The Application of Cluster-Based Heterogeneity Analysis on an Observational Study

A quantitative research into exploring cluster-based heterogeneous treatment effects using observational data

---

Author Name

Mike Lips

A paper presented for the Erasmus School of Economics

Supervisor

dr. Nuno Almeida Camacho

Second Assessor

dr. Michel van de Velden

Erasmus School of Economics

Data Science and Marketing Analytics

Master Thesis

October 18, 2023

Abstract

Cluster-based heterogeneity analysis in observational studies is an underdeveloped field of causal research. In this research, we attempt to extend the knowledge base of heterogeneity analysis by providing an empirics-first multi-step framework. This multi-step framework is applied to an observational study of 10,391 9th graders across 76 schools in the United States that were non-randomly assigned to growth mindset programs (Student Experience Research Network, 2023). The effectiveness of this growth mindset program (treatment) will be compared across different clusters containing different covariate distributions. The framework consists of propensity score estimation using Generalized Boosted Modelling (GBM), k-means clustering on institutional-level variables, propensity score matching, and G-computation. The common approach of causal research using observational studies is referred to as propensity score matching (PSM) in which the matched propensity scores are estimated using a logistic regression. We introduce the GBM model to solve the problem of correct functional form specification. This problem arises with the application of parametric models and comprises the dependence on correct functional form specification a priori. Moreover, we introduce data-driven clusters to compare the average treatment effects on the treated (ATTs). These ATTs are estimated using G-computation that utilizes marginal effects as an interaction effect between the treatment and all other covariates. We benchmark our multi-step framework to a reference approach in which the non-parametric GBM model is replaced by a parametric logistic model.

The results show positive and statistically significant treatment effects for each cluster. Furthermore, these ATTs differ up to around 50% across clusters, which could indicate heterogeneity. However, these differences are not statistically significant. At the same time, our results show that the GBM model does not show superior performance compared to the logistic model in terms of the average standardized absolute difference (ASAM), confidence interval, and standard error. We also compare our approach with the causal forest approach by Athey and Wager (2019). We find that our approach complements the approach of Athey and Wager (2019) by providing a framework to deeper understand the structure of the indicated heterogeneity. The causal forest algorithm stands out in indicating heterogeneity, whereas our approach allows fur further interpretation of this heterogeneity. Altogether, this research extends the knowledge base of heterogeneous treatment effects on observational data and provides an empirics-first framework for the exposure of cluster-based treatment heterogeneity.

Contents

1 Introduction **4**

2 Unhealthy relationship between theory and contribution **7**

3 Data **9**

4 Methodology **12**

4.1 Background on treatment effects 12

4.1.1 Different kinds of treatment effects 13

4.1.2 Underlying assumptions 13

4.1.3 The importance of propensity scores 14

4.2 The framework for our approach 15

4.2.1 The work of Athey & Wager (2019) 15

4.2.2 Limitations of Athey & Wager (2019) 15

4.2.3 Propensity score matching and its limitations 16

4.2.4 Identification issues of heterogeneity 17

4.2.4.1 General heterogeneity issues 17

4.2.4.2 Heterogeneity issues using PSM 17

4.2.5 Cluster-based solution 17

4.2.6 Visual representation of our approach 18

4.2.7 A comparison with other approaches 19

4.3 GBM for propensity scores 20

4.4 K-means clustering 22

4.5 Propensity score matching 23

4.5.1 The bridge to propensity score matching 23

4.5.2 How does matching work? 24

4.5.3 Balance requirements and implications 24

4.5.4 Application of matching 25

4.6 Estimating ATT 26

4.6.1 Functional form 26

4.6.2 Marginal effect estimation 26

4.6.3 Cluster-robust SEs 27

4.7 Comparison to parametric approach 28

4.8 Sensitivity analysis 28

4.9 Comparison with causal forest approach 29

5 Results **29**

5.1 Propensity score estimation using GBM 29

5.2 K-means clustering 31

5.2.1 Deciding on the number of clusters 31

5.2.2 Cluster composition . . . 32
5.2.3 Cluster interpretation . . . 33
5.3 Propensity score matching . . . 33
5.4 Estimating ATT . . . 36
5.5 Comparison to parametric approach . . . 38
5.6 Sensitivity analysis . . . 38
5.7 Causal forest as a benchmark . . . 39
5.7.1 Causal forest application . . . 39
5.7.2 Comparison with causal forest approach . . . 40
5.7.2.1 Subgroups . . . 40
5.7.2.2 Empirics-first . . . 40
5.7.2.3 Heterogeneity interpretations . . . 40

6 Conclusion/Discussion **41**

A Additional NSLM program details **50**
A.1 Selection of schools . . . 50
A.2 So, what does the program itself look like? . . . 50

B Details on the analysis **51**
B.1 Propensity score estimation . . . 51
B.2 K-means clustering . . . 51
B.3 Propensity score matching . . . 52
B.4 Estimating ATT . . . 53
B.5 Sensitivity analysis . . . 54

C Balance plots for all clusters **56**
C.1 Cluster 2 . . . 56
C.2 Cluster 3 . . . 57
C.3 Cluster 4 . . . 58
C.4 Cluster 5 . . . 59
C.5 Cluster 6 . . . 60

1 Introduction

"The gold standard for drawing inferences about the effect of a policy is a randomized controlled experiment. However, in many cases, experiments remain difficult or impossible to implement, for financial, political, or ethical reasons, or because the population of interest is too small" (Athey and Imbens, 2017). Furthermore, Rosenbaum (2002) suggests that observational studies are common in the majority of fields that aim to expose the treatment effects on people. These observational studies are an empiric investigation of treatments, policies, or exposures and their resulting effects, but it is different from an experiment due to the uncontrollable treatment assignment to subjects (Rosenbaum, 2002).

Exploring and measuring causal effects, a field which is widely sought after, but to this day, much remains unknown. In fact, companies tend to rely heavily on causal research for determining the impact of changes in policy, products, services, or features on key performance indicators (KPIs) (Qualtrics, 2023). However, we still lack a complete understanding of how to properly estimate cause and effect models.

Randomized controlled trials (RCTs) are the foundation of causal analysis; however, these experiments are rarely performed because of the disadvantage of being costly. Research into the application of causal analysis on observational studies has advanced over the past decade, but much more research is required to fully understand the playground. The mere estimation of treatment effects is a rather simple procedure when dealing with RCTs. In RCTs, the treatment assignment is random. This means that the group receiving the treatment is similar to the group not receiving the treatment, except for the treatment itself. Consequently, the difference in the outcomes between the two groups can be attributed to the treatment. Although, some researchers disagree with the validity of treatment effect attribution of randomization experiments. [1]

For observational studies, the treatment assignment is based on the observed covariates and non-random, which means that covariate distribution across the two groups is not similar. This implies that the difference in outcomes cannot be fully attributed to the treatment itself. The problem faced here seems to be a dealbreaker for the applicability of causal analysis. However, methods have been discovered that allow for the application of causal analysis by balancing the treatment- and control group as if they were formed during an RCT (Rosenbaum and Rubin, 1983).

This paper will further investigate the application of causal analysis on observational studies. More specifically, this paper will be complementary to the paper Athey and Wager (2019) in which the authors apply causal forest analysis to investigate heterogeneous treatment effects. The dataset used in the paper of Athey and Wager (2019) is the same dataset as used in this paper and originates from an RCT but has been changed into a semi-synthetic version to represent an observational study. This allows for extending the knowledge base about causal analysis using observational studies, and more specifically, cluster-based causal analysis.

Causal forest analysis, as demonstrated by Athey and Wager (2019), is applied to investigate heterogeneity.

[1] Pearl (2009) & Bollen and Pearl (2013) express concerns regarding RCTs as a cure-all approach for causality. Systematic errors in randomization as well as complex causal structures between the treatment and outcome lead to complexities that cannot be solved by randomization alone. This means that randomizing treatment assignment may not be enough to establish causality.

Similar to the random forest algorithm, the causal forest algorithm consists of an ensemble of decision trees. However, the purposes of the two algorithms diverge. Random forest combines decision trees to maximize prediction accuracy of an outcome variable. Meanwhile, in the causal forest algorithm, we maximize the difference in estimated treatment effects between subgroups (Athey et al., 2019). This enables the comparison of treatment effects across subgroups with distinct covariate distributions. In this regard, the causal forest algorithm offers a versatile means of modeling heterogeneous treatment effects.

Athey and Wager (2019) apply causal forest analysis to investigate heterogeneity across two groups of high- and low predicted out-of-bag conditional average treatment effects (CATEs). The out-of-bag CATEs represent average treatment effects, conditional on the characteristics of the subgroup, estimated on data points that are not used in the decision tree construction. Athey and Wager (2019) sought to determine if the difference in the average treatment effect (ATE) between the two subgroups is statistically significant.

In addition, Athey and Wager (2019) conducted an omnibus evaluation of the quality of the causal forest estimates via calibration. This assessment seeks to fit the CATE as a linear function of the out-of-bag estimates. By doing so, the calibration is able to determine the presence of heterogeneity. The group differences approach and calibration approach applied by Athey and Wager (2019) attempt to expose heterogeneity, but fail to do so.

By forming subgroups using low- and high predicted CATEs, the authors use a priori theorization about important differentiators for forming subgroups. This paper will take a different, more interpretable approach to potentially expose treatment heterogeneity across clusters and make the clusters tangible. More specifically, this paper will use a full data-driven approach for each step of the analytical framework. By doing so, this paper allows for the interpretability and comparability of data-driven clusters and their respective treatment effects. The question therefore is:

To what extent can heterogeneity across school-based clusters be exposed using a multi-step data-driven framework applied to an observational study?

This question will be answered by using a combination of multiple methods and leveraging the strengths of each method. As opposed to Athey and Wager (2019), this paper will prioritize the formation of data-driven clusters to expose heterogeneous treatment effects across these clusters. First, propensity scores will be estimated using Generalized Boosted Modelling (GBM), also referred to as Gradient Boosting. Propensity scores are balancing scores, which means that subjects with the same propensity score have similar covariate distributions and are therefore comparable. The problem with propensity score estimation using parametric models is that the correct functional form has to pre-specified, while this functional form is near impossible to determine (Drake, 1993). However, GBM does not require functional form assumptions for the estimation of propensity scores, unlike parametric models (McCaffrey et al., 2004; Stoltzfus, 2011). Furthermore, k-means clustering will be used to cluster the sample and to potentially expose heterogeneity across clusters (Likas et al., 2003; Kodinariya and Makwana, 2013).

Additionally, propensity scores will be matched within each cluster to obtain a balanced sample within each cluster that allows for the estimation of treatment effects. The way propensity score matching works is that the algorithm matches each treated observation with an untreated observation based on similar propensity scores (Morgan and Harding, 2006; Austin and Mamdani, 2006; Austin, 2011a, 2014b). This matching

procedure will be facilitated using k-nearest neighbor (KNN) and optimal matching, and the performance of these two matching methods will be compared. The treatment effect of each balanced cluster will be estimated using G-computation. The heterogeneity of these treatment effects across clusters will be checked using confidence intervals. Based on cluster-specific interpretations, potentially some conclusions can be drawn about the effectiveness of treatment effects on different subgroups. To benchmark our approach, we will replace the GBM model with a logistic model. Furthermore, we will compare our results to the results of Athey and Wager (2019) to discuss the differences in insights. Finally, the sensitivity of the treatment effect on hidden bias from unconfounded covariates will be investigated using sensitivity analysis using Rosenbaum's (2002) bounds.

First, we hypothesize that the GBM model outperforms the benchmark model in terms of ASAM, narrowness of the confidence interval, and standard error, which is in line with McCaffrey et al. (2004). Furthermore, we hypothesize that our approach will expose heterogeneity across clusters. If not, we expect to find cluster-specific interpretations that can be linked to the treatment effects. If neither of these hypotheses are true, this paper will at least provide a transparent framework and guideline for exposing heterogeneous treatment effects using observational studies in other fields. Furthermore, this paper will contribute to the broader understanding of causal inferences in non-random experiments.

The extent to which cluster-based heterogeneity in the causal inference field has been explored and applied is worrying. Unsurprisingly, the marketing field has not outpaced these advancements. Nevertheless, causal relationships and heterogeneity are of utmost importance in the marketing field to expose causal relationships and effect modifiers that would otherwise remain unknown. This paper will make further contributions to the still relatively unexplored cluster-based heterogeneity field. By extending the knowledge base of causal research using observational studies, we provide a solution for the problem of scarcity of cost-effective alternatives to RCTs (Duley et al., 2008; Djurisic et al., 2017).

The marketing contribution is providing marketers with a cost-efficient alternative for causal research, while also improving the knowledge base regarding this alternative (McDonald et al., 2011). Marketers are usually confronted with determining causal effects of marketing campaigns, and this paper gives an interpretable solution (Guadalupe, 2018). Additionally, heterogeneity in treatment effects across clusters can be crucial for tailoring marketing strategies (Punj and Stewart, 1983; Ascarza, 2018). Nevertheless, extant marketing literature overlooks this importance. Ascarza (2018) discusses customer heterogeneity regarding churn probabilities in anticipation of retention programs. However, the methodology used by the author focuses on individual-level heterogeneity and uses a churn probability threshold for participation in the retention program. The cluster-based heterogeneity gap within marketing literature will be filled by this paper.

Furthermore, most extant marketing literature emphasizes a theory-driven approach, while we prioritize a data-driven approach (Prasad, 2023). This allows for accurate specification of effect modifiers and can give a different perspective on how to approach marketing related data. Therefore, this paper will contribute to the marketing field by laying out clear instructions and providing other researchers with handles that can subsequently be used to adopt new insights on heterogeneity and effect modifiers.

The methodological contribution of this thesis is the novel approach of investigating cluster-based hetero-

geneous treatment effects. Athey and Wager (2019) made one of the first contributions to subgroup-based heterogeneity applied to observational studies, and expressed their concern for the limited research in this field. Heterogeneous treatment effects in observational studies have been researched before (Xie et al., 2012; Athey and Wager, 2019; Carnegie et al., 2019; Wendling et al., 2018; Athey and Imbens, 2017; Austin and Stuart, 2015). However, cluster-based heterogeneity research is still surprisingly limited.

This paper will contribute to the knowledge base of cluster-based treatment heterogeneity using observational data. Furthermore, this paper will extend the work of Athey and Wager (2019) by providing a multi-step, more interpretable framework. Athey and Wager (2019) applied causal forest analysis to the same dataset to investigate heterogeneous treatment effects. Causal forests allow for the direct and instant implementation of propensity score calculation, matching and treatment effect estimation. On the one hand, this makes the analysis more efficient as well as leveraging the strength of random forests for estimating the treatment effects. On the other hand, causal forests are a black-box method, which means that we cannot extract information about what exact steps are performed by the algorithm.

The framework outlined in this paper will separate the propensity score calculation, clustering, propensity score matching, and treatment effect estimation into different steps. This allows for the interpretation and understanding of each step, contributing to the knowledge base of cluster-based heterogeneity analysis on observational data. Furthermore, we emphasize the use of a data-driven methodology to let true effect modifiers come to light and to potentially expose cluster-based heterogeneity driven by data.

2 Unhealthy relationship between theory and contribution

As mentioned earlier, this paper will prioritize a data-driven approach. This chapter will further elaborate on the new perspectives that a data-driven approach can bring. Lately, discussion has been sparked about the theorization and institutionalization of journals and their respective papers with the deteriorating effect on knowledge production. Prasad (2023), former senior editor of multiple journals, expresses his worries about the emphasis on theoretical disguise in papers to be eligible for publication. The editorial decisions have, in the opinion of Prasad (2023), become rather a reproduction of a theoretical contribution-framework than judging the papers based on their knowledge contributions. Prasad (2023) claims that, especially in the discipline of management and organizational studies (MOS), journals have become overly obsessed with theoretical contributions. These theoretical contributions often seem to be either disguised as fancy, build upon theories that have not been empirically validated, or are just an end in and of itself instead of a means to an end (Prasad, 2023). This sometimes leads to the publication of papers that are just a fancy representation of an unvalidated theory and rejection of papers that would actually contribute to knowledge production.

This vision of impracticality and institutionalization can be extended to other fields as well, such as data science. The rise of the importance of data science in fields as marketing is getting undermined by the theory-driven research that is meant to enhance the pool of knowledge. Researchers tend to theorize beforehand which factors could potentially influence the outcome, based on other theories that are not even validated theories themselves (Prasad, 2023). This means that researchers will investigate superficial relationships while failing to uncover relationships that would contribute significantly to knowledge production. This urge to clothing the research in a theory-driven manner has limited to potential of research contributions.

Controversially, using a data-driven approach to empirically uncover unobvious relationships in the data could lead to contributions that are currently beyond imaginable. Researchers are forced to align their research with the substantial requirements of journals to make the papers eligible for publication. However, while increasing the odds of publication, researchers deteriorate the actual contribution of the paper. Schwarz (2023) confirms all these findings and explains that researchers buy into the debate about the same theories and reinvest their attention in the already existing theory. This stimulates the process of re-integration of the same theories. Schwarz (2023) refers to the abovementioned phenomenon as a scholarly Ponzi scheme. We tend to debate the same aspect with the same lens through different resources. We build upon what we already know and try to fill empty spaces that are forgotten or just overlooked for a while (Schwarz, 2023). Consequently, we become subject to this vicious circle, thinking that we contributed, while only replicating current states with different words.

Furthermore, John et al. (2012) emphasizes the prevalent concerns regarding questionable research practices (QRPs) in the field of psychology. One of the most prevalent QRPs is HARking, which means hypothesizing after the results are known (Kerr, 1998). This refers to the situation in which post hoc constructed hypotheses are reported is if they were constructed a priori. Kerr (1998) claims that HARking is widely practiced and a danger for the validity of research results.

This paper does not disapprove the use of theorization and the a priori hypotheses of effect modifiers. However, often we lack validated theories to determine, a priori, how treatment effects will vary across subpopulations. This implies that, in some cases, we should not pre-determine the importance of certain effect modifiers or assume drivers of heterogeneity. The approach that will be used in this paper will help in these kinds of situations, by obviating the need for theorizing a priori and by avoiding questionable practices such as HARking. This implies that we should not pre-determine the importance of certain effect modifiers or assume the drivers of heterogeneity, nor should we post hoc reconstruct the hypotheses. This paper will ensure that we do not theorize the contribution and prevent this research from being limited by what is already known and written. Therefore, no assumptions will be made about effect modifiers or drivers of heterogeneity.

In fact, we will follow an empirics-first approach, as suggested by Golder et al. (2023). In this empirics-first approach, we let models decide on the main drivers of treatment effects. We use it as a stepping-stone to theory without necessarily developing or testing a theory (Golder et al., 2023). Heterogeneity in this paper will not be influenced by the functional form dependencies of the parametric models. On the contrary, we follow an approach in which heterogeneity will be measured across clusters instead of within clusters. These clusters are data-driven, without human intervention or pre-specified assumptions. The heterogeneous treatment effects across clusters can be attributed to the cluster-characteristics, which can be shown by summarizing or plotting these clusters. The analysis consists of multiple steps and models, which are all data-driven without theorizing the outcome. Conclusions drawn at the end of the analysis are therefore purely data driven. This will potentially expose the 'true' effect modifiers and drivers of heterogeneity without theorization of these influential factors a priori.

3 Data

The data that will be used in this research is data from The National Study of Learning Mindsets (NSLM).[2] The NSLM study is a randomized controlled trial (RCT) that has an online growth mindset program as treatment which is randomly assigned to 10,391 9th graders across 76 schools in the U.S. In the rest of this paper, whenever we refer to "treatment", we refer to 9th graders who were selected to follow the growth mindset program. Whenever we refer to "control", we refer to the 9th graders that were not selected, and thus did not receive the growth mindset program.

So, within each selected school, 9th graders are randomly selected to either the treatment- or control group. The online growth mindset program teaches students that intelligence can be developed. This is in contrast to a fixed mindset, which is a belief that intelligence is a genetic, fixed trait (Student Experience Research Network, 2023). 139 U.S. high schools were selected using sophisticated sampling techniques. Out of these 139 schools, 76 schools agreed to participate, of which 65 schools provided all the requested records. All requested records include both survey data and administrative records for students, whereas the other 11 schools only provided survey data. The goal of this sample of 76 schools is to be a nationally representative probability sample of regular public U.S. high schools. Since the school selection was random as well, this original NSLM experiment allows for generalizations, an implication that most experiments lack. More information on the NSLM study can be found in Appendix A.

