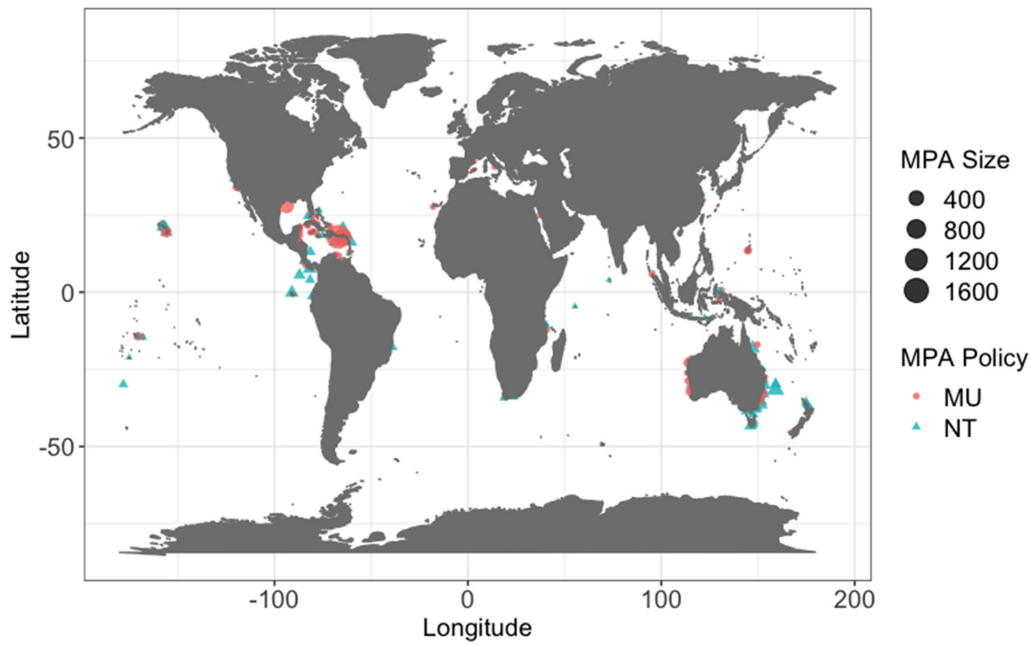

Figure 1: Map showing MPA location, size and policy type (MU = multi-use, NT = no-take); MPA policies are present by the majority within each MPA.

Confounding and clustering present two major challenges. We compare the covariate distributions under the two MPA policies for both unadjusted and adjusted samples. The unadjusted sample refers to the raw observation, while the adjusted sample results from multiple matching (one-to-three) using the Mahalanobis distance and with replacement. A hypothetical example to illustrate the applied multiple matching is provided in Supplementary Material S6, Figure S1.

Figure 2 (a) describes the covariate balance by calculating the standardized mean difference for unadjusted and adjusted samples. For the unadjusted sample, differences between two MPA policies among covariates suggest a nontrivial impact of confounding. After matching, many covariates become more balanced;

Table 1: Feature list in the MPA database with units in parentheses for continuous variables and number of levels in parentheses for categorical variables

	Site-level covariates	MPA-level covariates
Continuous (13)	Latitude (degree)	MPA size (km^2)
	Longitude (degree)	
	Depth (m)	
	Wave exposure (kW/m)	
	Distance to shoreline (km)	
	Distance to population center (“market”; km)	
	Coastal population (million/100 km^2)	
	Sample date (year)	
	Minimum sea surface temperature (°C)	
	Chlorophyll-A (mg/m^3)	
	Reef area within 15 km (km^2)	
	MPA age (years)	
Categorical (4)	Habitat type (16)	Country
	Marine ecoregion (56)	
	Sampling protocol (6)	

A detailed summary of the covariates is given by Gill et al. [6], Supplementary Table 5.

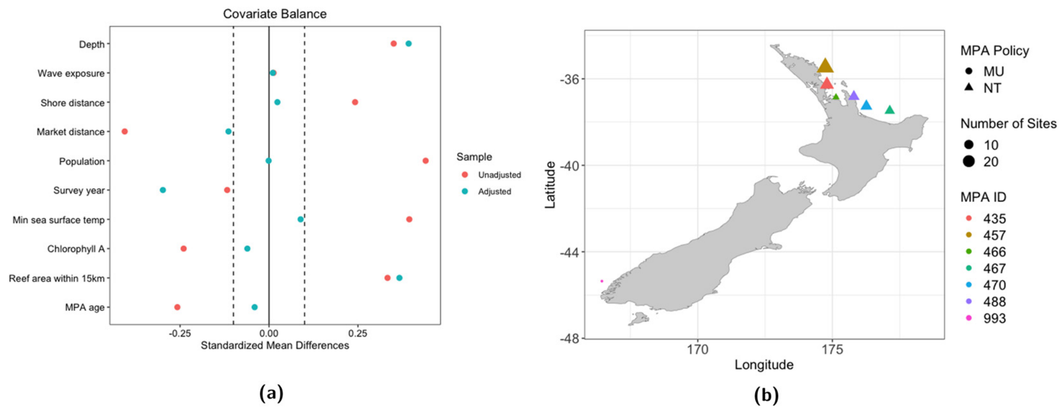

Figure 2: Challenges in MPA global dataset. (a) Covariate balance between unadjusted and adjusted situations and (b) MPAs in New Zealand.

however, due to the matching discrepancy (since matching a large dimension of covariates), several covariates still exhibit severe imbalance, requiring further adjustments for residual confounding bias. An example of seven selected MPA locations in New Zealand is plotted in Figure 2(b), where six are under the NT policy while the other is under the MU policy. Here, different MPAs contain different numbers of sites, ranging from 1 to 29. Figure 2(b) shows a type of clustering pattern in this MPA global dataset that nearby MPAs tend to follow the same policy. It is practically reasonable because similar regions are likely to share common environmental, geographical, and local governmental characteristics, which impact fish biodiversity and MPA policies' choice. Therefore, to explore the causal effect of MPA policies on fish biodiversity, methods that account for both the cluster-level and site-level confounding factors are desired.

3 Notation, assumptions, and estimands

To establish causal effect in clustered observational studies, we build our proposed matching strategy based on the potential outcomes framework [27–30], also known as the Neyman-Rubin causal model. Under the potential outcomes framework, the unit-level causal effect is defined as the difference between the outcomes under treatment and control in the same unit. Since we always observe only one of the outcomes for a single unit in reality, it is also considered a missing data problem.

Several works under the potential outcomes framework have been proposed for clustered data [11,13,31–35]. Distinguished from the existing literature, we focus on the setting where treatment is assigned at the cluster level and leverage the framework of matching estimators in the study by Abadie and Imbens [16] for causal inference.

Suppose for units $i = 1, \ldots, N$ in cluster $r = 1, \ldots, R$, where the sample size of each cluster is $n_1, \ldots, n_r$, respectively. Let Z_r be a vector of cluster-level covariates and A_r be the binary treatment indicator. To emphasize the role of cluster, we use $X_{i,r(i)}$ to denote the vector of unit-level covariates of the ith unit, where $r = r(i)$ indicates the cluster it belongs to and thus is a function mapping from $i \in \{1, \ldots, N\}$ to $r \in \{1, \ldots, R\}$. For simplicity, let $X_{ir} = X_{i,r(i)}$ denote the unit-level covariates, and similarly, let $Y_{ir}(a)$ be the potential outcome receiving the treatment $a \in \{0, 1\}$, and $Y_{ir} = Y_{ir}(A_r)$ be the observed outcome. Here, we suppose each unit belongs to only one cluster and clusters are not overlapped, i.e., $N = n_1 + \ldots + n_R$. Also, assume treatment is assigned at the cluster level, which implies that units within the same cluster are in the same treatment group. Supplementary Material Section 5 describes an extension to the case where the treatment varies within clusters.

