UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

A NOVEL POST-HOC MATCHING PROCEDURE USING STATISTICAL LEARNING METHODS

A THESIS

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

Degree of

MASTER OF SCIENCE

By

OLIVIA PERRET
Norman, Oklahoma
2016

Downloaded from SHAREOK. Persistent Handle URL: https://hdl.handle.net/11244/31207

A NOVEL POST-HOC MATCHING PROCEDURE USING STATISTICAL LEARNING METHODS

A THESIS APPROVED FOR THE

SCHOOL OF INDUSTRIAL AND SYSTEMS ENGINEERING

BY

_____

Dr. Charles Nicholson, Chair

_____

Dr. Suleyman Karabuk

_____

Dr. F. Hank Grant

© Copyright by OLIVIA PERRET 2016
ALL RIGHTS RESERVED.

Acknowledgments

First, I would like to express my sincere gratitude to my advisor, Dr. Charles Nicholson, for all his support, knowledge, advice, patience and for providing me with a great work environment to do my research. I could not have imagined having a better advisor and mentor for my Master's thesis. I appreciate the discussions on random topics before starting the serious work. He has taught me a lot over the last year and a half, has encouraged me the whole time and made me feel like I was doing great even when I thought the opposite. I would also like to thank him for his trust to have me as his TA for the graduate course "Intelligent Data Analytics", from which I have learned a lot. It was a great new experience.

I want to thank my family not only for their financial support, but most importantly for their emotional support that kept me going and made this experience possible.

I would also like to thank my friends for making this experience amazing and unforgettable; Alison Jalanti for being my workout partner and a great friend; Vera Wendler Bosco and Mikael Perrin for being amazing friends and always here for me to give me advice; and Cyril Beyney, with whom I have spent every day at our lab and who made this experience more enjoyable.

Thank you to the faculty and staff for all their help during my time at the University of Oklahoma, and a special thanks to Jennifer Covington.

Contents

1 Background **1**
1.1 Introduction 1
1.2 Literature review 2

2 Methods **6**
2.1 Estimating the propensity score using random forest 8
2.2 The proximity matrix method 10
2.3 R implementation 12

3 Experimentation **17**
3.1 Generated data 17
3.2 Random forest specifications 20
3.3 Different experiments according to the models 23

4 Results **25**

5 Conclusion **42**

References **44**

List of Tables

2.1 Example of the proximity matrix 12

3.1 Summary of the simulation study 24

4.1 Percentage of success, 600 obs 26
4.2 Percentage of success, 2,000 obs 26
4.3 Average treatment effect statistics: full sample, 600 observations . 28
4.4 Average treatment effect statistics: out-of-bag and bag data, 600 observations 29
4.5 Average treatment effect statistics: full sample, 2,000 observations 30
4.6 Average treatment effect statistics: out-of-bag and bag data, 2,000 observations 31
4.7 Summary of all methods, mean, 600 observations 40
4.8 Summary of all methods, mean, 2,000 observations 41

List of Figures

2.1 Overview of process 6
2.2 Classification tree 8

4.1 Average treatment effect density: full sample, 600 observations . . 35
4.2 Average treatment effect density: out-of-bag and bag data, 600 observations 36
4.3 Average treatment effect density: full sample, 2,000 observations . 37
4.4 Average treatment effect density: out-of-bag and bag data, 2,000 observations 38

Abstract

In this thesis, a statistical learning method is leveraged to create a novel measure for conducting post-hoc matching between a treatment group and a candidate set. Post-hoc matching is a necessary element in many non-random observational studies and arises in diverse fields such as economics, medicine, marketing, and others.

Post-hoc matching has been in use for many years and different methods have been used. A common measure to match the two groups, called the propensity score, can be estimated in a variety of ways. A recent method to estimate it was introduced in 2013 using random forests.

The method introduced in this work utilizes random forest to develop an alternative measure to the propensity score. The new measure, proximity matrix method, is intuitive and potentially captures more similarities between subjects. In order to compare the propensity score method with the novel post-hoc matching method, data sets are generated which logically reflect observational studies with various assumptions regarding treatment selection. Experiments are conducted to evaluate the average treatment effect between the treatment and the control group that are matched. The empirical analysis shows promising results for the proximity matrix method. In particular, the technique has superior results

when the treatment selection is made using complex rules, namely, a non-linear model, and when the bag is used to estimate the proximity matrix.

This study demonstrates significant potential of the novel method for both researchers and practitioners interested in matching candidates to a test set to estimate the average treatment effect within an observational study when there is an unknown, and possibly complex multivariate relationship with the initial treatment selection.

Chapter 1

Background

1.1. Introduction

In observational (non-randomized) studies, the goal is often to determine the effect of treatment on a test group. Applications arise in various fields and for different reasons and often are performed on people or animals (e.g., drug trials, marketing promotions). For convenience, the term "subject" will be used when referring to test, control, and candidate populations; however, the application of the work in this study is broadly inclusive of the type of objects included any observational study. In order to estimate the effect of a treatment on a group, a comparison between subjects exposed to a treatment and subjects in a control group needs to be made. In randomized studies, the assignment of subjects to a treatment group or control group is random and therefore the groups are statistically similar. In observational studies, the assignment rule of subjects to the treatment group is unknown. Therefore, the distribution of pre-treatment covariates can be different between the groups (treatment and non-treatment)

and comparisons may be misleading since their differences cannot be attributed to the treatment effect. A primary issue for observational studies is to determine an unbiased treatment effect when a randomized control group is non-existent.

The main problem is to find a subset of the candidate pool (non-treatment group) similar to the treatment group. Post-hoc matching is one approach to address this issue. A number of studies relating to post-hoc matching has been performed in a variety of areas, such as law (Epstein et al. 2005), economics (Abadie and Imbens 2006), statistics (Rosenbaum 2002, Rubin 2006), medicine (Rubin 1997), political science (Herron and Wand 2007) and others. The selected subset will become the control for the treatment group and allow for a better estimation of the treatment effect.

1.2. Literature review

For effective matching, there need to be some common support between the two groups, or else the matching will be done based on subjects that don't have any values in covariates in common. Common support between the treatment and the control group is defined as an overlap of every of the covariate distributions.

The matching procedure has existed for around 70 years but an actual technique was not developed until Cochran and Rubin (1973) and Rubin (1973). At this time, data sets with only one covariate were used. If there was more than one covariate, it was mostly a computational problem because it was harder to find a good match where all the covariates would have a close value between a treatment and a candidate. They used "nearest available" matching method by ordering treatment subjects randomly and then picking the closest subject (using

the Mahalanobis distance) to assign a match. The Mahalanobis equation is as follows:

$$D_{ij} = \sqrt{(X_i - X_j)^T C^{-1} (X_i - X_j)}$$

where X_i and X_j are the covariates for units i and j, C^{-1} is the inverse sample covariance matrix of X and T is the matrix transpose.

Ideally, we would want to have the same distributions of X (the covariates) in both groups, which means that each treated subject would be exactly matched on all of their covariates to a corresponding control. Rosenbaum and Rubin (1983) showed that matching on a balancing score is sufficient. In 1983, a new method was introduced: the propensity score. The propensity score was defined as the conditional probability of a subject being assigned to the treatment group given a set of covariates, that is

$$e(X) = Pr(Z = 1|X) \tag{1.1}$$

where X is the set of covariates for a subject and Z the binary treatment variable whether the subject was treated ($Z = 1$) or not ($Z = 0$). Subjects matched according to their propensity score will then have the same distribution. The true value of the propensity score cannot be known and must be estimated from the available data. It is estimated by a logistic regression with the treatment as the dependent variable and the covariates as the independent variables.

The optimal matching was introduced by Rosenbaum (1989), using distances to match the subjects. Distances can be defined in many ways that relates the covariates. In their research, two covariates are used to define the distance, where their values are replaced by their rank, and the distance between two

subjects is the sum of the two absolute differences in their ranks on the two covariates. Optimal matching is superior to greedy matching in that with greedy matching, the order in which the treated subjects are matched can change the quality of matches, but with the optimal matching this issue is avoided. With greedy matching, the lowest distance is considered first and so on until every treatment has a match, whereas with the optimal matching method, the overall sum of pair-wise distances is minimized. Although, the same controls are usually picked out with optimal matching, but the difference is that this technique does a better job at assigning each individual match to a treatment subject.

Another measure that can be use while matching is a caliper. This was introduced by Cochran and Rubin (1973). It is defined as a restricted subset of controls whose propensity score is within a specified amount of the treatment subject's propensity score. A caliper is used in order to avoid poor matches when matching is done without any restrictions. With this, the match subject will only be selected if it is within the caliper.

Among these methods, the propensity score is the one that is most commonly used to-date. When it was first designed, it was estimated via logistic regression. A variety of statistical learning methods are now being used, such as Classification and Regression Trees (Luellen et al. 2005), Random Forest (Lee et al. 2010), Neural Networks (Setoguchi et al. 2008), Generalized Boosted Modeling (McCaffrey et al. 2004) or even Support Vector Machines (Westreich et al. 2010).

A recent study from Cham (2013) showed how random forest (Breiman 2001), an ensemble learning method for classification regression and other task, was performing for estimating propensity score. With that technique, classification trees to predict the treatment group are built and the propensity score is average

over the multiple trees from the random forest from the terminal nodes.