The randomness of the school selection prevents the effectiveness of the intervention being subject to school characteristics that are more likely to participate in such a study, and therefore preventing biased results. The participating high schools are a mix of rural, urban, and suburban schools. The driving force behind performing this experiment is the potential value for mindset science of exposing heterogeneity effects of social psychological interventions in education. This dataset is supposed to expose the link between academic performance and a growth mindset. The use of an RCT allows for making causal inferences and exposes the exact heterogeneity that the mindset science is aiming to understand. By doing so, more understanding arises about which kind of students in what environment benefit the most from a growth mindset program. So, the question that the NSLM experiment is aiming to answer is: for whom and under what conditions can a growth mindset be effective for 9th graders?

This dataset is very privacy sensitive, which is why a semi-synthetic version of this dataset has been generated to facilitate the 2018 Atlantic Causal Inference Conference. The number of observations within this semi-synthetic version of the NSLM dataset is 10,391. The Atlantic Causal Inference Conference is a hotspot of the causal inference field in which eight great research teams attempt to make causal inferences using appropriate models (a.o. Athey and Wager & Carnegie et al.). The dataset that will be used in this paper has been modified in such a way that the random assignment within schools is not guaranteed anymore with the introduction of additional confounding (Carvalho et al., 2019). This turns the dataset into an observational study rather than an RCT. Evidence for this modification is shown in Figure 1.

[2]The National Study of Learning Mindsets is the "largest-ever randomized controlled trial of an online growth mindset program in the United States in K-12 settings, and uses a nationally representative probability sample" (Network, 2023).

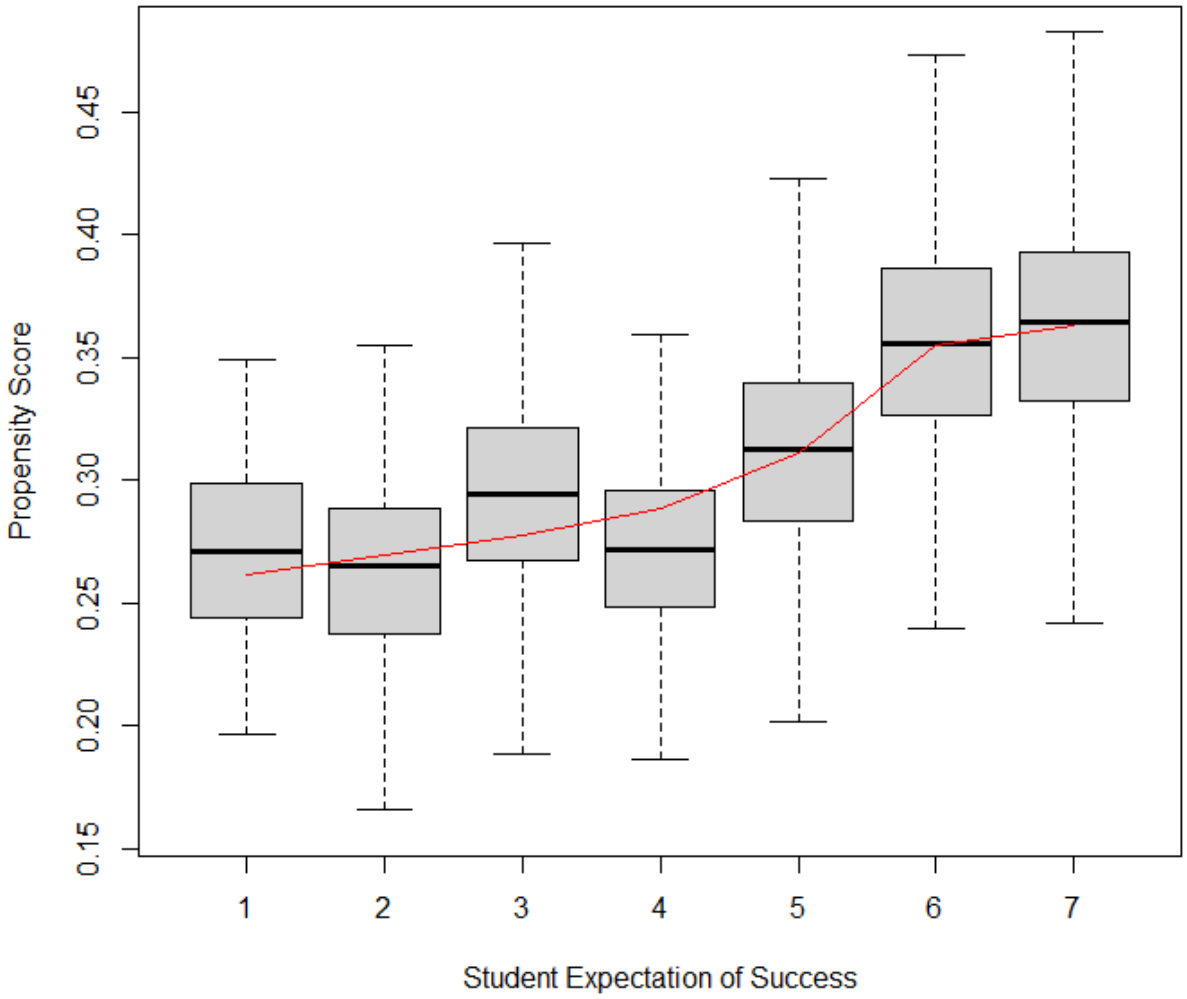


Figure 1: Proof of non-random treatment assignment

Figure 1 shows that students with a higher expectation of success have higher propensity scores as well. This means that students with higher expectations of success are more likely to be treated, which confirms that the treatment assignment is non-random. This transformation into an observational study requires more sophisticated models to identify causal effects. Thereby, we contribute to the broader understanding of the field of applied machine learning for causal inferences. To make causal inferences with this dataset, the control- and treatment group need to be balanced in terms of covariates. This means that the probability of treatment assignment needs to be the same across control- and treatment groups. For this, we calculate the propensity scores using Generalized Boosted Modelling (GBM) and use propensity score matching to ensure balance. One-hot encoding will be used from the next step onwards, since it can be counterproductive for the GBM algorithm (Ridgeway et al., 2022). Therefore, the only data preparation for propensity score estimation is transforming the categorical variables into factors and transforming the treatment variable into an integer, since this is necessary for the GBM algorithm.

Furthermore, cluster-analysis using k-means clustering will be performed to measure heterogeneity in the treatment effect across clusters of schools. More specifically, we generate school-level data-driven clusters and estimate the treatment effects individually for each of these clusters. The treatment effect comparisons between these clusters allow for the measurement of heterogeneity by comparing the confidence intervals of the treatment effect estimates.

We do not have the background information about what category of the student-level variables each value represents. Therefore, no student-level specific conclusions can be drawn. For instance, variable `C1`, representing the students' race/ethnicity, contains 14 different races/ethnicities. Due to the privacy-limiting transformation of the data, we do not know which value stands for what race. For school-level variables (`XC, X1, X2, X3, X4`, and `X5`), higher values indicate higher scores for that specific variable. To clarify, `X1` represents the school-level mean of students' fixed mindsets. Higher values of `X1` for subject i imply that the school of subject i has a higher mean of students' fixed mindsets. Cluster analyses will be performed based solely on school-level characteristics, meaning that cluster-specific interpretations can still be performed. Consequently, even though this dataset does not allow for student-level conclusions/interpretations, it does allow for school-level conclusions/interpretations. Additionally, the analytical framework that will be outlined in this paper can be applied to the real dataset, allowing for clear understanding of the driving factors of mindset growth effectiveness.

The variables that are included in the semi-synthetic version of the NSLM data are shown in Table 1.

Table 1: Data description

Variable	**Description**	**Type**
schoolID	Unique ID of student's school	Categorical (76)
Z	Treatment; growth mindset intervention	Binary
Y	Post-treatment outcome, a measurement of achievement	Continuous
S3	Students' self-reported expectations for success in the future	Categorical (7)
C1	Student race/ethnicity	Categorical (15)
C2	Student identified gender	Binary
C3	Student first-generation status; first person in family to go to college	Binary
XC	School-level variable of the urbanicity of the school (rural, countryside, town, suburban, city)	Categorical (5)
X1	School-level variable of the mean of students' fixed mindset (reported prior to treatment assignment)	Continuous
X2	School achievement level, as measured by test scores and college preparation for the previous 4 cohorts of students	Continuous
X3	School racial/ethnic minority composition (% black, latino, or native/american)	Continuous
X4	School poverty concentration (% of students who originate from families with incomes below federal poverty line	Continuous
X5	School size	Continuous

Table 2 shows the descriptive statistics for the continuous variables in the dataset (`Y, X1, X2, X3, X4, X5`). Table 2 also makes a distinction between treated and untreated observations and makes the statistics of each of these variables comparable between the two groups. As we can already see, the standard deviation for each school-level variable is close to one and the mean is close to zero. This means that the continuous variables in this dataset have been standardized. Standardization is most likely applied since the scales of the school-level covariates differ and thus this enhances the interpretability and comparability (Kim and Ferree, 1981). Standardization implies that for each data point in each variable, the sample mean is substracted from the original value of the data point and divided by the sample standard variation of that variable (Milligan

and Cooper, 1988).

On a similar note, the 2018 Atlantic Causal Inference Conference has perturbed the covariates from the original NSLM dataset. They have added random noise to the continuous variables that are sampled from multivariate normal distributions, while preserving the covariance structure. So, the variables are standardized and extra noise has been added, while maintaining the covariance structure and statistical patterns.

Table 2: Descriptive statistics of the continuous variables.

	Treatment (N=3384)				**Control (N=7007)**			
	Mean	Min	Max	SD	Mean	Min	Max	SD
Measurement of achievement (Y)	0.11	-1.88	2.20	0.64	-0.20	-2.10	1.86	0.62
School-level fixed mindset (X1)	-0.11	-3.09	2.84	0.97	-0.01	-3.09	2.84	1.00
School achievement level (X2)	0.09	-3.35	2.17	0.93	0.04	-3.35	2.17	0.93
School ethnic minority composition (X3)	-0.09	-1.58	2.36	0.96	-0.09	-1.58	2.36	0.97
School poverty concentration (X4)	-0.06	-1.93	2.82	0.97	-0.04	-1.93	2.82	0.97
School size (X5)	0.02	-1.81	1.89	1.02	-0.05	-1.81	1.89	1.01

4 Methodology

4.1 Background on treatment effects

In general, in the field of causal inferences, we want to estimate the average treatment effect. However, while researchers observe the outcome of a unit that received the treatment, researchers do not observe the outcome of the treatment that a unit did not receive (Athey and Imbens, 2017). This inability to observe the parallel outcome for either the treated or untreated is the reason why causal effects cannot directly be observed. Holland (1986) referred to this phenomenon as the fundamental problem of causal inference. Therefore, causal effects have to be estimated by comparing two groups; the treatment- and control group. However, this depends on the assumption of unconfoundedness which will be discussed further down the road.

In this research, Y is the continuous outcome variable, Z is the binary treatment variable and X is a vector of observed baseline covariates. Regarding the treatment effect, $Z = 1$ implies that the participant received the treatment (treatment group) and $Z = 0$ implies that the participant did not receive the treatment (control group). First of all, under the potential outcomes framework described by Imbens (2004), each subject has a pair of potential outcomes that can be described as: $Y_i^{(1)}$ and $Y_i^{(0)}$ (Austin, 2011b). $Y_i^{(0)}$ is the outcome under control and $Y_i^{(1)}$ is the outcome under treatment. The problem is, however, that only one of the potential outcomes is observed:

$$Y_i = Y_i(Z_i) = \begin{cases} Y_i^{(0)} & \text{if} \quad Z_i = 0 \\ Y_i^{(1)} & \text{if} \quad Z_i = 1 \end{cases} \tag{1}$$

4.1.1 Different kinds of treatment effects

Regarding causality, the treatment effect for an individual i would intuitively be $Y_i^{(1)} - Y_i^{(0)}$. However, as mentioned before, we never observe both $Y_i^{(1)}$ and $Y_i^{(0)}$ for the same person (Imbens and Angrist, 1994). This means that we rely on comparisons between individuals that got the treatment and people who did not, to estimate the average treatment effect (ATE). Imbens (2004) defines this estimation as:

$$\text{ATE} = \mathbb{E}[Y^{(1)} - Y^{(0)}] \tag{2}$$

However, this average treatment effect would imply the treatment effect for the whole population, preventing further distinction into heterogeneous treatment effects. Therefore, we aim to measure the average treatment effect conditionally on the sample distribution of covariates (Imbens, 2004). The conditional average treatment effect is widely defined in the literature as:

$$\text{CATE} = \mathbb{E}[Y^{(1)} - Y^{(0)} \mid X = x] \tag{3}$$

Using the CATE, one can estimate the average treatment effect for a certain set of covariates. For instance, one could investigate the average treatment effect for a certain race/ethnicity in the dataset. This treatment effect could also be measured specifically for people that received the treatment. In this case we compare the outcome and counterfactual outcome for people that received the treatment. This is referred to as the treatment effect on the treated (ATT). This research will specifically focus on the ATT to measure the treatment effect for people that were exposed to the growth mindset program. The notation of the ATT is:

$$\text{ATT} = \mathbb{E}[Y^{(1)} - Y^{(0)} \mid Z = 1] \tag{4}$$

4.1.2 Underlying assumptions

In randomized trials, X is known to include all the covariates that are used for treatment assignment and are related to the outcome (Rosenbaum and Rubin, 1983). This implies that for randomized trials, the treatment assignment Z is independent of the potential outcomes Y conditional on X (Wager and Athey, 2018). This means that all factors that are correlated with both potential outcomes and the treatment assignment are expected to be observed, implying that the treatment is as good as randomly assigned. This is also called the unconfoundedness assumption and can be notated as:

$$\left\{Y_i^{(1)}, Y_i^{(0)}\right\} \perp Z_i \mid X_i \tag{5}$$

Rosenbaum and Rubin (1983) suggest that conditional independence with observational data can also be obtained using the so called strong ignorability assumption. This assumption is often used interchangeably with the unconfoundedness assumption. It assumes that the treatment assignment is independent of potential outcomes, conditional on the probability of treatment, which is referred to as the propensity score. Instead of assuming that the treatment assignment is independent of outcomes conditional on covariates, we assume that the treatment assignment is independent conditional on the propensity score (Xie et al., 2012). The

propensity score is defined as the conditional probability of receiving the treatment (Austin and Stuart, 2015). This means that, conditional on the propensity score, the treatment assignment is independent of measured covariates (Athey and Imbens, 2017). The notation of the propensity score is:

$$e(x) = \text{PR}(Z_i = 1 \,|\, X_i = x) = \mathbb{E}[Z_i \,|\, X_i = x] \tag{6}$$

The strong ignorability assumption can therefore be defined as:

$$\left\{Y_i^{(1)}, Y_i^{(0)}\right\} \perp Z_i \,|\, e(X_i) \tag{7}$$

Rosenbaum and Rubin (1983) refer to the propensity score as a balancing score since units with the same propensity score will have similar covariate distributions. This shows that it is sufficient to condition on the propensity score $e(X_i)$ for causal inferences (Rosenbaum and Rubin, 1983). Even though the data used in this paper is derived from a randomized trial, the data has been slightly perturbed, and the outcome variable has been simulated. This means that this data should be considered as an observational study rather than an RCT. Furthermore, even if the data originates from a proper RCT, conditioning on propensity scores still enhances the reliability of the results (Hirano et al., 2003; McCaffrey et al., 2004). Rosenbaum (1987) claims that weighting by true propensity scores only compensates for systematic differences whereas weighting by estimated propensity scores corrects for both chance imbalances and systematic differences. Another assumption of this approach is the overlap assumption. The overlap assumption assumes that there is no observation that is deterministically assigned to the treatment or control group. This guarantees that for a large enough sample, enough treatment and control units will be presence near any test point x for local methods to work (Wager and Athey, 2018). This assumption can be tested by plotting the distribution of propensity scores for both the treatment and control group. The overlap assumption is denoted as:

$$0 < \mathbb{E}[Z_i = 1 \,|\, X_i = x] \leq 1 \tag{8}$$

4.1.3 The importance of propensity scores

Rosenbaum and Rubin (1983) lay the foundation of causal effects in observational studies using propensity scores. The authors claim that the disadvantage of non-randomized treatment assignment compared to randomized treatment assignment is that the propensity score function is unknown. In a two-armed randomized trial, the probability of a treatment assignment is aimed to be 0.5 which means that every subject has the same probability of being assigned to the treatment group. This means that the treatment and control group do not differ in any way other than the treatment assignment. This comparability allows for causal inferences. In observational studies, the assignment to the treatment group is often not based on a random assignment. This implies that the distribution of propensity scores of these studies does not follow a normal distribution with its peak at the 0.5 score. The absence of a normal distribution of treatment probabilities means that the treatment assignment is conditional on covariates. Consequently, the subjects in the treatment and control group differ in their characteristics which prevents us from comparing these groups to make causal inferences.

The solution to this problem, as mentioned earlier, is introduced by Rosenbaum and Rubin (1983). The authors claim that computation and matching of propensity scores can allow the distribution of covariates

to be the same across the control- and treatment group. The most used approach for calculating propensity scores is to estimate the treatment probability using logistic regression (Mansournia and Altman, 2016). In this logistic regression, the treatment variable will be used as outcome variable and the covariates will be used as predictors. The reason behind the application of logistic regression is that it is constrained to generate probabilities in the range of 0 to 1 (Westreich et al., 2010). Furthermore, logistic regressions are reliable and interpretable while not overcomplicating the computation process, which makes it easy to use. Westreich et al. (2010) claim that the disadvantage of using logistic regression is the underlying assumptions for using a logistic regression. The authors suggest using machine learning techniques with less structural assumptions to estimate the propensity scores. The implications of the difference between the application of parametric model and a machine learning model will be further elaborated on.

4.2 The framework for our approach

4.2.1 The work of Athey & Wager (2019)

Athey and Wager (2019) use each school as a cluster and use these clusters as grouped input into the causal forest analysis. By doing so, their conclusions can be generalized outside of the sample, since they robustly account for the sampling variability across these schools by using each school as a clustered input. Then, instead of just drawing k random samples as is common with the random forest algorithm, the random forest draws a subsample of clusters and then draws k samples from each of these clusters.

Consequently, Athey and Wager (2019) predict the treatment effect for each individual observations using causal forests which returns both an average treatment effect (ATE) for the whole sample, and individual conditional average treatment effects (CATEs) for each observation. To test heterogeneity, the authors split the observations into two subgroups, one with CATEs below the median and one with CATEs above the median. Subsequently, the ATEs for both subgroups can be estimated which in turn allows for testing whether these ATEs differ significantly.

Additionally, Athey and Wager (2019) use another tool to test heterogeneity, which is an omnibus evaluation of the quality of the causal forest estimates via calibration. This test calibration computes the best linear fit of the target estimand using the prediction on out-of-bag data as well as the mean forest prediction. This "best linear predictor" method seeks to fit the CATE as a linear function of the out-of-bag estimates (Athey and Wager, 2019). The test calibration returns two coefficients; the 'mean prediction' and the 'differential prediction'. The mean prediction refers to correctness of the average treatment effect estimate, whereas the differential prediction refers to the presence of heterogeneity. If the coefficient of the differential prediction is positive and significant, then the null hypothesis of no heterogeneity can be rejected.

4.2.2 Limitations of Athey & Wager (2019)

Athey and Wager (2019) extend the knowledge base of cluster-based heterogeneity research. However, their approach lacks interpretability and a transparent structure; only one model is applied in which all the steps are performed simultaneously. The causal forest approach of Athey and Wager (2019) only differentiates subpopulations by using a 50-50 split based on high- and low predicted out-of-bag CATEs.

Moreover, this approach uses a priori theorization on sample division. In fact, the causal forest approach can only really indicate the strength of heterogeneity without allowing for interpreting the differences between the two groups. Therefore, we propose a empirics-first multi-step framework to expose heterogeneity across clusters in which propensity score matching is the power source in preparing this observational study for causal research.

4.2.3 Propensity score matching and its limitations

PSM is an alternative solution to the application of causal forests to estimate treatment effects and expose potential heterogeneity. The idea behind PSM is that propensity scores are estimated for each observation in both the treatment- and control group. Subsequently, each treated observation is matched with an untreated observation based on similar propensity scores (Morgan and Harding, 2006). Consequently, a matched dataset is constructed in which each treated observation has formed a matched pair with an untreated observation. By doing so, the treatment- and control group are balanced in terms of covariate distribution since similar propensity scores represent similar covariate distributions. The advantage of this process is that this former observational study is artificially transformed into a randomized study. This means that we can now perform causal research on this dataset.

