Computer Age Sociology



Mark D. Verhagen
Nuffield College
University of Oxford

Word count: 47,221

A thesis submitted for the degree of
Doctor of Philosophy
Trinity 2022

Acknowledgements

I would like to start by thanking three institutions that have helped shape this thesis in the past years. Firstly, the vibrant community at Nuffield College, and Oxford more generally, which has provided me with the academic environment and support that has allowed me to develop the ideas in this thesis. Secondly, the hugely exciting, forward-thinking and high-paced research hub that is the Leverhulme Centre for Demographic Sciences of which I consider myself lucky to have become a part of during the process of writing this thesis. Thirdly, the Summer Institute in Computational Social Sciences to which I feel indebted by strengthening my convictions, as a participant of the 2019 edition in Oxford, of the potential of the digital era for social research. To all those who have provided me with so much inspiration and fruitful discussion at these institutions: thank you. My gratitude also goes out to the Economic and Social Research Council and Nuffield College for funding me throughout this process and making all this possible.

Separately, I would like to thank a couple of close friends with whom I could share and reflect on the experiences of the rollercoaster ride that writing a DPhil thesis has been. These are Arun Frey, Michael Garstka, Leo Huberts, and Julius Yam. Knowing that there are always friends nearby that have experienced a similar process has been hugely comforting.

I would also like to convey my sincere thanks to Charles Rahal and Dave Kirk, my supervisors during this process. Although constrained by limited N, I am convinced that I could not have wished for better supervisors. This has been the case in terms of helping me decide the substantive course of the thesis, providing insightful feedback on so many earlier drafts, as well as providing me with the best structure I could wish for to produce this work.

Finally, I would like to start by thanking my parents and brother for making me curious about the world and comfortable questioning it. A special thanks goes out to my mother, who has helped me copy-edit every single chapter in this work. All remaining typographical errors are likely last-minute changes to her edits that I should not have made. I would also like to thank my friends in Amsterdam who have seen me come and go in the past years, but have never ever felt far. Last but not least, I would like to thank Dominique. First, for letting me take care of her cat Poekie, which has helped keep me sane during this sometimes lonely process. Second, for truly letting me be myself.

Abstract

Quantitative sociology is dominated by $\hat{\beta}$-questions, which refer to the process of defining a model of the world and interpreting its estimated coefficients when fit to data. Unfortunately, the situations where we actually know how this model should look are few and far between, yet a monolithic focus on $\hat{\beta}$ remains. This leads to a number of issues, ranging from i) a lack of appreciation of our social theories' ability to reflect the actual world, ii) misrepresentation of the actual patterns of interest resulting in biased findings, and iii) free range for researcher degrees of freedom to muddy findings. Instead, we can learn much from $\hat{y}$-questions that embrace the fact that we do not know the true model of the world, and should approach sociological questions accordingly. In this state of the world, there is great benefit in exploiting ever more powerful computers to assist us in making sense of social life. In this thesis, I explore how two simple ideas from fields that embrace computational power can be used to assist sociologists in their empirical workflow. First, that prediction can be used as a broad model analysis tool that can help sociologists estimate and interpret their models in a better way. Second, that methods from the machine learning domain that excel in finding patterns in data can be used to improve sociological model-building. I illustrate both ideas throughout this thesis with a range of empirical examples and case studies. The goal throughout is to complement — rather than replace — the standard empirical workflow of sociologists with computational ideas and methods. Accordingly, the aim remains to distill interpretable findings about the social world from data. However, I propose we do so without shying away from the computational riches and associated exciting methodological developments that characterize this day and age.

Contents

List of Figures **vi**

List of Tables **viii**

List of Abbreviations **xi**

1 Computer Age Sociology **1**
- 1.1 Introduction . . . 1
- 1.2 The Two Statistical Cultures: $\hat{\beta}$ versus $\hat{y}$ . . . 3
- 1.3 The Potential of Computer Age Methods . . . 7
- 1.4 Looking ahead . . . 11

2 A Pragmatist's Guide to Using Prediction in the Social Sciences **14**
- 2.1 Introduction . . . 14
- 2.2 A New Perspective on Prediction for the Social Sciences . . . 19
- 2.3 The Three Virtues of Prediction . . . 28
- 2.4 Taking Stock . . . 42

3 To Predict and Explain: A Predictive Approach to Studying Teacher Bias **46**
- 3.1 Introduction . . . 46
- 3.2 The Case for Out-of-sample Prediction . . . 50
- 3.3 Empirical Application: Teacher Bias in Student Tracking in The Netherlands . . . 56
- 3.4 Data & Descriptive Statistics . . . 64
- 3.5 Results . . . 69
- 3.6 Discussion & Concluding Remarks . . . 80

4 Incorporating Machine Learning into Sociological Model-building **83**
- 4.1 Introduction . . . 83
- 4.2 Incorrect Models and Risks to Inference . . . 88
- 4.3 A Computational Framework to Improve Model-building . . . 92
- 4.4 Applying the Framework in Practice . . . 110

4.5 Discussion & Concluding Remarks . . . 123

5 Using Tree-based Machine Learning and Registry Data to Study the Equitability of the Dutch Decentralization of the Social Domain 127

5.1 Introduction . . . 127
5.2 Background . . . 131
5.3 A Heterogeneous Policy Effect Approach to Equity . . . 136
5.4 Empirical Case Study: The Dutch Decentralization of the Social Domain . . . 142
5.5 Data and Research Design . . . 146
5.6 Results . . . 156
5.7 Discussion & Concluding Remarks . . . 174

6 Discussion & Concluding Remarks 178

Appendices

A Chapter appendices . . . 188
A.1 Chapter 1 . . . 188
A.2 Chapter 2 . . . 189
A.3 Chapter 3 . . . 190
A.4 Chapter 4 . . . 193
A.4.1 Simulations . . . 199
A.5 Chapter 5 . . . 201
A.6 Chapter 6 . . . 217
T Technical appendices . . . 218
T.1 Misspecification bias . . . 218
T.2 Model fit statistics . . . 219
T.2.1 In-sample statistics . . . 220
T.2.2 Out-of-sample statisics . . . 222
T.3 The ordered probit model . . . 223
T.3.1 Multilevel setup . . . 225
T.4 The Jackknife and the Bootstrap . . . 226
T.5 Tree-based methods for prediction . . . 228
T.5.1 Decision tree . . . 229
T.5.2 Random forest . . . 232
T.5.3 Gradient boosting . . . 233
T.6 Super Learner . . . 235
T.6.1 Oracle result for cross-validated model selection . . . 236
T.7 Shapley values . . . 237

T.7.1 Kernel SHAP . . . 239
T.7.2 Tree SHAP . . . 241
T.7.3 Pairwise Shapley values . . . 242
T.8 Tree-based methods for effect heterogeneity . . . 244
T.8.1 Honest estimation . . . 244
T.8.2 Causal trees . . . 246
T.8.3 Causal forests . . . 247
T.9 Code and packages . . . 248

Bibliography **i**

List of Figures

2.1 Three different strategies to define an estimation and evaluation set . . . 25
2.2 The first virtue of prediction: improving insights into model fit . . . 31
2.3 The second virtue of prediction: providing a holistic benchmarking metric . . 36
2.4 The third virtue of prediction: generating substantive insights into complicated models . . . 41

3.1 Schematic overview of varying the estimation and evaluation sets to generate insights into effect sizes . . . 54
3.2 An example of the latent variable approach mapping continuous outcome y^* to five categories using cut-off points c_k, $k = 1, \ldots, 4$ . . . 59
3.3 Difference in assigned track levels and track levels as suggested by central test by parental education . . . 61
3.4 Educational structure and tracking moment in The Netherlands . . . 63
3.5 Coefficient estimates and predictive performance of a linear probability model and an ordered probit model estimated to data on tracking decisions . . . 71
3.6 Individual out-of-sample track predictions using a linear probability model and an ordered probit model . . . 73
3.7 Predictive performance and modes of random effects when estimating a hierarchical ordered probit model estimated to data on tracking decisions . . . 74
3.8 Predicted track levels when estimating models to separate groups in the data 77
3.9 Modes of school-level random effects in an ordered probit model estimated to data on tracking decisions . . . 79

4.1 Linear model approximations estimated to data from four different DGPs . . 90

4.2 Schematic illustration of the proposed benchmarking framework 93
4.3 Predicted values using two ML methods estimated to data from four different DGPs. 96
4.4 Schematic illustration of a Shapley explanation 103
4.5 Implied versus actual effects based on Shapley explanations estimated to simulated data 106
4.6 Out-of-sample fit statistics comparing four hypothesized functional forms with a GB model estimated to simulated wage data 114
4.7 Implied versus actual effects based on Shapley explanations using simulated data on wages 114
4.8 Out-of-sample fit statistics comparing a hypothesized hedonic regression and a GB model estimated to data on London house sales 117
4.9 Implied effects based on Shapley explanations for six variables using data on London house sales 118
4.10 Out-of-sample fit statistics comparing a hypothesized model and a GB model estimated to data on party affiliation in the United States 121
4.11 Implied direct and interacted effects based on Shapley explanations using data on party affiliation in the United States 122

5.1 Schematic illustration of the Heterogeneous Policy Effect approach and the standard 'Aggregated' approach to studying equity 136
5.2 Development of social assistance usage between 2016 (horizontal axis) and 2019 (vertical axis) for all 208 municipalities that submitted social assistance data to the registry from 2016 onwards 151
5.3 Coefficient estimates of a standard linear probability model estimated to data on social assistance use in 2016 153
5.4 Predicted versus observed social assistance usage at the municipality level based on a linear probability model estimated to 2016 data 153
5.5 Analytical approach to estimating effect heterogeneity 155
5.6 Causal tree results based on 500,000 individuals from the set of consistent municipalities for the years 2019 and 2016 159
5.7 Causal tree results based on 500,000 individuals from both the The Hague area and the Amsterdam area for the years 2019 and 2016 162
5.8 Causal tree results based on 500,000 individuals from the Rotterdam area for the years 2019 and 2016 165
5.9 Causal tree results based on 317,575 individuals from the Almere and 500,00 individuals from the Utrecht area for the years 2019 and 2016 167
5.10 Coefficient estimates from linear models relating heterogeneous effects to demographic characteristics for the set of consistent municipalities and the five alternate municipalities 169

A.3.1 Density plots of test scores for each of the three didactic areas included in the data and box-plots relating test scores with track levels 190
A.3.2 Representativity of the sample relative to the population 191
A.4.1 Bootstrapped distributions of the coefficient of interest estimated to simulated data for three functional forms 193
A.5.1 Social assistance use and by type in 2019 and 2016 for The Netherlands as a whole (top) and the difference between 2019 and 2016 across municipalities (bottom) 201
A.5.2 Correlation plots for all municipalities between the difference in various types of social assistance use between 2019 and 2016 202
A.5.3 Covariate balance of the data before and after balancing using propensity score matching and entropy balancing 203
A.5.4 Out-of-sample error of causal trees estimated as a function of tree depth . . . 203
A.5.5 Estimated policy effect between 2019 and 2016 compared to the prevalence in 2016 per heterogeneous group 204
A.5.6 Estimated policy effects through subset regression based on causal tree samples and the entire population 205
A.5.7 Pre-2015 trend of social assistance use 206

List of Tables

2.1 Number of articles mentioning `predict` or `prediction` in the abstract, and actual usage of prediction from six flagship journals in economics, political science, and sociology, 2010 to 2021. 16