Porro and Iacus (2009) used a different type of method than Cham (2013), called the Random Recursive Partitioning (RRP) method. This method is different in a way that regression trees are built and it is not to predict the treatment group but a fictitious response variable that is created for each new regression tree.

When building decision trees, such as with random forest, to evaluate the propensity score, some terminal nodes can have the same proportion of treated subjects, therefore the propensity score for all those subjects is the same even though their covariates were not the same since they were not in the same terminal node. Treated and candidates subjects could have a close propensity score and be matched on that and have completely different values from their covariates. The new method presented in this paper will take into account, before matching, the fact that two observations falls into the same node which means that they have similar covariates.

Chapter 2

Methods

This chapter describes existing methods to find control subjects to match to the treatment group. The novel approach is also detailed.

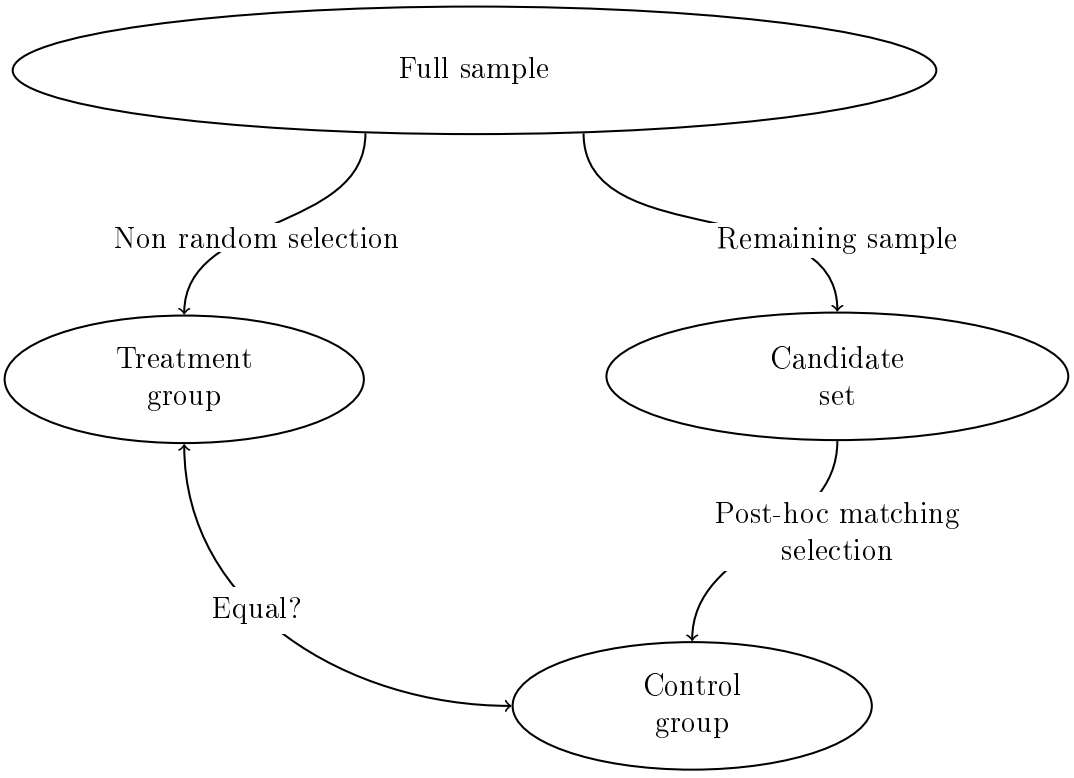

Figure 2.1: Overview of non-randomized study and post-hoc matching

All the methods follow the same type of process shown in Figure 2.1. First, the entire data set represents the full sample (FS). A non-random selection is then made from the full sample to define the treatment group. Non-random selection happens in different fields, such as economics, statistics, sociology, medicine and even law. For example in medicine, if the effect of exposure to a particular drug needs to be estimated and the subjects in the treatment group are typically older than the subjects not given the drug, the treatment assignment is not random. The remaining sample represents the candidate set. From this candidate set, a subset must be identified to match the treatment group. A variety of modeling techniques exist to enable this selection. This can be done with logistic regression, decision trees, or random forests, etc. Models may be built so as to produce a score to use for the next step called post-hoc matching. The matching process will find observations in the candidate set to match each observations in the treatment group based on the model score. After a control group is created from this step, a comparison between the two groups can be made in order to evaluate the matching process. That is, quantify relevant differences between the treatment group and the matched control group.

One well known approach for finding matches to the treatment group is to use the propensity score, the conditional probability listed in Equation 1.1. Over the years, multiple different methods have been used to estimate the propensity score. A common technique is to use logistic regression (Cox 1970) for the treatment Z,

$$\log\left(\frac{e\left(X\right)}{1-e\left(X\right)}\right)=\alpha+\beta^{T}f\left(X\right)$$

where α and β are parameters, $f(X)$ is a specified function and $e(X)$ is the propensity score.

Estimation of the propensity score using random forest is another possibility (Cham 2013), and will be detailed in Section 2.1. An alternative to the propensity score is the proximity matrix, a novel method, which will be introduced in Section 2.2.

2.1. Estimating the propensity score using random forest

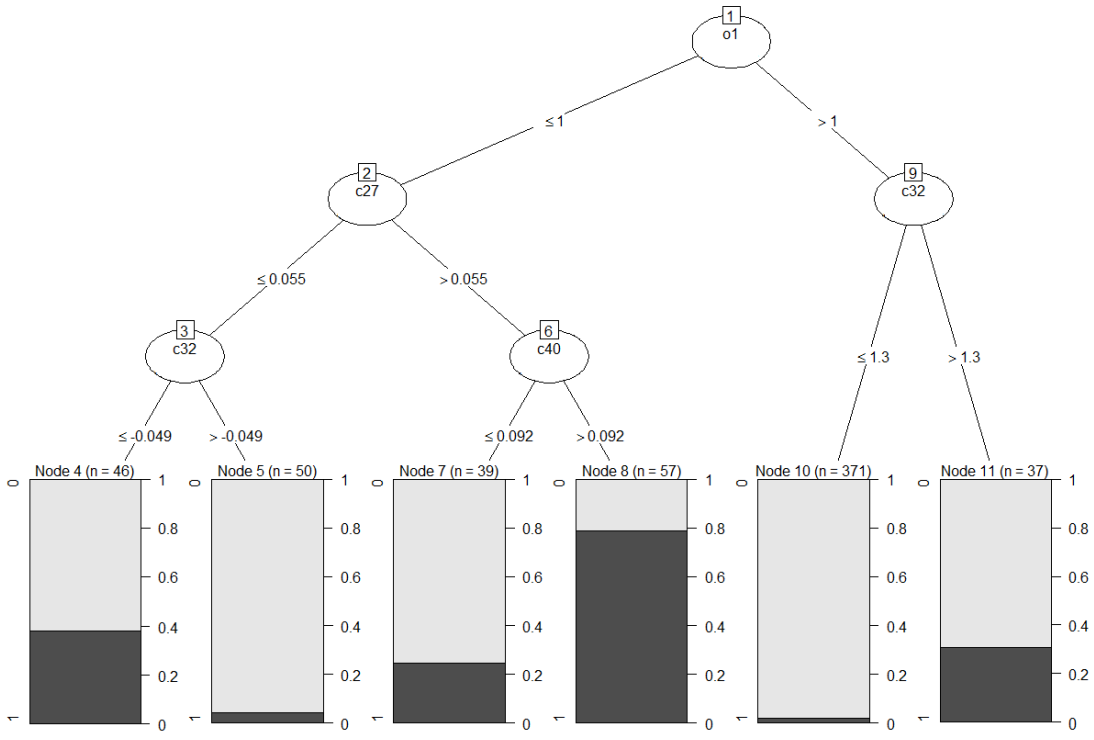

Figure 2.2: Classification tree

The propensity score was designed in order to have a representative value to match on subjects from the candidate set to the treatment group. Cham (2013) uses the random forest method to estimate the propensity score. The random

forest is the process used for the post-hoc matching selection in Figure 2.1.

The random forest method (Breiman 2001) is an ensemble learning method for classification, regression and other tasks. Given a data set $D = (X, Z)$ with $n > 0$ observations and where X is the set of covariates and Z is the response variable, a random forest induces multiple individual decision trees in an ensemble approach. Each of the trees is built from a bootstrap sampling of D. Bootstrapping is sampling with replacement in which each sample is of size n. Observations not selected in the bootstrap sample are called "out-of-bag" (OOB) sample. To decorrelate trees in the "forest", only a randomly chosen subset of covariates is considered for the split at each node.