However, regular PSM comes with its risks and limitations. First of all, regular PSM uses a logistic model to estimate the propensity scores. The use of a parametric model like a logistic model requires pre-specification of its functional form. However, neither the functional form nor the influential covariates for treatment selection are known in advance, which makes this process prone to mistakes and misspecifications (McCaffrey et al., 2004). Additionally, a logistic model has underlying assumptions that need to be verified which makes its application even less flexible. The solution to this problem is to separate use a non-parametric model for the propensity score estimation part. In fact, a non-parametric approach without functional form assumptions can potentially improve the obtained balance after matching and decrease the bias (Lee et al., 2010). The algorithm that will be used as a replacement for a logistic model is the GBM algorithm.

However, regular PSM comes with its risks and limitations. First of all, regular PSM uses a logistic model to estimate the propensity scores. The use of a parametric model like a logistic model requires pre-specification of its functional form. However, neither the functional form nor the influential covariates for treatment selection are known in advance, which makes this process prone to mistakes and misspecifications (McCaffrey et al., 2004). Additionally, a logistic model has underlying assumptions that need to be verified which makes its application even less flexible. The solution to this problem is to separate the propensity score estimation and propensity score matching procedure. Furthermore, a non-parametric approach without functional form assumptions can potentially improve the obtained balance after matching and decrease the bias (Lee et al., 2010). The algorithm that will be used as a replacement for a logistic model is the GBM algorithm.

4.2.4 Identification issues of heterogeneity

4.2.4.1 General heterogeneity issues

Individuals do not only differ in their covariate distribution, and thus propensity scores, but also differ in their reaction to treatment. Elwert and Winship (2010) express how, while controlling for interaction effects, heterogeneity has not been studied and understood well enough. Additionally, extant literature acknowledges the importance of heterogeneity across subpopulations as it increases the effectiveness of treatment assignment by maximizing average outcomes and balancing competing objectives (Holland, 1986; Winship and Morgan, 1999; Heckman, 2005; Rubin, 1974). Studies tend to investigate and report the main treatment effect, but not how this treatment effect differs across subpopulations. According to Xie et al. (2012) one of the reasons for this is the lack of accessible, ready-to-use statistical methods. With the rise of data accessibility and the scarcity of RCTs, we require frameworks that allow for estimating heterogeneity across subpopulations that can be applied on observational data.

Many researchers have tried to expose heterogeneity through statistical approaches by incorporating interactions between treatments and individual-level variables, such as Bayesian analysis (Gelman et al., 1995), meta-analysis (Hedges, 1982), and the latent class model (Heckman and Singer, 1984). Xie et al. (2012) emphasizes the importance of the exposure of treatment heterogeneity and estimates heterogeneous treatment effects as a function of treatment propensity. By doing so, the heterogeneity across different strata of propensity scores can be exposed. Consequently, this allows for making inferences about heterogeneous treatment effects across these propensity-score-differing strata. However, the limitation is the assumption of homogeneity within these strata as well as the dependence on specifying a global functional form.

4.2.4.2 Heterogeneity issues using PSM

Besides the limitation of poor model selection using PSM, the regular PSM approach does also not deal properly with heterogeneity. Even though treatment heterogeneity could be indicated using PSM, this approach does not allow for interpretability and comparability between subgroups. The absence of distinct subgroups makes it difficult to measure, as well as interpret heterogeneous treatment effects. Therefore, an alternative solution is desired that splits the dataset into data-driven subgroups with varying treatment effects. This allows for the interpretability of subgroup-specific treatment effects as well as the comparability between these subgroups which potentially exposes treatment heterogeneity.

4.2.5 Cluster-based solution

The discussion above shows the issues and limitations of current approaches in extant literature in both model specification and exposing heterogeneity. The solution we propose is a cluster-based solution, facilitated by the k-means clustering algorithm. More specifically, by clustering the data on school-level covariates, we generate clusters that are similar in their school-level covariates. So, our approach does not require functional form determination since we apply GBM. Furthermore, we make a clear distinction between subgroups using a data-driven clustering approach. Moreover, by applying propensity score matching after clustering, we obtain similarity across observations within clusters as well as across the treatment- and control group within these clusters. We use G-computation to estimate the treatment effects including their confidence interval. These

cluster-specific estimations allow us to determine the strength of the treatment effect for each specific cluster as well as comparing the treatment effects across clusters. This can potentially expose heterogeneity, and if not, still provide cluster-specific interpretations linked to differing treatment effects.

4.2.6 Visual representation of our approach

Figure 2 shows a visual representation of the framework of our approach. We can observe that the core of the analysis consists of four different layers: (1) propensity score estimation, (2) clustering, (3) propensity score matching, and (4) estimating treatment effects. We intentionally use this sequence. The reason for this is that we want to start with estimating the propensity scores using the GBM model on the full dataset. Furthermore, we want to cluster our data before making matched pairs since we do not want matched pairs across clusters. Then we use the propensity scores and data-driven clusters and apply propensity score matching within each cluster. Subsequently, we estimate the ATTs using G-computation. Figure 2 solely represents the core analysis of this paper without including the parametric benchmark model, the causal forest benchmark model, and sensitivity analysis. Figure 2 also includes the actions taken in each layer and the outcome of each layer that is further used in subsequent layers through data flows. Overall, Figure 2 provides a clear overview of the main analysis and describes how each layer contributes to the final estimation of cluster-specific treatment effects.

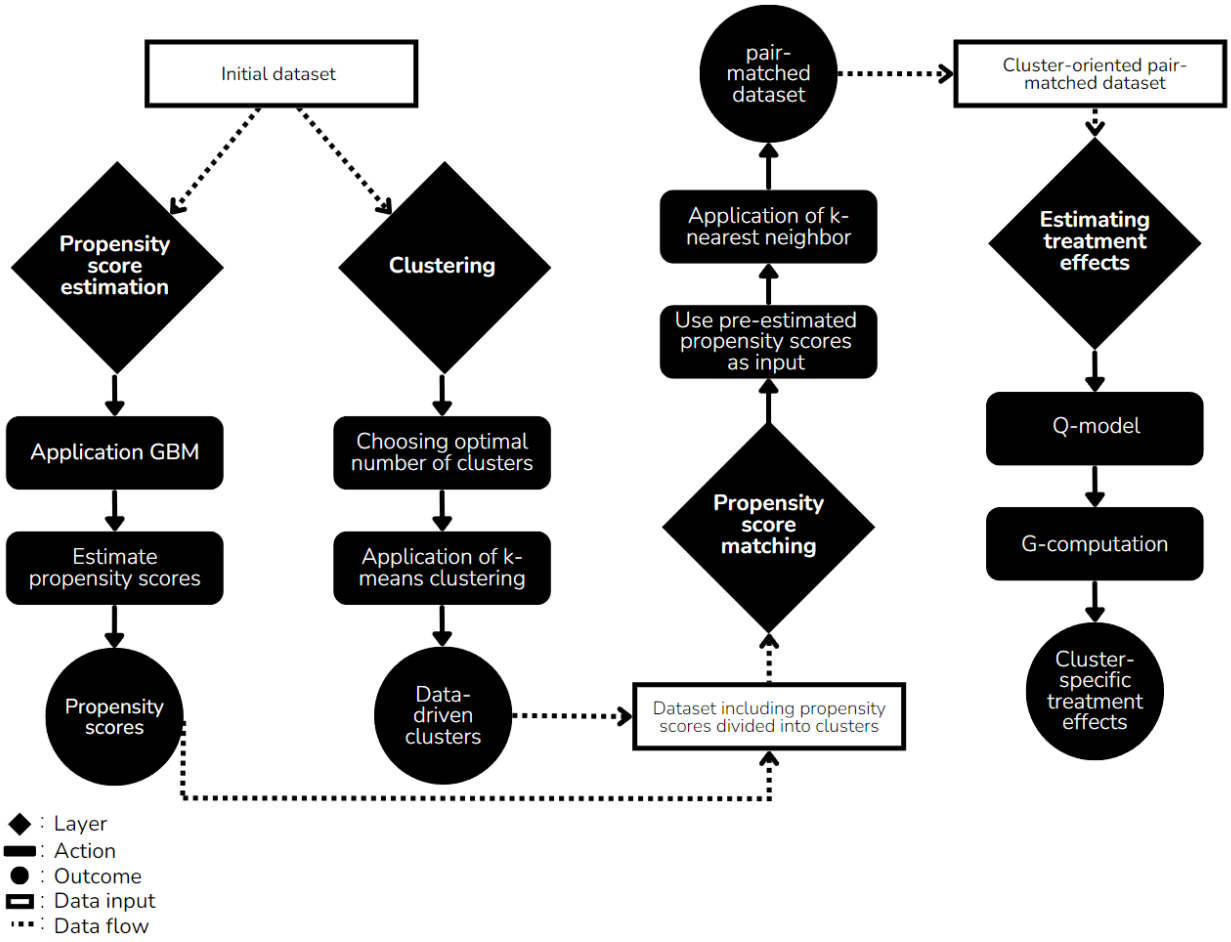


Figure 2: Methodological framework

4.2.7 A comparison with other approaches

Table 3 shows the comparison between our approach, the regular PSM approach, and the approach of Athey and Wager (2019). We compare the three approaches on their differences in interpretability, heterogeneity measurement, analysis flow, and flexiblity. Table 3 shows how our approach differs from the other two approaches and how it is superior in reaching its goals; providing a transparent, empirics-first framework to expose cluster-based heterogeneity.

Table 3: Comparison with other approaches

	Our multi-step framework	**Regular PSM**	**Athey and Wager (2019)**
Interpretability	Each subgroup is highly interpretable and has its own estimated treatment effect.	Besides the overall treatment effect, the interpretation is limited.	One algorithm is used for the whole analysis which limits the interpretability + interpretation of subgroups is not meaningful.
Heterogeneity	Across data-driven clusters based on school-level variables	Absence of distinct subgroups to measure heterogeneous treatment effects.	Across high-and low out-of-bag predicted CATEs.
Flow	The outcome of each layer in the framework can be inspected individually and models within each layer can be replaced (see Figure 2).	Each step can be inspected individually, but models are pre-set.	Only the final outcome can be interpreted, which is the treatment effect and the strength of heterogeneity.
Flexibility	The functional form of propensity score estimation does not have to be pre-specified, and the GBM model can deal with complicated forms + no model assumptions have to be validated.	The functional form must be pre-specified which makes it harder to get unbiased estimates + multiple assumptions need to be validated (assumptions on logistic regression).	The functional form does not have to be pre-specified, and the Causal Forest can deal with complicated forms + no model assumptions have to be validated.

In comparison with Athey and Wager (2019), we provide a complementary framework that enhances the interpretability of each layer within the analysis. This paper does not intend to outperform the analysis of Athey and Wager (2019) in terms of narrowness of the confidence intervals of treatment effects. As a matter of fact, the construction of our analysis differs significantly from Athey and Wager (2019) as it divides the data into subgroups in the second layer of the analysis. The treatment effects are specific for each subgroup, which makes these two approaches incomparable in terms of performance. However, we intend to compare the differences in insights and output and discuss these differences. Consequently, we want to extend the

methodological knowledge base regarding heterogeneous treatment effects.

Furthermore, we will benchmark our results using the a parametric model in the first layer. This parametric approach allows for the comparison of performance. In line with McCaffrey et al. (2004), we use the ASAM, confidence intervals of the treatment effects, and standard errors to compare our framework to the parametric approach. So, the difference in performance comes down to the difference in propensity score estimation techniques, which is the difference between GBM and logit.

4.3 GBM for propensity scores

We discussed the limitations of regular PSM and the need for a non-parametric approach. Clearly, more adaptive and flexible models are required to properly estimate propensity scores to get unbiased results. Even though this has received little attention in the literature, we build forward on the foundation set by McCaffrey et al. (2004) by re-introducing generalized boosted models (GBM), a form of Gradient Boosting that utilizes decision trees as base learners, to the causal inference field. GBM is a non-parametric, data adaptive modeling algorithm which can be referred to as general and automated. GBM is capable of estimating nonlinear relationships between a great number of covariates and the treatment variable to accurately estimate treatment assignment probabilities (McCaffrey et al., 2004).

The difference between parametric models and non-parametric models is that the functional form of the relationship between the predictors and the response variable does not have to be pre-specified. Even though our data does not have many covariates, it still allows for a flexible computation of propensity scores, even with nonlinear relationships. This means that both variable selection and manual incorporation of polynomials and interactions are not required. Quite the contrary, adaptive methods such as GBM automatically add terms and variables to the model based on statistical improvements (McCaffrey et al., 2004).

GBM will represent a sum of regression trees, allowing for the advantages of boosting while maintaining the advantages of regression trees. Trees can handle nominal ordinal, continuous, and missing independent variables while also having the ability to capture nonlinear effects and interaction terms (McCaffrey et al., 2004). Furthermore, the trees are not sensitive to transformations of the independent variables (i.e. log or squared).

Generally, traditional regression methods cannot handle a broad range of covariates which means that variable selection is a crucial step in modelling propensity scores. This means that GBM would be preferred in case of many covariates, since it adaptively accounts for the wide range of covariates. However, this paper will investigate the application of GBM in case only a small portion of covariates are in play.

So, GBM will be applied as a recursive tree-fitting algorithm that allows for the computation of treatment assignment probabilities. To simplify, GBM combines simple functions that individually lack the ability to estimate a smooth function but collectively manage to approximate the function of interest smoothly (McCaffrey et al., 2004). The simple functions in this case are represented by regression trees. Regression trees have a pre-determined depth and use recursive partitioning to split the data based on the input variables. The algorithm does this by minimizing the prediction error for each split (McCaffrey et al., 2004). Each split divides the data into two partitions, which means that a regression tree with a depth of two divides the data into four groups. The higher the depth of the tree, the more complex the tree becomes. For each group in the

leaf nodes of the tree, a prediction is made on the outcome, which in this case is the treatment probability.

GBM combines all these individual trees to get a smooth function of the treatment probability assignment and provides us with propensity scores. More specifically, instead of directly modelling propensity scores, GBM computes the log-odds of treatment assignment. This can be notated as: $g(x) = log(p(x)/(1-p(x)))$, in which $p(x)$ is the propensity score. McCaffrey et al. (2004) suggest that the algorithm initially sets $g(x)$ to: $log(\overline{z}/(1-\overline{z})$, in which $\overline{z}$ represents the average treatment assignment for the entire sample. Now, the model tries to find small adjustments that enhance the model's fit. Therefore, to get the best estimate of the propensity score, $p(x)$, we should examine the expected Bernoulli log-likelihood function (McCaffrey et al., 2004):

$$\mathbb{E}(\ell(p)) = \mathbb{E}(z\log(p(x) + (1-z)\log(1-p(x)) \,|\, x) \tag{9}$$

The probability of treatment assignment is retrieved by maximizing Equation 9. A logistic transformation of the propensity score will simplify the analysis:

$$p(x) = \frac{1}{1+\exp(-g(x))} \tag{10}$$

This ensures that $p(x)$ will always be between 0 and 1. By combining Equation 9 and 10, the expected log-likelihood is retrieved in terms of regression function $g(x)$, in which boosted regression trees will be used as $g(x)$. The expected Bernoulli log-likelihood becomes:

$$\mathbb{E}(\ell(g)) = \mathbb{E}(zg(x) - \log(1+\exp g(x)) \,|\, x) \tag{11}$$

Now, whenever the algorithm finds an $h(x)$ that improves the model's fit (Equation 12), $g(x)$ changes to $g(x) + h(x)$,which increases the log-likelihood. This process continues iteratively till the stopping rule is reached. By letting the process continue for too long, the model overfits the data. Therefore, the model needs to be instructed to stop at pre-determined conditions. McCaffrey et al. (2004) suggests using the minimized effect size in covariates as the stop condition. The effect size measures the balance in the means of individual covariates between the treatment- and control group. However, McCaffrey et al. (2013) suggests using the Kolmogorov-Smirnov (KS) statistic. The KS statistics compare the distributions of the covariates between the treatment- and control group. The advantage of the KS statistic is that it compares the entire distribution of the covariates rather than just the mean. However, the effect size is independent of sample size whereas the KS does depend on the sample size (McCaffrey et al., 2013).

The guideline for modest datasets is that threshold is around 0.10 for considering imbalance. Ridgeway et al. (2021) suggest using the *twang* package in the R environment that allows for the application of both the effect size and the KS statistic. A table can be constructed using the *twang* package that provides the information about pretreatment covariates before and after weighing using GBM. This table shows the absolute standardized mean difference for both the effect size approach and the KS statistic approach Ridgeway et al. (2021).

4.4 K-means clustering

Clustering is a data selection technique that assigns subjects with similar characteristics to the same group and dissimilar subjects to different groups out of multivariate data (Kodinariya and Makwana, 2013). The goal is to minimize the within-cluster sum of squares so that no movement of subjects between clusters reduces the within-cluster sum of squares further (Hartigan and Wong, 1979). Clusters can be used for further data analysis by making inferences about specific groups with similar characteristics rather than about individuals. This makes targeting more efficient and allows for group-level distinctions.

In this research, we will use the traditional and widely accepted k-means clustering approach. K-means clustering starts with a random initial partition and iteratively re-assigns observations to the cluster centers, that shift with each iteration (Jain et al., 1999). The observations are iteratively assigned to the clusters based on minimizing the within-cluster sum of squares till the convergence criterion is met. This convergence criterion, also referred to as clustering error, represents the situation in which no re-assignment of observations occurs because the within-cluster sum of squares cannot be further minimized (Jain et al., 1999). This within-cluster sum of squares is characterized by the sum of squared Euclidean distances between each observation x_i and centroid m_k of subset C_k which contains x_i (Likas et al., 2003). C_k represents a subset that is formed by the respective cluster, with $C_1, \ldots, C_K$ for each of the K different clusters (Likas et al., 2003). The clustering error can be formulated as:

$$\mathbb{E}(m_1, \ldots, m_K) = \sum_{i=1}^{N} \sum_{k=1}^{K} I(x_i \in C_k)|x_i - m_k|^2, \tag{12}$$

in which K is the number of clusters (Likas et al., 2003). The solutions found by the k-means algorithm are locally optimal solutions and depend on the arbitrarily placed initial starting points. Many other clustering techniques have been introduced, but k-means clustering still remains the prioritized technique (Likas et al., 2003). The drawback of k-means clustering compared to other clustering techniques, is that its performance depends on the initial starting conditions (Peña et al., 1999). However, this dependence is reduced by the introduction of multiple restarts (Likas et al., 2003). To come near the optimal solution, the algorithm needs to be run with multiple random starting points to reduce the sensitivity to outlying initial starting positions. Extant literature has generally agreed upon using 25 different initializations, which is why 25 different initializations will be used in this research as well (Sinaga and Yang, 2020; Lemenkova, 2019; Apon et al., 2006). According to Jain and Dubes (1988) the average of the final cluster centers must be taken to obtain the final result.

Secondly, the number of clusters must be pre-specified in order for the algorithm to run. The algorithm must know many centroids need to be placed initially before repetitively assigning objects to the most nearby centroid (Kodinariya and Makwana, 2013). Determining the optimal number of clusters can either be done by expert knowledge about the data set, which is the theoretical approach, or can be done by analyzing the data using the Elbow method and the Silhouette method. The Elbow method is visualized as a plot of the number of clusters against the cost function. The Elbow method assumes that the optimal number of clusters is the point before the cost function dramatically drops (Kodinariya and Makwana, 2013). This means that when a steep downward slope shifts to a significantly less steep slope, the cluster number that is positioned on that angle should be chosen. The rationale behind this method is that increasing the number of clusters

after a specific point does not significantly improve the distinction between the clusters. Once that point has been reached, one should decide to go with the number of clusters that are positioned on that turning point. The drawback of this method is that it can be hard to identify a real 'elbow' in the visualization of the data, which makes it an arbitrary decision on what number of clusters to go for (Kodinariya and Makwana, 2013).

The second method that can effectively used to decide on the optimal number of clusters is the Silhouette method. This method uses a well-balanced coefficient called the silhouette width to make a trade-off between the within-cluster difference and between-cluster difference (Kodinariya and Makwana, 2013). The silhouette width is first introduced by Kaufman and Rousseeuw (2009) and has gained popularity since. The silhouette width, referred to as $s(i)$ can be defined as:

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \tag{13}$$

Here, $b(i)$ is the minimum of the average distances between i and all the observations in each other cluster (Kodinariya and Makwana, 2013). Furthermore, $a(i)$ represents the average distance between i and all other observations within i's cluster.

So, on the one hand we have the between-cluster distance while on the other hand we have the within-cluster distance. The silhouette width is calculated by Equation (13) which says that the between-cluster distance minus the within-cluster distance should be divided by the maximum value of both distances. This gives a well-balanced coefficient that can be maximized to obtain the optimal number of clusters. Since we are looking for high between-cluster distances and low within-cluster distances, we tend to maximize the average silhouette width (Kaufman and Rousseeuw, 2009). As Kodinariya and Makwana (2013) claim, the largest average silhouette width, over different numbers of clusters, indicates the optimal number of clusters. So, the decision on the optimal number of clusters will be made using both the Elbow- and Silhouette method. These methods will be interpreted individually and in case they indicate contrasted outcomes, a well-supported decision will be made taking both outcomes into consideration.