3.1 Descriptive statistics of the data on tracking decisions 67
3.2 Regression results from linear probability models and ordered probit models estimated to data on tracking decisions. 70
3.3 Regression results from ordered probit models estimated to data on tracking decisions including various school-level characteristics. 75

4.1 Regression results for three linear additive models and a plot of the implied and true age effect based on simulated data 89
4.2 Super Learner performance of various models estimated to simulated data . . 98
4.3 Out-of-sample fit statistics comparing three linear additive models with a GB model estimated to simulated data 99

4.4 Descriptive statistics of the simulated wage data 112
4.5 Descriptive statistics of the London house price data 116
4.6 Descriptive statistics of the GSS data 120

5.1 Number of municipalities and the size of the population living in these municipalities across time for i) the total population in The Netherlands, ii) those municipalities submitting data to the registry, and iii) those municipalities submitting data in both 2016 and 2019 148
5.2 Descriptive statistics for 2019 and 2016 of all 208 municipalities who submitted social assistance data from 2016 onwards to the Dutch registry, and total population descriptive statistics for the year 2016 150
5.3 Descriptive statistics of the data by social assistance use 151
5.4 Estimated ATEs using various estimation methods estimated to data on social assistance use in 2019 and 2016 156
5.5 Variable importance scores for a causal forest estimated to data on social assistance use in 2019 and 2016 157

A.3.1 Correlation matrix of the data on tracking decisions 192
A.4.1 Super Learner performance, simulated data 194
A.4.2 Super Learner performance, simulated wage data for the models Linear-I (top) and Linear-II (bottom) 195
A.4.3 Super Learner performance, simulated wage data for the models Linear-III (top) and Linear-IV (bottom) 196
A.4.4 Super Learner performance, London house price data. 197
A.4.5 Super Learner performance, GSS data 198
A.5.1 Proportion of social assistance use between 2016 and 2019 by univariate demographic characteristics 207
A.5.2 Estimated ATEs using various estimation methods for the set of consistent municipalities and each of the alternate municipalities 208
A.5.3 Omnibus test of heterogeneity in the overall policy effect 209
A.5.4 Descriptive statistics of the heterogeneous groups identified for the consistent municipalities and the The Hague area 210
A.5.5 Descriptive statistics of the heterogeneous groups identified for the Amsterdam area and the Rotterdam area 211
A.5.6 Descriptive statistics of the heterogeneous groups identified for the Almere area and the Utrecht area 212
A.5.7 Regression results decomposing policy effects by demographics for the consistent municipalities and each of the alternate municipalities 213
A.5.8 Percentage of population by migrant background for The Netherlands and each of the alternate municipalities 214

A.5.9 ATE setting 2016 as the treatment year and 2015 as the control year using covariate balancing 215
A.5.10 Missingness of the education variable for the entire sample as well as for adults (18+), including the average age of the entire sample versus those with missing education 216
T.1 `R` (4.2.1) and `Python` (3.8.10) packages used throughout this thesis. . . . . .

List of Abbreviations

ABW . . . . . General assistance act ('Algemene Bijstands Wet')

AWBZ . . . . . Special care act ('Algemene Wet Bijzondere Ziektekosten')

ATE . . . . . Average treatment effect

CAK . . . . . Central administrative office AWBZ ('Stichting Centraal Administratie Kantoor AWBZ')

CBS . . . . . Statistics Netherlands ('Centraal Bureau voor de Statistiek')

CoE . . . . . Causes of effects

CV . . . . . Cross validation

DGP . . . . . Data generating process

EMSE . . . . . Expected mean squared error

EoC . . . . . Effects of causes

FFC . . . . . Fragile Families Challenge

GB . . . . . Gradient boosting

GSS . . . . . General social survey

HPE . . . . . Heterogeneous policy effects

LASSO . . . . Least absolute shrinkage and selection operator

LOESS . . . . Locally estimated scatterplot smoothing

LOO . . . . . Leave-one-out

LR . . . . . Likelihood ratio

LTC . . . . . Long-term care

MSE . . . . . Mean squared error

ML . . . . . Machine learning

OLS . . . . . Ordinary least squares

OOS . . . . . Out-of-sample

pp . . . . . Percentage point

PTS . . . . . Pupil tracking system

PW . . . . . Social care act ('Participatie Wet')

RF . . . . . Random forest

RMSE . . . . . Root mean squared error

SES . . . . . socioeconomic status

X-AI . . . . . Explainable artificial intelligence

WLS . . . . . Weighted least squares

WMO . . . . . Social assistance act ('Wet Maatschappelijke Ondersteuning')

"[T]he avalanche of digital data that is being generated both by online activities and also parallel efforts to digitize existing archival records, *combined with massively faster computational methods and methodological advances in statistics and econometrics* is opening up many exciting opportunities for sociologists to explore questions of long-standing interest with better tools and data." (Emphasis added)

Duncan Watts, 2014

Chapter 1

Computer Age Sociology

1.1 Introduction

The title of this DPhil thesis is inspired by the 2015 book 'Computer Age Statistical Inference' by Bradley Efron and Trevor Hastie. Their book was dedicated to incorporating some of the exciting methods that had been developed in fields on the periphery or outside of statistics, which were tackling topics very similar to those classical statistics had for many centuries. The methods described by Efron and Hastie heavily relied on computational resources, delegating parts of the statistical process typically reserved for the researcher to the computer instead. This DPhil thesis attempts to achieve a similar goal: to incorporate computational methods into sociological work. Just as quantitative sociological research has heavily relied on classical statistics to establish regularities in the social world, it can benefit from computational methods developed with similar goals in mind. This idea lies at the core of 'computer age sociology', the central theme of this thesis. In this introductory chapter, I outline a vision for computational methods to complement the typical empirical workflow of sociologists and help them address a number of practical issues that are prevalent in quantitative sociology.

This thesis falls within a broader turn towards computational methods that can be observed within the social sciences (Lazer et al., 2009, 2020). The digital era has led to new sources and scales of data, like digital trace, social media or digitalized archival data (Salganik, 2019). These new types of data require new methods and approaches to draw inferences from them. In parallel, social scientists

have been using the newfound riches of computational power to improve on classical methods, which were designed with ease-of-estimation and parsimony in mind (Efron and Hastie, 2016, Chapter 5). A case in point is the use of specification curves and related visualizations of model uncertainty, which estimate a large number of 'plausible' models similar to that specified by a researcher to assess robustness of a functional form and uncertainty in model selection (Sala-i Martin, 1997; Young and Holsteen, 2017; Simonsohn et al., 2020). This thesis falls within the second type of computational social sciences, but goes a step further in fully embracing methods and ideas from computational domains like Machine Learning (ML), which often have their roots in computer science, to complement quantitative sociological work.

The direct motivation for this thesis is the development of exciting methods to find patterns in data that heavily rely on the exponential increases in computational power available to researchers in this day and age.[1] These methods have already revolutionized many aspects of industry and research (Rahal et al., 2022; Salganik, 2019, Chapter 1). The crucial paradigm shift of ML has been to let algorithms find patterns in data instead of relying on humans to formulate the relationships amongst variables in the form of a model. As is often the case with technological developments, this sounds more sinister than it really is. Take the example of the set of popular tree-based ML methods, which are also implemented in the second half of this thesis. At the core of tree-based methods is a simple search algorithm that is provided with a basic goodness-of-fit metric similar to those optimized in a linear regression. Instead of formulating the model a-priori, tree-based methods simply search across all explanatory variables for some value by which the goodness-of-fit metric improves the most when splitting the data in two parts and assigning both parts the average of the outcome as fitted values. This process is then repeated on the subsequent partitions of the data and can lead to highly non-linear functional forms. This approach seems both simple and obtuse, but is hugely effective and, crucially, cheap to perform in an age of computational riches, but would have been

[1] Moore's law stating that the number of transistors on a microprocessor chip doubles approximately every two years — and by extension its computational power — has held true for the majority of the last half century (Waldrop, 2016).

computationally burdensome when constrained by 20th century computer power. The goal of this thesis is to bridge the current gap between these kinds of methods and related ideas, and the standard empirical workflow in sociology.

1.2 The Two Statistical Cultures: $\hat{\beta}$ versus $\hat{y}$

At the heart of the current divide between computational methods like ML and standard empirical work are two distinct approaches to statistical inference. In quantitative sociology, we are typically interested in $\hat{\beta}$-questions or, as Leo Breiman would say, in *data models* (Breiman, 2001, p. 199).[2] The goal is to identify empirical regularities in the social world using data and statistical models representing how data relate to one another. Given some theory — for example that students from a disadvantaged background receive disproportionately unfavorable assessments by their teachers, which is the topic of Chapter 3 and a classic sociological field of inquiry (Bourdieu and Passeron, 1979) — a model is hypothesized reflecting the way in which teachers might formulate opinions of students. This model incorporates the possibility for bias towards different groups of students — i.e., favouring some types of students over others. This model could include an ability assessment of students by their teacher as an outcome of interest, and an indication of their 'true' ability (like test scores) as well as demographic characteristics as explanatory variables. Crucial to the $\hat{\beta}$-approach is that the sociologist defines the exact way in which these variables affect one another. Then, after collecting data on the variables and estimating the model, we can assess evidence in favor of or against the hypothesis by assessing the estimated parameters. This framework of hypothesizing about some social regularity, formulating a model capturing this regularity, collecting data, estimating the model, and assessing evidence in favor of or against the regularity can be observed throughout quantitative sociology.

Computational methods like those from ML are developed with a slightly different goal in mind, instead emphasizing $\hat{y}$-questions and *algorithmic models* (Breiman,

[2]The $\hat{\beta}$ nomenclature broadly refers to the parameters in a model, typically coefficients in a linear regression model, which are to be estimated from the data.

2001, p. 199).[3,4] Rather than explicitly specifying a model that allows for a direct inferential focus in the association amongst variables — e.g., whether and how demographic characteristics affect a teacher's evaluation — the emphasis lies on predicting the outcome well. In this process, and in contrast to the sociologist, we would only determine a set of variables that may affect the outcome of interest, which could be the very same variables as the sociologist had included in their model. However, the exact way in which these variables affect the outcome is left unspecified and for the computer to decide. In what follows, the computer might very well conclude that the model originally formulated by the sociologist fits the data best or, instead, that different patterns amongst the explanatory variables and outcome are appropriate. This seems like a strict improvement on the sociologist's approach, particularly when the algorithmic approach leads to a much better fitting model. This improved pattern recognition comes at a price, though. We typically don't know why the ML model fits the data well and we wouldn't even know whether the algorithm lead to the exact same model as the sociologist had originally hypothesized. We simply have a (possibly complex) model $\bar{f}(\cdot)$ that fits the outcome accurately when provided with explanatory variables. This model might fit the data better than the sociologist's model, but it needn't be informative for the regularity we are interested in.

The difference between the two approaches has been dubbed the 'two cultures in statistical modeling' by Leo Breiman (2001); the first assumes that the data is given by some known stochastic model whereas the second assumes this process is unknown and uses algorithms to discern patterns describing it. More recently, Galit Shmueli has rephrased the two approaches above as the question whether 'to predict or to explain' (2010).[5] Sociologists overwhelmingly subscribe to explanation

[3]I interchange ML, algorithmic modeling, and computational methods throughout this thesis. With these terms, I broadly refer to the set of methods to find patterns in data that rely heavily on computational power and the broader set of ideas and concepts used to assist in this process.