The goal is to build multiple classification trees in order to classify a percentage of observations according to their covariates in a terminal node. An example of a classification tree from the random forest can be seen on Figure 2.2. The first split occurs at node 1 with the covariate $o1$ and the value 1. All observations with a value equal or less than 1 will go to the left node and the others will go to the right node. A total of 4 more splits will occur before having the 6 final terminal nodes. Subjects in the same terminal node will have a particular propensity score, that will represent the proportion of treatment subject in that terminal node, for example in the node 5, the propensity score value for all 50 observations will be 0.3845. This operation is repeated until the number of classification trees requested has been reached. A different sample of subjects is chosen for each classification tree. The propensity score obtained at the end of each classification tree is then averaged for each subject, and will be their final propensity score. The next step is the matching process. Cham (2013) uses the nearest neighbor method to find a control match for each of the treatment subjects. The nearest neighbor method is computed using the propensity score as

the distance to find the best possible match. That is, a subject from the candidate set which has a similar propensity score as a subject in the treatment group will be matched. This is done to find only one match for each treatment subject and without replacement. A caliper of 0.25 times the standard deviation of the estimated propensity scores is used during the matching process in order to avoid matching observations that are too different, but this will result in potentially not matching all the treated subjects. Finally, a control group has been found and a comparison between the two groups can be made in order to evaluate the matching process.

2.2. The proximity matrix method

The way the propensity score is estimated with classifications trees only takes into account in which terminal node a particular subject is and does not look at the other subjects in the same terminal node. This information could be important and relevant to the matching process because if a treatment and a control subject are in the same terminal node of a classification tree, it means that they were subjected to the same criteria at a splitting node, and therefore have close values for the covariates that were used in the classification tree. This new method takes into account this information. In order to apply this method, a series of classification trees needs to be computed. The model will be built with the treatment variable (1 for the treated group and 0 for the candidate set) as the response variable against all the covariates of the data set. Essentially the idea is to use the stochastic nature of the random forest classification supervised learning approach to consider observations (subjects) from a variety of perspectives with respect to the treatment class. Observations which occur in multiple leaf nodes

together across multiple trees are likely to be inherently similar with respect to classification.

A variation of this approach is presented in Porro and Iacus (2009), however their technique relies on regression trees and a fictitious response variable drawn from a uniform distribution. Similar to Cham (2013), they also employ calipers and additionally filter observations based on a common support threshold. Doing a preliminary reduction of the data and setting a caliper for matching can result in not matching all of the treated observations.

The way the proximity matrix is built is as follows: it takes information from each computed tree by looking if an observation i is in the same terminal node as an observation j. The matrix at the end represents the fractions of trees where observations were in the same terminal node. The values are between 0 and 1. The closest the value is to 1, the more times observations i and j were in the same terminal node, meaning their covariates are similar with respect to the likelihood of being chosen as a treatment observation. Only the rows representing the treatment observations and the columns representing the candidates are kept, since the information necessary for this study is the similarity between treatment and candidates observations. The proximity matrix can be built using the bag data or the FS. Since for each tree we only take a random sample of the data, the bag data represents all the observations that were use for the particular tree. Each bag sample is assign to a terminal node and the proximity matrix is then build according to the outcomes from each tree. An excerpt from an example proximity matrix is depicted in Table 2.1.

Table 2.1: Example of the proximity matrix

	Candidate cases				
Treatment	**8**	**9**	**10**	**11**	**12**
18	0.00515	0.01648	0.04278	0	0.00510
30	0.01064	0.01081	0	0.00518	0
34	0.00510	0.02083	0.01562	0.00985	0.00490
36	0.01005	0.01657	0.05208	0.00526	0.04000
39	0.01538	0	0.00510	0.01546	0.00995

The matching process is done with this proximity matrix. Looking at the highest value in the proximity matrix, the control is matched to the corresponding treatment observation. This procedure is repeated until each treatment has the requisite number of pre-determined controls assigned. The matching can be done with or without replacement, meaning that a control can be matched only once or more than once to treatment observations.

2.3. R implementation

Empirical analysis will be performed using R. Several packages and functions are available to do this.

To build a random forest, two functions are available from different packages; the function *randomforest* from the R package **randomforest 4.6-10** (Liaw and Wiener 2015) and the *cforest* function from the R package **party 1.0-23** (Hothorn et al. 2015). The difference between the two is the covariate and split value selection criterion. There are two possible choices, the Gini index or the conditional significance test decision rule. The function *randomforest* uses the former whereas the function *cforest* uses the latter. In our preliminary testing, we found condi-

tional significance test to outperform the Gini index. The default function is the following:

```
cforest(formula, data=list(), subset=NULL, weights=NULL,
    controls=cforest_unbiased(),
    xtrafo=ptrafo, ytrafo=ptrafo, scores=NULL)
```

From this function, the parameters used for the experiments are the following:

- *formula*: description of the model to be fit.
- *data*: the input data.
- *controls*: list of parameters to control the aspect and the creation of the trees. The default parameters are the following:

```
cforest_control(teststat = "max",
            testtype = "Teststatistic",
            mincriterion = qnorm(0.9),
            savesplitstats = FALSE,
            ntree = 500, mtry = 5, replace = TRUE,
            fraction = 0.632, trace = FALSE, ...)
```

From these different parameters, the following will be used:

- *ntree*: number of trees to grow in the random forest.
- *mtry*: number of covariates randomly selected for potential splitting at each node.

– *replace*: logical value to determine if the sampling of observations is done with or without replacement.

– *fraction*: the fraction of observations to draw from the sample without replacement.

– *minbucket*: minimum number of observations required in a terminal node.

Hence, an object of class *RandomForest-class* will be created, containing each single tree that was computed. This will store information on each single tree, for example in which terminal node each observation was, the response variable (either a class identification for classification or a prediction for regression), etc. This information is necessary to create the proximity matrix. It can be created from the following function from the same package:

$$proximity(object, newdata = NULL) \tag{2.1}$$

In this function, the *object* represents the object computed from the *cforest* function. It will produce an proximity matrix using the bag data. There is also the possibility to build the proximity matrix using the FS, by using the argument *newdata* in the *proximity* function.

After obtaining the proximity matrix, a matching process is required. The goal is to find a match in the candidate set for each treatment. It can be done with or without replacement. With replacement means that a particular candidate can be matched more than once to a treatment. In our preliminary testing, we found matching with replacement to outperform matching without replacement. The algorithm to describe the matching process is detailed in Algorithm 1.

The solution will be returned as a list of controls for each treatment. In

Algorithm 1 Match each treatment to C controls

Require: M (proximity matrix), C (number of controls needed)
Ensure: $ListC$ (list of matched controls to each treatment)

```
Create an empty list for the controls to match each treatment, ListC, with
treatment names as the rownames
Create a vector, V, of values from M by decreasing order
col ← −99 (initialization)
k ← 0
for i = 1 to (number of row in M * C) do
    repeat
        bool = false
        k ← k + 1
        maxVal = Value in M corresponding to V[k] (may be more than one)
        row = Treatment id from maxVal
        col = Candidates id from maxVal
        for l = 1 to length(row) do
            if treatment row[l] doesn't have C controls yet then
                bool = true
                exitforloop
            end if
        end for
    until bool = true
    Control found, col[l], corresponding to treatment row[l] added to ListC
end for
```

order to find the best possible match for each treatment, the algorithm will find the highest value in the proximity matrix, $M[i, j]$, meaning that the candidate j and the treatment i were the most often in the same terminal node. The process then matches treatment i to candidate j and adds this pair to $ListC$ if treatment i has less than C controls already assigned. If treatment i already has all the required controls, the algorithm will look for the next highest value in the proximity matrix and repeat the process. When a matched is found, the algorithm continues looking at the next highest value in the proximity matrix until all treatments have exactly C matched controls each. This process is done with replacement, in order to optimize the results. Indeed, if a candidate has been in the same terminal node of numerous treatments, it may be a good match for all of them.

Chapter 3

Experimentation

This chapter describes how the data sets are generated and how the different experiments will be conducted. The simulated data in this study is constructed based on Cham (2013), which in turn was constructed based on an empirical example Im et al. (2013) and simulation studies from Austin (2012), Lee et al. (2010), and Setoguchi et al. (2008). Most of the simulated covariates and models are kept and created the same way, but an important modification is introduced with respect to the balance of the test and candidate pool.

3.1. Generated data

A series of data and models are created. An overview of the different steps follow:

- Create 64 covariates for two data sets, of size 600 and 2,000 observations.
- Generate the treatment group based on different models:
  - Linear propensity score model

  - Nonlinear propensity score model

- Compute two response outcomes based on the two different models. They are created with a particular average treatment effect:
  - 0.0 average treatment effect
  - 0.73 average treatment effect

Covariates

In the data sets, a total of 64 covariates are randomly generated. Out of those 64 covariates, 16 are binary ($b1$ through $b16$), 40 are continuous ($c1$ through $c40$) and 8 are ordered categorical ($o1$ through $o8$). The binary covariate are dummy coded 0.0 and 1.0 with a mean of 0.245. The continuous covariates are generated as a standard normal distribution. The ordered-categorical covariates are generated based on discretization of normally distributed random variable, with 7 categories (0 to 6) and are approximately symmetrically distributed. The covariates are manipulated to have two levels of correlation, low and high, and values of correlations betweens those levels are pre-determined identically to that of Cham (2013).