4.5 Propensity score matching

4.5.1 The bridge to propensity score matching

In Chapter 4.2., we briefly discussed the implications of propensity score matching and the advantages and the limitations of a regular PSM approach. Now, we will further elaborate on the application of propensity score matching in our approach. So, the clusters generated using k-means clustering are used as input in the propensity score matching process. In normal practice, existing R software is used as effortless integration of both propensity score calculation and matching in one process. However, these R packages do not incorporate the ability to apply GBM for the calculation of propensity scores, which is why the propensity score calculation and matching process are separated. These propensity scores have been calculated using GBM in advance and these calculated propensity scores in combination with the cluster analysis are the foundation of the propensity score matching process.

The propensity scores will be matched for each cluster separately, allowing for the estimation of various different ATTs. As discussed earlier, both the standardized effect size (es.mean) and the Kolmogorov-Statistic (ks.max) will be used as stop conditions for propensity score estimation using GBM. Propensity score

matching will be performed using both stop conditions and the relative performance will be evaluated using the ASAM. Moreover, both nearest neighbor matching and optimal matching will be used in combination with these stop conditions. This means that a superior combination will be chosen that is characterized by a combination between a stop condition and a matching method. The performance of each matching combination will be evaluated using the ASAM. The matching combination with the overall most frequent superior performance across the clusters will be chosen as the superior matching method to proceed with. The relative performance of these matching methods will be visualized in a table.

4.5.2 How does matching work?

The matching process consists of matching untreated and treated observations with similar propensity scores into matched pairs (Austin, 2014b). One-to-one matching is the most common approach and will be applied in this research as well. One-to-one matching entails forming matched pairs between each treated observation and one untreated observation (Austin, 2011a). This means that for every treated observation, one untreated observation will be found with the best matching propensity score. Furthermore, we apply matching without replacement, which means that once an untreated observation is matched with a treated observation, this untreated observation can no longer be used as a pair with another treated observation (Rosenbaum, 2002; Austin, 2011a). Matching treated to untreated observations based on the propensity score allows for the estimation of the treatment effect by comparing the outcomes between the observations in the matched pair (Austin, 2014b).

As mentioned earlier, Rosenbaum and Rubin (1983) suggest that observations with equal propensity scores tend to have the same distribution of observed baseline covariates. These covariates are related to both the outcome and treatment assignment and therefore correlation exists in the pair matched outcomes (Austin, 2014b). This means that the outcome of two observations within a matched pair are likely more similar than the outcome of a randomly chosen pair. By finding a matched pair for each treated observation, we obtain a dataset that only contains matched pairs with similar propensity scores. Since the observations are matched based on propensity scores, the covariate distribution between the treatment- and control group should be similar as well. The covariate balancing between both groups is important and will be elaborated on further.

4.5.3 Balance requirements and implications

After effectively applying propensity score matching, we obtain this dataset with matched pairs. Consequently, based on this matched dataset, we can check how the covariate balance of the dataset changed after matching. This can be done both by constructing a balance table as well as plotting either the covariate balance or the propensity score distribution. For examining balance, we use the standardized difference. The standardized difference compares the means of covariates between treatment- and control groups. The standardized difference measures the difference in the sample mean of a covariate between the treatment- and control group and divides it by the sample variance (Mansournia and Altman, 2016). The standardized difference for continuous variables can be defined as:

$$d = \frac{(\overline{x}_{treatment} - \overline{x}_{control})}{\sqrt{\frac{s^2_{treatment} + s^2_{control}}{2}}} \tag{14}$$

The standardized difference for binary variables can be defined as:

$$d = \frac{(p_{treatment} - p_{control})}{\sqrt{\frac{P_T(1-P_T) + P_C(1-P_C)}{2}}} \tag{15}$$

To draw conclusions about overall balance, we need to set a threshold on what is considered to be unbalanced. Extant literature has agreed upon a standardized difference threshold of 0.10 (Austin, 2007; Normand et al., 2001; Austin and Mamdani, 2006). This means that when the standardized difference of a covariate exceeds this threshold, it is considered to be imbalanced, whereas every value below 0.10 is considered to be properly balanced.

For the overall balance, we use the average standardized absolute mean difference (ASAM) to compare the relative performance of the different methods (optimal vs. greedy matching; es.mean vs. ks.max). The ASAM averages the standardized differences over all covariates. This metric allows for the comparison of the obtained balance across different methods, as the ASAM is calculated in the same way for each method.

4.5.4 Application of matching

Two different implementations of matching are greedy matching and optimal matching (Austin, 2007). Greedy matching is an application of k-nearest neighbor (KNN) and implies that the algorithm one by one assigns an untreated unit to a treated unit with the smallest distance. So, the algorithm randomly selects a treated unit and assigns the untreated unit with a propensity score closest to the propensity score of the treated unit to become a matched pair (Austin, 2011a). This is a repeated process until all treated units have been matched. The greediness comes from the ignorance of any form of optimization by not looking at the distance trade-off between other potential matches.

On the contrary, optimal matching minimizes the total within-pair difference of the propensity score (Austin, 2011a). However, Gu and Rosenbaum (1993) argue that optimal matching generally does not create better matched samples than greedy matching does. To test this finding, greedy matching and optimal matching will be compared through matching balance to see which matching method is superior. The superior matching method will be used for the matching process.

Furthermore, matching will be performed without replacement, implying that matched untreated observations are withdrawn from the sample and cannot be matched again (Rosenbaum, 2002; Austin, 2011a). A specification that will be used in the greedy matching procedure is the caliper distance. This caliper distance states that the absolute difference between the propensity scores of the matched units should not be above the specified threshold (Austin, 2011a).

Consequently, some treated units might be left out as well since, under the maximum distance restrictions, these treated units cannot find a match amongst the untreated units. While there might be no 'best' standard for the caliper value, extant literature has agreed upon a caliper width of 0.2 * the standard deviation (SD) of the logit of the propensity score (Austin, 2011b,a, 2014a; Zhao et al., 2021). The caliper width of 0.2 * SD seems to minimize the mean squared error (MSE) of the estimated treatment effect and to be the best

performing value for enhanced matching results (Austin, 2011b). Furthermore, Austin (2011a) claims that using 0.2 * SD of the logit of the propensity score as the caliper width eliminates about 99% of measured confounders bias. Therefore, a caliper width of 0.2 * SD will be used as an extra specification in the KNN matching algorithm. This matching procedure is also referred to as caliper matching (Morgan and Harding, 2006).

In contrast to greedy matching, optimal matching attempts to minimize the sum of absolute pair distances in the matched sample (Greifer, 2023). Instead of just looking for the nearest observation, optimal matching uses optimization to potentially allow for an overall balance. So, both greedy (caliper) matching and optimal matching will be used to balance the sample. KNN will be used to assess, for each treated observation within a cluster, which untreated observation is the nearest neighbor in terms of propensity scores. The threshold for the maximum distance is 0.2 * SD of the logit of the propensity score, meaning that every untreated observation with a greater distance to the treated observation will not be included for pair matching. By effectively applying KNN, each treated observation that meets the caliper width condition is paired with an untreated observation which is meant to form a balanced sample.

4.6 Estimating ATT

4.6.1 Functional form

Matching based on the covariates does not exclude these covariates in the outcome model. Ho et al. (2007) claims that matching reduces the dependence on correct form specification in the outcome model. However, this reduction of dependence does not mean that the covariates should be excluded in the outcome model. In fact, Greifer (2022) claims that including covariates in the outcome model can increase the precision of the treatment effect estimate, reduce bias from residual imbalance, and make the treatment effect estimate "doubly robust". This means that the estimate is consistent if either the outcome model is correct, or imbalance is sufficiently reduced using matching.

Greifer (2022) advices generally not to repeatedly alter the functional form to optimize the results since this can invalidate the results and make them complicated to replicate. While including the covariates in the outcome model, the individual coefficients of these covariates do not represent causal effects and should not be interpreted. The application of matching combined with estimating the ATT using a linear model allows for the causal interpretation of only the treatment effect. Incorrectly interpreting the coefficients of other covariates in the outcome model is referred to in extant literature as the Table 2 fallacy (Westreich and Greenland, 2013; Greifer, 2022). To avoid falling for this fallacy, we use an approach that only shows the treatment effect of the outcome model using the G-computation method.

4.6.2 Marginal effect estimation

Snowden et al. (2011) explain the application of G-computation and the difference to a traditional regression approach. G-computation fits a regression model (referred to as Q-model) on the treatment variable and remaining covariates and uses this as foundation for the computation of marginal effects. Just as with traditional regression, the Q-model needs to be correctly specified to reduce estimation bias. However, by combining G-computation with matching, the functional form does not have to be correctly specified to

reduce estimation bias (Greifer, 2022). The Q-model is used to predict counterfactual outcomes for both the treatment- and control group by imputing $z = 0$ for every observation as well as $z = 1$ for the same observations (Snowden et al., 2011). By generating counterfactual outcomes (potential outcomes) for each observation in both treatment settings, the researcher has Y_1 and Y_0 for each observation (Snowden et al., 2011). From this point, we could merely take the difference between Y_1 and Y_0 and average it across the observed distribution of confounders to get the treatment estimate (Snowden et al., 2011).

However, this also allows for the application of a marginal structural model (MSM) to extract the ATT in case of a functional form beyond singular effects. Incorporating interaction effects in the Q-model is a safe way to ensure that the treatment effect is fully captured. In the case of interaction effects in the Q-model, G-computation permits the estimation of the single, marginal effect estimate which is averaged across the observed distribution of all the interactions with the treatment variable (Snowden et al., 2011). This computation of a single, marginal effect estimate simplifies the interpretation and allows for a clear picture of the treatment effect.

So, the advantage of using G-computation over traditional regression is that the focus is purely on the causal treatment effect by decoupling the estimate of the treatment effect from adjustment for confounding and nuisance effect modification (Snowden et al., 2011). The purpose of this research is to look at the heterogeneity of treatment effects across clusters rather than within clusters. By capturing the treatment effect within clusters by a single, marginal treatment estimate, we increase the interpretability which allows for a better foundation to answer the central research question.

4.6.3 Cluster-robust SEs

For the application of G-computation with matched samples, the estimated potential outcomes under each treatment level ($z = 0$ and $z = 1$) must take the matching weights into account (Greifer, 2022). Additionally, for targeting the ATT, we should only estimate potential outcomes for the treatment group since we want to measure the treatment effect on the treated. Furthermore, the estimation of standard errors and computation of the confidence intervals play an important role in causal analysis using G-computation. Using cluster-robust standard errors (SE) has been shown to perform well after matching as uncertainty estimation (Greifer, 2022).

Consequently, cluster-robust SE could be a valuable tool in this research for estimating uncertainty for the ATTs of each cluster. Greifer (2022) suggests that in order to compute SEs after G-computation, a method called delta method can be used. This method computes the SEs of the expected potential outcomes and their counterfactuals from the coefficients' variance of the outcome model (Greifer, 2022). This means that the variance of the coefficients should be estimated correctly, since the delta method is reliant on these variances. Liang and Zeger (1986) suggest that normally you would assume independence across observations to consistently estimate variance. However, after applying matching techniques, a certain level of dependence is introduced within each matched pair.

In extension, the data has been clustered as well as matched, which could increase the dependence across observations even more. Cluster-robust SEs instead of standard SEs can be used to account for this dependence. Cluster-robust SEs take the correlation between matched pairs into consideration and allow for more accurate calculations of uncertainty. The validity of the use of cluster-robust SEs after matching has

been confirmed by several researchers, referring to it as the difference between "robust" and "naïve" methods (Lin and Wei, 1989; Gayat et al., 2012; Abadie and Spiess, 2022; Austin, 2009, 2013; Austin and Small, 2014; Wan, 2019)

4.7 Comparison to parametric approach

As mentioned in Chapter 4.2.6., we suggest a new perspective on exposing heterogeneous treatment effects. We will compare our results to a benchmark approach which consists of a logistic model instead of a GBM model to estimate propensity scores. Note, we already expand the regular PSM approach with clustering and G-computation, but this allows us to generate comparable results. Table 3 of Chapter 4.2.6. already captures the differences between our approach, the benchmark approach, and the causal forest approach of Athey and Wager (2019).

Both our approach and the benchmark approach generate cluster-specific treatment effect estimates including the standard error and confidence intervals. The metrics that will be used for measuring the relative performance are: (1) the ASAM, (2) confidence intervals of the treatment effects, and (3) standard errors. These metrics will be summarized in a table for both approaches. The metrics will be compared, and conclusions will be drawn about whether or not our approach is superior to the benchmark approach based on the proposed metrics.

4.8 Sensitivity analysis

Initially, we assume that our propensity score matching is based on all relevant characteristics without the presence of unobserved confounders that account for differences across treatment- and control groups (Keele, 2010). The reasonability of this assumption, however, is tested using the Wilcoxon's signed rank test following Rosenbaum's (2002) approach. According to Rosenbaum (2002), observational studies vary significantly due to hidden bias. The importance of the degree of sensitivity of outcomes is captured by the difference of either reflecting hidden bias, or representing the effect as a direct result of treatment. As a matter of fact, the reflection of hidden bias is merely an indicator of the capability of hidden bias to alter the results. The presence of hidden bias cannot be proved using sensitivity analysis (Rosenbaum, 2002). However, we can still uncover the sensitivity of the results to hidden bias to conclude to degree of cautiousness when interpreting the average treatment effect.

Rosenbaum (2002) suggests using the Wilcoxon's signed rank test to obtain statistical significance levels for increased odds of treatment. The Wilcoxon's signed rank test obtains the differences between treated and untreated observations of each matched pair. Subsequently, if the differences are positive, the ranks of the absolute differences are summed (Keele, 2010; Rosenbaum, 2002). Afterwards, p-values are calculated that inform us how likely we would observe the treated outcome due to chance (Keele, 2010).

These p-values are then calculated different altered odds of treatment assignment, referred to as Gamma. More specifically, an increased Gamma represents an increase in the odds of treatment of one person in a matched pair that is due to unobserved covariates. Assume a Gamma of 1.1, this implies that the odds of a unit to be assigned to a treatment is 1.1 times higher than someone else with the same covariates, due to differences in unobserved covariates. So, for a Gamma of 1.1, the p-value is given for the statistical significance

of the contribution of the treatment to the outcome if one unit is 1.1 times as likely to be treated as another unit with the same recorded covariates (Zhao et al., 2021). By studying the respective change in the p-value, one can draw conclusions, for each cluster, about the sensitivity of the outcome to unobserved covariates, referred to as hidden bias. The sensitivity of the outcome to unobserved covariates will be measured for both our approach and the benchmark approach and the results will be compared.

4.9 Comparison with causal forest approach

Even though the results of our approach and the causal forest approach are not one-to-one comparisons, we will still give some indication of differences in insights and output. We expect that our approach allows for more extensive interpretations of heterogeneity. On the contrary, we expect that the causal forest approach is better at exposing heterogeneity by using both test calibrations and statistical group differences.

The test calibrations report the mean prediction as well as the differential prediction. The differential prediction indicates heterogeneity if the coefficient is significant and positive. Furthermore, the statistical group differences indicate heterogeneity if the confidence interval of the difference between the estimates of the subgroups does not include zero. This will be compared with our approach in which each cluster has its own interpretation, ATT estimate and confidence interval. The implications of the using these different insights will be discussed.

5 Results

In this section, the most important results from the analysis will be discussed. Additionally, this section will provide a broad sense of how the analysis has been performed. For further details regarding the analysis and reproducibility, see Appendix B. Only the most important results for each layer will be exhibited. This chapter will start with propensity score estimation using GBM. Subsequently, school-based clustering using k-means clustering is conducted, followed by cluster-specific propensity score matching. Further, the ATTs are estimated for each cluster using the matched dataset. Moreover, the ATTs of our approach will be compared to the ATTs of the parametric benchmark approach to compare the performance. In addition, sensitivity analysis is performed for both our approach and the parametric benchmark approach in which the confidence intervals of the p-values are calculated for different odds of treatment assignment. To finalize, our approach will be compared with the causal forest approach of Athey and Wager (2019) to discuss differences in insights and outcomes.

5.1 Propensity score estimation using GBM

The GBM algorithm requires pre-specification of hyperparameters. Details regarding the application of the GBM algorithm using R software can be found in Appendix B. The hyperparameters have been tuned carefully, while adapting *n.trees* to the close-bounds warning. Eventually, the optimal set of the hyperparameters (*n.trees*, *interaction.depth*, *shrinkage*) turned out to be (20,000; 2; 0.001). The optimal number of iterations per stopping condition has been visualized as well as directly extracted from the algorithm.

Figure 3 shows the balance measures plotted against the number of iterations for both stopping conditions (es.mean and ks.max). This resulted in a smooth line with 17,097 iterations minimizing the average effect size difference and 17,917 iterations minimizing the largest KS statistics computed for the covariates (Ridgeway et al., 2022). The optimal number of iterations does not differ much between these two stopping conditions, which means that the stopping rules are compatible (Ridgeway et al., 2022). This means that the results of the final analysis should not be very sensitive to the stopping condition used. However, further down the analysis we will still compare the balancing outcomes of the two stopping conditions and choose the superior one for estimating the treatment effect.