[4]The $\hat{y}$ nomenclature broadly refers to the predicted values of the outcome using an estimated model.

[5]Although Shmueli is less committed to either of the two approaches, Breiman spares little love for the culture of data models. As he states '[t]he statistical community has been committed to the almost exclusive use of data models' continuing that '[t]his commitment has led to *irrelevant theory*, *questionable conclusions*, and has *kept statisticians from working on a large range of*

and $\hat{\beta}$-questions to model the social world. In doing so, they put considerable trust in the ability of the researcher to specify a data model that correctly defines the ways in which explanatory variables affect the outcome. This is a crucial leap of faith, as statistical inference into $\hat{\beta}$ heavily relies on correctly specified models.[6] Conversely, machine learners do not share this blind faith in the researcher's ability to specify the model of the world a-priori and use the computer to identify patterns. Their models typically fit the outcome well, but tend to be limited in providing insights into the patterns found — although exciting developments are being made in this domain (Samek et al., 2019). This lack of inferential ability has led many social scientists to conclude that ML can only help sociologists in purely predictive tasks rather than explanatory work (Mullainathan and Spiess, 2017; Molina and Garip, 2019). For example, to use ML to better impute missing data, perform auxiliary regressions or to replace otherwise tedious human tasks like labeling images. These are all examples where the revolutionary potential of ML for sociology is undisputed. However, restricting the use of ML to such tasks is shortsighted and the fact that ML models excel in identifying patterns holds relevance for empirical work more broadly. There is a wealth of ways in which ML methods can complement sociologists beyond replacing less precise predictive models.[7]

To appreciate the potential of algorithmic modeling, we need to embrace the fact that we live in a world of messy data and fundamentally unknown data generating processes (DGPs). This means that in practice, sociologists know their models are likely incorrect representations of the real world and iteratively improving these incorrect models is a fundamental part of the scientific process.[8] That we frequently find ourselves in situations where we do not know the true model would suggest

interesting current problems' (emphasis added) (Breiman, 2001, p. 199).

[6]Misspecification of a functional form, for example by assuming that an explanatory variable affects the outcome linearly whilst the effect is not, leads to omitted variable bias and will typically bias $\hat{\beta}$ (see Section T.1).

[7]There is also an increasing sense that predictive ability is an important property of social theories more generally (Hofman et al., 2017; Salganik et al., 2020). Some even go as far as argue that most, if not all quantitative sociologists (perhaps unknowingly) would indeed subscribe to the goal of making predictive statements about the world (Watts, 2014).

[8]This process can be observed both in terms of developing methods that are robust to certain types of violations of model assumptions (White, 1981; Buja et al., 2019b), as well as improving the alignment of our models with the real world.

that we can learn much from the algorithmic modeling culture that takes this very same uncertainty as a starting point. Instead, sociologists overwhelmingly tend to approach empirical work from the $\hat{\beta}$-perspective and make inferences about the social world based on their self-hypothesized models, as if these were correct representations of the true underlying processes. Practically, this leads to a number of problems. Three of these problems are discussed in more detail below, and form key topics to which I will frequently return throughout this thesis.

First, mis- and under-specification are considerable risks to empirical analysis and $\hat{\beta}$-inference. For example, one of the most frequently implemented models of the world is a linear additive functional form estimated using ordinary least squares (OLS), where explanatory variables are assumed to affect the outcome independently from one another and to a similar degree irrespective of other explanatory variables or distributional location. Although such stylized relationships might hold true in the natural world, it seems unlikely to be as clearcut in the social world (Freedman, 1991; Abbott, 1988). This view is repeatedly echoed through the richness and depth in the work of our qualitative colleagues. Even though it is well known that misspecification can lead to serious bias in inference and invalidate results (Section T.1), quantitative sociology overwhelmingly remains wedded to simple functional forms (Hindman, 2015). Tellingly, even in cases where we know that our models ought to be more complex, social scientists still tend to prefer the ease-of-estimation of parsimonious models over more complex, but more appropriate alternatives (Liddell and Kruschke, 2018). An example of this phenomenon is illustrated in Chapter 3.

Second, besides correct model specification, standard statistical inference tools also require that the model is formulated before having seen the data and not tinkered with since or at least estimated on a sample of the data not used in model-building (Leamer, 1983; Ioannidis, 2005; Cameron and Trivedi, 2005, p. 286). The 'crisis in science' stems from frequent violations of these conditions, as researchers are able to repeatedly estimate different models and thus incorporate researcher degrees of freedom into the estimation process that invalidate the reported uncertainty estimates (Gelman and Loken, 2013; Young, 2019). This risk of repeated estimation

is further compounded by publication pressure within academia and a preference by journal editors to publish surprising findings (Kasy, 2021; Ioannidis, 2005).

Third, a re-appreciation for the predictive potential of sociological knowledge has shown that our social theories, even when assuming these are correctly modeled and analyzed, can be poor predictors of the outcomes they should supposedly determine (Salganik, 2019). In general, a focus on $\hat{\beta}$ has distracted attention away from the relevance of $\hat{y}$ and can lead to statistically significant but substantively irrelevant mechanisms to be identified and published (Watts, 2014).[9]

These three practical issues facing quantitative sociology all stem from a preference for data models and $\hat{\beta}$-questions, whilst we reside in a world where we are highly uncertain about the true underlying patterns in the data. The point is not that inferential questions aren't the right ones to ask, quite to the contrary. But given that we live in a state of the world where we do not know the true model a-priori, we can probably be assisted by methods developed with a similar level of uncertainty in mind.

1.3 The Potential of Computer Age Methods

There are numerous ideas and methods in the ML domain that speak directly to the three issues outlined above. First, prediction is central to most ML methods and sociologists using these methods will have to face up to their model's ability to accurately predict the outcomes we are interested in. Even when estimating standard statistical models, a re-appreciation of prediction can provide many complementary insights to standard empirical work. Second, repeated specification searches and researcher degrees of freedom can be partly neutralized when using algorithms to find the best fitting solution given the data, instead of the data *and* the researcher. Especially as concerns regarding overfitting a model to the sample

[9] A similar point is made by Ed Leamer when discussing the difference between 'statistical significance' — or just 'significance', as it is often used in practice — on the one hand, and the actual practical relevance of a coefficient on the other. Leamer suggests we replace (statistically) significant by 'measurable' in order to minimize confusion that something statistically significant is also meaningfully relevant in the real world (`https://www.econtalk.org/leamer-on-the-state-of-econometrics/`, accessed September 30th, 2022).

at hand is central to virtually every ML method, and out-of-sample validation and bootstrap / Monte Carlo inspired resampling are commonly applied tools. Third and relatedly, when using ML models to determine the complexity of the patterns that are modeled, we limit the risk that our models are overly simplistic and do not accurately reflect the patterns we are interested in. The central aim of this thesis is to incorporate these computationally inspired ideas and approaches into the empirical workflow of the quantitative sociologist.

Concretely, this thesis consists of four substantive chapters, which can be divided into two thematic parts. Each part consists of a broad introduction of a computational method or idea, as well as a concrete application. The first part (Chapters 2 and 3) is dedicated to incorporating prediction into standard model analysis, and the second part (Chapters 4 and 5) explores the potential of actively using computational models within the sociological workflow.

Specifically, in Chapter 2 I discuss the potential of prediction as a tool to complement model analysis in a broad sense. I showcase how the simple process of making predictions using an estimated model can lead to enhanced insights at virtually no cost to the researcher.[10] At the same time, I show how these insights can meaningfully change conclusions or prompt further scrutiny of the data. I show how a predictive perspective can bring $\hat{y}$-relevance into empirical work — one of the three recurring themes in this thesis — without requiring any change of models or methods. I also show how prediction can be a useful tool when we do want to incorporate more complex methods, which may be more appropriate but defy straightforward interpretation — another of the recurring themes in this thesis. Finally, as is common in the computational domain, I illustrate how repeatedly estimating predictive accuracy on resampled datasets can improve robustness of the reported results — the third and final recurring theme in this thesis.

Chapter 3 puts this predictive perspective into practice by studying teacher bias in assigning primary school students to track levels in The Netherlands, an extremely consequential moment in the life course that dictates the level of further education

[10] Note that prediction as understood in this chapter is as a tool to improve model understanding. Optimizing predictive accuracy for prediction's sake is not the goal.

accessible to a student. I use prediction as a way to generate understanding into more appropriate, but less interpretable methods. An improved alignment that leads to new substantive findings on the presence of teacher bias. Throughout, I also consider improvements in $\hat{y}$ when introducing new modeling components and possible mechanisms through which teacher bias might manifest itself, leading to various surprising findings.

This brings us to the second part of this thesis, which actively explores how ML models can directly benefit the sociological workflow. Chapter 4 discusses the potential of tree-based methods to discover patterns that may not have been hypothesized by the researcher and outlines a benchmarking framework where researcher-hypothesized models are compared with computational alternatives. When the former strongly underperforms the latter, we can establish a lack of ability to accurately model the data in our hypothesized models, which implies we need to allow for different associations amongst variables. In case a difference in fit is found, methods from the field of Explainable A.I. (X-AI) are used to provide insights into the type of patterns that are missing in our own models so that they can be improved accordingly. Using ML methods to benchmark researcher-hypothesized models not only improves alignment with the patterns in our data, it also provides new transparency into the research process. The fit potential of ML methods is a function of the data, rather than the researcher, and should improve detection of dubious specification searches.

In Chapter 5, I put the intuition that ML methods can find patterns in data that we may not have hypothesized a-priori into practice by studying the Dutch decentralization of the social domain using registry data. Of particular interest, and an ongoing topic of discussion in the literature, is the potential for decentralization to lead to heterogeneous effects for different demographic groups and / or regions. However, rather than hypothesizing a-priori which demographic groups or regions might exhibit heterogeneous effects — which can suffer from missing relevant heterogeneous effects — or simply trying all possibilities — which quickly becomes intractable but also risks p-hacking and spurious findings — I implement tree-based

ML models to identify such heterogeneous groups. Using the data to find patterns of interest rather than hypothesizing them a-priori provides a complete picture of the impact of the intervention, which is particularly relevant when there are no strong prior hypotheses on groups that might show heterogeneous effects. Again, prediction, correct model specification, and robustness to researcher degrees of freedom are key topics throughout the second part of the thesis.

Although this thesis can be broadly divided into the two thematic parts described above, the boundaries between them are not particularly strongly drawn. Each chapter is written as a substantive article that should be able to stand on its own, so some concepts and ideas might be repeated across chapters. This naturally includes discussing elements from either part of the thesis at certain points in time. For example, a discussion of the potential of ML methods to benchmark our own models (Chapter 4) requires a discussion of the use of prediction (the topic of the first part of the thesis). As a consequence, there will also be a level of intrinsic redundancy across chapters due to the stand-alone nature of each chapter as a self-contained article. Whenever possible, I refer the reader to earlier chapters throughout introductory sections in order to limit this redundancy somewhat.