Propensity score models

Two different propensity score models are created based on simulation study designs by Austin (2012). The *linear propensity score model* is created so that all of the covariates are linearly related to the propensity score in the form of the following logistic regression:

$$\ln\left(\frac{e(X)}{1-e(X)}\right) = \gamma_0 + \gamma_1 b_1 + ... + \gamma_{19} c_1 + ... + \gamma_{64} o_8$$

$e(X)$ is the propensity score, and the regression coefficients of the covariates (γ_1 to γ_{64}) are set at two levels, low and high, separately for each type of covariate. As for the intercept γ_0, Cham (2013) set it so that the average propensity score was about 0.38 for the linear model and 0.45 for the nonlinear model, which would give approximately the same as of the proportion of the treatment group. In this study, γ_0 is decreased to adjust the balance between the treatment group and the candidate pool. Hence, the proportion of the treatment group is reduced to 0.15 for the linear model and 0.22 for the nonlinear model, thereby increasing the relative size of the candidate pool. This represents a variety of real-world scenarios in which a relatively small control group must be selected from a large subset of the population. On the other hand, if 50% of the group were assigned as a treatment group, and a one-to-one match with the candidate pool is desired, then no matching strategy is necessary. The entire candidate pool would be used as the control. However, when required to select out a one-to-one match when 85% of the cases are available as candidates, then there are many possible choices and the matching method will need to be more accurate.

The *nonlinear propensity score model* is created from the linear propensity score model but with other terms describing nonlinear relationships between the covariates. Three different types of nonlinear relationships are added to a subset of the predictors: two-way interactions, quadratic effects, and two-way interaction with quadratic effects.

The two propensity score models are created to decide what the treatment assignment Z for all observations will be. Each case has a random number that follows a uniform distribution between 0.0 and 1.0 assigned. If that random number is smaller than the corresponding propensity score, the observation is assigned to the treatment group ($Z = 1$). Otherwise, the observation is assigned

to the candidate set ($Z = 0$).

Treatment-outcome model

The continuous outcome Y for this study is based on Austin (2012), Lee et al. (2010) and Setoguchi et al. (2008). The outcome is created such that the treatment and the covariates are linearly related to it, as follows:

$$Y = \beta_0 + \beta_1 b_1 + ... + \beta_{17} c_1 + ... + \beta_{64} o_8 + (ATE)Z + \epsilon$$

Z is the treatment assignment (1 for treatment, 0 for candidates), ϵ is randomly generated normally distributed residual, β_0 is the intercept and the regression coefficients (β_1 to β_{64}) are manipulated to have two levels, low and high, for each type of covariate Cham (2013). The average treatment effect (ATE) is manipulated to have two levels: a null ATE (0.0) and a non-zero ATE (0.73) according to Cohen (1988) guidelines. The data sets will have two outcome variables depending on the level of the ATE: ATE0 and ATE073.

Sample size

Two data sets with different sizes are generated; one with 600 observations and another one with 2,000 observations.

3.2. Random forest specifications

When building a random forest, certain modeling parameters must be specified. These parameters and the values used in this investigation are now detailed.

Covariate and split value selection

When building the tree, a decision needs to be made at each node to decide on what covariate the node will be split and on which value. This decision will need to be made starting at the root node and for all subsequent nodes until all observations have been classified into a terminal node. A common focus to find the best covariate and its split value is the impurity of a node. When we split it, the less impurity the better, which means that the split value that clearly separates the binary treatment into two distinct groups will be chosen. To calculate the impurity of the node, the Gini index can be use, represented by $2p(1-p)$ (Berk 2008, Hastie et al. 2001) where p represents the proportion of the treatment group participants in the node. It will select the covariate and its split value simultaneously. When the data contains different types of covariates such as binary, categorical and continuous, this method has a disadvantage and it is biased when selecting a covariate at a node towards categorical and continuous variables instead of binary variables.

Another approach is the conditional significance test (Hothorn et al. 2006) in which the choice of the covariate and its split value are not done simultaneously. The first step consists of choosing which covariate will be used for splitting. To do that, for each node, a statistical test is conducted for each covariate. The covariate selected is the one that has the smallest p-value. This implies that it is statistically significant that the covariate is associated with the treatment group. Two different test statistics exist to do this. No splitting will occur if the smallest p-value is greater than the pre-specified nominal level α (Hothorn et al. 2006). Once the covariate has been selected in the first step, a second permutation test can be performed to determine the split value of that covariate. The covariate value with the largest test statistic will be selected, meaning that the proportions

of treatment group participants are equal between the two nodes.

In our experimentation, the conditional significance test was found to outperform the Gini index, since separated tests are used for selecting the covariate and the split value which reduces bias.

Methods for estimating the propensity score and the proximity matrix This specifications has two levels; the use of the FS or the OOB data to estimate the propensity score and the FS or the bag data to estimate the proximity matrix. In Cham's study, the out-of-bag is preferred to the full sample due to the fact that it reduces the tendency for propensity scores to be estimated that are biased toward the extreme values of 0.0 and 1.0 (Berk 2008, Strobl et al. 2009). The full sample to estimate the propensity score shows better results in this experimentation since some modifications regarding the proportion of the candidate group was made. With the proximity matrix method, the goal is to see which observations from both the treatment group and the candidate set are in the same terminal node to create the proximity matrix reflecting the connection between the observations. The bag data is used to estimate the proximity matrix since each tree from the random forest is built using that data and this is the information needed. The out-of-bag data is not relevant in this case because we are not looking at new data to estimate the proximity matrix but at the data use to create each single tree from the random forest.

To find a control match to each treatment, the nearest neighbor matching is used for the propensity score method. For the proximity matrix, Algorithm 1 is use.

Default arguments

All of the other random forest arguments are set to their default values. Those arguments included are the following:

- Number of classification trees, set to 500;
- Number of covariates randomly selected for potential splitting at each node, set to 8 (square root of the total number of covariates, 64);
- Minimum number of observations required in a terminal node, set to 5;
- Fraction of observations to draw from the sample without replacement, set to 0.632;

3.3. Different experiments according to the models

From all those different models for the data sets, multiple experiments have been done in order to address the primary research question. For both of the data sets, two propensity score models and 2 different outcomes were created. Either the full sample or the out-of-bag sample can be used to estimate the propensity score, and the full sample or the bag data is used to estimate the proximity matrix. Finally, either 1 or 2 controls can be matched to a subject from the treatment group. This gives a total of 32 different experiments. Each of those experiments are replicated 200 times. A summary of the simulation study design is presented in Table 3.1 along with notation that will be used in later chapters.

Table 3.1: Summary of the simulation study

Between-subject (Replication) factors	**Levels and Notations**
1. Propensity Score Models	Linear (L), Nonlinear (N)
2. Average Treatment Effect	0, 0.73
3. Sample Size	600, 2000
Within-subject (Replication) factors	
4. Benchmark Methods	Propensity Score Matching without replacement (PS), Proximity Matrix Matching with replacement (PM)
5. Covariate and Split value selection	Conditional Significance Test
6. Methods to Estimate Propensity Scores or Proximity Matrix	Full Sample (FS), Out-Of-Bag Sample (OOB), Bag sample
7. Number of controls variables to be matched	1 control (1C), 2 controls (2C)

Chapter 4

Results

In order to answer the primary question, the results from the simulation study are presented in the following sections.

The goal of the matching process is to find a good match for the treatment group that produces an unbiased estimate of the average treatment effect (ATE). Two different ATEs have been created in order to have either a 0 average treatment effect or a 0.73 average treatment effect. For the experimentation in which ATE = 0, the more successful matching process will produce a smaller difference between the average outcomes of treatment and control. With respect to the experimentation for ATE = 0.73, the average treatment outcome minus the average control outcome should approach 0.73.

Tables 4.1 and 4.2 represent the percentage of success of the proximity matrix (PM) method over the propensity score (PS) method according to the type of data used to estimate either of those methods (FS or OOB data for the propensity score and FS or bag data for the proximity matrix). Success is defined by producing an estimated ATE closer to the true value (ATE = 0 or ATE = 0.73).

Each row corresponds to a different problem. For example, ATE0-L-1C denotes experimentation on the data which has ATE = 0 generated using a linear propensity score model and the matching requires 1 control per treatment observation. Whereas ATE073-N-2C denotes experimentation on the data which has ATE = 0.73 generated using a nonlinear propensity score model and the matching process requires 2 controls per treatment observation.

Table 4.1: Table of the percentage of success for the data set containing 600 observations.

Problems	**Percentage FS**	**Percentage OOB/bag**
ATE0-L-1C	36.5	96.5
ATE0-L-2C	95.5	100.0
ATE0-N-1C	61.5	100.0
ATE0-N-2C	100.0	100.0
ATE073-L-1C	8.0	65.5
ATE073-L-2C	49.5	97.0
ATE073-N-1C	17.5	94.5
ATE073-N-2C	100.0	100.0

Table 4.2: Table of the percentage of success for the data set containing 2,000 observations.