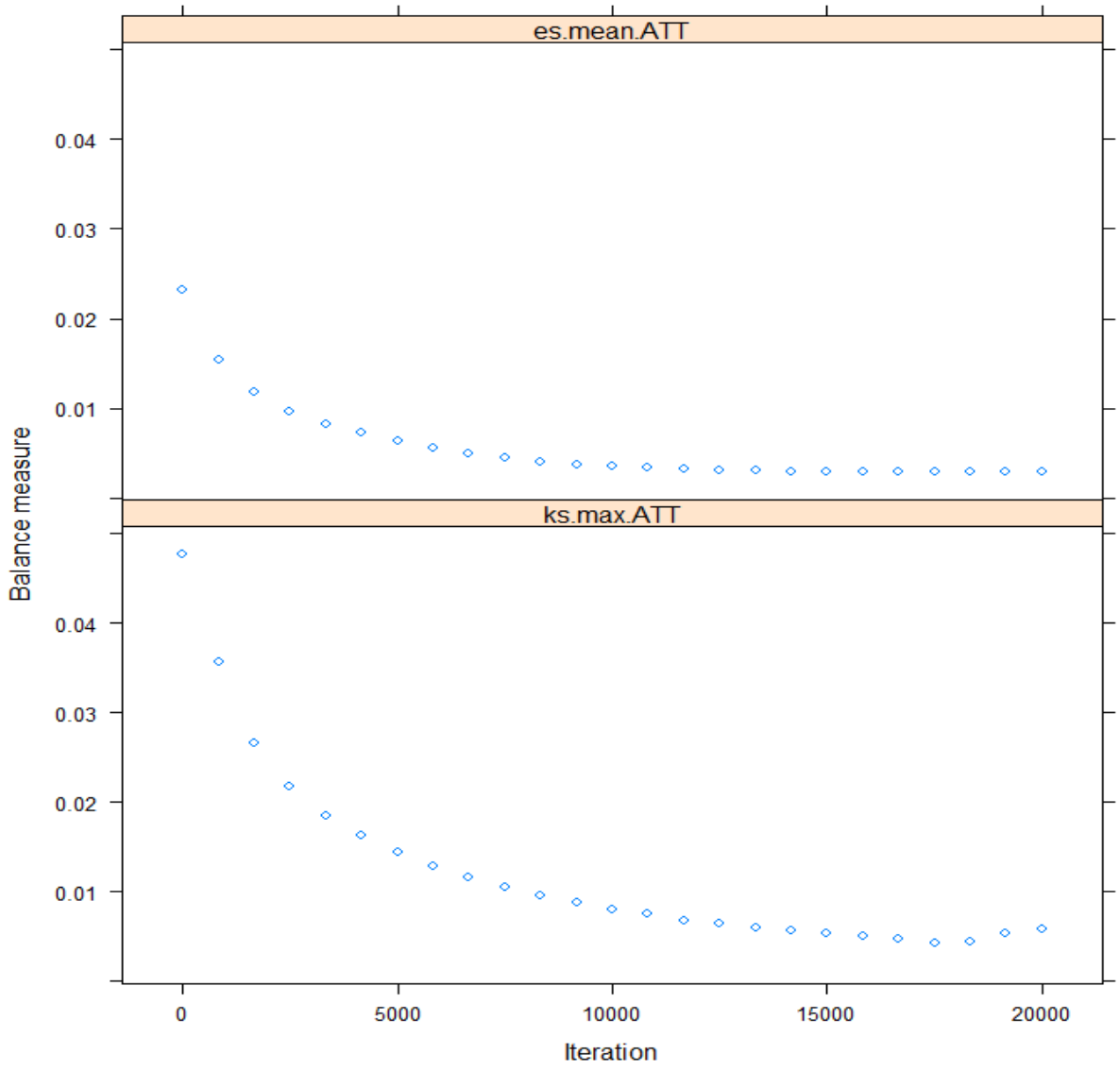


Figure 3: Balance measures as a function of the number of iterations

5.2 K-means clustering

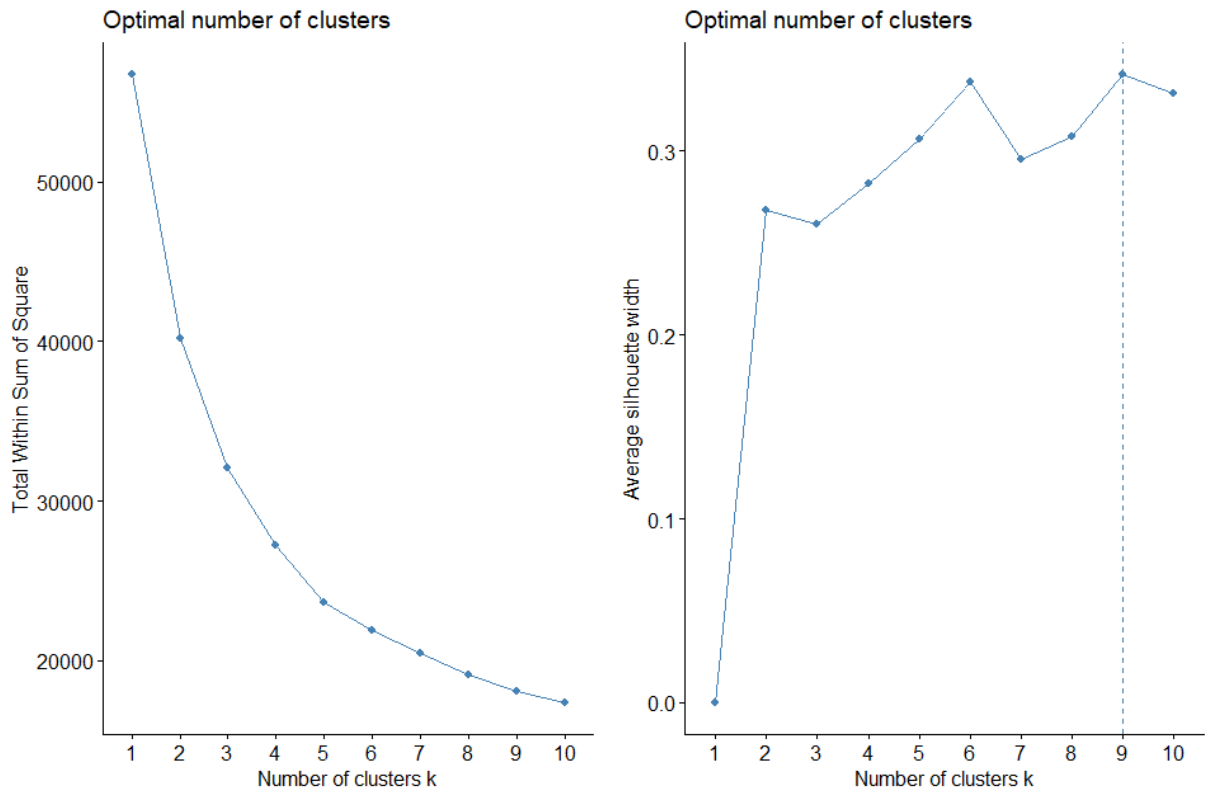
5.2.1 Deciding on the number of clusters

We applied GBM to estimate the propensity scores using both ks.max and es.mean. These estimated propensity scores were added to the dataset, after which k-means clustering was performed based on school-level characteristics (School-level fixed mindset (`X1`), School achievement level (`X2`), School ethnic minority composition (`X3`), School poverty concentration (`X4`), School size (`X5`), and Urbanicity (`XC.0-XC.4`)). We applied the Elbow method and the Silhouette method to decide on the optimal number of clusters (Kodinariya and Makwana, 2013). Details on the facilitation of this process can be found in Appendix B.

The results of the Elbow method and Silhouette method are visualized in Figure 4a and 4b, respectively. The way the Elbow method works is that the optimal number of clusters is the point before an elbow-like turn happens in the line. This means that when the line changes from being steep to being flat, this infliction point indicates the optimal number of clusters. Figure 4a shows a smooth line with no obvious infliction point. This means that based on the Elbow method, we cannot undisputably determine the optimal number of clusters. Nevertheless, we observe a small infliction point at 5 clusters and a slightly smaller infliction point at 6 clusters.

Figure 4b shows the average silhouette width for each number of clusters. As mentioned earlier, the average silhouette width is a balanced coefficient for the trade-off between the within-cluster difference and between-cluster difference. Since we are looking for high between-cluster distances and low within-cluster distances, we tend to maximize the average silhouette width. Therefore, based on figure 4b, the optimal number of clusters would normally be 9. However, we observe an average silhouette width for 6 clusters that is approximately as high as the average silhouette width for 9 clusters. For the sake of interpretability, we prefer going with fewer clusters, and since the average silhouette width is nearly the same for both 6 and 9 clusters, we prefer to proceed with 6.

The Elbow method does not clearly indicate what the optimal number of clusters is, since there is no real infliction point in Figure 4a. Nevertheless, the line seems to change direction the most around 5 clusters. The Silhouette method shows that there is only a slight difference between 6 and 9 clusters. Due to the advantage of interpretability of using fewer clusters, 6 clusters is the preferred option for the Silhouette method. Similarly, 6 clusters is closer to the infliction point for the Elbow method as well. Therefore, we choose to proceed the analysis with 6 clusters.



(a) Elbow method

(b) Average silhouette method

Figure 4: Metrics to determine the optimal number of clusters

5.2.2 Cluster composition

The means of each covariate across clusters was calculated, which allowed for the interpretation of the individual clusters. Table 4 shows the content of each cluster regarding the mean of each of its school-based covariates. Now, we will use this table to interpret each of these clusters. The variables `X1-X5` are all standardized with a mean around zero and a standard deviation of one. This means that we cannot directly interpret what these values represent, but we do know that higher values correspond to high levels for that variable.

Dummies `XC.0-XC.4` represent the urbanicity categories of the school. We observe that the first cluster is characterized as the most rural, with all schools in this cluster located in places that are characterized as either rural, countryside, or town. Meanwhile, the sixth cluster is characterized as pure urban. The second- and fifth cluster score high in urbanicity as well whereas for the other clusters it is more equally distributed.

In terms of school-level mean of fixed mindsets (`X1`), we observe that the second cluster has the highest mean of school-level fixed mindsets, whereas the sixth cluster has the lowest mean of school-level fixed mindsets. Moreover, the first cluster scores relatively low as well.

In terms of the school achievement level (`X2`), we observe that cluster 1 has the highest mean of school achievement, whereas cluster 2 has the lowest mean of school achievement. Cluster 6 scores relatively high as well.

The racial/ethnic minority composition is represented by the percentage of students that is Black, Latino,

or Native American (`X3`). Cluster 2 has the highest mean of minority composition, whereas cluster 3 and 4 have a relatively low mean of minority composition.

In terms of school poverty concentration (`X4`), we observe that cluster 2 and 6 score relatively high on school poverty, whereas cluster 1 and 3 have relatively low poverty concentrations.

In terms of school size (`X5`), we observe that schools in cluster 6 have the greatest size on average, whereas schools in cluster 4 have the smallest size on average.

Table 4: Cluster-specific content containing the mean of covariates

Cluster	**1**	**2**	**3**	**4**	**5**	**6**
School-level fixed mindset (X1)	-0.883	1.250	-0.421	0.107	0.434	-1.160
School achievement level (X2)	1.170	-1.220	0.395	-0.072	-0.383	0.749
School ethnic minority composition (X3)	-0.227	1.320	-0.915	-0.767	0.649	0.032
School poverty concentration (X4)	-0.989	1.140	-0.839	0.189	-0.387	1.070
School size (X5)	1.060	-0.673	0.140	-1.050	-0.101	1.600
Rural (XC.0)	0.199	0.100	0	0.083	0.114	0
Countryside (XC.1)	0.307	0.095	0.437	0.301	0.086	0
Town (XC.2)	0.494	0	0.438	0.063	0.092	0
Suburban (XC.3)	0	0.287	0	0.433	0	0
City (XC.4)	0	0.518	0.124	0.120	0.708	1

5.2.3 Cluster interpretation

The six clusters can be interpreted individually using the mean values captured in Table 4. For this, we constructed another table that emphasizes high-scoring variables that differentiate a cluster from the rest of the clusters. Table 5 shows an overview of the characteristics of each cluster in which a comprehensive interpretation of each cluster is included. For the continuous variables, the zero, plus sign, and minus sign mean high values, average values and low values, respectively. For the categorical variables, zero means that the category is absent in that cluster. The number of stars is positively related to the proportion size of that category in the respective cluster.

So, for instance, cluster 1 scores low school poverty and fixed mindsets and scores high on achievement and size. Furthermore, the schools are only located in towns, countryside, and rural areas. The interpretation of cluster 1 therefore is; big rural schools with high achievements and low poverty, which has been given the name: "Big and Rich Rural Achievers". The cluster-specific interpretations can be used when comparing the ATTs of the different clusters to potentially assign a label to the heterogeneity.

5.3 Propensity score matching

We applied propensity score matching with two matching methods using two different propensity score distances. The difference in performance of these matching combinations have been tested using the ASAM after matching for each cluster. Based on the ASAM of each matching method across the clusters, the superior

Table 5: Overview of high-scoring variables for each cluster

Cluster	**1**	**2**	**3**	**4**	**5**	**6**
School-level fixed mindset (X1)	-	+	0	0	0	-
School achievement level (X2)	+	-	0	0	0	+
School ethnic minority composition (X3)	0	+	-	-	+	0
School poverty concentration (X4)	-	+	-	0	0	+
School size (X5)	+	-	0	-	0	+
Rural (XC.0)	*	*	0	*	*	0
Countryside (XC.1)	**	*	**	**	*	0
Town (XC.2)	**	0	**	*	*	0
Suburban (XC.3)	0	**	0	**	0	0
City (XC.4)	0	***	*	*	***	****
Name assigned to the cluster based on cluster-specific interpretations	Big and Rich Rural Achievers	Struggling Urbans	White Country-side	Widespread White Tinies	Multicultural Cities	Big City's Potential

Proportions of categories (XC.0-XC.4): $^{*} < 0.25$, $^{**}(0.25, 0.5)$, $^{***}(0.5, 0.75)$, $^{****} > 0.75$

Value of continuous variables (X1-X5): $^{+} > 0.5$, $^{-} < -0.5$, $^{0}(-0.5, 0.5)$

matching method has been chosen to proceed with in the analysis. Table 6 shows the ASAM per cluster for the different matching methods. The method with the lowest ASAM for each cluster has been indicated using a checkmark. From Table 6, we can extract that caliper matching using es.mean for propensity score calculation has the most frequent superior performance (3 checkmarks). Therefore, caliper matching using es.mean will be used to proceed the analysis with.

Table 6: Mean effect sizes for different clusters.

	ASAM			
	caliper.es	caliper.ks	optimal.es	optimal.ks
cluster1	0.0090	0.0114	0.0114	0.0093
	✓	×	×	×
cluster2	0.0110	0.0110	0.0105	0.0113
	×	×	✓	×
cluster3	0.0092	0.0089	0.0101	0.0095
	×	✓	×	×
cluster4	0.0081	0.0095	0.0106	0.0107
	✓	×	×	×
cluster5	0.0104	0.0133	0.0127	0.0141
	✓	×	×	×
cluster6	0.0116	0.0105	0.0128	0.0111
	×	✓	×	×

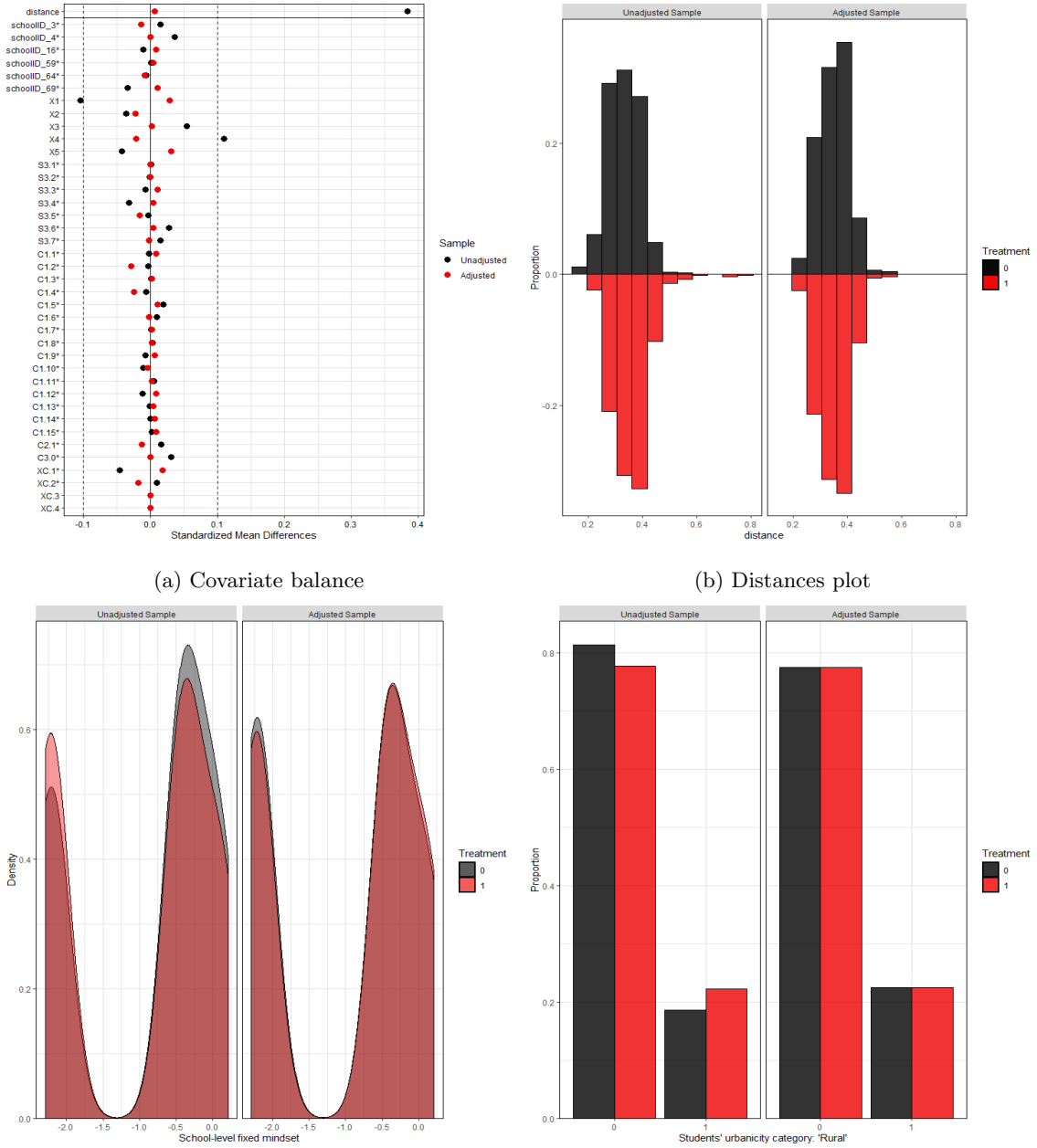
The covariate balance obtained by the es.mean method between the control- and treatment group can be visualized using both balance plots and love plots. Four such plots are showcased for cluster 1 in Figure 5. All these plots are only showcased for cluster 1 to keep it clear. Figures for the rest of the clusters can be found in Appendix C.

Figure 5a is a love plot of the covariate balance of each of the covariates. The black dots represent the standardized mean difference of the unadjusted sample (before matching) and the red dots represent the standardized mean difference of the adjusted sample (after matching). The love plot contains a zero-line with two dashed lines at the 0.10 threshold that, if surpassed, indicate imbalance. As we can see, some of the covariates were imbalanced before matching, which has been dissolved by matching the two groups. Furthermore, for each covariate, the mean standardized difference has diminished, which means that the balance for each covariate has improved.

Figure 5b shows the distributional balance of the propensity scores across the control- and treatment group. The unmatched sample shows differences in propensity score distribution while the matched sample dissolved most of these dissimilarities.[3] Consequently, since the propensity score distribution is similar between these two groups, the rest of the covariates' distribution is likely to be similar as well.

Figure 5c shows the distributional balance of the school-level mean of fixed mindsets. The school-level mean of fixed mindset is an average score across all students of a school for their belief in fixed mindsets, as opposed to growth mindsets. As we can see, the surface of the control- and treatment group are not aligned in the unmatched sample, whereas this has mostly been dissolved in the matched sample. Once again, this means that the matching method matched the two groups properly in this analysis. Figure 5d shows the distributional balance of the first dummy of the urbanicity dummy variable. This dummy represents the rural area. As we can see, the students with schools in rural areas differed between the control- and treatment group in the unmatched sample. However, after matching, the distribution is as good as similar between these two groups. Overall, Figure 5 shows that the applied matching method properly balanced the dataset. Figure 5 only shows the results of the first cluster, but the same conclusions can be drawn for the 5 other clusters captured in Appendix C.

[3]Examples of these differences are shown in the peak as well as the tails. The left side of the plot represents the unadjusted sample and shows a peak at a propensity score of around 0.3 for the control group and a peak at a propensity score around 0.4 for the treatment group. The peak in the adjusted sample for both groups is around 0.4. Furthermore, the unadjusted sample ranges from 0.15 to 0.5 for the control group and from 0.2 to 0.8 for the treatment group. After adjustment, this ranges from 0.2 to 0.5 for both groups.



(a) Covariate balance

(b) Distances plot

(c) Balance of school-level fixed mindset

(d) Balance plot of urbanicity category "Rural"

Figure 5: Plots of the differences between the adjusted and unadjusted sample for cluster 1

5.4 Estimating ATT

First of all, to estimate the ATT based on the previous steps, we will fit a linear regression to the data, consisting of interaction effects between the treatment and all covariates. This model is used as the Q-model, as mentioned earlier. By applying G-computation, we can effectively calculate the ATTs for each cluster with cluster-robust SEs and a confidence interval. By doing so, heterogeneity across clusters can be extracted and the significance of the ATT difference can be checked accordingly.

For estimating the ATT based on marginal effects using a linear model, we need to ensure the absence of multicollinearity. Multicollinearity in linear models can cause issues since some variables (almost) perfectly predict each other. Therefore, we tested the correlation between each covariate and deleted the covariates

that were highly correlated with other covariates. For further details on the correlation matrix, navigate to Appendix B. Consequently, we decided to remove `schoolID`, as well as the urbanicity dummy variables (`XC.0-XC.4`). Moreover, the first dummy of every categorical variable has been deleted from the linear model to serve as reference category (`S3.1, C1.1, C2.1, C3.0`). The rest of the covariates will be used to fit the linear model, including interaction terms with the treatment variable.

By following the G-computation approach, we obtained an estimate for each cluster, including a confidence interval of this estimate. Table 7 shows the ATT estimates for each cluster. Here we can observe a difference between the estimates, with a maximum difference of almost 50% (cluster 6 vs. Cluster 4). The significance of these results is not questionable, given that each estimate is significant at the 0.001 level. However, to expose heterogeneity, we should further investigate the confidence intervals.

With 95% certainty we can claim that the estimate will be within the range of the lower- and upper bound. For a cluster-based ATT to be significantly different from the ATT of another cluster, the confidence intervals should not overlap. For instance, if the actual estimate of cluster 4 is in the right tail around 0.270 and the actual estimate of cluster 6 is in the left tail around 0.250, then the ATT of cluster 4 exceeds the ATT of cluster 6, even though the estimated ATT of cluster 6 is approximately 50% higher. Heterogeneity can be exposed if the ATT across subgroups differs significantly. However, here we see that this is not the case and that, for each cluster, there is at least one other cluster of which the confidence interval overlaps its own confidence interval. Therefore, we cannot claim school-based heterogeneity based on this data.