There are also numerous aspects related to computational methods, which can be beneficial to quantitative sociology but are not actively explored here. These include the use of widely shared datasets for which quantitative benchmarks can be set, like predictive performance. Such common benchmarks have helped computational researchers to work together on improving their models, and have for instance led to some of the astronomic improvements in speech and text recognition models today. More generally, open science is a much more central topic to ML-based research from which sociologists could take note, although we are already witnessing improvements across the social sciences in this area (Open Science Collaboration, 2015). Collaborative work — particularly through the Common Task Framework, where many research teams attempt to find the most appropriate solution to a well-defined problem (Donoho, 2017) — is frequently observed in ML as well, and is slowly being introduced into sociology. The important example of the Fragile

Families Challenge being a case in point (Salganik et al., 2020). Although the above topics are less actively explored in this thesis, all code underlying the empirical results presented throughout, as well as instructions to obtain the data when publicly available are shared for others to use.[11]

1.4 Looking ahead

As mentioned earlier, this thesis is not about trying to initiate wholesale changes to quantitative sociology, nor is it about taking sides in heated debates like those that sometimes arise between proponents of prediction or explanation. The approach I take throughout this thesis is fundamentally pragmatic and introduces methods and ideas that are exclusively complementary to the standard empirical workflow of sociologists. Applying these methods does not require 'picking sides' between $\hat{y}$ or $\hat{\beta}$ or other seeming dichotomies along which the use of computational versus conventional methods are usually set.[12] We can still have the ambition to estimate data models representing the social world, but this does not mean we should shy away from the tools of algorithmic modeling.

Another possible misconception about the goals of this thesis might be that incorporating computational methods is effectively aiming to take the sociologist out of the scientific process. The goal is quite to the contrary. Improved pattern recognition is just a small (albeit important) part of the empirical workflow when compared to theorizing about social regularities, determining appropriate explanatory variables and data, interpreting results, placing these results within existing theory, and / or translating them into policy recommendations (Pigliucci, 2009; Sampson et al., 2013). An analogy can be made with the difference between *counterfactual-based* causality and *mechanistic* causality (VanderWeele, 2009, p.

[11] Code underlying the analyses, and instructions to obtain the data can be found at `https://github.com/MarkDVerhagen/DPhil_Thesis_Computer_Age_Sociology`.

[12] Similarly, the use of computational methods transcends the dichotomy between "causes of effects" (CoE) and "effects of causes" (EoC). Although purely predictive approaches might seem more fitting to CoE-questions, estimating correct functional forms and using predictive tools to better understand how estimated models work is just as crucial to research emphasizing EoC (Heckman, 2008, p. 4-5).

222). The former is considered to be simple "black box" causality that applies methods from the causal inference domain to estimate the causal effect of some intervention. The latter is the considerably more complex ambition of "look[ing] inside the black-box" and "to understand the mechanisms governing the various processes which give rise to particular outcomes."[13] Similarly, computational methods can broadly improve how we build a model of the world. However, arguably the most important part of the empirical process, which is making sense of an estimated model, remains squarely within the domain of the sociologist. Calls that computational power is ushering in 'the end of theory' (Anderson, 2008) strongly misrepresent the vision of how computational methods will change sociology that is presented here. In fact, the methods and ideas outlined here should help stimulate an era of renewed theory-building, albeit one that is more pragmatic and cooperative in nature.

The enduring relevance of the researcher should also be reflected by the substantive findings that are reported throughout this thesis. Both Chapter 3 concerning the existence of teacher bias towards student characteristics, and Chapter 5 on inequity due to decentralization require extensive understanding into the local context and existing theory to design, estimate and meaningfully interpret model results. Computational methods simply replace certain steps in the workflow that would otherwise be skipped or simplified. Importantly, as both chapters illustrate, these methods considerably enhance our understanding of the substantive topics under study as well. For teacher bias, likely under-estimation of teacher bias towards disadvantaged students was found as well as the importance of the institutional level (i.e., schools) in determining track levels. With regards to decentralization, various unexpected patterns and consequences of decentralization were identified like local capture, providing policymakers with actionable insights into the consequences of decentralization.

[13] As VanderWeele notes, a similar distinction has been made by Heckman (2008) who distinguishes "statistical causality" from a "scientific model of causality", which requires considerably more cognitive effort.

To summarize, this DPhil thesis is dedicated to improving the tried and tested empirical workflow in quantitative sociology by incorporating methods and ideas from more computational domains, particularly ML. This does not require any re-inventing of the wheel or drastic changes to the way in which we do research. Instead, it requires a pragmatic perspective and willingness to tackle a number of enduring issues that plague the social sciences. These include being overly reliant on $\hat{\beta}$-questions and losing track of the relevance of $\hat{y}$, but also remaining wedded to models that stem from an age of computational constraints. Perhaps most importantly, computational resources can help fight the crisis in science that stems from interference of researcher degrees of freedom and p-hacking in quantitative work. My central thesis is that computational methods can greatly complement our empirical work by providing tools, ideas and methods to help sociologists address these practical issues facing quantitative sociology. A re-appreciation of the value of prediction and an acceptance that our own hypothesizing may not always lead to the correct model should help pave the way for a natural symbiosis between computational methods — that excel in these themes and have already revolutionized so many aspects of daily live and research — with quantitative sociological work. In other words, in this thesis I argue that it is high time to fully embrace the potential of computer age sociology.

Chapter 2

A Pragmatist's Guide to Using Prediction in the Social Sciences

Chapter abstract: Prediction is an underused tool in the social sciences, often for the wrong reasons. Many social scientists confuse prediction with unnecessarily complicated methods or with narrowly predicting the future. This is unfortunate. When we view prediction as the simple process of evaluating a model's ability to approximate an outcome of interest, it becomes a more generally applicable and disarmingly simple technique. For all its simplicity, the value of prediction should not be underestimated. Prediction can address enduring sources of criticism plaguing the social sciences, like a lack of assessing a model's ability to reflect the real world, or the use of overly simplistic models to capture social life. I illustrate these benefits with empirical examples that merely skim the surface of the many and varied ways in which prediction can be applied, staking the claim that prediction is a truly illustrious "free lunch" that can greatly benefit social scientists in their empirical work.[1]

2.1 Introduction

Social scientists should start using prediction more often. Prediction is the process of generating predicted values of a dependent variable by applying an estimated model to a set of explanatory variables, bringing a unique analytical perspective to empirical work. Prediction can also help address enduring sources of criticism facing the social sciences. Examples are a general lack of assessing research findings in terms of their real-world relevance, and the use of overly simplistic models to study the complexities of social life. In this chapter, I address common misconceptions about prediction and provide a simple definition that addresses existing barriers to adoption. I then discuss and illustrate some of the many benefits that prediction can bring when

[1]This chapter has been published in *Socius* (Verhagen, 2022).

used as a complement to traditional empirical methods. I argue that prediction can and should become a fundamental part of the social scientist's empirical toolkit, but that this first requires us to look beyond the current dichotomy between prediction and explanation and instead view the two as complementary to one another.

The current lack of prediction in the social sciences stems from a seeming incompatibility between wanting to explain and wanting to predict, effectively forcing the researcher to choose between the two approaches. A case in point is the much-cited paper by Galit Shmueli — aptly titled 'To Explain or to Predict' — which outlines how a social scientist's empirical workflow differs in terms of data processing, modeling, and post-estimation diagnostics when choosing to either predict or explain (Shmueli, 2010). Naturally, the paper assumes that a researcher would not normally attempt to do both. This is an accurate reflection of social science research. The apparent need to dogmatically choose between either approach means that, in practice, social scientists tend to stick to explanation almost exclusively. Illustratively, the terms "predict" and "prediction" are mentioned in less than 5% of abstracts over the last ten years in various flagship journals in economics, political science and sociology, and of the papers mentioning either term, only 13% proceed to generate actual predictions of the outcome variable (Table 2.1).[2,3]

So why does this dichotomy exist? In many cases, an unnecessarily narrow interpretation of prediction is to blame. For example, the type of prediction discussed by Shmueli refers to the practice of maximizing predictive power, which is a subset of the more general practice of making predictions. As a result, prediction is often conflated with the use of complex non-linear models like those from the domain of ML, which have their own unique set of challenges (Athey and Imbens, 2019). There is no reason to transfer these challenges to the general process of making predictions,

[2]In most of the articles that mention the term 'predict' or 'prediction', the authors use the commonplace, conceptual meaning term — e.g., 'we predict that' or 'our theory makes several predictions'. The actual process of making predictions of the outcome variable is virtually non-existent in the literature cited. Note that the term "explain" or "explanation" only features in about 13% of abstracts, although this proportion likely does not reflect the proportion of work that is explanatory. Explanation is the default approach to empirical work, making it less relevant to explicitly mention the term in the abstract.

[3]The single sociological paper making predictions in fact generated mortality forecasts (Miech et al., 2011).

Table 2.1: Number of articles mentioning `predict` or `prediction` in the abstract, and actual usage of prediction from six flagship journals in economics, political science, and sociology, 2010 to 2021.

Journal	Total articles[†]	Mentions prediction	Actually makes predictions[*]
American Economic Review	2414	85	12
Quarterly Journal of Economics	458	47	3
American Journal of Political Science	800	61	14
American Political Science Review	743	33	4
American Journal of Sociology	394	8	0
American Sociological Review	523	24	1

[†]Data were collected using the Scopus API.

[*]Understood as generating predictions of the outcome variable (including forecasting). Papers generating predicted probabilities by setting explanatory variables to their mean or median were excluded as such predictions don't reflect actual observation in the data.

which can be done with any type of model. Prediction also tends to be narrowly understood as forecasting, which is but one of numerous examples of making predictions (Watts, 2014).[4] The biggest culprit, however, is the enduring discussion whether prediction and explanation are conceptually the same, and whether the latter should imply the former. This philosophical debate, although interesting, is ultimately irrelevant to applying prediction in explanatory research. When viewing prediction as a simple tool to evaluate a model's ability to approximate the outcome of interest, it can be applied without exception to most social science questions, rendering a dogmatic choice between prediction or explanation unnecessary.

The more relevant question is what prediction might bring to the table. To illustrate just one dimension, take the Fragile Families challenge (FFC), a case

[4]Prediction is more commonly encountered in those social science domains that put emphasis on forecasting and projecting. Typical examples are the analysis of (financial) time-series, but can also include the prediction of conflicts or rare events (Cederman and Weidmann, 2017), network science (Cheng et al., 2014), and demography (Booth, 2006).

which I will return to throughout this chapter (Salganik et al., 2019b). During the FFC, 160 research teams around the globe were asked to predict a number of important early-stage life outcomes of general interest to social scientists (e.g., eviction and material hardship). The idea was to evaluate the general predictability of these outcomes through a Common Task Framework (Donoho, 2017). This setup mirrored the popular competition website Kaggle, where datasets with some outcome of interest and a number of possible explanatory variables are published online. Participants are challenged to estimate models that can accurately predict the outcome. These models are then tested on a partition of the data that is kept secret. Similar to Kaggle, the FFC made available a rich dataset to generate predictive models, while storing an evaluation set against which each team's predictive model was scored. The organizers encouraged the use of prediction-focused algorithms, rather than the explanatory methods already applied in hundreds of peer-reviewed articles using the same data.

The conclusions were telling for a number of reasons. First, many teams applied methods using flexible functional forms and variable selection techniques not often seen in the social sciences. Second, most models were nonetheless poorly able to predict life outcomes, although some did improve on benchmark models including a curated number of explanatory variables in a standard linear model. Poor overall predictability was thus a feature of both predictive and explanatory techniques. Third, and most important, the FFC was a rare occasion where the onus was truly on prediction rather than explanation. As a consequence, it put into sharp focus the fact that decades of explanatory research into the outcomes of interest had not led to much predictive ability. This somewhat awkward finding led the organizers to conclude that researchers had to 'find a way to reconcile a widespread belief that understanding has been generated by these data ... with the fact that the very same data could not yield accurate predictions of these important outcomes' (Salganik et al., 2020, p. 8402).