Problems	**Percentage FS**	**Percentage OOB/bag**
ATE0-L-1C	0.0	73.5
ATE0-L-2C	11.0	100.0
ATE0-N-1C	77.0	100.0
ATE0-N-2C	100.0	100.0
ATE073-L-1C	0.0	81.0
ATE073-L-2C	21.5	100.0
ATE073-N-1C	66.5	99.5
ATE073-N-2C	100.0	100.0

Table 4.1 reports the results for all eight problems using data with 600 obser-

vations each. The column "Percentage FS" reports the percentage of success of the proximity matrix over the propensity score method on the full sample across all 200 replications. Similarly, the "Percentage OOB/bag" column reports the similar results but on the out-of-bag or bag sample. The results demonstrate that when estimating the propensity score and the proximity matrix with the full sample, the proximity matrix method performs better (more than 50%) 4 times out of the 8 different problems. In 2 problems, PM outperforms PS in 100% of the instances. For the ATE073-L-2C, the results are mixed in that PM outperforms PS only about 50% of the time. On contrary, when evaluating success on the OOB/bag data, the proximity matrix method is clearly better than the propensity score method outperforming PS in all eight problems. PM has more than 65% success and has a 100% success rate for 4 of the experiments.

From Table 4.2 representing the 2,000 observations data, the results are similar to Table 4.1. When the FS is use to estimate the propensity score and the proximity matrix, the PM method performs better (more than 50%) 4 times out of the 8 different problems with ATE0-N-2C and ATE073-N-2C outperforming PS in 100% of the instances. When the OOB/bag data is used, the PM method is, again, outperforming PS for the 8 problems. The same 4 problems from Table 4.1 (ATE0-L-2C, ATE0-N-1C, ATE0-N-2C and ATE073-N-2C) have a 100% success rate, and ATE073-L-1C, ATE073-L-2C and ATE073-N-1C have a higher success rate than with the 600 observations data.

Tables 4.3-4.6 are also arranged in 8 rows for each of the problems and 3 columns. The first column represents the average of the mean of the ATE for all of 200 replications, and the associated standard deviation. The second column has the same type of values but for the proximity matrix method. A paired t-test is computed and the p-value is listed in the last column.

Table 4.3: Table of the ATE Mean, Standard Deviation and p-value for the experimentation on the 600 observations data and the matching process using the FS.

Problems	Propensity Score	Proximity Matrix	p-value
ATE0-L-1C	0.31668	0.34029	1.7181E-4
	(0.05074)	(0.08015)	
ATE0-L-2C	0.50808	0.36085	6.27E-64
	(0.02876)	(0.08190)	
ATE0-N-1C	0.24674	0.23135	0.00032
	(0.02511)	(0.05838)	
ATE0-N-2C	0.59106	0.26972	3.79E-145
	(0.01519)	(0.06209)	
ATE073-L-1C	1.01195	1.12839	2.44E-54
	(0.03777)	(0.07271)	
ATE073-L-2C	1.13306	1.13286	0.96979
	(0.02648)	(0.07346)	
ATE073-N-1C	0.92469	0.97965	1.98E-30
	(0.02164)	(0.05462)	
ATE073-N-2C	1.22409	1.02492	3.21E-108
	(0.01247)	(0.06073)	

Table 4.4: Table of the ATE Mean, Standard Deviation and p-value for the experimentation on the 600 observations data and the matching process using the OOB/bag data.

Problems	Propensity Score	Proximity Matrix	p-value
ATE0-L-1C	0.32637	0.12399	9.19E-75
	(0.06506)	(0.08221)	
ATE0-L-2C	0.51936	0.11860	2.20E-147
	(0.02949)	(0.07482)	
ATE0-N-1C	0.26600	0.06018	5.04E-114
	(0.03697)	(0.04806)	
ATE0-N-2C	0.59049	0.06081	2.64E-207
	(0.01774)	(0.04601)	
ATE073-L-1C	1.02496	0.99375	2.70E-05
	(0.06116)	(0.07957)	
ATE073-L-2C	1.13942	0.98247	3.74E-71
	(0.02886)	(0.07390)	
ATE073-N-1C	0.94230	0.79835	1.12E-60
	(0.03665)	(0.07702)	
ATE073-N-2C	1.22660	0.80171	4.47E-159
	(0.01533)	(0.06935)	

Table 4.5: Table of the ATE Mean, Standard Deviation and p-value for the experimentation on the 2,000 observations data and the matching process using the FS.

Problems	Propensity Score	Proximity Matrix	p-value
ATE0-L-1C	0.35717	0.51481	4.45E-98
	(0.04250)	(0.04158)	
ATE0-L-2C	0.45363	0.50614	6.81E-40
	(0.01700)	(0.04277)	
ATE0-N-1C	0.33292	0.30429	8.88E-25
	(0.01258)	(0.03266)	
ATE0-N-2C	0.69069	0.30202	2.35E-216
	(0.00644)	(0.03299)	
ATE073-L-1C	1.19625	1.34119	3.07E-102
	(0.03718)	(0.03847)	
ATE073-L-2C	1.31131	1.33853	6.12E-17
	(0.01573)	(0.04035)	
ATE073-N-1C	1.09727	1.08144	8.62E-11
	(0.01165)	(0.03071)	
ATE073-N-2C	1.39691	1.08659	2.19E-200
	(0.00671)	(0.03088)	

Table 4.6: Table of the ATE Mean, Standard Deviation and p-value for the experimentation on the 2,000 observations data and the matching process using the OOB/bag data.

Problems	Propensity Score	Proximity Matrix	p-value
ATE0-L-1C	0.36490	0.32987	4.87E-14
	(0.04074)	(0.05436)	
ATE0-L-2C	0.46021	0.31945	4.14E-93
	(0.01537)	(0.05279)	
ATE0-N-1C	0.34718	0.14644	4.51E-127
	(0.01775)	(0.04319)	
ATE0-N-2C	0.69058	0.14992	9.29E-219
	(0.00852)	(0.04496)	
ATE073-L-1C	1.21193	1.16595	3.53E-20
	(0.03729)	(0.05331)	
ATE073-L-2C	1.31906	1.16601	1.43E-101
	(0.01577)	(0.05028)	
ATE073-N-1C	1.10578	0.93222	9.29E-115
	(0.01650)	(0.04278)	
ATE073-N-2C	1.39839	0.93996	7.43E-213
	(0.00807)	(0.04002)	

The success of the proximity matrix over the propensity score method is notable in Tables 4.3-4.6. From Table 4.1 where the proximity matrix method success was over 50% for all 8 test data with the OOB/bag data, the corresponding mean also shows better results in Table 4.4. For ATE0-L-1C, ATE0-L-2C, ATE0-N-1C and ATE0-N-2C where the ATE should be closer to 0, the PM technique outperforms the PS technique by being closer to 0 with 0.12399, 0.11860, 0.06018 and 0.06081 respectively, whereas the propensity score is doing poorly with a relative high ATE of 0.32637, 0.51936, 0.266 and 0.59049 respectively. On that same table, for the problem where the ATE should be 0.73, the proximity matrix method is also doing much better by having an ATE between 0.79835 (ATE073-N-1C) and 0.99375 (ATE073-L-1C) for all 4 test data whereas the propensity score

method have an ATE between 0.94230 (ATE073-N-1C) and 1.2266 (ATE073-N-2C). Regarding the experimentation using the 2,000 observations data and the OOB/bag data where on Table 4.2 the proximity matrix method success was over 50% for all 8 problems, the related average ATE differences and statistical tests are reported in Table 4.6. The results are not as good as for the ones with the 600 observations using the OOB/bag data, but the ATE for the proximity matrix method for ATE0-L-1C, ATE0-L-2C, ATE0-N-1C and ATE0-N-2C are closer to 0 than the ATE for the propensity score method. The same can be observed regarding the ATE of 0.73; for the proximity matrix method the results are between 0.93222 (ATE073-N-1C) and 1.16601 (ATE073-L-2C) whereas for the propensity score the ATE goes up to 1.39839 (ATE073-N-2C). From these two tables, the standard deviation for each of the test data is always smaller for the PS method and is usually around 0.1, whereas for the proximity matrix method it is more around 0.3 or 0.6.

Although the propensity score method has a smaller standard deviation for each problems, the results in Table 4.4 and Table 4.6, the PM method outperforms the PS method consistently. The associated p-values indicates the difference in means is highly significant. By having a really low p-value and having more than 50% of success out of the 8 test data for both of those experiments, it is clear than the proximity matrix method is outperforming the propensity score method.

On Tables 4.3 and 4.5 where the experiments are performed on the full sample data, results are mixed. For the 600 observation data, as said previously, the proximity matrix success was over 50% for 3 of the test data (ATE0-L-2C, ATE0-N-1C and ATE0-N-2C) regarding the ATE of 0, which can be seen in the average mean from Table 4.3. The p-values indicates that the performance of PM is statistically different than that of PS in all cases except for ATE073-L-2C. Note

while the difference in performance is statistically significant, the absolute value of the difference is rather small. Compare this for example to the problem ATE0-L-2C where the PM method success is 95.5% with an average mean of 0.36085 to an average mean of 0.50808 for the PS method with a small p-value of 6.27×10^{-64}. As for the problems where the ATE should be 0.73, the PM method having a low success for the problems ATE073-L-1C and ATE073-N-1C is seen with an average mean of 1.12839 and 0.97965 respectively compared to an average mean of 1.01195 and 0.92469 for the PS method. With the problem ATE073-N-2C, the PM outperforms the PS method with 100% success and the average mean is clearly better. Regarding the experimentation using the 2,000 observations data and the FS from Table 4.5, the success rate of the PM method for the problems ATE0-L-1C, ATE0-L-2C, ATE073-L-1C and ATE073-L-2C is low and goes from 0.0 to 21.5%, but the PS is not outperforming the PM method by much for the problems ATE0-L-2C and ATE073-L-2C by having an average mean not too different. Whereas for the problems ATE0-N-2C and ATE073-N-2C where the success rate of the PM method is 100%, it is clearly outperforming by having a much better average mean. The p-value is consistently small, showing that the difference in means between the two methods is highly statistically different. Only problem ATE073-L-2C in Table 4.3 has a larger p-value of 0.96979. From those two tables, the standard deviation for each of the problems is always smaller for the PS method and ranges from 0.00671 to 0.05074, whereas for the PM method ranges from 0.03071 to 0.08190.