To define the potential outcomes model, we introduce additional notation. Given $X_{ir} \in \mathbb{X}, Z_r \in \mathbb{Z}$ and $a \in \{0, 1\}$, we denote $\mu(x, z, a) = \mathbb{E}[Y_{ir} | X_{ir} = x, Z_r = z, A_r = a]$, $\mu_a(x, z) = \mathbb{E}[Y_{ir}(a) | X_{ir} = x, Z_r = z]$, $\sigma^2(x, z, a) =$

$\mathbb{V}[Y_{ir}|X_{ir} = x, Z_r = z, A_r = a]$, and $\sigma_a^2(x, z) = \mathbb{V}[Y_{ir}(a)|X_{ir} = x, Z_r = z]$. Assume the following relationship between the potential outcome $Y_{ir}(a)$ and observed covariates X_{ir} and Z_r,

$$Y_{ir}(a) = \mu_a(X_{ir}, Z_r) + \alpha_r + \varepsilon_{ir}, \tag{1}$$

where α_r represents unobserved cluster-level random effect with mean 0, ε_{ir} independently follows a distribution with mean 0 and variance $\sigma^2(X_{ir}, Z_r, a)$ for $i = 1, \ldots, N$ and $r = 1, \ldots, R$. By including the cluster-level random effects, we impose a dependence structure for the error terms. Our goal is to estimate two aggregate estimands, i.e., the ATE and the average treatment effect for the treated (ATT). The ATE is defined as $\tau = \mathbb{E}\{Y(1) - Y(0)\}$, and the ATT is $\tau^t = \mathbb{E}\{Y(1) - Y(0)|A = 1\}$.

We adopt the matching estimator proposed in Abadie and Imbens [16] and extend to clustered data. Each unit is matched to the closest M units in the opposite treatment group, where matching is done based on the Euclidean distance of the covariates. Denote $\mathcal{J}_M(i, r)$ as the indices of the M matched units for unit i in the cluster r, and $K_M(i, r)$ be the number of times that unit i in cluster r is matched, i.e., $K_M(i, r) = \sum_{l=1}^{N}\sum_{k=1}^{R}\mathbb{I}\{(i, r) \in \mathcal{J}_M(l, k)\}$. Similar to X_{ir}, $r = r(i)$ in $\mathcal{J}_M(i, r)$ and $K_M(i, r)$ denotes the cluster to which unit i belongs. In the double summation, we consider only the valid terms and disregard any undefined terms with incorrect cluster indices. Then the missing outcome for the unit i in cluster r can be imputed as follows:

$$\hat{Y}_{ir}(0) = \begin{cases} Y_{ir}, & \text{if } A_r = 0, \\ \dfrac{1}{M}\displaystyle\sum_{(j,k)\in\mathcal{J}_M(i,r)} Y_{jk}, & \text{if } A_r = 1, \end{cases} \quad \text{and} \quad \hat{Y}_{ir}(1) = \begin{cases} \dfrac{1}{M}\displaystyle\sum_{(j,k)\in\mathcal{J}_M(i,r)} Y_{jk}, & \text{if } A_r = 0, \\ Y_{ir}, & \text{if } A_r = 1. \end{cases}$$

To establish a valid causal effect in clustered data, we visit the necessary assumptions. We first retain the stable unit treatment values assumption (SUTVA), that is, the potential outcomes for each unit are not influenced by the treatment assigned to other units. While the SUTVA is a strong assumption, it is plausible in our case where most sites within one MPA received the same policy and MPAs are in general geographically separated from each other. We then modify the strong ignorability assumption in our setting, where treatments are assigned at the cluster level. Under the SUTVA and modified strong ignorability assumptions, $\mu(x, z, a) = \mu_a(x, z)$ and $\sigma^2(x, z, a) = \sigma_a^2(x, z)$.

Assumption 1. (i) $\{Y_{ir}(0), Y_{ir}(1)\} \perp\!\!\!\perp A_r|Z_r$; (ii) $\eta < P(A_r = 1|Z_r = z) < 1 - \eta$ almost surely, for some $\eta > 0$.

Assumption 1 suggests conditional independence between treatment A_r and the potential outcome $Y_{ir}(a)$ when accounting for the cluster-level covariates Z_r. It also requires overlap between treatment groups in the observed covariate distributions. The assumption of strong ignorability in (i) is untestable in practice. However, it is reasonable to proceed under this condition with enough knowledge on the treatment assignment mechanism and sufficient cluster-level covariates in the analysis. When treatment assignment occurs at the cluster level, individual-level covariates are not needed to achieve the strong ignorability assumption. For example, in the MPA dataset, sites within the same MPA typically receive the same policy, and the policy is determined by the MPA's characteristics rather than site-specific covariates. Researchers can determine whether incorporating summary statistics of individual-level covariates is necessary

The matching estimator for the ATE in clustered data is

$$\hat{\tau}_{mat} = \frac{1}{N}\sum_{i=1}^{N}\sum_{r=1}^{R}\{\hat{Y}_{ir}(1) - \hat{Y}_{ir}(0)\},$$

which can be rewritten as follows:

$$\hat{\tau}_{\text{mat}} = \frac{1}{N}\sum_{i=1}^{N}\sum_{r=1}^{R}(2A_r - 1)\left\{1 + \frac{K_M(i, r)}{M}\right\}Y_{ir}. \tag{2}$$

Note that the denominator is N instead of NR, because only one $r = r(i)$ as the cluster index of unit i is included. Consequently, the total number of terms in the summation equals the total sample size N. Following a similar routine in unstructured data, we decompose the estimator $\hat{\tau}_{\text{mat}}$ as follows

$$\hat{\tau}_{\text{mat}} = \frac{1}{N}\sum_{i=1}^{N}\sum_{r=1}^{R}(2A_r - 1)\left\{1 + \frac{K_M(i, r)}{M}\right\}Y_{ir} = \overline{\tau(X, Z)} + E_M + B_M, \tag{3}$$

where $\overline{\tau(X,Z)} = N^{-1}\sum_{i=1}^{N}\sum_{r=1}^{R}\{\mu_1(X_{ir},Z_r) - \mu_0(X_{ir},Z_r)\}$ represents the average conditional treatment effect, $E_M = N^{-1}\sum_{i=1}^{N}\sum_{r=1}^{R}(2A_r - 1)\{1 + K_M(i,r)/M\}\{Y_{ir} - \mu_{A_r}(X_{ir},Z_r)\}$ is a weighted average of the residuals, and the conditional bias relative to $\overline{\tau(X,Z)}$ is $B_M = N^{-1}\sum_{i=1}^{N}\sum_{r=1}^{R}(2A_r - 1)[M^{-1}\sum_{(j,k)\in\mathcal{J}_M(i,r)}\{\mu_{1-A_r}(X_{ir},Z_r) - \mu_{1-A_r}(X_{jk},Z_k)\}]$.

For the ATT in clustered data, the strong ignorability assumption can be relaxed as in the unstructured data as well.

Assumption 2. (i) $\{Y_{ir}(0)\} \perp\!\!\!\perp A_r|Z_r$; (ii) $P(A_r = 1|Z_r = z) < 1 - \eta$ almost surely, for some, $\eta > 0$.