If prediction had been a more natural tool for social scientists, the main take-away of the FFC would likely not have taken so long to materialize. An earlier realization

of the predictive limits of our knowledge might have stimulated a rigorous evaluation of the mechanisms hypothesized, the methods employed, and/or data collected at an earlier stage in the dataset's rich academic career and throughout life course research more generally. Importantly, it is unlikely that poor predictability is only a feature of life course research. Assessing the ability of our research findings to meaningfully predict outcomes we are interested in will most likely spur important debate in many other domains as well. The point is that our traditional preference for in-sample diagnostics means that we often don't assess our models ability to approximate the outcomes we care about. Prediction can, amongst other things, solve this problem.

The FFC is but a single example how prediction can shine a different light on empirical work and represents one of many approaches to making and evaluating predictions. More generally, this chapter argues that prediction can bring the following three key virtues to the table of the social scientist:

1. Prediction provides improved insight into model fit.
2. Prediction provides a benchmarking tool across modeling domains.
3. Prediction can help generate insight into the behavior of complicated models.

These key virtues come in addition to other benefits. Some examples are an improved alignment of research findings and policy (Athey and Imbens, 2019), providing a metric to focus scientific efforts towards (Forster, 2002; Hofman et al., 2017; Cranmer and Desmarais, 2017), and improving transparency and the ability to scrutinize estimated models (Greenhill et al., 2011). Viewing prediction as a complement to classical methods would also ease the incorporation of prediction-focused methods from ML into the social sciences (Rahal et al., 2022).

To summarize, with this chapter I aim to increase the use of prediction in explanatory research by challenging the unnecessary dichotomy between prediction and explanation, and illustrating the many benefits prediction brings when applied as a complement to explanatory analysis. Hopefully, this chapter can serve as a pragmatic guide to the varied ways in which prediction can be successfully applied in the social sciences. The remainder of this chapter is structured as follows. First,

I will discuss several reasons why prediction is currently being underutilized, and provide a definition of prediction that should address these obstacles to adoption. I then provide a number of ways in which prediction can be operationalized, dependent on the case at hand. To showcase the benefits of prediction, I will then present three sets of empirical examples — in line with the three virtues outlined above — to illustrate the application of prediction. I conclude the paper with a summary and discussion of the main claims and findings.

2.2 A New Perspective on Prediction for the Social Sciences

The social sciences are currently dominated by a focus on explanation. This often boils down to estimating models reflecting some explanatory mechanism and assessing the in-sample coefficient estimates of these models. Prediction — which broadly reflects an interest in how well the models we estimate are able to approximate the dependent variable — plays, at best, an auxiliary role.[5] Below, I identify three reasons why a predictive focus in the social sciences is lacking. Then, I will provide a simple definition of prediction that should not suffer from such barriers to adoption.

Before doing so, it is appropriate to briefly reflect on the intriguing philosophical debate whether explanation and prediction are conceptually the same. Some authors have forcefully claimed that causal explanation should always have predictive implications (Hempel and Oppenheim, 1948; Freedman, 1991; Watts, 2014) whereas others (equally strongly) qualify this viewpoint (Hedström and Ylikoski, 2010). This chapter does not seek to wade into this debate for two reasons. First, because the debate has been documented extensively elsewhere (Hofman et al., 2021; Watts, 2014). Second, and more importantly, because the formal (in)equality between prediction and explanation is not strictly required to apply prediction for explanatory

[5]Whenever prediction is applied, it is usually in the form of an auxiliary regression, e.g. Heckman selection methods, 2SLS or matching methods from the causal inference literature. These predictions should not be considered as pure predictions given that they are meant to support standard in-sample evaluation methods and the predictions typically are not assessed substantively.

purposes. Therefore, I do not aim to support or assume either view going forward and encourage others to take a similarly pragmatic approach when considering to use prediction in their work. In that respect, none of the examples I use in this chapter requires a strong position on the above.

Prediction is often misperceived as deterministic forecasting

Prediction is underutilized in the social sciences in part due to misperceptions of what prediction actually is. Prediction is often understood to deal with predicting outcomes or events in the future — i.e., outside of the time frame on which we have current data — and to be intrinsically deterministic — i.e., as making statements with certainty. This type of prediction is at best a small subset of the general process of making predictions.[6]

In fact, predictions need not be made on future events, nor does prediction have to exclusively concern time-varying data. A mechanistic theory describing the effect of some variable X on an outcome y via some model $y = f(X)$ can lead to predictions in future, current, and past cases as long as the data used for prediction is similar to that used in estimating the model. For example, predictions can be made for a small partition of the dataset collected to study some mechanism, which is set aside and not used for model estimation but purely for predictive evaluation. This is the typical approach to prediction observed in the field of ML (Donoho, 2017). Tellingly, most ML applications do not concern time-varying events at all (Hastie et al., 2009). The only thing conceptually required to predict is a set of data similar to that used in estimation.

Accordingly, predictions are made using estimated models and should thus be viewed from a probabilistic perspective, just like classic techniques like (logistic) regression are inherently probabilistic in nature, too. What makes prediction different is an explicit focus on the outcome variable. There is no reason to assume

[6]Similar points have been made within the prediction versus explanation debate in sociology, for example by Watts (2014). In this particular work, the author implies another possible reason why prediction might be underutilized by explanatory researchers, noting that 'explanations will also become less satisfying' when forced to be predictive (p. 313). In other words, prediction might also be actively avoided by researchers as it restricts the types of explanations one can plausibly argue for.

497 determinism any more when making predictions using some model, than determinism
498 is involved when evaluating the estimated coefficients of that very model.

499 Historical limitations limit the use of prediction in the present

500 Historically, there were considerable limitations on both data and computational
501 resources available to researchers. This still affects the use of prediction in the
502 present. A parallel can be drawn to the enduring imbalance between Bayesian and
503 Frequentist approaches to inference in the social sciences. Bayesian statistics require
504 a relative abundance of computational resources, compared to a Frequentist approach.
505 This made the use of fairly simplistic linear models — plugged into exponential
506 family probability distributions with computationally convenient properties — the
507 preferred methodological approach for social scientists during the latter half of the
508 20th century (Efron and Hastie, 2016, Chapter 5). This dominance persists up to
509 this very day. Such choices were reasonable and necessary at the time, but have
510 led to an analytical mono-culture today (Hindman, 2015).

511 Making predictions is similarly expensive: in some cases a part of the dataset
512 has to be put aside for evaluation or models have to be estimated many times
513 for robust inferences into the predictive performance of a model. Limits on data
514 and computational resources have thus strengthened a (historical) preference for
515 in-sample inferential methods (Efron and Hastie, 2016, p. 53). In a day and age
516 of ever larger datasets and computational power, these issues are a problem of the
517 past. Just as the increases in both data and computational resources have led to
518 a burgeoning growth of methods using Bayesian approaches, the use of prediction
519 should no longer be held back by practical concerns. Even in small N settings,
520 techniques are available that still allow prediction to be applied.[7]

[7] Increases in data size are a key feature of the past decade, although some of the larger datasets available to social scientists needn't be on par in terms of data quality (Salganik, 2019). However, for prediction small N settings are not restrictive, as Leave-One-Out prediction — discussed below — still allows a predictive perspective to be pursued in such cases.

Prediction is conflated with the use of convoluted models

More recently, prediction is approached with hesitance due to the astronomic rise of techniques from the domain of ML which place a strong emphasis on prediction (Shmueli, 2010; Hofman et al., 2017). This has led to the risk that the limitations of ML methods are blindly transferred to prediction in general. To illustrate, reviews discussing the potential of ML for the social sciences have appeared in various important journals (Mullainathan and Spiess, 2017; Hofman et al., 2017; Molina and Garip, 2019; Grimmer et al., 2021). All these reviews discuss the benefits of ML — e.g., increased model complexity and the lack of reliance on pre-specified functional forms — as well as the key difference: ML's focus on prediction.

ML methods have various limitations and risks associated with them, most notably highly convoluted models with a profound lack of interpretability (Athey, 2017).[8] These risks have little to do with the general process of making predictions. Decoupling prediction from black-box methods is crucial to break the misperception that predictive accuracy is something that is naturally maximized at the cost of interpretability. Predictions can be made as easily using an additive linear model as with the complicated non-linear algorithms commonly applied in ML. That researchers within ML almost exclusively predict doesn't mean that prediction is exclusive to ML.

A simple definition of prediction

Prediction, understood as the process of *evaluating a model in terms of its ability to accurately approximate the outcome*, should not suffer from the definitional confusion outlined above. Prediction simply calls for a renewed emphasis on our model's ability to model the dependent variable in our data. Based on this definition, making predictions consists of the following simple steps:

[8]Note that considerable developments in the field of 'Explainable A.I.' are advancing the interpretability of complex model spaces (Ribeiro et al., 2016; Samek et al., 2019), and even allow ML methods to inform functional form development in typical exponential family models, as is pursued further in Chapter 4.

1. Define an `estimation set` to fit the model, and an `evaluation` set to generate predictions for;
2. Estimate the `model` using the `estimation set`;
3. Make `predictions` of the outcome using the `model` and the data in the `evaluation set`;
4. Evaluate the performance of the `predictions` against the `observed outcome`.

Clearly, the above subsumes the more narrow definitions of prediction like forecasting, or the use of ML, which fall within the confines of this broader definition.

To make the above more concrete, assume we estimate some functional form $f_\mu(\cdot)$ in order to find evidence for the association of years of education, A, on wages, y — an example I will return to later. We include work experience, B, as a control variable leading to the model $y = f_\mu(y, A, B)$. Typically, $f_\mu(\cdot)$ is a linear additive model plugged into an exponential family probability distribution with parameter-vector μ, although more complicated algorithms can be applied without loss of generality. Prediction is as simple as estimating $\hat{f}_\mu(\cdot)$ using information on $y^{\text{estimation}}$, $A^{\text{estimation}}$, and $B^{\text{estimation}}$ from some dataset $\mathcal{D}^{\text{estimation}}$ and generating predictions using information on $A^{\text{evaluation}}$, and $B^{\text{evaluation}}$ from some dataset $\mathcal{D}^{\text{evaluation}}$:

$$\hat{y}^{\text{evaluation}} = \hat{f}_\mu(A^{\text{evaluation}}, B^{\text{evaluation}} | \mathcal{D}^{\text{estimation}}). \tag{2.1}$$

The predictions $\hat{y}^{\text{evaluation}}$ can then be evaluated, for example by comparing them against the actually observed $y^{\text{evaluation}}$. There are many summary metrics of fit available for this purpose — e.g., the Root-Mean-Squared-Error or F1-score — but one can also compare (sets of) individual predictions against observed outcomes. Of interest is the broad ability of $\hat{f}_\mu(\cdot)$ to accurately model the outcome.

Based on this definition, the only decision a researcher has to make is how to define the set used for estimating the model, and how to define the set used to evaluate its predictive performance. I identify three general approaches, which I discuss below.