Table 7: Average treatment effect (ATT) estimates for each cluster

Cluster	**Estimate**	**Std. Error**	**2.5%**	**97.5%**
1	0.234***	0.032	0.171	0.297
2	0.234***	0.035	0.166	0.302
3	0.277***	0.026	0.225	0.328
4	0.221***	0.028	0.167	0.275
5	0.266***	0.030	0.207	0.324
6	0.317***	0.036	0.247	0.387

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Nevertheless, suggestions can still be made about the heterogeneity implications in case these confidence intervals would not have overlapped. We notice the biggest difference in ATT estimates between cluster 6 and cluster 4. These clusters are characterized by Widespread White Tinies and Big City's Potential, respectively. The biggest difference between the two clusters is the average school size with cluster 4 containing, on average, the smallest schools of all clusters and cluster 6 containing, on average, the biggest schools of all clusters. If the confidence intervals of clusters 4 and 6 would not have overlapped, this would mean a heterogeneous treatment effect across these two clusters. More specifically, this would imply that big schools in the cities with students that come from relatively poor families experience a higher treatment effect than small schools across the country with mainly white students in it.

Although we cannot draw this conclusion based on our own analysis, this would be exceptionally interest-

ing to study on the real dataset, in which more variables are available. Therefore, this framework carries great potential to be used for the actual NSLM study, as well as other studies, both randomized and observational.

5.5 Comparison to parametric approach

Up until now, we have only discussed the results of our approach. However, to draw any conclusions about the relative superiority of this approach, we need to compare it against a benchmark approach. In this case, we use a parametric approach as our benchmark. The difference between these two approaches comes down to the use of either GBM or logit for propensity score estimation. The ASAM, confidence interval and standard error will be used to compare the two approaches.

Table 8 shows the comparison between using our approach with a GBM model and using the parametric approach which means using a logistic model. For each metric, the results of the GBM and logit model are reported side-by-side to facilitate interpretation. We observe that in neither of these metrics, a significant difference is present between the two approaches. Consequently, this means that propensity score estimation using a logistic model generated very similar results to propensity score estimation using a GBM model. We can therefore not claim that the GBM model outperforms the logistic model using this dataset.

Table 8: Comparison between the different approaches

	ASAM		**ATT estimate**		**Confidence interval**		**Std. Error**	
	GBM	Logit	GBM	Logit	GBM	Logit	GBM	Logit
Cluster 1	0.009	0.013	0.234***	0.239***	(0.171, 0.297)	(0.177, 0.302)	0.032	0.032
Cluster 2	0.011	0.009	0.234***	0.245***	(0.166, 0.302)	(0.175, 0.314)	0.035	0.035
Cluster 3	0.009	0.007	0.277***	0.299***	(0.225, 0.328)	(0.245, 0.353)	0.026	0.028
Cluster 4	0.008	0.008	0.221***	0.220***	(0.167, 0.275)	(0.166, 0.274)	0.028	0.027
Cluster 5	0.010	0.010	0.266***	0.239***	(0.207, 0.324)	(0.180, 0.297)	0.030	0.030
Cluster 6	0.012	0.009	0.317***	0.323***	(0.247, 0.387)	(0.255, 0.391)	0.036	0.035

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

5.6 Sensitivity analysis

Sensitivity analysis has been performed as an additional tool to check the stability of the obtained treatment effects. Rosenbaum's (2002) bounds have been used to determine at what Gamma the treatment effect would still be significant (upper bound of p-value < 0.05). As a reminder, Gamma represents the artificially increased odds of treatment assignment for a person with the same covariate distribution. The sensitivity of both our approach and the benchmark approach will be analysed and compared. The results have been summarized in Table 9 in which, for both approaches, the highest significant Gamma has been reported plus its associated confidence interval of the p-value.

We can see that Gamma differs between these six clusters. For our approach, Gamma ranges from 1.4 to 2.0. The interpretation of a Gamma of 2.0 is when the odds of treatment for one person as twice as high

due to hidden bias as another person with the same covariate distribution, the positive treatment effect is still significant. For a Gamma of 1.4, this means that if the odds of treatment are 1.4 times as high due to hidden bias, that the positive treatment effect is still significant. Part of this difference can intuitively be explained by the fact that the sixth cluster has the highest estimated treatment effect, whereas the fourth cluster has the lowest estimated treatment effect. The distance to zero is higher for the sixth cluster than for the fourth cluster, which is why more hidden bias should be introduced to get its confidence interval to include zero. Other than that, a Gamma of 1.4 would still mean that hidden bias should account for a 40% increase in the odds of treatment assignment for any of the treatment effects to lose its significance. This would require some significantly important treatment assignment predictors to be left out. We can conclude that the treatment effects of our approach are significant and only mildly sensitive to hidden bias.

In Chapter 5.5. we saw that the GBM model did not outperform the logit model on any of the three metrics. Now, Table 9 shows that the Gamma is on average even higher for the logit model than for the GBM model. This means that the estimated treatment effects using the benchmark approach are less sensitive for hidden bias than the estimated treatment effects using our approach. So, the significance- and sensitivity of the treatment effects are not subject to propensity score model choice.

Table 9: Sensitivity analysis using Rosenbaum's bounds

	Significant at Gamma		**Confidence interval of p-value**	
	GBM	Logit	GBM	Logit
Cluster 1	1.6	1.6	(0,0.0249)	(0,0.0143)
Cluster 2	1.5	1.6	(0,0.0426)	(0,0.0340)
Cluster 3	1.8	2.1	(0,0.0140)	(0,0.0468)
Cluster 4	1.4	1.6	(0,0.0485)	(0,0.0202)
Cluster 5	1.8	1.7	(0,0.0318)	(0,0.0167)
Cluster 6	2.0	2.2	(0,0.0286)	(0,0.0481)

5.7 Causal forest as a benchmark

5.7.1 Causal forest application

Athey and Wager (2019) predict the CATE of each observation using causal forest analysis. The authors use two approaches to test heterogeneity, which are: (1) statistical group differences using a confidence interval of the difference in ATEs, and (2) using test calibrations to see if the heterogeneity effect is significant. As a reminder, the first approach divides the sample into two subgroups, one with high predicted CATEs and one with low predicted CATEs. The ATE of both groups is calculated and the difference between the two ATEs is statistically tested. The 95% confidence interval for this difference is **(-0.02,0.128)**.

We observe that the 95% confidence interval includes zero which means that there is no statistical proof that the ATE of the two groups differ significantly. This absence of statistical difference implies that no heterogeneity is found between the group with low predicted CATEs and the group with high predicted CATEs.

The second approach Athey and Wager (2019) use to test heterogeneity is the test calibration. The test calibration seeks to fit the CATE as a linear function of the out-of-bag estimates and returns a mean prediction and differential prediction. Table 10 shows the results of the test calibration. We can see that the estimate of the mean prediction is approximately 1 and statistically significant. A coefficient of 1 suggests that the average forest prediction is accurate, which means that the predictions are well-calibrated. A coefficient close to 1 for the differential prediction implies that the variability of heterogeneity of the forest predictions are well-calibrated. The p-value for the differential prediction indicates the presence of heterogeneity.

Table 10 shows an differential prediction estimate of 0.231, which is not close to 1. This means that the predictions are not reliable for different subgroups in the data. Furthermore, the p-value equals 0.322, which is greater than 0.05 and therefore the estimate is not statistically significant. This means that we cannot reject the null hypothesis that there is no heterogeneity.

Table 10: Average treatment effect (ATT) estimates for each cluster

	Estimate	**Std. Error**	**p-value**
Mean prediction	1.003	0.082	0.000***
Differential prediction	0.231	0.498	0.322

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

5.7.2 Comparison with causal forest approach

5.7.2.1 Subgroups

We want to highlight the differences between the outcomes of our approach and the outcomes of the causal forest approach by Athey and Wager (2019). Our empirics-first approach defines six different, individually interpretable subgroups with their own ATT estimate and confidence interval. The causal forest approach has two subgroups that are based on the distinction between high- and low predicted CATEs. This means that our approach allows for the comparison across six different groups, whereas the causal forest approach allows for the comparison across two different groups.

5.7.2.2 Empirics-first

The subgroup division has been performed arbitrarily to some extent by Athey and Wager (2019). The authors used a 50-50 split on what they thought was the correct division. Our approach uses data-driven clustering to divide the sample into different subgroups. This empirics-first approach ensures that the subgroups are distinctive based on the used covariates with high between-group sum of squares and low within-group sum of squares.

5.7.2.3 Heterogeneity interpretations

Athey and Wager (2019) leverage two approaches to potentially expose heterogeneity. Our approach only uses G-computation and looks at the confidence intervals of the ATT estimate of each cluster. The statistical group differences is a flexible tool for measuring heterogeneity, since the group division can be performed

arbitrarily. Athey and Wager (2019) could also have chosen to split the sample into four subgroups and see if any of the groups significantly differ. Furthermore, test calibration is a convenient way to directly extract whether heterogeneity is present or not. The significance of the differential prediction coefficient immediately shows whether any heterogeneity is present.

In contrast, our approach makes indicating heterogeneity harder. However, in case signals of heterogeneity are present in the data, our approach allows for more in-depth interpretation of these heterogeneous treatment effects. We leverage the interpretability of data-driven clusters for the underlying effect modifiers. This means that, in the presence of heterogeneity across clusters, our empirics-first multi-step framework allows for more convenient and efficient targeting across these clusters based on their heteregeneous treatment effects.

6 Conclusion/Discussion

This research paper discusses the application of an empirics-first multi-step framework to potentially expose cluster-based heterogeneity. The field of cluster-based heterogeneity is still underdeveloped. We introduce a framework that extends the current knowledge base and provides a foundation for future researchers to build upon. The common approach for doing causal research on observational studies is regular PSM, in which a parametric model is used to estimate propensity scores. Subsequently, these propensity scores are matched after which the ATT can be extracted. However, we provide an empirics-first multistep framework and the theoretical contribution is twofold.

Firstly, we suggest using a GBM model instead of a parametric model like a logistic model to overcome shortcomings such as functional form determination. Secondly, we introduce data-driven clustering on institutional-level variables to allow for group-level comparisons. The ATTs are estimated for each individual cluster using G-computation in which the clusters are the foundation of potentially exposing heterogeneous treatment effects. This complements the work of Athey and Wager (2019) by providing data-driven subgroups and allowing for more in-depth interpretations of heterogeneity.

Regarding our analysis, we find that the es.mean method generates the best covariate balance. Consequently, our approach is capable of properly balancing the control- and treatment group within each cluster with an average ASAM across the clusters of around 0.01 (0.10 is the threshold for balance). Furthermore, we find that the treatment effect is positive and significant for each of the six clusters. This means that students who follow a growth mindset program have, on average, higher school achievements than students who do not follow the growth mindset program. We do not find heterogeneity across any of the clusters, even though the estimates differ up to 50%. Nevertheless, using sensitivity analysis, we find that the treatment effects generally remain significant, even if the odds of treatment assignment due to unobserved covariates increases to 1.5.

In addition, we also compare the results to a parametric benchmark model to measure the relative performance and check the superiority of our approach. Unfortunately, we find that the application of a GBM model as an alternative to a logistic model does not outperform the logistic model using three different metrics. The logistic model even seems to be less sensitive to hidden bias. Nevertheless, the parametric benchmark model does also not expose heterogeneity between clusters.

One of the reasons that our approach does not outperform the parametric benchmark approach could be

due to a simple correct functional form. Overall, the main advantage of the GBM model is that the functional form does not have to be pre-specified and is determined by the algorithm itself. However, apparently, the main effects-only logistic model is close enough to the actual functional form to generate proper estimates. The GBM model does not manage to come any closer to the actual functional form to make it the superior model in its application of estimating ATTs.

These findings are remarkable, since other researchers find that a GBM model does outperform a logistic model in terms of estimating unknown propensity scores (McCaffrey et al., 2004; Austin, 2012; Lee et al., 2010). First of all, Drake (1993) suggests that propensity model misspecification can substantially bias the results due to the fact that the correct functional form is near impossible to determine. Furthermore, Lee et al. (2010) find that, even though a main effects-only logistic model generally provides adequate covariate balance, GBM is able to provide significant bias reduction. Both Austin (2012) and Lee et al. (2010) suggest that GBM is superior, especially in cases with nonlinearity and interaction effects.

Apparently, the correct functional form is not complex enough for a model like GBM to flourish. In a case with more covariates and a more complex functional form including non-linear effects and interaction effects, GBM is more likely to outperform a limited, parametric model like a logistic model (McCaffrey et al., 2004). So, this research shows that the GBM model does not always outperform the logistic model in estimating propensity scores.

We also compare our approach with the causal forest approach of Athey and Wager (2019). The output of both methods are not one-to-one comparisons, however, we find interesting insights into the complementary application. We suggest that the advantage of the causal forest approach is identifying heterogeneity by using two different approaches that both leverage the full dataset. At the same time, our approach leverages the interpretability and applicability of heterogeneity signals. Even though it might be harder to indicate heterogeneity since the data points are divided into subgroups, the interpretation and applicability of the heterogeneity enhances. We manage to set up a framework that allows for intra-cluster interpretations as well as inter-cluster treatment comparisons. This framework can be applied by other researchers that have indicated heterogeneity and want to retrieve actionable insights regarding the heterogeneity in their data.

A practical limitation of this research is that while applying propensity score matching, we use a caliper of 0.2 times the standard deviation to enhance the balance between the treatment- and control group (Austin, 2011b; Austin and Small, 2014). This has been agreed upon in extant literature to be a sweet spot for enhanced matching results. However, this leaves some of the observations in the treatment group unmatched. Consequently, the estimate of the treatment effect no longer one-to-one corresponds to the ATT (Greifer, 2022, 2023; Mahmood, 2018). In our research, we still refer to the treatment estimate as the ATT for simplicity. In future research, one could examine the difference between a treatment estimate with unmatched or discarded treated observations and one without unmatched or discarded treated observations. This implies the difference between a treatment estimate using a caliper (or other balancing tools like common support restriction) and the ATT. The difference between the two can be explored and properly defined to prevent misconceptions in the future.

For further research, it would also be interesting to see how this framework performs on other observational studies with more covariates and more complex functional forms. Additionally, even though we do not find

cluster-based heterogeneity using this dataset, this framework could be applied to other datasets in which signs of heterogeneity are present. Athey and Wager (2019), Carnegie et al. (2019), and many other researchers do not find heterogeneous treatment effects either using this dataset. Therefore, applying this framework to other datasets could still provide a transparent way to expose cluster-based heterogeneity. At the same time, in case heterogeneity has been indicated, this empirics-first multi-step framework can be applied to get further familiarized with the roots of the indicated heterogeneity.

References

Abadie, A. and Spiess, J. (2022). Robust post-matching inference. *Journal of the American Statistical Association*, 117(538):983–995.

Apon, A., Robinson, F., Brewer, D., Dowdy, L., Hoffman, D., and Lu, B. (2006). Initial starting point analysis for k-means clustering: A case study.

Ascarza, E. (2018). Retention futility: Targeting high-risk customers might be ineffective. *Journal of Marketing Research*, 55(1):80–98.

Athey, S. and Imbens, G. W. (2017). The state of applied econometrics: Causality and policy evaluation. *Journal of Economic perspectives*, 31(2):3–32.

Athey, S., Tibshirani, J., and Wager, S. (2019). Generalized random forests.

Athey, S. and Wager, S. (2019). Estimating treatment effects with causal forests: An application. *Observational studies*, 5(2):37–51.

Austin, P. C. (2007). Propensity-score matching in the cardiovascular surgery literature from 2004 to 2006: a systematic review and suggestions for improvement. *The Journal of thoracic and cardiovascular surgery*, 134(5):1128–1135.

Austin, P. C. (2009). Type i error rates, coverage of confidence intervals, and variance estimation in propensity-score matched analyses. *The international journal of biostatistics*, 5(1).

Austin, P. C. (2011a). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate behavioral research*, 46(3):399–424.

Austin, P. C. (2011b). Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies. *Pharmaceutical statistics*, 10(2):150–161.

Austin, P. C. (2012). Using ensemble-based methods for directly estimating causal effects: an investigation of tree-based g-computation. *Multivariate behavioral research*, 47(1):115–135.

Austin, P. C. (2013). The performance of different propensity score methods for estimating marginal hazard ratios. *Statistics in medicine*, 32(16):2837–2849.

Austin, P. C. (2014a). A comparison of 12 algorithms for matching on the propensity score. *Statistics in medicine*, 33(6):1057–1069.

Austin, P. C. (2014b). The use of propensity score methods with survival or time-to-event outcomes: reporting measures of effect similar to those used in randomized experiments. *Statistics in medicine*, 33(7):1242–1258.

Austin, P. C. and Mamdani, M. M. (2006). A comparison of propensity score methods: a case-study estimating the effectiveness of post-ami statin use. *Statistics in medicine*, 25(12):2084–2106.

Austin, P. C. and Small, D. S. (2014). The use of bootstrapping when using propensity-score matching without replacement: a simulation study. *Statistics in medicine*, 33(24):4306–4319.

Austin, P. C. and Stuart, E. A. (2015). Moving towards best practice when using inverse probability of treatment weighting (iptw) using the propensity score to estimate causal treatment effects in observational studies. *Statistics in medicine*, 34(28):3661–3679.

Bollen, K. A. and Pearl, J. (2013). Eight myths about causality and structural equation models. In *Handbook of causal analysis for social research*, pages 301–328. Springer.

Carnegie, N., Dorie, V., and Hill, J. L. (2019). Examining treatment effect heterogeneity using bart. *Observational Studies*, 5(2):52–70.

Carvalho, C., Feller, A., Murray, J., Woody, S., and Yeager, D. (2019). Assessing treatment effect variation in observational studies: Results from a data challenge. *Observational Studies*, 5(2):21–35.

Djurisic, S., Rath, A., Gaber, S., Garattini, S., Bertele, V., Ngwabyt, S. N., et al. (2017). Barriers to the conduct of randomised clinical trials within all disease areas. *Trials*, 18:1–10.

Drake, C. (1993). Effects of misspecification of the propensity score on estimators of treatment effect. *Biometrics*, pages 1231–1236.

Duley, L., Antman, E., Arena, J., Avezum, A., Blumenthal, M., Bosch, J., and Yusuf, S. (2008). Specific barriers to the conduct of randomized trials. *Clinical Trials*, 5(1):40–48.

Elwert, F. and Winship, C. (2010). Effect heterogeneity and bias in main-effects-only regression models. In *Heuristics, probability and causality: A tribute to Judea Pearl*, pages 327–336.

Gayat, E., Resche-Rigon, M., Mary, J. Y., and Porcher, R. (2012). Propensity score applied to survival data analysis through proportional hazards models: a monte carlo study. *Pharmaceutical statistics*, 11(3):222–229.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (1995). *Bayesian data analysis*. Chapman and Hall/CRC.

Golder, P. N., Dekimpe, M. G., An, J. T., van Heerde, H. J., Kim, D. S., and Alba, J. W. (2023). Learning from data: An empirics-first approach to relevant knowledge generation. *Journal of Marketing*, 87(3):319–336.

Greifer, N. (2022). Estimating effects after matching. MatchIt: Nonparametric Preprocessing for Parametric Causal Inference. Available online: https://cran.r-project.org/web/packages/MatchIt/vignettes/estimating-effects.html.

Greifer, N. (2023). Matching methods. https://cran.r-project.org/web/packages/MatchIt/vignettes/matching-methods.html.

Gu, X. S. and Rosenbaum, P. R. (1993). Comparison of multivariate matching methods: Structures, distances, and algorithms. *Journal of Computational and Graphical Statistics*, 2(4):405–420.

Guadalupe, M. (2018). Why firms should conduct randomized controlled trials. INSEAD Knowledge.

Hartigan, J. A. and Wong, M. A. (1979). Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):100–108.

Heckman, J. and Singer, B. (1984). A method for minimizing the impact of distributional assumptions in econometric models for duration data. *Econometrica: Journal of the Econometric Society*, pages 271–320.

Heckman, J. J. (2005). The scientific model of causality. *Sociological methodology*, 35(1):1–97.

Hedges, L. V. (1982). Fitting categorical models to effect sizes from a series of experiments. *Journal of Educational Statistics*, 7(2):119–137.

Hirano, K., Imbens, G. W., and Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71(4):1161–1189.

Ho, D. E., Imai, K., King, G., and Stuart, E. A. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political analysis*, 15(3):199–236.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, 81(396):945–960.

Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics*, 86(1):4–29.

Imbens, G. W. and Angrist, J. D. (1994). Identification and estimation of local average treatment effects. *Econometrica: journal of the Econometric Society*, pages 467–475.

Jain, A. K. and Dubes, R. C. (1988). *Algorithms for clustering data*. Prentice-Hall, Inc.

Jain, A. K., Murty, M. N., and Flynn, P. J. (1999). Data clustering: a review. *ACM computing surveys (CSUR)*, 31(3):264–323.

John, L. K., Loewenstein, G., and Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological science*, 23(5):524–532.

Kaufman, L. and Rousseeuw, P. J. (2009). *Finding groups in data: an introduction to cluster analysis*. John Wiley Sons.