The estimator of the ATT can be written as follows:

$$\hat{\tau}^t_{\text{mat}} = \frac{1}{N_1}\sum_{A_r=1}\{Y_{ir} - \hat{Y}_{ir}(0)\} = \frac{1}{N_1}\sum_{i=1}^{N}\sum_{r=1}^{R}\left\{A_r - (1 - A_r)\frac{K_M(i,r)}{M}\right\}Y_{ir}, \tag{4}$$

where N_1 is the number of subjects receiving treatment 1. Similar to the decomposition for the ATE estimator,

$$\hat{\tau}^t_{\text{mat}} - \tau^t = \{\overline{\tau(X)}^t - \tau^t\} + E^t_M + B^t_M, \tag{5}$$

where $\overline{\tau(X)}^t = N_1^{-1}\sum_{i=1}^{N}\sum_{r=1}^{R}A_r\{\mu(X_{ir},Z_r,1) - \mu_0(X_{ir},Z_r)\}$, $E^t_M = N_1^{-1}\sum_{i=1}^{N}\sum_{r=1}^{R}\{A_r - (1 - A_r)K_M(i,r)/M\}\{Y_{ir} - \mu_{A_r}(X_{ir},Z_r)\}$, and $B^t_M = N_1^{-1}\sum_{i=1}^{N}\sum_{r=1}^{R}A_r M^{-1}\sum_{(j,k)\in\mathcal{J}_M(i,r)}\{\mu_0(X_{ir},Z_r) - \mu_0(X_{jk},Z_k)\}$.

Consistent with the findings in unstructured data, the matching discrepancies of the ATE and ATT estimators come from the conditional bias terms B_M and B^t_M, which need careful consideration when matching on multiple covariates. According to Abadie and Imbens [16], $B_M = O_p(N^{-1/k})$ such that the asymptotic distribution of the ATE estimator can be dominated by the bias when matching on $k > 1$ covariates. Similar rates apply for the ATT estimator, where $B^t_M = O_p(N_1^{-r/k})$ for some $r \geq 1$. Since we consider matching on both cluster-level and unit-level covariates, such bias terms are non-negligible. For instance, in Figure 2(a), although matching reduces covariate imbalance to a large extent, certain covariate imbalance still persists. Following Abadie and Imbens [17], we work on the bias-corrected estimators instead. Denote $\hat{\tau}$ and $\hat{\tau}^t$ as the bias-corrected estimators for ATE and ATT, we have

$$\begin{aligned}\hat{\tau} &= \hat{\tau}_{\text{mat}} - \hat{B}_M\\ &= \hat{\tau}_{\text{mat}} - \frac{1}{N}\sum_{i=1}^{N}\sum_{r=1}^{R}(2A_r - 1)\left[\frac{1}{M}\sum_{(j,k)\in\mathcal{J}_M(i,r)}\{\hat{\mu}_{1-A_r}(X_{ir},Z_r) - \hat{\mu}_{1-A_r}(X_{jk},Z_k)\}\right],\end{aligned}$$

and $\hat{\tau}^t = \hat{\tau}^t_{\text{mat}} - \hat{B}^t_M = \hat{\tau}^t_{\text{mat}} - N^{-1}\sum_{i=1}^{N}\sum_{r=1}^{R}A_r M^{-1}\sum_{(j,k)\in\mathcal{J}_M(i,r)}\{\hat{\mu}_0(X_{ir},Z_r) - \hat{\mu}_0(X_{jk},Z_k)\}$, where $\hat{\mu}_{A_r}(X_{ir},Z_r)$ is a consistent estimator of $\mu_{A_r}(X_{ir},Z_r)$.

4 Large sample properties

In our clustered data setting, matching on cluster-level covariates is sufficient to account for the confounding variables. However, motivated by blocked randomized experimental designs, matching on both cluster-level and unit-level covariates may improve efficiency. To compare the two matching schemes, we denote S as a unified variable representing either the cluster-level covariates or both cluster-level and unit-level covariates to be matched on. We explore the asymptotic normality under the fixed M matches for the bias-corrected matching estimators. We then illustrate the benefits of matching on both cluster-level and unit-level covariates based on the asymptotic variance. Following Abadie and Imbens [16], we extend necessary assumptions for independent and identically distributed data to clustered data (details can be found in Supplementary Material S1).

To illustrate the role of the matching scheme, for a general random variable S, we further define $\mu_a(S) = \mathbb{E}[Y_{ir}(a)|S]$ and $\tau(S) = \mu_1(S) - \mu_0(S)$ as the conditional mean outcome and treatment effect functions. Denote the variance terms $V^{\tau(S)} = \mathbb{E}[(\tau(S) - \tau)^2]$, $V^{\tau(S),t} = \mathbb{E}[(\tau(S)^t - \tau^t)^2]$, $V^E = \text{plim}[N^{-1}\sum_{i=1}^{N}[\sum_{r=1}^{R}\{1 + M^{-1}K_M(i,r)\}^2]\mathbb{V}[Y_{ir} - \mu_{A_r}(S_{ir})]]$, and $V^{E,t} = \text{plim}\,[N_1^{-1}\sum_{i=1}^{N}[\sum_{r=1}^{R}\{A_r - (1 - A_r)M^{-1}K_M(i, r)\}^2]\mathbb{V}[Y_{ir} - \mu_{A_r}(S_{ir})]]$, we establish the asymptotic normality for bias-corrected estimators in Theorems 1 and 2 (see proofs in Supplementary Material S2).

Theorem 1. *Suppose* Assumption 1, *and Assumptions S1 and S2 in Supplementary Material Section S1 hold. Suppose that the cluster sample size we have, for* $r = 1, \ldots, R$, *satisfies the condition that* $\min_{1 \le r \le R} n_r \to \infty$ *and* $\sup_{1 \le r \le R} n_r = O(N^{1/2})$. *Then*

$$\{V^E + V^{\tau(S)}\}^{-1/2}\sqrt{N}(\hat{\tau} - \tau) \xrightarrow{d} N(0, 1),$$

where V^E *and* $V^{\tau(S)}$ *are finite.*

Theorem 2. *Suppose* Assumptions 2, *and S1 and S2 in Supplementary Material Section S1 hold. Suppose that the cluster sample size we have for* $r = 1, \ldots, R$, *satisfies the condition that* $\min_{1 \le r \le R} n_r \to \infty$ *and* $\sup_{1 \le r \le R} n_r = O(N^{1/2})$. *Then*

$$\{V^{E,t} + V^{\tau(S),t}\}^{-1/2}\sqrt{N_1}(\hat{\tau}^t - \tau^t) \xrightarrow{d} N(0, 1),$$

where $V^{E,t}$ *and* $V^{\tau(S),t}$ *are finite.*

The asymptotic variances for the two estimators involve the variances of residual terms, i.e., V^E and $V^{E,t}$. Specifically, we have

$$V^E = \text{plim}\left[\frac{1}{N^2}\sum_{i=1}^{N}\left[\sum_{r=1}^{R}\{1 + M^{-1}K_M(i, r)\}^2\right]\mathbb{V}[Y_{ir} - \mu_{A_r}(S_{ir})]\right],$$

$$V^{E,t} = \text{plim}\left[\frac{1}{N_1^2}\sum_{i=1}^{N}\left[\sum_{r=1}^{R}\{A_r - (1 - A_r)M^{-1}K_M(i, r)\}^2\right]\mathbb{V}[Y_{ir} - \mu_{A_r}(S_{ir})]\right].$$

Consider two matching schemes, i.e., $S = (X, Z)$ and $S = Z$, we note that for $\mathbb{V}[Y_{ir} - \mu_{A_r}(X_{ir}, Z_r)]$ and $\mathbb{V}[Y_{ir} - \mu_{A_r}(Z_r)]$, when unit-level covariates provide valuable information to the outcome Y_{ir}, matching on both cluster-level and unit-level covariates results in lower residual variance than matching on cluster-level covariates. Despite the increased number of covariates for matching, the bias correction helps remove the conditional bias. Hence, we recommend that matching on both cluster-level and unit-level covariates is more efficient when the unit-level information is available.