From both Tables 4.3 and 4.5, the proximity matrix method is performing better 8 times out of the 16 problems and is close for one of the problems. For the other problems where the PS method has a better success rate, it is not outperforming the PM method by much.

Figures 4.1 and 4.2 shows the density of the different ATE magnitude (0 and 0.73) for the propensity score and the proximity matrix method for the 600 observation data set and the type of data used to estimate the PS and the PM (FS, OOB and bag data). Figures 4.3 and 4.4 represent the same information on the 2,000 observations data set.

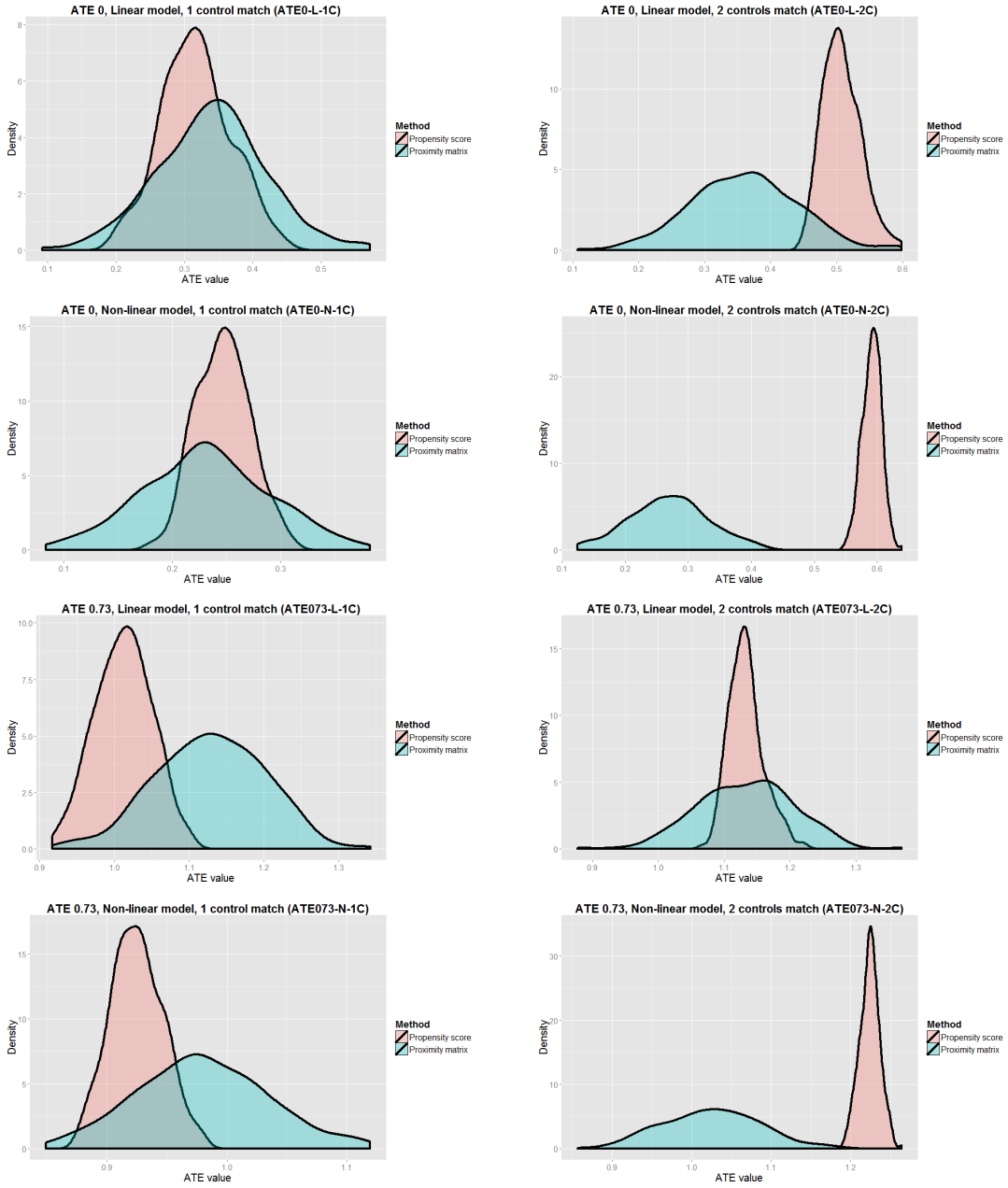

Figure 4.1: Density of the ATE value using the FS data and the 600 observations data set.

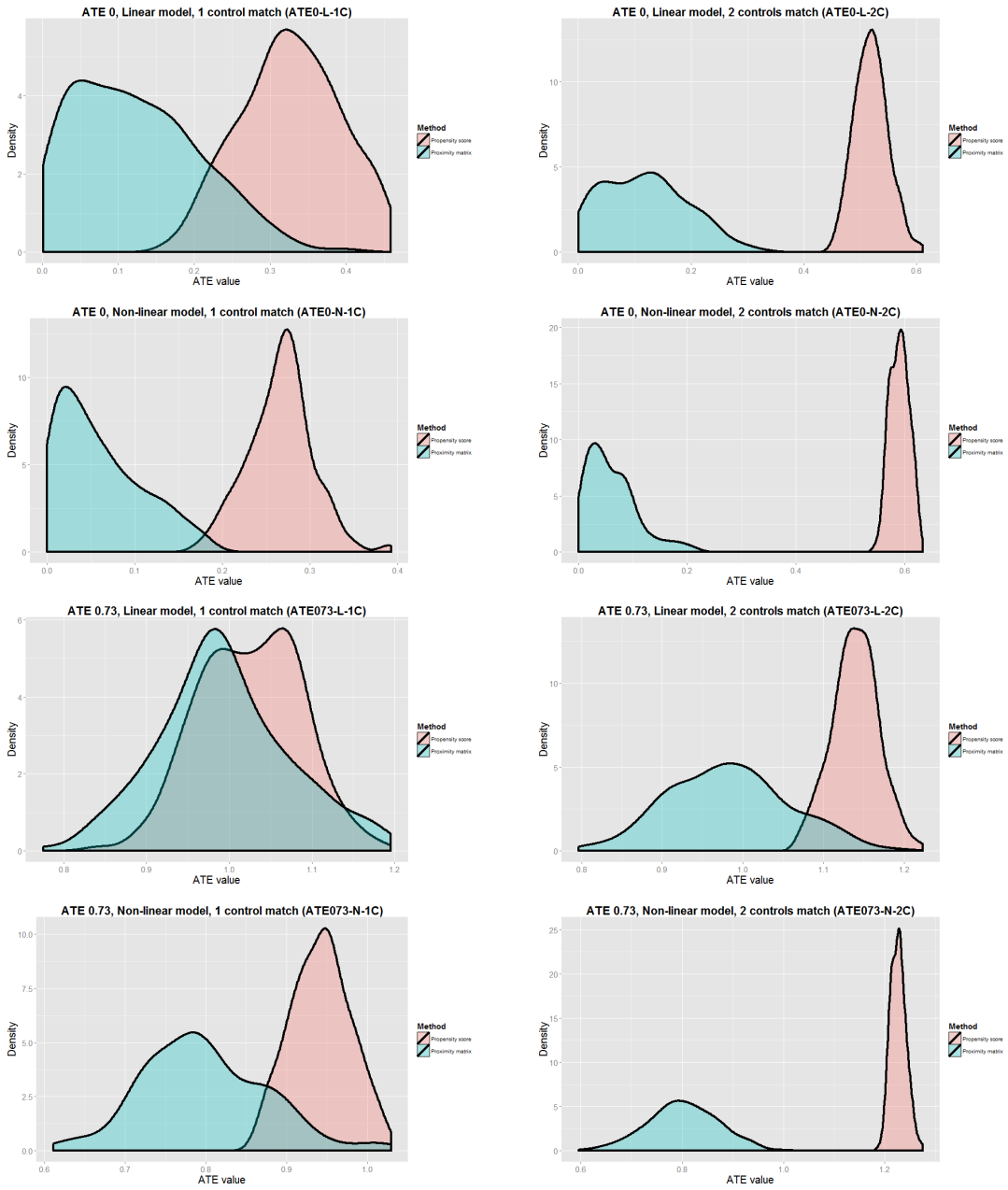

Figure 4.2: Density of the ATE value using the OOB/bag data and the 600 observations data set.