In-sample evaluation A first option is to simply use the same data used for estimation to make predictions (Figure 2.1, Panel A). This choice would effectively lead to an in-sample assessment of model fit, and the well-known R^2 is an example of aggregate fit under this choice of evaluation set. A form of in-sample prediction is sometimes applied by researchers interpreting coefficients in non-linear models where the coefficient estimates lack straightforward interpretation, like categorical outcome models (Breen et al., 2018; Hanmer and Ozan Kalkan, 2013).[9]

In effect, in-sample evaluation boils down to assessing the fitted values of the model estimated in step 2, above. For parametric models with sufficient sample sizes, this is an efficient approach as it uses all the available data for both estimation and evaluation. The downside is the risk that in-sample predictions can be overfit — leading some to argue that predictions should be made exclusively out-of-sample (Cranmer and Desmarais, 2017). Overfitting is the reason why aggregated in-sample fit metrics like the Adjusted R^2 or information criteria are scaled downwards based on the degrees of freedom in a model (see Section T.2.1).[10] When evaluating predictions at a lower level of aggregation — e.g., for a subset of the data — or when models become more parametrized — e.g., multilevel models or when using regularization techniques — using separate estimation and evaluation sets is strongly advised.

Cross Validated evaluation A second approach is to partition the existing dataset into disjunct estimation and evaluation sets. This is the typical approach often observed in ML and has the added benefit that any risk of overfitting the data is explicitly addressed.[11] Predictions are only ever made for data, which

[9]When researchers use prediction in the context of categorical outcome variables it is more common to perform *simulated prediction*, where many covariates are set to their mean or median values. This approach does not necessarily reflect the model's ability to approximate the outcome, as the data used need not be representative of the underlying population.

[10]In the case of parametrized models without shrinkage terms, the correction term of the unexplained residuals is $\frac{N-1}{N-1-p}$ where p is the number of degrees of freedom in the model. This correction term converges to one quickly for moderately-sized datasets. Therefore practically, in-sample predictions might suffice for moderate N and small p. Note, however, that this correction is meant to be applied to the aggregated fit metric and not individual predictions.

[11]Overfitting relates to the issue that a model is too closely fit to a single sample that the estimated parameters no longer reflect the underlying processes, but are effectively fitting noise. This can lead to overly confident statement of model fit, or, when combined with repeated specification searches, to spurious findings (Sala-i Martin, 1997).

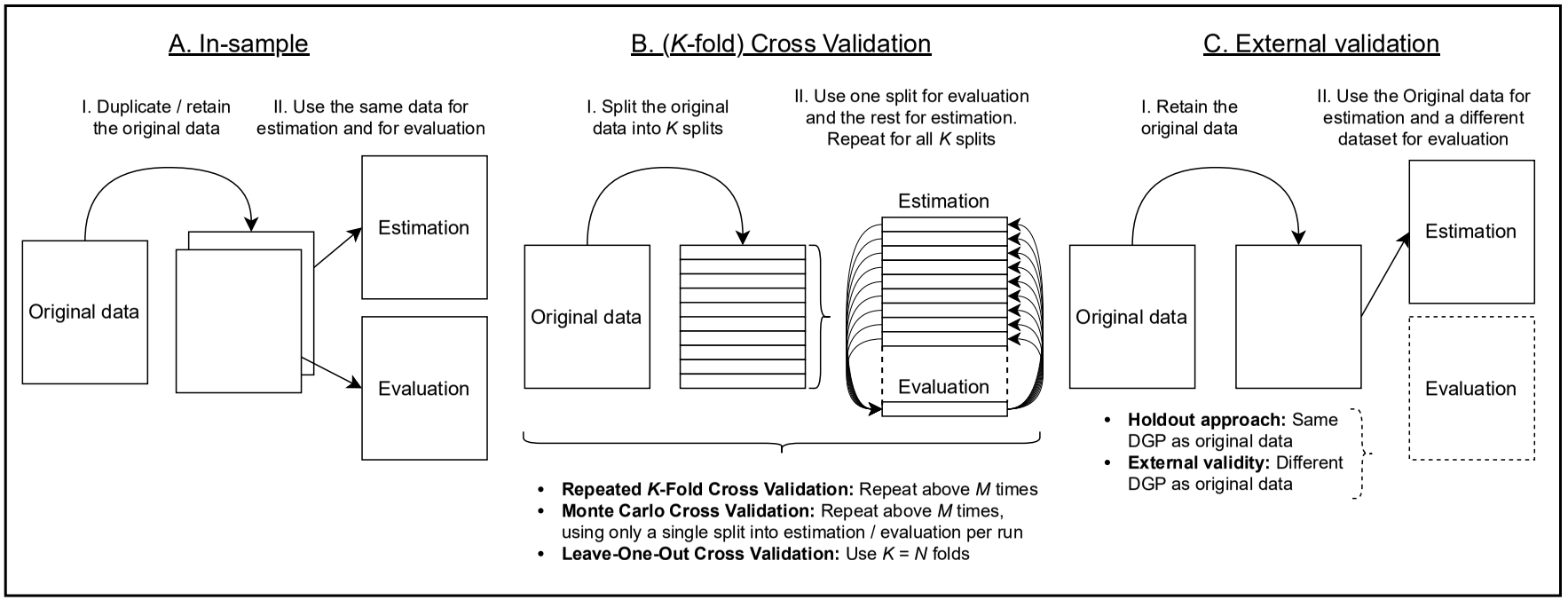


Figure 2.1: Three different strategies to define an estimation set and an evaluation set. The first strategy (A) uses the original data as both the estimation set as well as the evaluation set. The second strategy (B) splits the original data into K splits. Each split is used once as an evaluation set, to evaluate the model estimated to the remaining $K-1$ splits. The model is thus estimated K times. This step can be repeated M times, leading to repeated K-fold cross validation, or Monte Carlo cross validation in case only a single estimation / evaluation cycle is done per run instead of K (Yousef, 2020). The third strategy (C) uses the original data to estimate the model, and uses a different dataset to evaluate the model, which can be a holdout set from the same DGP but partitioned off prior to analysis, or collected separately.

was not used to estimate the model. The most common such approach is K-fold cross validation, which consists of dividing the dataset into K equal-sized 'folds' — typically K is set to 5 or 10. The model is then estimated K times, each time omitting one of the folds from the estimation process and using the omitted fold as the evaluation set (Figure 2.1, Panel B). This ensures that predictions are generated for every observation in the dataset, thus maximizing the number of predictions given the available data. It does make the routine computationally more expensive as the model has to be fit K times. K-fold cross validation also means that estimation is only ever done on $n - n_k$ data points per run, which can lead to a loss in efficiency and precision of the estimates.

There are various alternatives or variations to implementing cross validation. For example, in low N situations or when efficiency in estimation is paramount, one could use Leave-One-Out (LOO) cross validation which is the special case of K-fold cross validation where K is set to N (Stone, 1974). As the number of folds K approaches N, the loss in efficiency decreases although the computational cost of performing the cross validation increases. For relatively straightforward estimation like OLS, the overall increase in computational time is negligible, but it can become prohibitive if the time required to perform a single estimation of the model is already considerable. Beyond varying the number of folds K, additional robustness to random variation in splitting the data into folds can be incorporated by repeating the entire routine multiple times.[12] Overall, K-fold cross validation remains the most commonly applied approach.

External evaluation A third and final choice for the evaluation set can be a set of data which is completely 'unseen' by the researcher (Figure 2.1, Panel C).

[12] Common examples include repeating the cross validation routine M times — so-called Repeated cross validation. Another variant is Monte Carlo cross validation, where again M runs of cross validation are done, but each run only uses a single split of the data into estimation and evaluation sets. Many of these approaches tend to converge to the same results in the limit, see Yousef (2020) for a review. For most social science applications, the number of Monte Carlo simulations M can be relatively low as the evaluation set is typically about 20%-30% in order to accurately reflect the original data. Therefore, with M around 100 the impact of assigning data to folds should be approximated well.

For example, by immediately partitioning off a part of the data into a holdout set which is kept separate from the entire estimation process or, ideally, never even shared with the researcher(s), which is the typical approach in Kaggle-style competitions. This is called the 'Holdout' approach and provides the most truthful assessment of a model's predictive performance. Unfortunately, it is expensive as a sufficient number of observations are required to make predictions and these observations cannot be used for estimation. Thus leading to both reduced efficiency in estimation, and a reduced number of predictions to evaluate.

External validation can also be done by assessing model predictions on a completely new set of collected data. For example, similar data that was collected at a different time — e.g., separate waves of a survey — or place — e.g., regional comparisons. The choice of an external evaluation set speaks directly to calls for increased attention to the external validity of research findings in empirical work (Simonsohn et al., 2020; Young, 2018), and can be particularly useful to assess the transferability of research findings outside of the sample used for estimation. By choosing an external validation set, prediction provides a simple and direct framework to assess model fit outside of the sample at hand.

In practice, the choice of splitting the data into estimation and evaluation sets will be made on a case-by-case basis. If a parametrized model is estimated with few coefficients and a considerable data size — e.g., $N > 500$ — the risk of overfitting will generally be low and in-sample prediction could be considered. When the number of parameters in a model increases, it is advisable to use some form of cross validation, either LOO in case the number of observations is limited, or K-Fold with K typically about 10 (Kohavi et al., 1995). If N is sufficiently large that setting aside a portion of the data does not meaningfully affect model estimation, the holdout approach can be applied where typically 20%-30% of the data is partitioned off as the holdout set.[13]

[13]Fundamentally, the holdout set needs to be large enough to capture the intricacies of the original data well. Therefore, for low dimensional data with limited variation a small holdout set might already suffice. Conversely, high dimensional data or clustered data might require considerably more observations in the holdout set to accurately reflect the data of interest. The same rationale holds when selecting the size of the single evaluation fold in Monte Carlo cross

When using prediction to improve model understanding — as most of the examples in this chapter do — LOO cross validation is generally attractive if computationally feasible. Prediction is out-of-sample and the maximal number of data points ($n-1$) are used to estimate the models used for each prediction. However, when the goal is to select an optimal predictive model to deploy amongst a set of candidate models, one might be interested in the expected prediction error on a completely new observation and the variability of this prediction error. In this case, K-Fold cross validation is typically a more efficient estimator, although some properties of the estimand are not yet completely understood (Efron and Hastie, 2016, Chapter 12).[14] The Holdout / External validation approach will give the most precise assessment of a model's ability to accurately predict the outcome of new observations, although it is clearly expensive as one has to completely set aside a part of the data for evaluation or collect a new dataset. These questions are less relevant when using predictions to improve our understanding of explanatory models, as is the focus in this chapter.

2.3 The Three Virtues of Prediction

As outlined in the introduction, complementing traditional methods with prediction brings three key virtues to the social scientist's table. In what follows, I illustrate these in kind using examples of novel reproductions and examples from latter chapters.

validation.

[14]The general approach to estimating the predictive error of a model is to assess the error of all the N predictions made using cross validation against their observed values and reporting the mean and standard deviation. Interestingly, cross validation has been shown to consistently estimate the expected error of the model fit to a random dataset drawn from the same underlying distribution as the training set, and not the expected error of the estimated model. In addition, the approach can lead to overly narrow confidence intervals (Bates et al., 2021; Bengio and Grandvalet, 2004). The reason is that the errors are not independent, as each observation is used in both estimation and evaluation. This problem will be minor when the impact of omitting a specific fold from the estimation process on the estimated model is small — i.e., the model is stable across the omission of folds — but can have a serious impact otherwise. Some solutions have been suggested to correctly scale the CI's of predictive error for certain families of models, but research remains ongoing (Bates et al., 2021).