5 Variance estimation

Variance estimation for matching estimators has been investigated in both parametric and nonparametric ways. In the study by Abadie and Imbens [16], an analytic form for the large sample variance is proposed with consistency achieved. Although the bootstrap method [36] is widely used to calculate standard errors for estimators with complicated forms, Abadie and Imbens [20] showed that the standard bootstrap is invalid for matching estimators. The bootstrap method does not preserve the distribution of $K_M(i)$, which follows a Binomial distribution. To overcome this challenge, statistical tools, including wild bootstrap [37] and weighted bootstrap methods [19], are developed to estimate asymptotic variance for matching estimators. The weighted bootstrap method [19] can be used when matching is directly performed on covariates, while the wild bootstrap [37] is applied based on the estimated propensity score. Both methods apply to unstructured data. As for clustered data, there are two main bootstrap strategies: (i) two-stage bootstrap, which is to resample entire clusters at first and then resample subjects within the selected clusters, and (ii) cluster bootstrap, which resamples entire clusters and includes all subjects from the selected clusters. Davison and Hinkley [18] discussed different bootstrap methods and showed that the latter is preferable theoretically in clustered data. Recent studies suggested that allowing the number of matches to grow with the sample size can address the invalidity of the bootstrap [38]. However, it remains challenging to determine how many matches are sufficient to ensure the bootstrap's validity in a given finite sample dataset.

Due to the complex analytic form of matching estimators in clustered data, we employ resampling methods to estimate the variance of the matching estimator and leverage the general procedure of the weighted bootstrap. To account for the dependence in the linear terms in the weighted bootstrap, we incorporate the cluster bootstrap for variance estimation. We describe the extension of weighted bootstrap in clustered data for the ATE first and present the procedure for our cluster weighted bootstrap method. A similar procedure for the ATT is summarized in Supplementary Material S4.

Theorem 1 states that under certain conditions, the bias-corrected estimator $\hat{\tau}$ is asymptotically normal: $\sqrt{N}(\hat{\tau} - \tau)/\sigma \xrightarrow{d} N(0, 1)$, where $\sigma^2 = \sigma_1^2 + \sigma_2^2$. The asymptotic variance σ^2 consists of two parts, i.e.,

$$\sigma_1^2 = \mathbb{E}[\{\mu(S_{ir}, 1) - \mu(S_{ir}, 0) - \tau\}^2], \quad \sigma_2^2 = \frac{1}{N}\sum_{i=1}^{N}\sum_{r=1}^{R}\{1 + M^{-1}K_M(i, r)\}^2\sigma^2(S_{ir}, A_r),$$

where σ_1^2 measures the variability of $\overline{\tau(S)} - \tau$ and σ_2^2 captures the variability of weighted average of the residuals E_M. By following Otsu and Rai [19], we write our bias-corrected estimator $\hat{\tau} = \hat{\tau}_{\text{mat}} - \hat{B}_M$ as a linear form in clustered data,

$$\begin{aligned}
\hat{\tau} &= \frac{1}{N}\sum_{i=1}^{N}\sum_{r=1}^{R}\{\hat{\mu}_1(S_{ir}) - \hat{\mu}_0(S_{ir})\} + \frac{1}{N}\sum_{i=1}^{N}\sum_{r=1}^{R}(2A_r - 1)\left\{1 + \frac{1}{M}K_M(i, r)\right\}\{Y_{ir} - \hat{\mu}_{A_r}(S_{ir})\}\\
&= \frac{1}{N}\sum_{i=1}^{N}\sum_{r=1}^{R}\left[\{\hat{\mu}_1(S_{ir}) - \hat{\mu}_0(S_{ir})\} + (2A_r - 1)\left\{1 + \frac{1}{M}K_M(i, r)\right\}\{Y_{ir} - \hat{\mu}_{A_r}(S_{ir})\}\right]\\
&= \frac{1}{N}\sum_{i=1}^{N}\hat{\tau}_i,
\end{aligned}$$

where $\hat{\tau}_i = \sum_{r=1}^{R}\{\hat{\mu}_1(S_{ir}) - \hat{\mu}_0(S_{ir})\} + (2A_r - 1)\{1 + M^{-1}K_M(i, r)\}\{Y_{ir} - \hat{\mu}_{A_r}(S_{ir})\}$.

On the basis of this form, we express the ith residual as $\hat{\tau}_i - \hat{\tau} = \hat{\xi}_i + \hat{e}_i$ with $\hat{\xi}_i$ and $\hat{e}_i$ are estimated values of $\xi_i = \sum_{r=1}^{R}\{\mu_{A_r}(S_{ir}) - \mu_{1-A_r}(S_{ir})\} - \tau$ and $e_i = \sum_{r=1}^{R}(2A_r - 1)\{1 + M^{-1}K_M(i, r)\}\{Y_{ir} - \mu_{A_r}(S_{ir})\}$, respectively. This form is the key aspect to obtain the estimated variance for matching estimators. A cluster-robust variance estimator $\hat{V}^{\text{CRV E}}$ can be computed as follows:

$$\hat{V}^{\text{CRV E}} = \frac{1}{N^2}\sum_{r'=1}^{R}n_{r'}^2\hat{\sigma}_{r'}^2,$$

where

$$\hat{\sigma}_{r'}^2 = \frac{1}{n_{r'} - 1}\sum_{i:r(i)=r'}(\hat{\tau}_i - \bar{\hat{\tau}}_{r'})^2,$$

and $\bar{\hat{\tau}}_{r'} = n_{r'}^{-1}\sum_{i:r(i)=r'}\hat{\tau}_i$ is the cluster mean. This estimator accounts for within-cluster dependence and provides a consistent variance estimate for the matching estimator.

Alternatively, we propose a cluster-level weighted bootstrap method following Otsu and Rai [19]. The key idea is to treat $\{\hat{\tau}_i\}_{i=1}^{N}$ as observations, ensuring that the distribution of $K_M(i)$ is preserved. Bootstrapping on $\{\hat{\tau}_i\}_{i=1}^{N}$ therefore provides a valid variance estimate. While the weighted bootstrap estimator ultimately aligns with the cluster-robust variance estimator based on the standard sample variance structure [39], we recommend using bootstrap methods to enhance finite-sample performance. Prior research has demonstrated the strong finite-sample properties of wild bootstrap methods in classical settings such as linear regression [40]. Other studies have also shown that bootstrap-based approaches perform well in situations with a small number of clusters or unbalanced cluster sizes [41–43]. On the basis of the literature, we believe that the proposed cluster bootstrap procedure is likely to exhibit good finite-sample performance in practice. Moreover, the bootstrap estimator is both convenient and easy to implement.