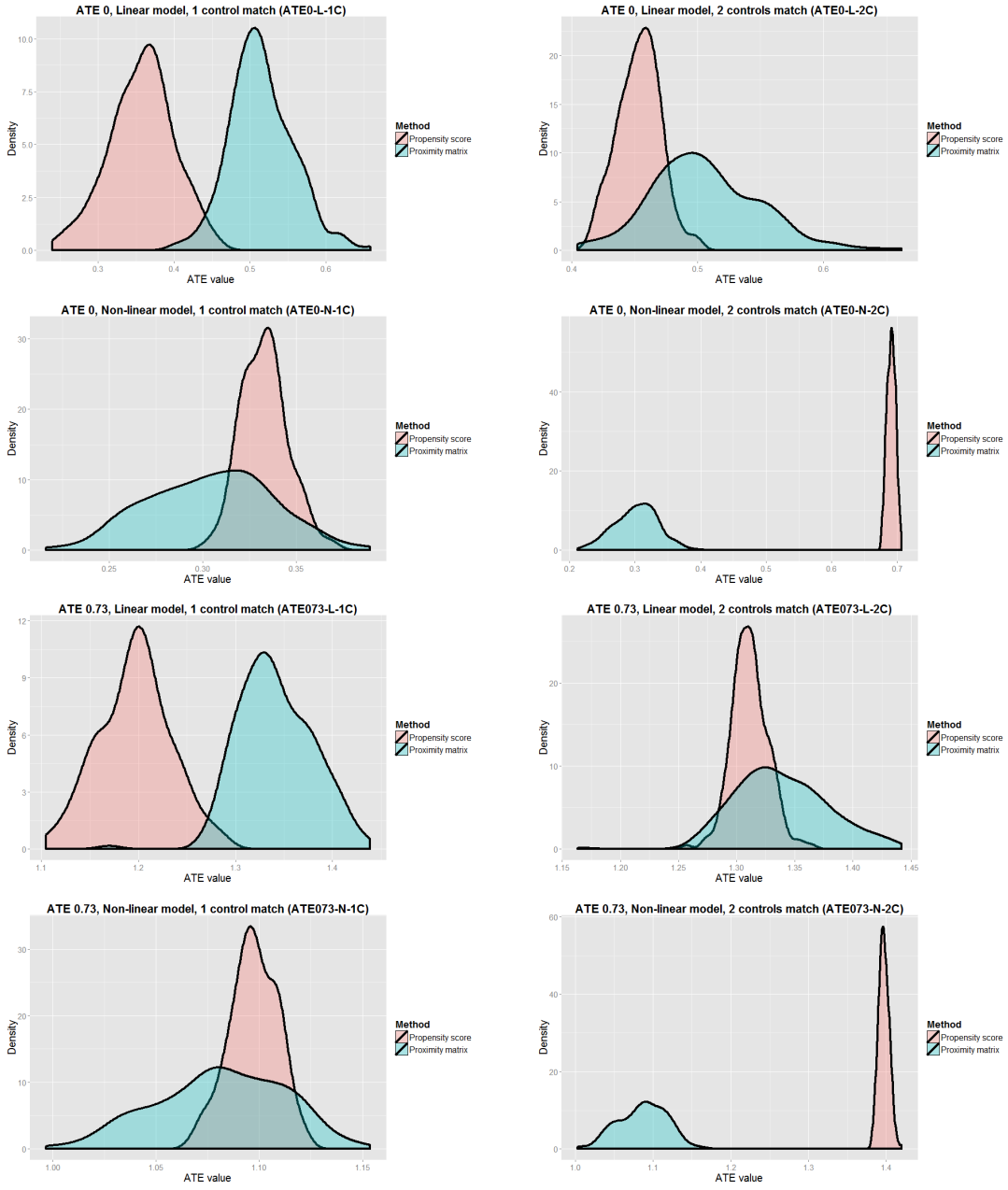

Figure 4.3: Density of the ATE value using the FS data and the 2,000 observations data set.

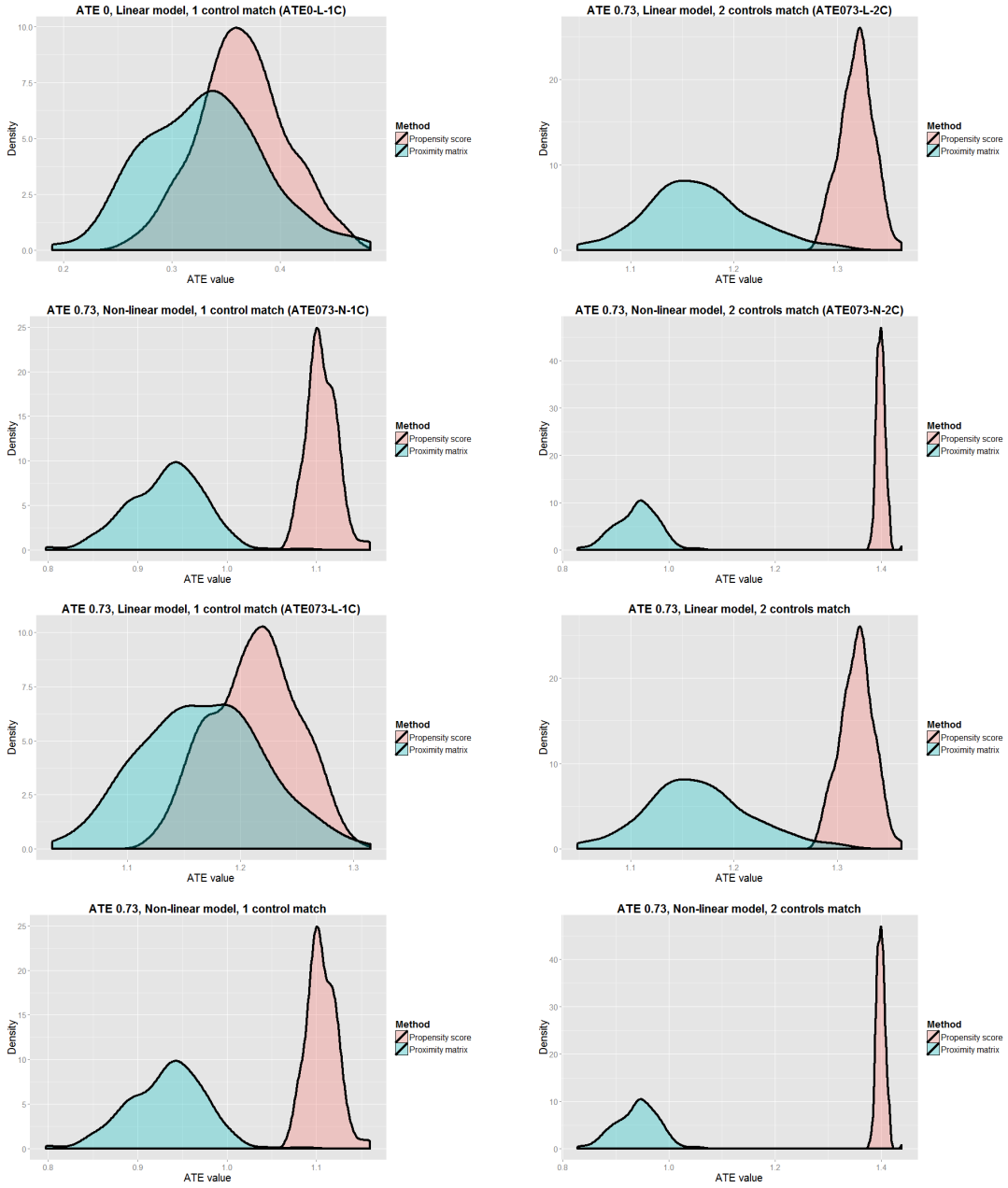

Figure 4.4: Density of the ATE value using the OOB/bag data and the 2,000 observations data set.

Figures 4.1-4.4 correspond to results from Tables 4.3-4.6. Each graph, going from top to bottom and from left to right, represents a different problem corresponding to each row of the previous tables. From Table 4.1 where the PM method success was over 50% for all 8 problems using the OOB/bag data, those results can be seen on Figure 4.2. Indeed, the density curve of the ATE value for all 8 graphs are much closer to 0 and 0.73 for the PM method than for the PS method. A larger range of values is also observed for the PM method. For the problem ATE073-L-1C, the density for both methods is close, but since the PM method has a wider range, it has more values close to 0.73 than the PS method. For the problems ATE0-L-2C, ATE0-N-2C and ATE073-N-2C, PM outperforms PS by having an ATE value much closer to 0. The two density curves are also not overlapping at all, meaning that the PS method is doing poorly. As for the problems ATE0-L-1C, ATE0-N-1C, ATE073-L-2C and ATE073-N-1, PM also outperforms PS but the two curves are overlapping showing some common values.

As for the 8 problems using the FS data corresponding to Figure 4.1, the results are mixed. PM clearly outperforms PS method for the problems ATE0-N-2C and ATE073-N-2C by having closer values to 0 and 0.73 and not overlapping the PS curve. The problems ATE0-L-2C and ATE0-N-1C shows that PM also outperforms PS, and for ATE073-L-2C the results are more tight. The PM densities have greater spread compared to the PS densities, but also have some better values than the PS method. The other problems show that the PS approach is doing better than the PM method.