Virtue I: Prediction provides improved insight into model fit

In its most simple form, prediction provides a distinct way to assess model fit on the level of the actual outcome variable. Such a perspective provides a renewed focus on what one could call predictive consciousness: an understanding how well our models are actually able to fit the outcome variable of interest. In practice, a model's fit is often left undiscussed, leading to a broad lack of predictive consciousness in empirical work. For example, whether the models we estimate are able to accurately predict 0.1%, 1% or 50% of the variation in the outcome. Model fit, if it is discussed at all, is typically assessed at the aggregate level only.[15] We are often left guessing what elements in a model contribute most to its ability to predict well. Nor do we know whether a model is able to predict all of the data equally well, or just parts of it. A renewed appreciation for prediction will improve our assessment of model fit, as predictions are made for every single observation in the evaluation set. As a consequence, a disaggregated assessment of fit on the scale of the outcome follows naturally.

Predictions are also intuitive to understand and will help the implementation of research findings in the real-world. Its natural interpretation should promote acceptance and understanding of research findings by both policymakers and the public.[16] Academically, a general sense of predictive accuracy is equally important to further a research agenda: if predictive performance (strongly) underperforms expectations, it prompts reflection of whether we are actually missing important determinants or perhaps our preferred functional form is not able to capture the mechanisms in operation. Finally, making and evaluating predictions beyond the aggregate level can also provide additional transparency into the academic process. By reporting on model fit at lower levels of aggregation, consumers of empirical

[15] Typically, the in-sample R^2 has been used for the purpose of evaluating explanatory power. The measure has various limitations in often encountered empirical setups, for instance when modeling ordered outcomes or estimating other non-linear models. In those cases, information criteria are typically reported although these tend to defy an intuitive interpretation of the model's ability to fit the outcome.

[16] A re-appreciation of prediction also echoes calls for the social sciences to become more solution-oriented by evaluating social theories with respect to their predictive power (Watts, 2017).

research become better able to critically assess what a model can and cannot do in terms of fitting the data.

As a first example of the above, consider the FFC that was introduced earlier (Salganik et al., 2019b). The FFC challenged research teams to accurately predict life outcomes at age 15 based on a rich set of data using the Holdout approach — i.e., setting aside a partition of the data.[17] As part of the challenge, the organizers calculated a benchmark performance using models constructed by domain experts. As a consequence, the low predictive power of the models typically estimated in this domain already became quite obvious from the outset (Figure 2.2, panel A). For many of the life outcomes of interest measured at age 15, a model including the hand-picked variables and a lagged version of the variable at age 9 did not substantially improve predictive accuracy relative to a null model predicting the overall mean of the outcome. This is a sobering finding putting into perspective the supposedly large amount of understanding that had been generated regarding these outcomes (Salganik et al., 2020).

Perhaps the outcomes of interest to the FFC are inherently noisy and difficult to predict, as the organizers also note (Salganik et al., 2020; Hofman et al., 2017). Regardless of the question why predictive ability was low within the FFC, the very insight and subsequent discussion it provoked are essential for the field to develop. The FFC illustrates how predictive consciousness can be crucial to instigate a critical reflection on the state of knowledge in a field and can spur important debate. Accordingly, the key take-away of the FFC is that the type of discussion which a predictive focus triggered should not have taken so long to materialize. Likely, a predictive consciousness is equally relevant for many other social science research fields. More consistent reporting of predictive accuracy would guarantee that the type of reflection resulting from the FFC becomes more likely to occur throughout the social sciences.

Predictions can also diagnose model performance at different levels of detail or aggregation. As an example, take the individual predictions rather than the overall

[17] The Holdout approach was chosen due to the competitive nature of the challenge: organizers were interested in finding the best predictive model amongst the participants.

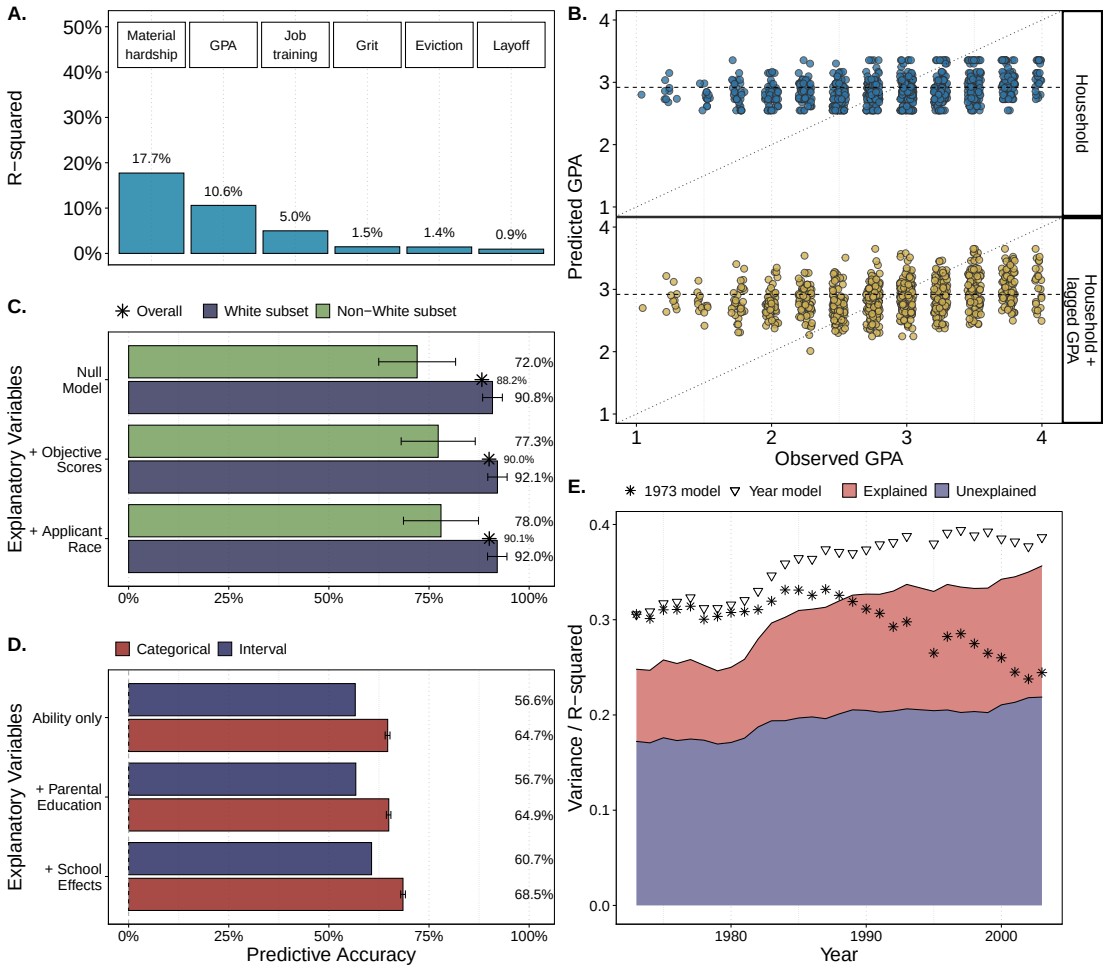


Figure 2.2: Panel A shows the predictive R^2 of the linear benchmarks chosen by domain experts relative to a null model for the FFC. Each model includes four explanatory variables and a lagged version of the outcome (Salganik et al., 2020). Panel B shows individual predictions of the GPA outcome in the FFC using only the domain expert variables (top) and when including the lagged version of the outcome (bottom). Panel C shows the performance of logistic regression models predicting whether a mortgage application was successful using various sets of explanatory variables. Performance is shown disaggregated for White (green) and Non-white (blue) applicants, showing that the performance is considerably lower for the latter. Panel D shows the performance of models predicting students' track levels using various sets of explanatory variables and different models. Performance of the categorical model is substantially higher than the linear model. The performance of both models strongly increases when including school variation. Panel E shows the performance of various models in explaining hourly wage in the United States. Predictive power is assessed for each year using the same model but re-estimated to that year's data (triangles) or using the model estimated in the year 1973 (stars). Initially, the latter model performs well on the first couple of subsequent waves, but deteriorates from 1983 onwards. Confidence intervals, where present, reflect 95% confidence bounds of the estimated predictive accuracy across the various evaluation folds.

predictive accuracy of the benchmark model for the GPA outcome in the FFC. We can visualize the predicted GPA against the observed GPA for each observation in the evaluation set and can do so for different models (Figure 2.2, panel B). Taking such a disaggregated approach to model fit, we observe that both models struggle to structure the outcome well, but this is especially so for students performing below average. A nuance which provides pointers for future avenues of research. In other words, prediction's ability to interrogate model fit on a disaggregated level provides a different vantage point than summary metrics of model fit.

In addition to individual predictions, they can also be assessed i) at the group level or ii) using completely different models and/or sets of explanatory variables. As an example, consider a reproduction of the influential 1996 study from the Federal Reserve Bank of Boston regarding discrimination in mortgage lending (Munnell et al., 1996). The authors — amongst other things — found evidence of discrimination against Non-whites based on a logistic regression including a Race dummy and conditioning on various objective characteristics of the application. By complementing their analysis with a predictive perspective, additional nuances emerge (Figure 2.2, panel C).[18]

For example, most individuals are successful in their mortgage application and a null model already correctly predicts 88% of the data (by predicting successes for everyone). Including variables like objective score measures and household characteristics further increases the model's performance. However, aggregate fit is somewhat misleading, as there is a considerable gap in the model's ability to predict outcomes for Non-white applicants compared to White applicants (77% versus 92%). The inclusion of a Race indicator only marginally improves the gap. In other words, Non-whites are modeled considerably less accurately than Whites. This could imply that additional sources of heterogeneity are present for the former — for example if bias is multimodal and depends on other factors like

[18]Repeated K-fold stratified cross validation was applied, ensuring similar proportions of Whites and Non-whites in each fold, with $K = 5$ and $M = 100$. Predictive accuracy was thus estimated for a total of 500 folds.

the employee reviewing the application — or some other reason is present why Non-whites are modeled considerably less well.

Another illustration where prediction provides additional understanding of model fit is the study described in Chapter 3 of this thesis, where I assess teacher bias in educational tracking — the process of assigning students to ability levels — in the Netherlands. In this study, I apply prediction to understand the relative importance of different sets of explanatory variables as well as modeling assumptions (Figure 2.2, panel D).[19] This predictive perspective led to a number of important nuances to the existing knowledge on teacher bias. First, a predictive approach showed that commonly studied bias factors like parental education — although statistically significant — mattered little for the model's fit of the data, improving the predictive R^2 by a mere 0.1 percentage points (pp). When allowing for separate intercepts per school — typically perceived as a control variable — the improvement in model fit was considerably more impactful, increasing the predictive R^2 by almost 3 pp. Second, using a non-linear categorical model strongly improved the model's fit of the data, when compared to the simpler linear model traditionally estimated in the field.[20]

Both nuances have important substantive implications for research on teacher bias, which were not picked up in pre-existing work that focused on traditional inference. For example, school effects were typically evaluated through the estimated variance term of the random intercept. They were not typically compared to the other variables in the model with respect to their substantive ability to model the outcome. As a consequence, a considerable source of variation in tracking had been neglected. Similarly, traditional fit metrics would only indicate an objective preference for the categorical model, but did not provide a normative reflection of the extent to which model fit improves. Importantly, changing to the categorical

[19] I apply stratified Monte Carlo cross validation with $M = 250$. The evaluation set represented a stratified 25% of the total data, ensuring similar proportions of students from each school in the estimation and evaluation sets. Predictive accuracy was thus estimated for a total of 250 folds.