Considering the clustered data where unobserved cluster-level confounding covariates could exist across and within clusters, we adopt the cluster bootstrap method [18] to account for the variance from potential cluster-level covariates. The cluster bootstrap suggests resampling on the cluster levels first and then including all observations within selected units. By incorporating the weighted bootstrap algorithm, we propose our cluster-weighted bootstrap method as follows. The weights can be generated from multinomial random variables $\{M_i^*\}_{i=1}^N$.

- **Step 1**: Obtain the weighted bootstrap samples $\{\hat{\tau}_i\}_{i=1}^N$ based on the matching estimator framework.
- **Step 2**: For clustered data with n observations and R non-overlapped clusters, sample R clusters with replacement.
- **Step 3**: Include all $\{\hat{\tau}_i\}_{i=1}^N$ within selected clusters and calculate their corresponding weights $\{W_i^*\}_{i=1}^N$. One option of generating weights is to set $W_i^* = M_i^*/\sqrt{N}$, where $(M_1^*, \dots, M_N^*)$ is a vector from a multinomial distribution with equal probability.
- **Step 4**: Obtain a bootstrap replicate as $\hat{\tau}_b^* = \sum_{i=1}^N W_i^*(\hat{\tau}_i - \hat{\tau})$.
- **Step 5**: Repeat the Step 1–4 B times. Compute the bootstrap variance estimator for the bias-corrected matching estimator $\hat{\tau}$ as the empirical variance of $\{\hat{\tau}_b^*\}_{b=1}^B$.

Theorem 3. *Suppose that* Assumption 1 *and S1, S3 in Supplementary Material S3 hold. Suppose that the cluster sample size we have, for* $r = 1, \dots, R$, *satisfies the condition that* $\min_{1\le r\le R} n_r \to \infty$ *and* $\sup_{1\le r\le R} n_r = O(n^{1/2})$. *Then*

$$\mathbb{E}\left[\left\{\sum_{i=1}^N W_i^*(\hat{\tau}_i - \hat{\tau})|(\boldsymbol{Y}, \boldsymbol{A}, \boldsymbol{S})\right\}^2\right] \xrightarrow{p} \sigma^2,$$

where σ^2 *is the asymptotic variance of cluster matching estimator* $\tilde{\tau}$.

Proof for Theorem 3 is detailed in Supplementary Material S3.

6 Analysis of conservation policy effects on marine biodiversity

The proposed matching framework is applied to the Gill et al.'s [6] MPA dataset to investigate the causal relationship between MPA policies and fish biodiversity. We consider MU as the treatment while the NT serves as the control. Two estimands are of interest, i.e., $\text{ATE} = \mathbb{E}[Y(\text{MU}) - Y(\text{NT})]$ and $\text{ATT} = \mathbb{E}[Y(\text{MU}) - Y(\text{NT})|A = \text{MU}]$. We impute the missing counterfactual for each site by matching with replacement and with a fixed number of matches. The Mahalanobis distance metric is calculated to measure the similarity of sites under different policies. Three nearest control sites are matched to the treatment site. Due to the skewed covariates distributions between two groups, we transform the non-negative continuous covariates by the Box-Cox transformation. The outcome of interest is the log transformation of fish biomass (g/100 m^2). Following our recommendation, matching is conducted on both MPA-level and site-level covariates, and the performance is compared to matching on MPA-level covariates. The mean outcome functions are estimated by three methods: (i) the spline regression [44], (ii) the sieve method [45], and (iii) the regression forest [46]. Because of the large number of first-order and second-order terms among covariates, we apply the LASSO method [47] to select a subset of covariates before matching [48,49]. The regularization parameter is chosen by fivefold cross validation and the bias-corrected estimators for the ATE and ATT are derived. Corresponding variances are computed by the proposed cluster-weighted bootstrap method.

Table 2 summarizes the estimated ATE and ATT with 95% confidence intervals under different matching strategies. When matching only on MPA-level covariates, all three methods result in negative point estimates for ATE and ATT. However, none detect significant differences between the MU and NT policy on fish biomass. When matching on both MPA-level and site-level covariates, the sieve method shows significantly different impacts of MPA policies on fish biomass. The ATE under the sieve method is significantly negative, suggesting that the NT policy is more beneficial to fish biodiversity. The result of the ATT by using the sieve method is

Table 2: Summary of the ATE and the ATT with estimated standard errors in parentheses when comparing the MU policy and NT policy in MPAs where MU is considered as treatment group; Response is $\log(\text{Fish Biomass})$

	ATE		ATT	
	Point estimate	**95% CI**	**Point estimate**	**95% CI**
Matching on MPA-level covariates				
Sieve method	−0.49 (0.31)	(−1.10, 0.13)	−0.67 (0.50)	(−1.64, 0.30)
Smooth spline	−0.27 (0.25)	(−0.77, 0.23)	−0.19 (0.38)	(−0.93, 0.56)
Regression forest	−0.57 (0.35)	(−1.26, 0.12)	−0.82 (0.49)	(−1.77, 0.13)
Matching on all covariates				
Sieve method	−0.41 (0.17)	(−0.76, −0.07)	−0.58 (0.26)	(−1.10, −0.06)
Smooth spline	−0.34 (0.30)	(−0.93, 0.26)	−0.41 (0.41)	(−1.22, 0.40)
Regression forest	−0.70 (0.32)	(−1.32, −0.07)	−1.03 (0.53)	(−2.06, 0.00)

consistent with the ATE result, which shows a positive impact of using the NT policy. Comparing the procedures between matching on MPA-level covariates and matching on all relevant covariates, we find a reduced estimated variance under the sieve method as expected when including individual covariates that are relevant to the outcome.

In these MPA data, several site-level covariates such as distance to markets, reef area within 15 km, and neighboring human population size are relevant to the fish biodiversity under either the MU or NT policy (Figure S2 in Supplementary Material S6). Certain regions at the habitat, country and ecoregion levels also show an important association with the fish biodiversity. Taking these covariates into account for matching helps reduce the matching variance. Using the regression forest method also reveals the beneficial effect of the NT policy for the estimated ATE, while the effect is barely detected for the ATT.

7 Simulation study

We conduct a simulation study to evaluate the finite-sample performance of matching estimators when matching on cluster-level covariates only or matching on both cluster-level and unit-level covariates, and compare the proposed cluster weighted bootstrap method with the original weighted bootstrap on variance estimation. Table 3 outlines the two data-generating processes – balanced and unbalanced designs – and the 12 matching estimators, which vary based on the choice of nuisance function estimation, covariates used for matching, and method for variance estimation. Table 4 summarizes the results under different scenarios for the ATE. When matching on the cluster-level covariate, the small biases for both balanced and unbalanced cluster size settings suggest that matching on the cluster-level covariate is sufficient to remove estimation bias for the treatment effect. It is also consistent with the theoretical findings by Abadie and Imbens [16] that the conditional bias is ignorable when matching on a single covariate. When comparing three approaches to approximate the conditional outcome mean functions, the sieve method usually outperforms the other two with the lowest bias. The linear spline method shows advantages in reducing the bias for unbalanced data but suffers from underestimated variance. The regression forest method has the largest absolute bias in most cases.

As for variance estimation, the standard weighted bootstrap always results in a much smaller variance than the cluster-weighted bootstrap and thus lower coverage for the 95% confidence interval. By ignoring the cluster effect, the weighted bootstrap method resamples on the unit level such that it fails to approximate the true distribution of matching estimators. On the contrary, by taking clusters into consideration and resampling at the cluster level, our proposed cluster-weighted bootstrap method results in high coverage for the 95% confidence interval.