Results from Table 4.2 can be seen in Figures 4.3 and 4.4 representing the FS and OOB/bag data respectively. Similar patterns can be seen between Figure 4.3 and Figure 4.1 for problems ATE0-N-1C, ATE0-N-2C, ATE073-L-2C and

ATE073-N-2C. Between Figure 4.4 and Figure 4.2, a clear pattern can be observed between each of the 8 problems. Only for the problem ATE0-L-1C a closer range is seen and for ATE073-N-1C there is a wider range between the two density curves.

Tables 4.7 and 4.8 are arranged in 8 rows for each of the problems and 4 columns, each representing one of the methods; propensity score method using the full sample or the out-of-bag data and the proximity matrix method using the full sample or the bag data. The mean of the ATE for all of the 200 replications from the previous tables is represented.

Table 4.7: Summary table of the mean of all the methods for the 600 observations data set

Problems	**PS with FS**	**PS with OOB**	**PM with FS**	**PM with bag**
ATE0-L-1C	0.31668	0.32637	0.34029	**0.12399**
ATE0-L-2C	0.50808	0.51936	0.36085	**0.11860**
ATE0-N-1C	0.24674	0.26600	0.23135	**0.06018**
ATE0-N-2C	0.59106	0.59049	0.26972	**0.06081**
ATE073-L-1C	1.01195	1.02496	1.12839	**0.99375**
ATE073-L-2C	1.13306	1.13942	1.13286	**0.98247**
ATE073-N-1C	0.92469	0.94230	0.97965	**0.79835**
ATE073-N-2C	1.22409	1.22660	1.02492	**0.80171**

Table 4.8: Summary table of the mean of all methods for the 2,000 observations data set

Problems	PS with FS	PS with OOB	PM with FS	PM with bag
ATE0-L-1C	0.35717	0.36490	0.51481	**0.32987**
ATE0-L-2C	0.45363	0.46021	0.50614	**0.31945**
ATE0-N-1C	0.33292	0.34718	0.30429	**0.14644**
ATE0-N-2C	0.69069	0.69058	0.30202	**0.14992**
ATE073-L-1C	1.19625	1.21193	1.34119	**1.16595**
ATE073-L-2C	1.31131	1.31906	1.33853	**1.16601**
ATE073-N-1C	1.09727	1.10578	1.08144	**0.93222**
ATE073-N-2C	1.39691	1.39839	1.08659	**0.93996**

The success of the proximity matrix method using the bag data over the other 3 methods can clearly be seen in tables 4.7 and 4.8. For the first 4 problems regarding the ATE0, the mean of the ATE is much more closer to 0 than the other methods. The same results can be observed regarding the ATE073 where the mean of the ATE is much more closer to 0.73 when the matching is done with the proximity matrix method using the bag data than the other methods. In summary, for all 16 problems over the 2 different data sets, the proximity matrix method using the bag data is outperforming all other methods.

Chapter 5

Conclusion

A primary issue for observational studies is the lack of an identified statistically equivalent control group to help determine an unbiased treatment effect. This issue can be addressed by matching a control group similar to the treatment group based on a particular score that will define their similarities. The propensity score is a well known and popular score used to match subjects and has been in use for many years. The random forest technique is an effective method to estimate the propensity score (Cham 2013). The propensity score being the conditional probability of a subject being assigned to the treatment group given a set of covariates, and representing the value of the proportion of treatment in a terminal node of a decision tree, subjects could have the same value and be in a different node if the proportion of treatment subjects in different terminal nodes are the same. This means that subjects could be matched on their propensity score but have different values for their covariates. The new method introduced in this thesis takes into account when treatment and candidate subjects fall into the same terminal node of a decision tree.

The new approach uses a proximity matrix M, where M_{ij} represents the frac-

tion of trees where subjects i and j were in the same terminal node. Control subjects are found to match each treatment subject from this matrix. This approach is compared to the propensity score method from Cham (2013). To compare the methods, the data from Cham (2013) is reproduced logically with a change regarding the proportion of the treatment and the candidate set. The proportion of subjects representing the treatment group was around 0.38, which does not leave a lot of choices regarding the candidates observations. This was changed to have approximately a proportion of the treatment group of 0.15, which gives a better representation of reality and forces the matching method to be more accurate since there will be more candidates to choose from.

The experimentation conducted on the different data sets shows good performance of the novel post-hoc matching method, compared to the competing propensity score method. The proximity matrix method clearly outperforms the propensity score method when the out-of-bag or the bag data is used to create the matching scores. This is consistent for each of the 8 problems and for both data sets. When the full sample is used, the proximity matrix method outperforms 8 out of 16 times, and the two methods are essentially equivalent for 1 problem. The proximity matrix method does much better for the problems using the non-linear propensity score model and when 2 controls are matched to each treatment. With these results, we conclude that the proximity matrix method has advantages in the matching process and does a better job than the propensity score method most of the time. Finally, when comparing the proximity matrix using the bag data to the other 3 methods, it is clearly outperforming for all of the problems.

The goal of this research being to find matches to the treatment group that have the same average treatment effect, the similarities between the distributions

of the covariates was not evaluated. In other studies, such as in medicine, this type of information can be relevant and therefore comparison between the covariates of each group could be done. Additionally, a caliper could be use before finding control matches for each subject of the treatment group. The aim of the research was to find a match for every subject in the treatment group. Therefore, the subject matched from the candidate set to a subject in the treatment group can be really different. By adding a caliper to the matching process, it will prevent matching subjects that are too different and have no common support, and will also result in not matching every subject from the treatment group. In future, such modifications can be studied to further improve the high quality results we observe from proximity matrix matching.

References

Abadie, A., G. Imbens. 2006. Large sample properties of matching estimators for average treatment effects. *Econometrica* **74** 235–267.

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

Berk, R. 2008. *Statistical learning from a regression perspective*. Springer Science & Business Media.

Breiman, L. 2001. Random forests. *Machine learning* **45** 5–32.

Cham, H. 2013. Propensity score estimation with random forests. Ph.D. thesis, Arizona State University.

Cochran, W., D. Rubin. 1973. Controlling bias in observational studies: A review. *Sankhyā: The Indian Journal of Statistics, Series A* 417–446.

Cohen, J. 1988. Statistical power analysis for the behavioral sciences .

Cox, D. 1970. The analysis of binary data. *London: Methuen* .

Epstein, L., D. Ho, G. King, J. Segal. 2005. Supreme court during crisis: How war affects only non-war cases, the. *NYUL rev.* **80** 1.

Hastie, T., R. Tibshirani, J. Friedman. 2001. *The elements of statistical learning: data mining, inference and prediction*. Springer.

Herron, M., J. Wand. 2007. Assessing partisan bias in voting technology: The case of the 2004 new hampshire recount. *Electoral Studies* **26** 247–261.

Hothorn, T., K. Hornik, C. Strobl, A. Zeileis, M. Hothorn. 2015. Package 'party'. *Package Reference Manual for Party Version 1.0-23* 39.

Hothorn, T., K. Hornik, A. Zeileis. 2006. Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical statistics* **15** 651–674.

Im, M., J. Hughes, O. Kwok, S. Puckett, C. Cerda. 2013. Effect of retention in elementary grades on transition to middle school. *Journal of school psychology* **51** 349–365.

Lee, B., J. Lessler, E. Stuart. 2010. Improving propensity score weighting using machine learning. *Statistics in medicine* **29** 337–346.

Liaw, A., M. Wiener. 2015. Package 'randomforest'. *Package Reference Manual for randomForest Version 4.6-12* 29.

Luellen, J., W. Shadish, M. Clark. 2005. Propensity scores an introduction and experimental test. *Evaluation Review* **29** 530–558.

McCaffrey, D., G. Ridgeway, A. Morral. 2004. Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological methods* **9** 403.

Porro, G., S. Iacus. 2009. Random recursive partitioning: A matching method for the estimation of the average treatment effect. *Journal of Applied Econometrics* **24** 163–185.

Rosenbaum, P. 1989. Optimal matching for observational studies. *Journal of the American Statistical Association* **84** 1024–1032.

Rosenbaum, P. 2002. *Observational studies*. Springer.

Rosenbaum, P., D. Rubin. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* **70** 41–55.

Rubin, D. 1973. Matching to remove bias in observational studies. *Biometrics* 159–183.

Rubin, D. 1997. Estimating causal effects from large data sets using propensity scores. *Annals of internal medicine* **127** 757–763.

Rubin, D. 2006. *Matched sampling for causal effects*. Cambridge University Press.

Setoguchi, S., S. Schneeweiss, M. Brookhart, R. Glynn, E. Cook. 2008. Evaluating uses of data mining techniques in propensity score estimation: a simulation study. *Pharmacoepidemiology and drug safety* **17** 546–555.

Strobl, C., J. Malley, G. Tutz. 2009. An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological methods* **14**.

Westreich, D., J. Lessler, M. Funk. 2010. Propensity score estimation: neural networks, support vector machines, decision trees (cart), and meta-classifiers

as alternatives to logistic regression. *Journal of clinical epidemiology* **63** 826–833.