Keele, L. (2010). An overview of rbounds: An r package for rosenbaum bounds sensitivity analysis with matched data. *White Paper. Columbus, OH, 1, 15*.

Kerr, N. L. (1998). Harking: Hypothesizing after the results are known. *Personality and social psychology review*, 2(3):196–217.

Kodinariya, T. M. and Makwana, P. R. (2013). Review on determining the number of clusters in k-means clustering. *International Journal*, 1(6):90–95.

Lee, B. K., Lessler, J., and Stuart, E. A. (2010). Improving propensity score weighting using machine learning. *Statistics in medicine*, 29(3):337–346.

Lemenkova, P. (2019). K-means clustering in r libraries cluster and factoextra for grouping oceanographic data. *International Journal of Informatics and Applied Mathematics*, 2(1):1–26.

Liang, K.-Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22.

Likas, A., Vlassis, N., and Verbeek, J. J. (2003). The global k-means clustering algorithm. *Pattern recognition*, 36(2):451–461.

Lin, D. Y. and Wei, L. J. (1989). The robust inference for the cox proportional hazards model. *Journal of the American statistical Association*, 84(408):1074–1078.

Mahmood, S. (2018). The performance of largest caliper matching: A monte carlo simulation approach. *arXiv preprint arXiv:1806.02149*.

Mansournia, M. A. and Altman, D. G. (2016). Inverse probability weighting. *Bmj*, 352.

McCaffrey, D. F., Griffin, B. A., Almirall, D., Slaughter, M. E., Ramchand, R., and Burgette, L. F. (2013). A tutorial on propensity score estimation for multiple treatments using generalized boosted models. *Statistics in medicine*, 32(19):3388–3414.

McCaffrey, D. F., Ridgeway, G., and Morral, A. R. (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological methods*, 9(4):403.

McDonald, A. M., Treweek, S., Shakur, H., Free, C., Knight, R., Speed, C., and Campbell, M. K. (2011). Using a business model approach and marketing techniques for recruitment to clinical trials. *Trials*, 12(1):1–12.

Milligan, G. W. and Cooper, M. C. (1988). A study of standardization of variables in cluster analysis. *Journal of classification*, 5:181–204.

Morgan, S. L. and Harding, D. J. (2006). Matching estimators of causal effects: Prospects and pitfalls in theory and practice. *Sociological methods & research*, 35(1):3–60.

Network, S. E. R. (2023). National study of learning mindsets - student experience research network.

Normand, S.-L. T., Landrum, M. B., Guadagnoli, E., Ayanian, J. Z., Ryan, T. J., Cleary, P. D., and McNeil, B. J. (2001). Validating recommendations for coronary angiography following acute myocardial infarction in the elderly: a matched analysis using propensity scores. *Journal of clinical epidemiology*, 54(4):387–398.

Pearl, J. (2009). *Causality*. Cambridge university press.

Peña, J. M., Lozano, J. A., and Larranaga, P. (1999). An empirical comparison of four initialization methods for the k-means algorithm. *Pattern recognition letters*, 20(10):1027–1040.

Prasad, A. (2023). What's up with our obsession with the theoretical contribution: A means to an end or an end in and of itself? *Organization*.

Punj, G. and Stewart, D. W. (1983). Cluster analysis in marketing research: Review and suggestions for application. *Journal of marketing research*, 20(2):134–148.

Qualtrics (2023). Causal research: Definition, examples and how to use it.

Ridgeway, G., McCaffrey, D. F., Morral, A. R., Cefalu, M., Burgette, L., Pane, J., and Griffin, B. A. (2021). Toolkit for weighting and analysis of nonequivalent groups: A guide to the twang package. *Vignette, July, 26*.

Ridgeway, G., McCaffrey, D. F., Morral, A. R., Cefalu, M., Burgette, L. F., Pane, J. D., and Griffin, B. A. (2022). *Toolkit for weighting and analysis of nonequivalent groups: a tutorial for the R TWANG package*. Santa Monica, Calif: Rand.

Rosenbaum, P. R. (1987). Model-based direct adjustment. *Journal of the American statistical Association*, 82(398):387–394.

Rosenbaum, P. R. (2002). *Overt bias in observational studies*. Springer New York.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688.

Schwarz, G. M. (2023). Down the garden path: Modelling theory won't fix the theory crisis. *Organization Studies*.

Sinaga, K. and Yang, M.-S. (2020). Unsupervised k-means clustering algorithm. *IEEE access*, 8:80716–80727.

Snowden, J. M., Rose, S., and Mortimer, K. M. (2011). Implementation of g-computation on a simulated data set: demonstration of a causal inference technique. *American journal of epidemiology*, 173(7):731–738.

Stoltzfus, J. C. (2011). Logistic regression: a brief primer. *Academic emergency medicine*, 18(10):1099–1104.

Wager, S. and Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242.

Wan, F. (2019). Matched or unmatched analyses with propensity-score–matched data? *Statistics in medicine*, 38(2):289–300.

Wendling, T., Jung, K., Callahan, A., Schuler, A., Shah, N. H., and Gallego, B. (2018). Comparing methods for estimation of heterogeneous treatment effects using observational data from health care databases. *Statistics in medicine*, 37(23):3309–3324.

Westreich, D. and Greenland, S. (2013). The table 2 fallacy: presenting and interpreting confounder and modifier coefficients. *American journal of epidemiology*, 177(4):292–298.

Westreich, D., Lessler, J., and Funk, M. J. (2010). Propensity score estimation: machine learning and classification methods as alternatives to logistic regression. *Journal of clinical epidemiology*, 63(8).

Winship, C. and Morgan, S. L. (1999). The estimation of causal effects from observational data. *Annual review of sociology*, 25(1):659–706.

Xie, Y., Brand, J. E., and Jann, B. (2012). Estimating heterogeneous treatment effects with observational data. *Sociological methodology*, 42(1):314–347.

Zhao, Q., Luo, J., Su, Y., Zhang, Y., Tu, G., and Luo, Z. (2021). Propensity score matching with r: conventional methods and new features. *Annals of translational medicine*, 9(9).

A Additional NSLM program details

The details outlined in this section are extracted from the official studentexperiencenetwork.org website. The National Study of Learning Mindsets was an initiative started in 2013 by leading experts on mindset science in anticipation of the novel founding of Carol Dweck. This professor/researcher found that people with a growth mindset instead of a fixed mindset, have relatively higher academic performance. To further investigate the relationship between growth mindset and academic performance and its subsequent implications, the NSLM study was invented. The purpose of this study was to better understand heterogeneity effects regarding social psychological interventions focused on education. This turned out to be the largest-ever controlled trial of a growth mindset program in the US in K-12 settings. The primary goal while designing this experiment was to make the sample nationally representative of regular US high schools. This has been accomplished by random sampling of schools, which makes the sample nationally representative. This design allows for studying effect modifiers, heterogeneity, and much more and generalize based on these findings.

A.1 Selection of schools

The sample selection was performed using the US Education Department's Common Core of Data (NCES). Some schools were intentionally left-out, such as private schools, schools that were too small, and schools with specialized missions. 139 public high schools in the US were invited to participate in the program. Of these 139 schools, 76 schools agreed to participate, of which 65 schools provided all the requested records (both survey data and administrative records) and 11 schools only provided the survey data. The survey data consisted of background information, beliefs, and classroom and school contexts. The administrative records consisted of student-level academic performance and demographics.

A.2 So, what does the program itself look like?

The program consists of an online growth mindset training randomly sampled across 16,000 9th graders across the 76 participating public US high schools. The experiment consists of random assignment to either an online growth mindset training or a control activity for two 25-minute sessions. Treated subjects were exposed to reading about and listening to scientific evidence on brain activity and the possibility to grow intellectual abilities over time. Consequently, students were encouraged to consider reasons for growing their own brain to improve things that matter to them. This could be friends, family, community, or just anything. Subsequently, students were asked to write a short letter about the transition to high school for future 9th graders. This stimulated active learning by applying the ideas and information that they just received. After the 16,000 9th graders participated in either the treatment- or control group of the growth mindset program, the experiment was still far from over. The students were followed throughout the rest of their academic trajectory to examine the short- and long-term impacts on academic performance and life outcomes. This means that a longitudinal dataset is generated which follows students throughout their academic life and careers.

B Details on the analysis

B.1 Propensity score estimation

The GBM model can be applied using the gbm package in the R environment. The gbm package allows for random subsampling as well as optimizing the number of iterations that minimizes the standardized difference (McCaffrey et al., 2004). The ps function from the gbm package can be used to estimate the propensity scores. The only data preperation that is needed for the GBM propensity score estimation is to transform the categorical variables into factors and change the treatment variable into an integer. The ps function estimates propensity scores and can only handle an outcome variable with type integer for its estimation. We specify that the treatment needs to be estimated by all other covariates, similar as when fitting linear models.

The parameters n.trees, interaction.depth, and shrinkage must be pre-specified. N.trees represents the maximum number of iterations that the GBM algorithm will run (Ridgeway et al., 2022). Shrinkage controls the amount of shrinkage with small values below 0.01 yielding smooth fits. However, these fits are only adequate with greater values of n.trees. The ps function estimates the optimal number of iterations and issues a warning if this number is too close to the bounds selected as n.trees. This indicates that balance can potentially be improved by using more complex models. For this, either n.trees should be increased, or shrinkage should be decreased, which means higher values for the shrinkage argument (Ridgeway et al., 2022). We will start increasing the n.trees parameter if the warning appears and will subsequently alter shrinkage to obtain the optimal set of parameters. Additionally. The interaction.depth controls the depth of the trees which is the number of interactions allowed per iteration (Ridgeway et al., 2022).

The hyperparameters are tuned by minimizing the mean standardized effect size. Subsequently, the balance measures can be plotted against the number of iterations. The optimal number of iterations is extracted from the lowest point in the plot. Moreover, the estimand argument is used to differentiate between estimating the average treatment effect (ATE) or the average treatment effect on the treated (ATT). Since our analysis incorporates propensity score matching, discarding observations from the control group, we are interested in calculating the ATT. Moreover, the stop.method argument requires specification on which stopping conditions should be used for assessing balance across pretreatment covariates. For this, the mean of the standardized effect size (es.mean) and the maximum KS statistics will be used (ks.max). The es/ks part refers to the covariate balance metrics and the mean/max part refers to the method for summarizing across balance metrics (Ridgeway et al., 2022). So, es.mean uses the standardized effect size to assess balance and summarizes this across covariates with the mean. Meanwhile, ks.max uses the KS statistics to balance the covariates and uses the maximum across covariates to summarize. Consequently, the algorithm selects the optimal number of iterations based on minimizing the differences between the control- and treatment group using these stopping conditions (Ridgeway et al., 2022).

B.2 K-means clustering

Before apply k-means clustering, the school-level covariates will be isolated since we are interested in school-level based clusters. Moreover, one-hot encoding will be used to cluster the data, since k-means clustering is

not able to handle categorical variables. In fact, the rest of the analysis requires one-hot encoded variables either way. K-means clustering will be applied using the kmeans function from the stats R package. Before using k-means clustering, we need to determine the optimal number of clusters using the Elbow- and Silhouette method. The fviz_nbclust function from the factoextra package in R will be used to apply the Elbow- and Silhouette method.

For the Elbow method, the method "wss" needs to be specified, which means 'within sum of squares'. Based on this plotted within sum of squares, the Elbow method can be applied for the first 10 clusters (Kassambara, 2018). For the Silhouette method, the method "silhouette" needs to be specified which calculates the loss function for the first 10 clusters. The fviz_nbclust function provides visualizations from which the optimal number of clusters can be extracted. Subsequently, the kmeans function will be used to perform the cluster analysis. The number of centroids will be specified based on the Elbow- and Silhouette method and the number of different starting points will be specified to be 25 (Sinaga and Yang, 2020; Lemenkova, 2019; Apon et al., 2006).

B.3 Propensity score matching

The machit function from the MatchIt package will be used to facilitate the propensity score matching process using both optimal matching and greedy (caliper) matching. The use of optimal matching will be specified by including "optimal" as the method. Moreover, greedy (caliper) matching will be specified by including "nearest" as the method. Moreoever, we specificy a caliper of 0.2 since this has been proven to mimimize the MSE of the ATT estimation (Austin, 2011b,a, 2014a; Zhao et al., 2021). Traditionally, GBM cannot be used in combination with the matchit function. The reason being that the matchit function only allows propensity score calculation using logistic regression, generalized additive models, CART, and single-hidden-layer neural networks (Zhao et al., 2021). However, we will use the twang package to calculate the propensity scores using GBM beforehand. Subsequently, we will use these calculated propensity scores as input in the distance parameter in the matchit function. By doing so, we can effectively match observations across treatment- and control groups without having the ability to do so through standard models. We combine optimal matching and caliper matching with both the calculated propensity scores using es.mean and ks.max. This results in four different matching combinations; caliper matching with es.mean and ks.max, and optimal matching with es.mean and ks.max.

The combination and integration of GBM into propensity score matching potentially allows for an even more accurate propensity matching process, by leveraging the strengths of both GBM propensity score estimation and nearest neighbor or optimal matching. The outcomes of the matching procedure will be analyzed using balance visualizations which also allows for the comparison between optimal matching and greedy (caliper) matching. First of all, the propensity score distributions across treatment- and control group can be visualized by plotting the propensity score distribution in histograms for the before- and after matching situation. This provides a clear overview of how the propensity distribution changed due to matching. Furthermore, the bal.tab function from the cobalt package can be used to generate a balance table. This balance table shows the standardized difference for each covariate before- and after matching and allows for both comparisons and absolute judgments on balance. Another way to clearly visualize the

covariate balance is turning the covariate table into a love plot by using the love plot function. By using a combination of the abovementioned covariate visualizations, we can draw conclusions about the balancing performance.

B.4 Estimating ATT

First of all, to estimate the ATT based on the previous steps, we will fit a linear regression to the data, consisting of interaction effects between the treatment and all covariates. This model is used as the Q-model, as mentioned earlier. Subsequently, the avg_comparisons function from the marginaleffects package in R in will be used to perform the G-computation. This function estimates the cluster-robust SEs, treatment effect, and its confidence intervals. Note, the matched sample will be used as input since we already performed propensity score matching. The arguments needed for the avg_comparisons function are: the Q-model, the name of the treatment variable, subclass to request cluster-robust SEs, and the matched dataset only containing the treated observations to estimate the ATT. The weights do not have to be specified since our approach of propensity matching (one-to-one) ensures that each weight is equal to one. By applying the avg_comparisons approach as described above, we can effectively calculate the ATTs for each cluster with cluster-robust SEs and a confidence interval. By doing so, heterogeneity across clusters can be extracted and the significance of the ATT difference can be checked accordingly.

For estimating the ATT based on marginal effects using a linear model, referred to as Q-model, we need to ensure the absence of multicollinearity. Multicollinearity in linear models can cause issues since some variables (almost) perfectly predict each other. Therefore, we plot a correlation matrix using the corrplot function from the corplott R package which uses the Pearson method as the default. Figure 6 shows the correlation matrix for cluster 1 with the correlation between each of the variables. We observe that schoolID is highly correlated to most of the school-level variables, which causes issues while estimating the coefficients of the covariates. Furthermore, the XC.0-XC.4 variables are highly correlated to the other school- level variables. Additionally, XC.0-XC.4 are not present in each of the clusters, since it represents urbanicity and only some of the rural to urban categories are represented in each cluster. This also explains the question marks in Figure 6. Therefore, both schoolID and XC.0-XC.4 will be removed from the linear model. Moreover, the first dummy of every categorical variable will be deleted to be used as reference model (S3.1, C1.1, C2.1, C3.0). The rest of the covariates will be used to fit the linear model, including interaction terms with the treatment variable.

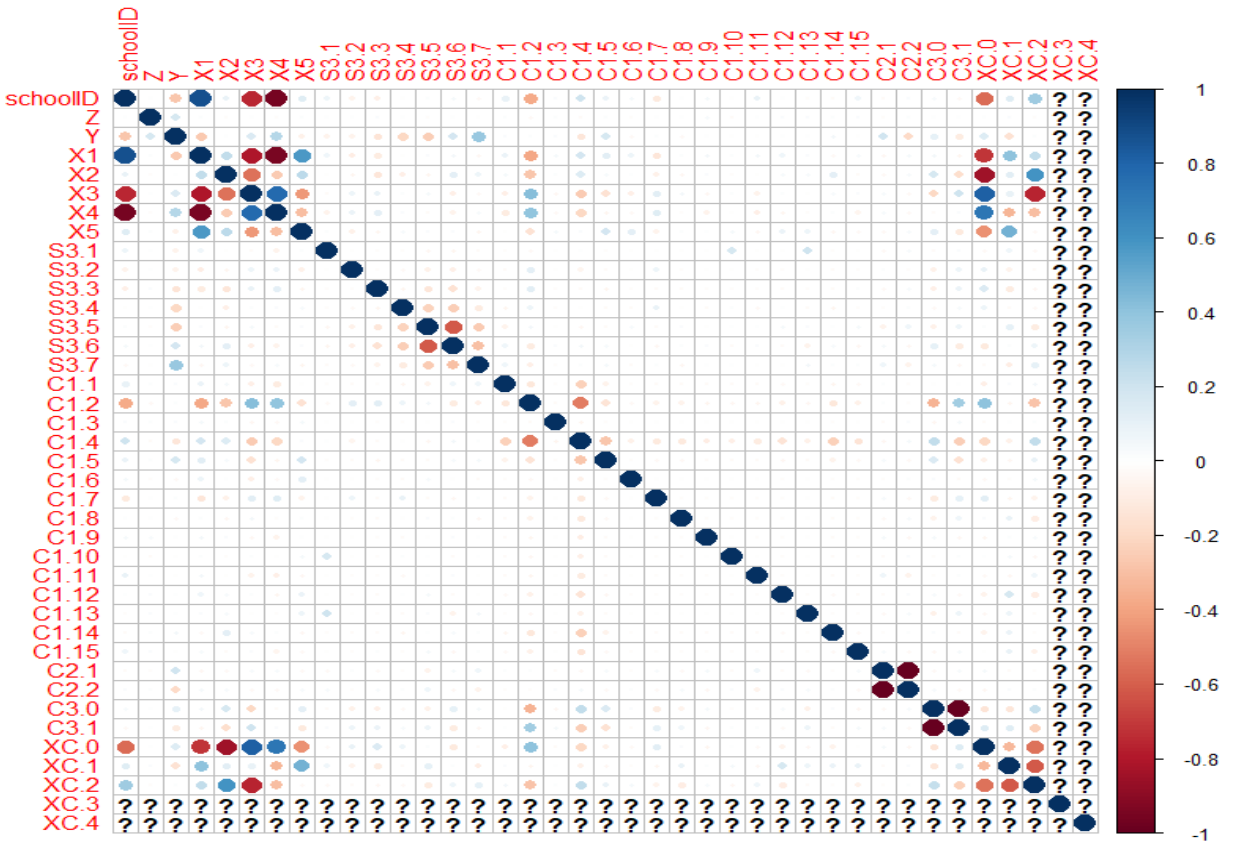


Figure 6: Correlation matrix of cluster 1

B.5 Sensitivity analysis

Rosenbaum's (2002) sensitivity test can be performed in R using the psens function from the rbounds package. Psens provides the bounds from the p-values of Wilcoxon's signed rank test for increasing values of Gamma (Zhao et al., 2021). So, for each cluster, psens provides the lower and upper bounds as well as the p-value for each specified value of Gamma, starting with Gamma is 1 (equal treatment odds). This calculation shows for which Gamma, the significance level exceeds the 0.05 threshold. In other words, for what Gamma the treatment effect is not significant anymore. More specifically, this shows sensitivity of the treatment effect on unobserved covariates by increasing the treatment odds for another observation with the same observed covariate distribution.

For cluster 1, this has been visualized in Table 10. The lower and upper bound represent the confidence interval of p-values for each Gamma. As we can see in Table 10, the upper bound of the confidence interval gradually increases. Our objective while performing this sensitivity analysis using psens is to determine for what Gamma the treatment effect loses its statistical significance. Table 10 shows that the upper bound of the p-value exceeds 0.05, which is the threshold, for Gamma = 1.7. This means that the significance of the treatment effect for cluster 1 is questionable when Gamma is 1.7 or higher. This is, when the odds of receiving treatment due to hidden bias are 1.7 times higher for one subject than for another subject with

the same covariate distribution. This also means that for a subject with 1.6 times higher odds of treatment, the treatment effect is still significant. This phenomenon captures the sensitivity of the treatment effect on hidden bias.