Comparing procedures 1 and 3, the estimated variances when using both cluster and unit covariates for matching are always smaller than those using only the cluster-level covariate for matching. Though matching on the cluster-level covariate leads to small biases, the consistently reduced estimated variances suggest matching on both cluster-level and unit-level covariates is beneficial and can achieve high coverage for the 95% confidence interval.

When estimating the ATT (Table 5), the sieve method still achieves the most accurate estimation and usually the highest coverage probability. When matching on both cluster and unit covariates, the estimated variance is lower than matching on the cluster-level covariate, and the coverages for the 95% confidence interval are as close to 95% as those when matching on the cluster-level covariate.

8 Discussion

In this article, we consider matching in a nonparametric way and discuss the matching estimators in clustered observational studies. Large sample properties for two estimands of interest are explored. For variance estimation, we propose a cluster-weighted bootstrap method that avoids the failure of the standard bootstrap and adjusts for the cluster effect. When the treatment assignment occurs at the cluster level, balancing on

Table 3: Comprehensive descriptions and specifications of the simulation study, including the data generating process and the matching schemes employed

Goal: evaluate the finite-sample performance of matching estimators in various matching schemes		
Simulated data: $\{Y_{ir}, A_r, X_{ir}, Z_r\}^n, R_i = 1, r = 1$		
Cluster	Number of clusters be $R = 50$	
	Balanced clustered data	Unbalanced clustered data
	Cluster size $n_r \in \{10, 50, 100\}$	Cluster size n_r from a discrete uniform distribution Unif{20,100} independently across clusters
Confounding variables	Unit-level confounding variables $X_{ir} \in \mathbb{R}^6$ are generated independently from a uniform distribution $\mathrm{Unif}(-1, 1)$	
	One cluster-level covariate $Z_r \sim \mathrm{Unif}(0, 1)$	
Nonlinear transformations	$X_1^* = g(X_1)g(X_2)$	$X_2^* = g(X_1) + g(X_2)$
	$X_3^* = 3\max(X_3, 0)$	$X_4^* = 3\max(X_4, 0)$
	$X_5^* = 3\max(X_5, 0)$	$X_6^* = 2X_6 - 1, Z_r^* = g(Z_r)$
	The nonlinear function $g(x) = 1 + [1 + \exp\{-20(x - 1/3)\}]^{-1}$, and the transformed variables are standardized with mean 0 and variance 1.	
Potential outcomes	The potential outcomes are generated from a mixed-effects model	
	$Y_i r(a) = \boldsymbol{X}^*\boldsymbol{\beta} + Z_r^* + \gamma + a\alpha_r + \varepsilon_{ir}$, where regression coefficients $\boldsymbol{\beta} = (1, 1, 1, 1, 1, 1)^{\mathrm{T}}$, true treatment effect $\gamma = 2$, random effect $\alpha_r \overset{\text{iid}}{\sim} N(0, 1)$, and random error $\varepsilon_{ir} \overset{\text{iid}}{\sim} N(0, 1)$	
Treatment	The treatment indicator A_r is assigned at cluster level following a Binomial distribution $\mathrm{Bin}\{\pi(Z_r)\}$, where $\pi(Z_r) = (1 + f(Z_r; 2, 4))/4$ and $f(z; 2, 4)$ is the probability density function for the Beta distribution.	
Matching estimators performed through the R package Matching, number of matches $M = 3$		
Methods to estimate nuisance functions	$\hat{\tau}_{\text{sieve}}$: the sieve method	
	$\hat{\tau}_{\text{ls}}$: linear b-spline method	
	$\hat{\tau}_{\text{rf}}$: regression forest	
Matching procedure	Procedure 1: match on the cluster-level covariate and estimate variance by the proposed cluster-weighted bootstrap method	
	Procedure 2: match on the cluster-level covariate and estimate variance by the standard weighted bootstrap method	
	Procedure 3: match on both cluster-level and unit-level covariates and estimate variance by the proposed cluster-weighted bootstrap method	
	Procedure 4: match on both cluster-level and unit-level covariates and estimate variance by the standard weighted bootstrap method	
Performance metrics	Bias, average variance, and coverage of 95% confidence intervals over 1000 simulated datasets with $B = 1000$ replicates for variance estimation	

cluster-level covariates is sufficient to remove confounding biases. However, we recommend that matching on both cluster-level and unit-level covariates is more efficient. The efficiency gain from including individual-level covariates may depend on whether they are correlated to the outcome and can provide additional information other than cluster-level covariates.

Another disadvantage of using cluster-level covariates alone is that when a unit is matched to a cluster with the opposite treatment, the distance to all units within that cluster is identical in the cluster-level covariate space. This can result in a large number of ties, which may vary based on the cluster size. Ties can be handled either deterministically or probabilistically. In our simulation study, we resolve ties by randomly selecting one sample. Alternatively, equal weights could be assigned to each observation within the cluster. However, both approaches may introduce significant bias or variance. To address this, we recommend incorporating unit-level covariates into the matching procedure to help break these ties. This further supports the case for matching on both cluster-level and unit-level covariates, rather than relying solely on cluster-level covariates.

We demonstrate the advantages of matching on both cluster-level and unit-level covariates in theory and simulation. Simulation results show reduced variance when matching on both cluster-level and unit-level covariates in various settings. Compared to the standard weighted bootstrap method widely applied in unstructured data, the proposed cluster-weighted bootstrap outperforms the standard bootstrap with a much higher coverage of 95% confidence interval. Three methods are utilized in constructing the bias-corrected estimators. In our result, the sieve method frequently achieves higher 95% confidence interval coverages and lower biases than the others. However, this may not be a general result, and we recommend evaluating the goodness of fit or conducting model diagnostics when selecting models for the nuisance functions in practice.

We apply the recommended matching strategy to study the ecological effects of different MPA policies on fish biodiversity. When matching on both MPA-level and site-level covariates, the sieve method with cluster-

Table 4: Summary of bias ($\times 10^3$), average variance ($\times 10^3$) and coverage (%) of 95% confidence intervals under different number of clusters R and different cluster size n_r when estimating the ATE based on 1,000 Monte Carlo samples; matching is performed by using cluster-level covariate only or using both cluster-level and unit-level covariates; variance is estimated by the cluster-weighted bootstrap or the standard weighted bootstrap method

(R, n_r)	(50, 10)			(50, 50)			(50, 100)			(50, [20, 100])		
	bias	**var**	**cvg**	**bias**	**var**	**cvg**	**bias**	**var**	**cvg**	**bias**	**var**	**cvg**
	Procedure 1: matching on cluster covariate only & cluster-weighted bootstrap											
$\hat{\tau}_{\text{sieve}}$	−15	215	96.1	17	170	95.2	−4	158	94.7	−15	172	95.4
$\hat{\tau}_{\text{ls}}$	33	177	93.4	56	148	90.8	28	136	92.3	13	145	92.8
$\hat{\tau}_{\text{rf}}$	−87	281	97.0	−35	143	91.8	−49	114	89.9	−69	138	92.7
	Procedure 2: matching on cluster covariate only & standard weighted bootstrap											
$\hat{\tau}_{\text{sieve}}$	−15	97	85.2	17	21	53.2	−4	10	40.6	−15	20	54.2
$\hat{\tau}_{\text{ls}}$	33	66	77.5	56	14	43.3	28	7	34.4	13	13	45.3
$\hat{\tau}_{\text{rf}}$	−87	127	88.4	−35	15	48.1	−49	6	33.3	−69	14	48.1
	Procedure 3: matching on both cluster and unit covariates & cluster-weighted bootstrap											
$\hat{\tau}_{\text{sieve}}$	−113	159	94.0	−59	119	92.6	−64	112	92.3	−76	129	93.9
$\hat{\tau}_{\text{ls}}$	33	125	91.8	65	104	91.0	47	100	90.2	39	110	91.1
$\hat{\tau}_{\text{rf}}$	−243	184	93.2	−111	99	89.4	−91	85	89.0	−128	103	90.3
	Procedure 4: matching on both cluster and unit covariates & standard weighted bootstrap											
$\hat{\tau}_{\text{sieve}}$	−113	73	80.0	−59	15	50.1	−64	8	35.1	−76	13	47.6
$\hat{\tau}_{\text{ls}}$	33	49	73.6	65	10	41.4	47	5	29.9	39	9	42.5
$\hat{\tau}_{\text{rf}}$	−243	87	79.4	−111	11	42.5	−91	5	29.1	−128	9	40.6

Table 5: Summary of bias (×10^3), average variance (×10^3) and coverage (%) of 95% confidence intervals under different number of clusters R and different cluster size n_r when estimating the ATT based on 1,000 Monte Carlo samples; matching is performed by using cluster-level covariate only or using both cluster-level and unit-level covariates; variance is estimated by the cluster-weighted bootstrap or the standard weighted bootstrap method

(R, n_r)	(50, 10)			(50, 50)			(50, 100)			(50, [20, 100])		
	bias	var	cvg	bias	var	cvg	bias	var	cvg	bias	var	cvg
	Procedure 1: matching on cluster covariate only and cluster-weighted bootstrap											
$\hat{\tau}_{\text{sieve}}$	50	305	96.8	17	231	96.3	11	226	95.9	15	248	97.0
$\hat{\tau}_{\text{ls}}$	33	227	93.8	37	189	92.7	33	183	91.9	30	203	93.8
$\hat{\tau}_{\text{rf}}$	−126	358	96.6	−101	169	91.7	−99	141	90.3	−108	178	92.9
	Procedure 2: matching on cluster covariate only and standard weighted bootstrap											
$\hat{\tau}_{\text{sieve}}$	50	138	89.4	17	29	52.2	11	15	42.3	15	29	53.5
$\hat{\tau}_{\text{ls}}$	33	89	80.1	37	19	42.7	33	10	33.7	30	19	44.8
$\hat{\tau}_{\text{rf}}$	−126	168	89.0	−101	20	46.5	−99	8	32.0	−108	19	47.7
	Procedure 3: matching on both cluster and unit covariates and cluster-weighted bootstrap											
$\hat{\tau}_{\text{sieve}}$	228	195	92.8	108	139	93.4	63	131	93.9	88	152	95.5
$\hat{\tau}_{\text{ls}}$	46	141	90.6	8	114	90.7	−7	110	89.1	−3	122	90.0
$\hat{\tau}_{\text{rf}}$	−243	216	97.2	−93	107	90.8	−89	92	88.9	−104	113	91.2
	Procedure 4: matching on both cluster and unit covariates and standard weighted bootstrap											
$\hat{\tau}_{\text{sieve}}$	228	92	79.7	108	18	50.9	63	9	39.8	88	16	48.6
$\hat{\tau}_{\text{ls}}$	46	58	74.5	8	12	41.7	−7	6	30.1	−3	10	41.8
$\hat{\tau}_{\text{rf}}$	−243	105	89.8	−93	13	44.5	−89	6	29.3	−104	10	42.3

weighted bootstrap successfully detects different ecological impacts between the MU and NT policies. Consistent with the results in the study by Gill et al. [6], we find that the NT policy positively affects the fish population compared to the MU policy. However, there remain several undiscussed aspects. First, the causal effect is estimated under strong assumptions. The SUTVA may be violated when there are multiple versions of MU and NT policies. Second, as discussed in the study by Gill et al. [6], conservation outcomes are closely related to the MPA management processes. Variability in MPA management effectiveness may cause different conservation impacts for the same MPA policy. In future studies, covariates that measure the adequacy and appropriateness of management should be included when comparing the relative causal effects of different MPA policies.

There are several issues that warrant future research. In cases with extremely high-dimensional X, matching on X even with bias correction may not be effective. One can use the double score matching idea [50] to conduct dimension reduction prior to matching. On the other hand, different weights should be assigned to different covariates during matching, and using 0–1 weights is equivalent to selecting a subset of variables. One criterion is to assign larger weights to covariates that are more strongly correlated with the outcome. The previous literature, such as Zhang et al. [48], suggests that removing instrumental variables and noise from the matching process can improve the finite sample efficiency of propensity score matching and double score matching estimators. Another consideration is to avoid excluding cluster-level covariates from the matching procedure to prevent confounding bias, as these covariates are more likely to be important confounders, particularly when the number of cluster-level covariates is limited. Variable selection should only be performed when there is a large number of cluster-level covariates. In such cases, we recommend selecting only the strongest predictors from the set of unit-level covariates.

Our current analysis and recommendation are based on the clustered data with a binary treatment assignment. This method could be generalized to a complicated data structure involving multilevel clusters with more than two treatment groups or continuous treatments. The benefits of matching estimators by different layers of observed covariates remain unknown. Nonparametric methods for variance estimation in multilevel observational studies would be complicated. Our proposed cluster-weighted bootstrap may be extended to these settings. As with most causal inference methods, we require the SUTVA and strong ignorability of treatment assumptions, which may be violated in some circumstances. It is of interest to study matching estimators in clustered data under relaxed assumptions and develop sensitivity analysis [51] to assess the robustness of the study conclusions to key assumptions.

Acknowledgments: The authors are grateful to the reviewers and editor for their valuable comments, which have significantly improved the manuscript.

Funding information: This study was supported by the NIH-NIEHS under grant number 1R01ES031651.

Author contributions: All authors accept responsibility for the entire content of this manuscript and have approved its submission. Can Cui developed the main methodology and implemented the simulation and application code. Yunshu Zhang prepared the manuscript, coordinated submissions and revisions, and contributed to writing with input from all co-authors. Shu Yang and Brian Reich contributed to the design of the estimators and the development of the main theoretical results. David Gill provided the dataset and contributed to the application component.

Conflict of interest: The authors declare no conflict of interest.

Ethical approval: This research does not involve human participants or animal subjects.

Data availability statement: The datasets generated and/or analyzed during the current study are available from the corresponding author upon reasonable request.