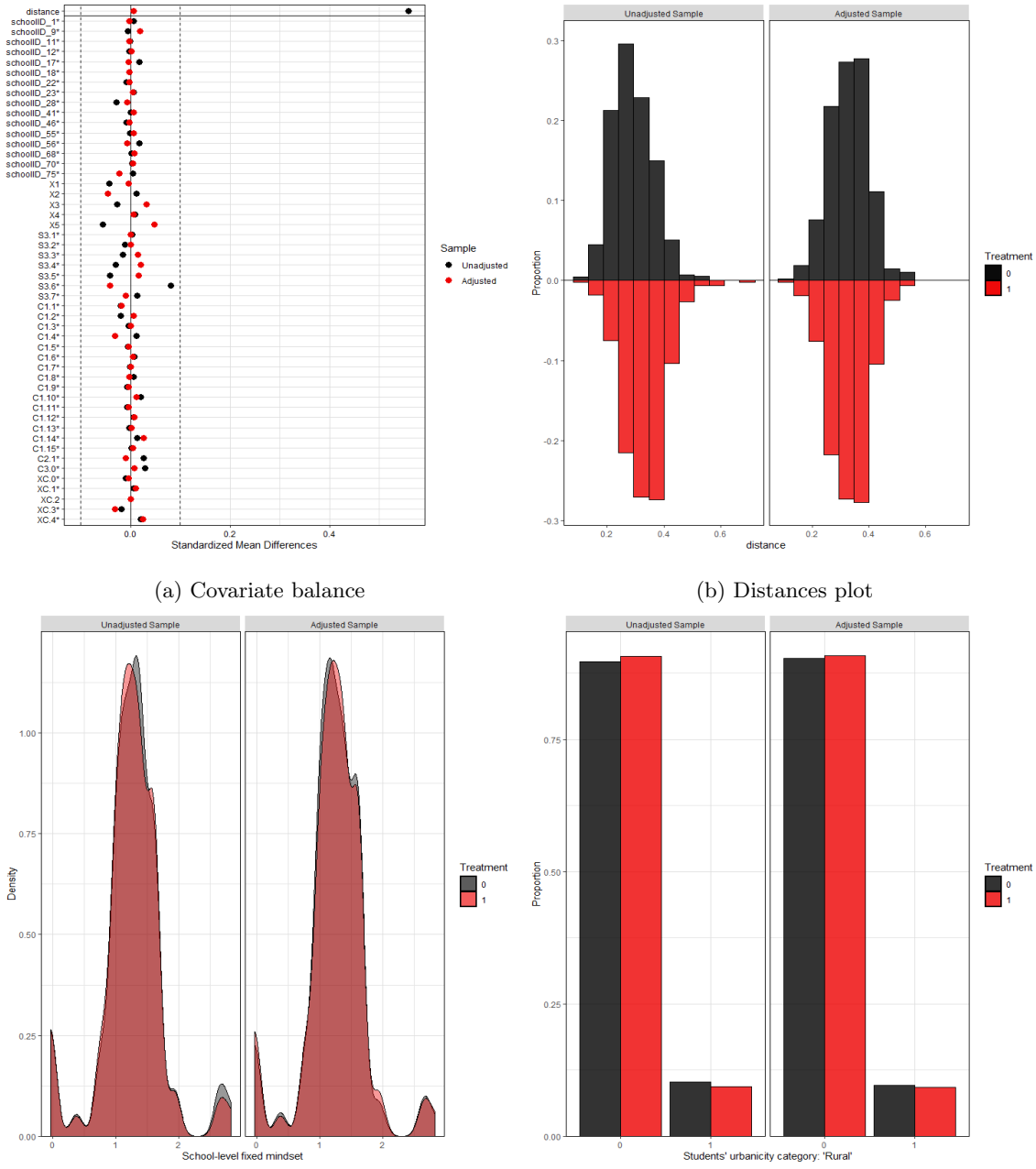
Table 11: Sensitivity Analysis Cluster 1 - Rosenbaum's Bounds

Gamma	Lower Bound	Upper Bound
1.0	0	0.0000
1.1	0	0.0000
1.2	0	0.0000
1.3	0	0.0001
1.4	0	0.0007
1.5	0	0.0054
1.6	0	0.0249
1.7	0	0.0791
1.8	0	0.1857
1.9	0	0.3427
2.0	0	0.5232

C Balance plots for all clusters

This section contains balance plots for the clusters 2-6, which have not been reported in the main text. We observe that Figures 8d and 11d, which represent the urbanicity category "Rural", are empty. The reason for this is that cluster 3 and 6 do not have any observations in the urbanicity category "Rural".

C.1 Cluster 2



(a) Covariate balance

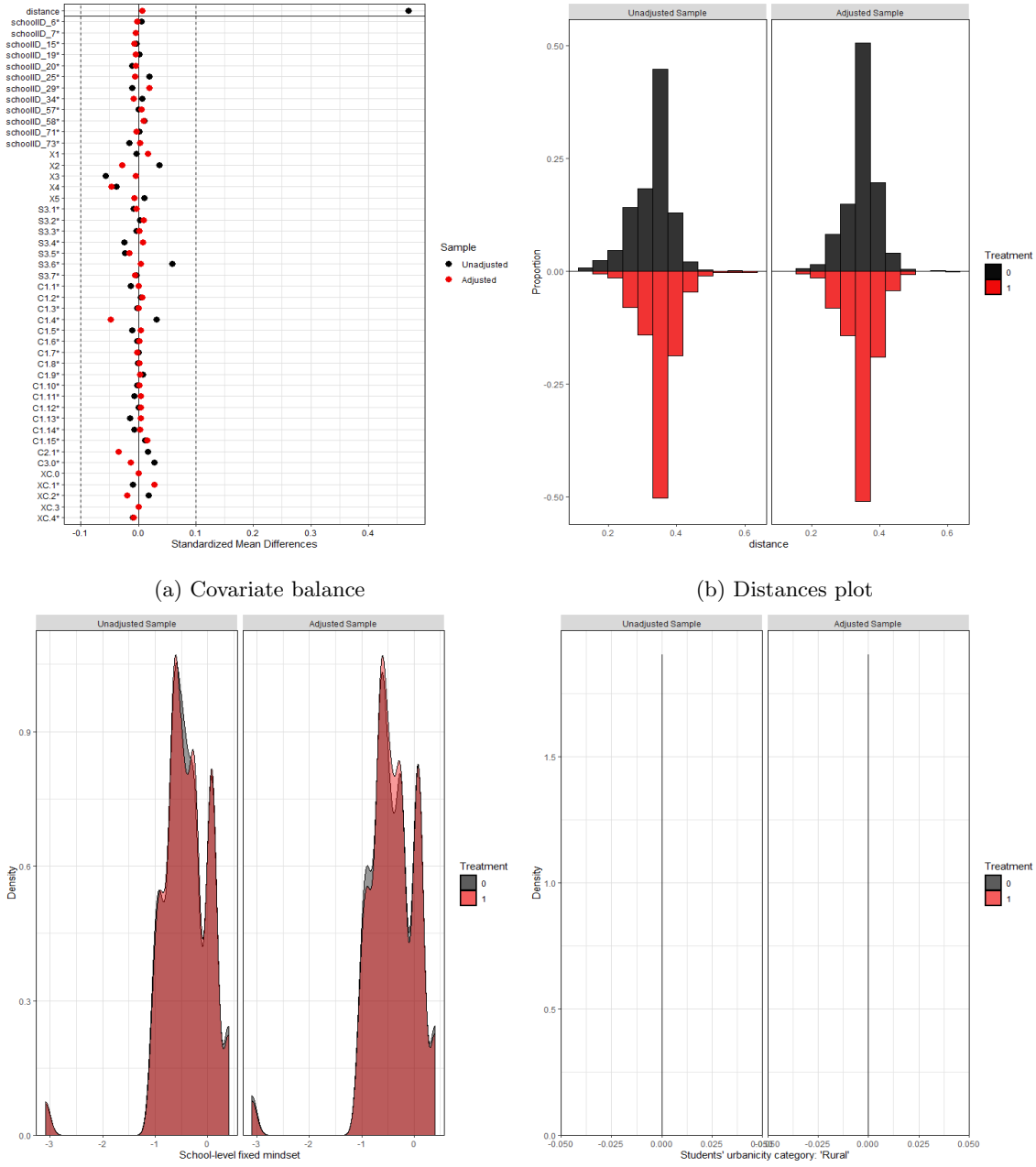
(b) Distances plot

(c) Balance of school-level fixed mindset

(d) Balance plot of urbanicity category "Rural"

Figure 7: Plots of the differences between the adjusted and unadjusted sample for cluster 1

C.2 Cluster 3



(a) Covariate balance

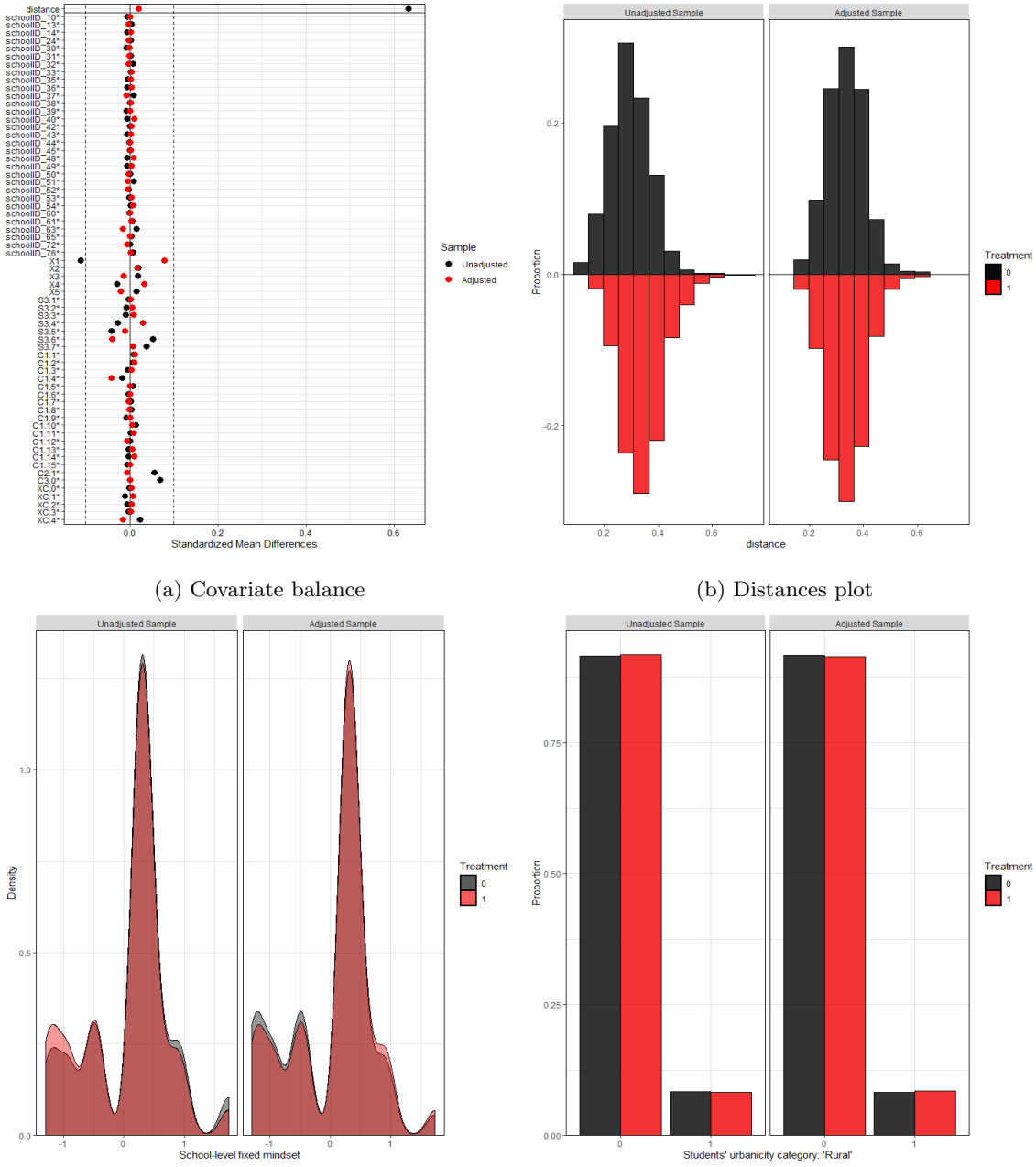
(b) Distances plot

(c) Balance of school-level fixed mindset

(d) Balance plot of urbanicity category "Rural"

Figure 8: Plots of the differences between the adjusted and unadjusted sample for cluster 1

C.3 Cluster 4



(a) Covariate balance

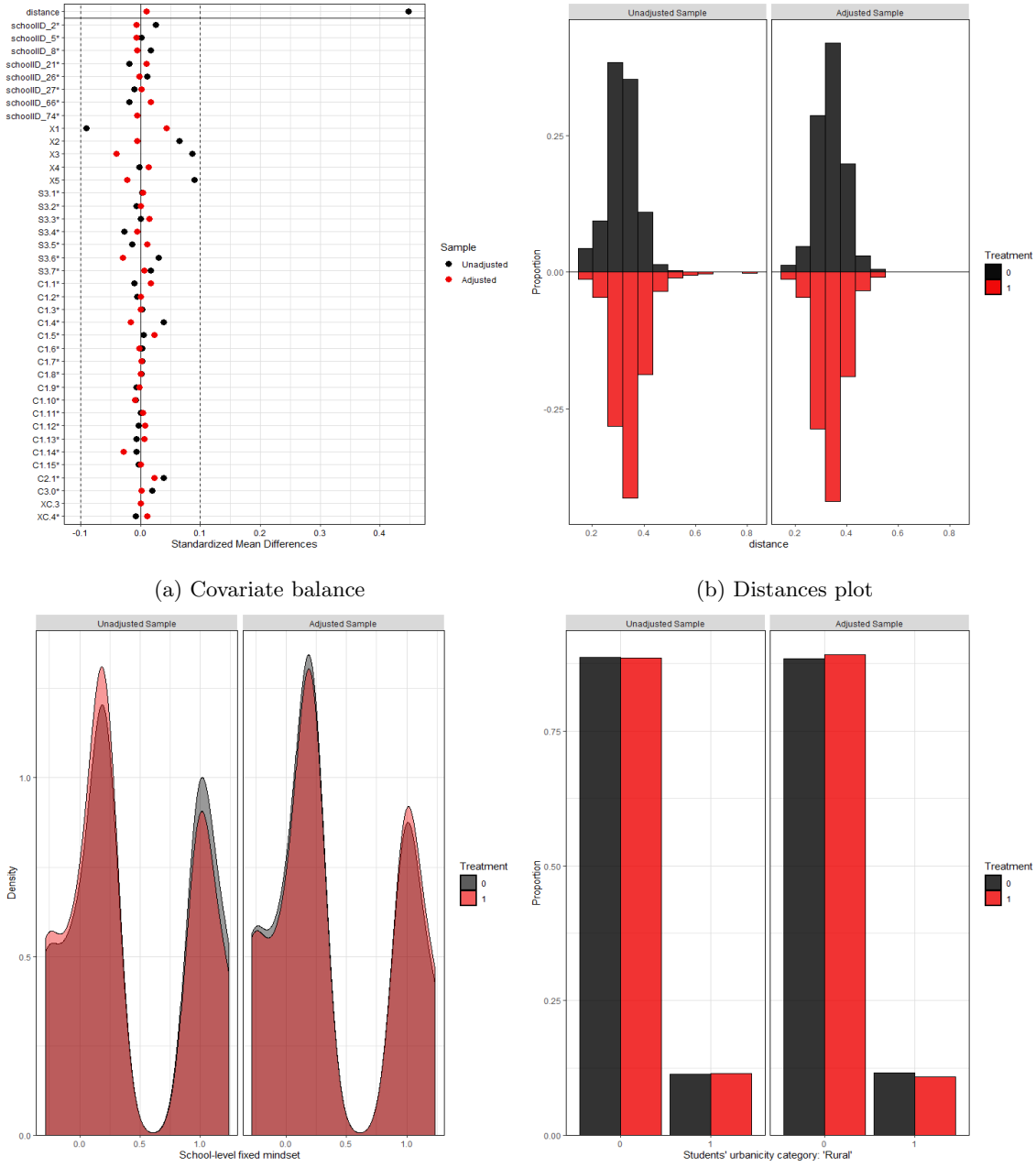
(b) Distances plot

(c) Balance of school-level fixed mindset

(d) Balance plot of urbanicity category "Rural"

Figure 9: Plots of the differences between the adjusted and unadjusted sample for cluster 1

C.4 Cluster 5



(a) Covariate balance

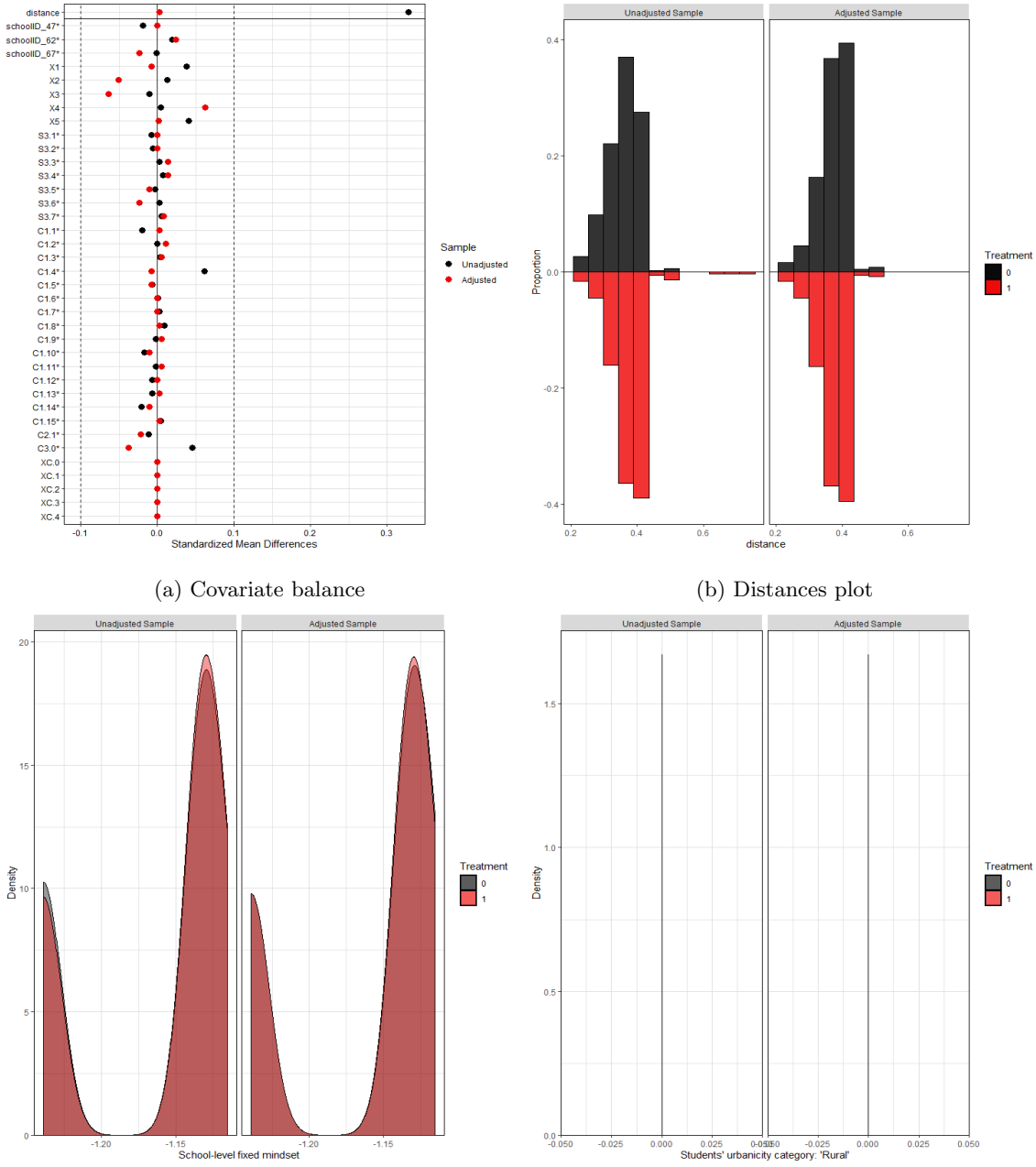
(b) Distances plot

(c) Balance of school-level fixed mindset

(d) Balance plot of urbanicity category "Rural"

Figure 10: Plots of the differences between the adjusted and unadjusted sample for cluster 1

C.5 Cluster 6



(a) Covariate balance

(b) Distances plot

(c) Balance of school-level fixed mindset

(d) Balance plot of urbanicity category "Rural"

Figure 11: Plots of the differences between the adjusted and unadjusted sample for cluster 1