References

[1] Grorud-Colvert K, Sullivan-Stack J, Roberts C, Constant V, e Costa BH, Pike EP, et al. The MPA guide: A framework to achieve global goals for the ocean. Science. 2021;373(6560):eabf0861.
[2] UNEP-WCMC, IUCN, NGS. Protected Planet Live Report 2021. Cambridge UK; Gland, Switzerland; and Washington, D.C., USA; 2021.
[3] Ferraro PJ, Sanchirico JN, Smith MD. Causal inference in coupled human and natural systems. Proc Nat Acad Sci USA. 2019;116(12):5311–8.
[4] Kamat V. The ocean is our farm: Marine conservation, food insecurity, and social suffering in Southeastern Tanzania. Human Organization. 2014;73(3):289–98.
[5] Bennett NJ, Dearden P. Why local people do not support conservation: Community perceptions of marine protected area livelihood impacts, governance and management in Thailand. Marine Policy. 2014;44:107–16.
[6] Gill DA, Mascia MB, Ahmadia GN, Glew L, Lester SE, Barnes M, et al. Capacity shortfalls hinder the performance of marine protected areas globally. Nature. 2017;543(7647):665–9.
[7] Pynegar EL, Gibbons JM, Asquith NM, Jones JPG. What role should randomized control trials play in providing the evidence base for conservation? Oryx. 2021;55(2):235–44.
[8] Stuart EA. Matching methods for causal inference: A review and a look forward. Stat Sci Rev J Instit Math Stat. 2010;25(1):1–21.
[9] Cafri G, Wang W, Chan PH, Austin PC. A review and empirical comparison of causal inference methods for clustered observational data with application to the evaluation of the effectiveness of medical devices. Stat Meth Med Res. 2019;28(10-11):3142–62.
[10] Zubizarreta JR, Keele L. Optimal multilevel matching in clustered observational studies: a case study of the effectiveness of private schools under a large-scale voucher system. J Am Stat Assoc. 2017;212(518):547–60.
[11] Hong G, Raudenbush SW. Evaluating Kindergarten retention policy: a case study of causal inference for multilevel observational data. J Am Stat Assoc. 2006;101(475):901–10.
[12] Arpino B, Mealli F. The specification of the propensity score in multilevel observational studies. Comput Stat Data Anal. 2011;55(4):1770–80.
[13] Yang S. Propensity score weighting for causal inference with clustered data. J Causal Inference. 2018;6(2):20170027.
[14] Chang TH, Stuart EA. Propensity score methods for observational studies with clustered data: A review. Stat Med. 2022;41(18):3612–26.

[15] King G, Nielsen R. Why propensity scores should not be used for matching. Politic Anal. 2019;27(4):435–54.
[16] Abadie A, Imbens GW. Large sample properties of matching estimators for average treatment effects. Econometrica. 2006;74(1):235–67.
[17] Abadie A, Imbens GW. Bias-corrected matching estimators for average treatment effects. J Business Econ Stat. 2011;29(1):1–11.
[18] Davison AC, Hinkley DV. Bootstrap Methods and their Application. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge: Cambridge University Press; 1997.
[19] Otsu T, Rai Y. Bootstrap inference of matching estimators for average treatment effects. J Am Stat Assoc. 2017;112(520):1720–32.
[20] Abadie A, Imbens GW. On the failure of the bootstrap for matching estimators. Econometrica. 2008;76(6):1537–57.
[21] Ahmadia G, Glew L, Provost M, Gill D, Hidayat N, Mangubhai S, et al. Integrating impact evaluation in the design and implementation of monitoring marine protected areas. Phil Trans R Soc B. 2015;370:20140275.
[22] Brewer TD, Cinner JE, Green A, Pressey RL. Effects of human population density and proximity to markets on coral reef fishes vulnerable to extinction by fishing. Conservat Biol. 2013;27(3):443–52.
[23] Edgar GJ, Stuart-Smith RD, Willis TJ, Kininmonth S, Baker S, Barrett N, et al. Global conservation outcomes depend on marine protected areas with five key features. Nature 2014;506:216–220.
[24] Campbell SJ, Darling ES, Pardede S, Ahmadia G, Mangubhai S, Amkieltiela, et al. Fishing restrictions and remoteness deliver conservation outcomes for Indonesia's coral reef fisheries. Conservation Letters. 2020;13(2):e12698.
[25] Toth LT, van Woesik R, Murdoch TJT, Smith SR, Ogden J, Precht WF, et al. Do no-take reserves benefit Florida's corals? 14 years of change and stasis in the Florida Keys National Marine Sanctuary. Coral Reefs. 2014;33:565–77.
[26] Karr KA, Fujita R, Halpern BS, Kappel CV, Crowder L, Selkoe KA, et al. Thresholds in Caribbean coral reefs: implications for ecosystem-based fishery management. J Appl Ecol. 2015;52(2):402–12.
[27] Rubin DB. Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies. J Educ Psychol. 1974;66(5):688–701.
[28] Holland PW. Statistics and causal inference. J Am Stat Assoc. 1986;81(396):945–60.
[29] Splawa-Neyman J, Dabrowska DM, Speed TP. On the application of probability theory to agricultural experiments. Essay on principles. Section 9. Stat Sci. 1990;5(4):465–72.
[30] Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika. 1983;70(1):41–55.
[31] Stuart EA. Estimating causal effects using school-level data sets. Educ Res. 2007;36(4):187–98.
[32] VanderWeele TJ. Ignorability and stability assumptions in neighborhood effects research. Stat Med. 2008;27(11):1934–43.
[33] Li F, Zaslavsky AM, Landrum MB. Propensity score weighting with multilevel data. Stat Med. 2013;101(19):3373–87 doi: 10.48550/arXiv.1808.01647.
[34] He Z. Inverse conditional probability weighting with clustered data in causal inference; 2018.
[35] Lee Y, Nguyen TQ, Stuart EA. Partially pooled propensity score models for average treatment effect estimation with multilevel data. J R Stat Soc Ser A (Stat Soc). 2021;184(4):1578–98.
[36] Efron B. Bootstrap methods: another look at the Jackknife. Ann Stat. 1979;7(1):1–26.
[37] Huber M, Camponovo L, Bodory H, Lechner M. A wild bootstrap algorithm for propensity score matching estimators. Faculty of Economics and Social Sciences, University of Freiburg/Fribourg Switzerland; 2016. p. 470.
[38] Lin Z, Ding P, Han F. Estimation based on nearest neighbor matching: from density ratio to average treatment effect. Econometrica. 2023;91(6):2187–217.
[39] Ding P. A first course in causal inference. New York, NY, USA: CRC Press; 2024.
[40] Mammen E. Bootstrap and wild bootstrap for high dimensional linear models. An Stat. 1993;21(1):255–85.
[41] Cameron AC, Gelbach JB, Miller DL. Bootstrap-based improvements for inference with clustered errors. Rev Econ Stat. 2008;90(3):414–27.
[42] Cameron AC, Miller DL. A practitioner's guide to cluster-robust inference. J Human Resources. 2015;50(2):317–72.
[43] MacKinnon JG, Nielsen MØ, Webb MD. Fast and reliable jackknife and bootstrap methods for cluster-robust inference. J Appl Econ. 2023;38(5):671–94.
[44] Schoenberg IJ. Contributions to the problem of approximation of equidistant data by analytic functions: Part A. on the problem of smoothing or graduation. a first class of analytic approximation formulae. Quarter Appl Math. 1946;4(1):45–99.
[45] Chen X. Chapter 76 Large sample sieve estimation of semi-nonparametric models. In: Handbook of econometrics. vol. 6. Elsevier; 2007. p. 5549–632.
[46] Athey S, Tibshirani J, Wager S. Generalized random forests. Ann Stat. 2019;47(2):1148–78.
[47] Tibshirani R. Regression shrinkage and selection via the Lasso. J R Stat Soc Ser B (Methodological). 1996;58(1):267–88.
[48] Zhang Y, Yang S, Ye W, Faries DE, Lipkovich I, Kadziola Z. Practical recommendations on double score matching for estimating causal effects. Stat Med. 2022;41(8):1421–45.
[49] Zhao H, Yang S. Outcome-adjusted balance measure for generalized propensity score model selection. J Stat Plan Inference. 2022;221:188–200.
[50] Yang S, Zhang Y. Multiply robust matching estimators of average and quantile treatment effects. Scand J Stat. 2023;50(1):235–65.
[51] Yang S, Lok JJ. Sensitivity analysis for unmeasured confounding in coarse structural nested mean models. Stat Sin. 2018;28(4):1703–23.