[20] In practice, most researchers studying tracking in The Netherlands have assumed the outcome to be continuous and estimate a simple linear model (van Leest et al., 2021). As the authors point out, this is predominantly a convenience assumption, as it yields easer-to-interpret coefficient estimates.

model also considerably affected the size of estimated biases in tracking (a finding that is discussed in more detail in Chapter 3).

A final advantage of using prediction as a measure of model fit is that it can be used as an approach to address questions of external validity. A recurring question in the social sciences is the persistence of research findings outside of the particular sample used to estimate a model. Prediction makes such assessments more natural than in-sample methods do. For example, consider the following puzzle in labour economics. A growing literature is studying the reasons underlying an increase in the amount of residual variance over time when explaining logged hourly wages using a similar set of explanatory variables: education, age, and their interactions (Lemieux, 2006b). This example lends itself well for an illustration of how the external validity of a model can be assessed from a predictive perspective. By estimating the model on one of the survey years — the first wave, 1973, in this illustration — the performance of the original model in terms of fitting the data can be explicitly assessed for datasets collected at later waves.

As the results show (Figure 2.2, panel E), the 1973 model tracks the performance of models that are retrained to each year quite closely for the first 5 to 10 years, but then starts to deviate.[21] This is insightful for two reasons. First, it provides an indication of the stability of the findings from the 1973 model outside of that sample. Second, it points at a shift in estimated model coefficients from the year 1983 onward, possibly providing additional pointers into the original puzzle. As the models use many interactions, leading to more than 100 coefficients, these differences in the model's fit to separate datasets would be close to impossible to learn from studying the in-sample coefficients of each model in isolation.

Virtue II: Prediction provides a benchmarking tool across modeling domains

Although social life is known to be complex to study, simple linear additive models are still the bread-and-butter methods used throughout the social sciences for this

[21] The outcome variables were normalized such that mean differences in average wages across time would not distort the predictive performance of the model fitted in 1973.

very purpose. The reason might be that we have grown accustomed to fitting such models for so long now, that we are reluctant to believe more complicated functional forms are appropriate. A more likely reason is that simple models allow for a more straightforward interpretation of results, which is usually not the case in complicated non-linear models even if they are objectively better at capturing reality. A key problem is that we often don't know whether our models are in fact too simple, prolonging the use of simplistic models in practice. Through benchmarking, prediction provides a way to assess whether the level of complexity in our models is appropriate, as predictive accuracy can be used as a holistic metric of model fit for any type of empirical model (Hindman, 2015). Therefore, it can be used to compare parametrized models with flexible alternatives (this idea is expanded on further in Chapter 4).

For example, model complexity was a key motivation of the FFC and many research teams heeded this call by innovating extensively on the methods applied.[22] In other words, the heterogeneity in modeling approaches was considerable. As a consequence, conventional model diagnostics would not have sufficed to allow comparisons of the various modeling approaches chosen by the research teams.[23] Opting for prediction on a holdout set solved this problem. As an illustration, the predictive performance of every single submission to the FFC has been visualized in Figure 2.3 (Panel A).[24] For some outcomes, considerable improvements in predictive accuracy were obtained relative to the benchmark models — e.g., for the GPA outcome — although for most, improvements were negligible — e.g., for the Job Training outcome.

[22]Note that the data was mainly appropriate for methods exploiting some form of variable selection. Methods like neural networks should not be recommended as the FFC contained only 4,000 observations, but more than 12,000 variables. This means that the 'curse of dimensionality' would be a serious issue without variable selection or regularization techniques (Bishop, 2006, Chapter 1). More generally, limited N might be an important reason for the lack of predictive improvement observed in the FFC.

[23]Traditional information criteria and goodness-of-fit measures are often dependent on predefined functional forms to correct for the degrees of freedom in the model. When including models from different paradigms of modeling, information criteria lack comparability (see Section T.2).

[24]By default, teams submitted the predictions of a null model unless they submitted their own predictions for an outcome, which is why there is considerable density around an improvement of 0 pp as many teams choose to focus on a select number of outcomes.

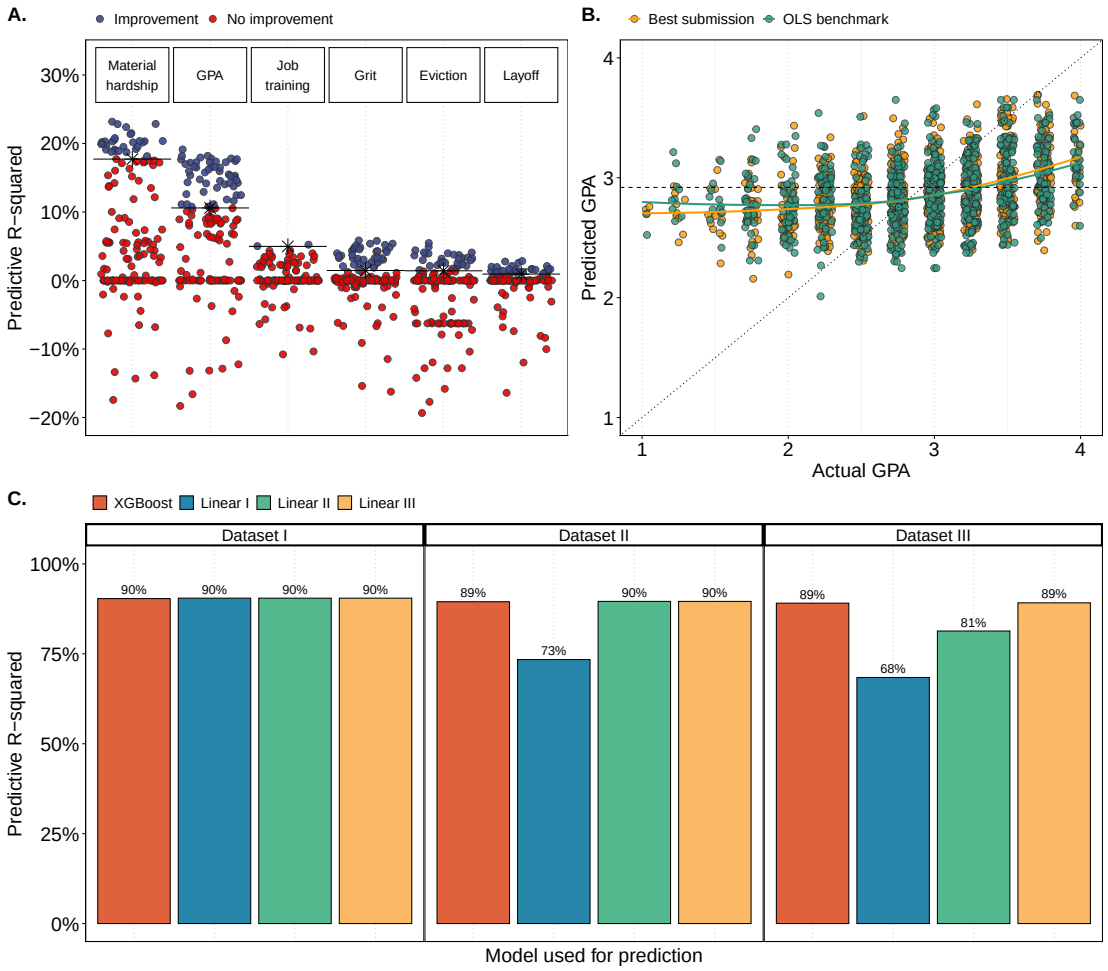


Figure 2.3: Panel A shows the predictive performance of all submissions to the FFC. The horizontal lines reflect the performance of the benchmark. Blue submissions outperformed the benchmark. Predictive accuracies of 0% indicate that a team did not submit a submission for that specific outcome (Salganik et al., 2020). Panel B shows the individual predictions of the best performing submission to the FFC for the GPA outcome (yellow) versus the individual predictions of the OLS benchmark (green). GPA was observed in 0.25 point intervals, and have been spread horizontally for illustrative purposes. The dashed horizontal line indicates the mean of the outcome and the dotted diagonal line indicates perfect predictions. The R^2 of the best submission was nearly double (0.19) that of the benchmark (0.11). Panel C shows the predictive accuracy for three simulated datasets using three pre-specified linear functional forms and a flexible non-linear algorithm (XGBoost). The flexible algorithm converges on the true functional form for all three datasets, whereas only the Linear II and Linear III model had the appropriate complexity to fit Dataset II well, and only Linear III had the appropriate complexity to fit Dataset III well.

Many of the top performing submissions made use of complex, flexible models but prediction made them directly comparable to the linear benchmarks (Rigobon et al., 2019). This is true at the level of the various submissions (Fig 3, Panel A), but a direct comparison can also be done for a single submission (Fig 3, Panel B). Here, the individual predictions of the top submission for the GPA outcome are visualized together with those from the benchmark model. As can be seen from the fit of locally estimated scatterplot smoothing (LOESS), the top submission is slightly better in predicting the extremes of the distribution correctly. That said, the plot also shows that both models still struggle to predict the low-end of the distribution well.

The role of benchmarking is arguably most important to identify functional form misspecification (Section T.1). Comparing the fit of a functional form hypothesized by a researcher with that of a flexible alternative fit to the same data provides an assessment whether the model might miss complexity. This can for example be used to verify whether the assumption of linear additivity is reasonable. As a concrete example, consider the simulation study discussed in more detail in Chapter 4, assessing the predictive performances of various Mincerian wage equations.[25] In these models, the effect of additional years of schooling on wages is of central interest. These so-called 'returns to schooling' are typically estimated by relating information on years of education to observed wages while controlling for years of work experience (Lemieux, 2006a).

The Mincerian wage equation is an interesting case because the complexity of the functional form has been innovated upon over the past decades. The functional form started out as a simple linear additive model where log wages were regressed on years of work experience and years of education. In the 80s and 90s, higher order terms on the effect of work experience were proposed and, more recently, a step function in the effect of education has been included into the functional form (Lemieux, 2006a; Heckman et al., 2018). In other words, the functional relationship between outcome and explanatory variables was found to be underspecified and lacking in complexity. Benchmarking can help identify such lack of complexity by

[25]This example is taken from the more elaborate analysis in Chapter 4, which includes a fourth functional form as well. The models are otherwise identical to those estimated there.

comparing a model's performance to that of a flexible alternative, which does not constrain the functional form in a particular way. If a flexible model using the exact same covariates strongly outperforms a linear additive model, there is likely a lack of functional form complexity (Efron and Hastie, 2016; Chapter 4).[26]

This rationale is illustrated in Figure 2.3 (Panel C). Three datasets were simulated that include the same explanatory variables on age, years of education and years of work experience. However, the outcome variable — log wages — is simulated according to a different functional form for each dataset. Specifically, the outcome of the first dataset follows a linear additive function (Linear I), the second a linear additive function including a squared term for the effect of work experience (Linear II), and the third further includes a step function for the effect of education (Linear III).[27] All three outcomes included white noise proportional to about 10% of the total variation — thus capping the potential R^2 at 0.9. These three datasets reflect the functional form development of the Mincerian wage equation observed over the past decades. Four models were fit to each of the datasets, the first being a linear additive model, the second allowing a squared effect for work experience, and the third included the step function. In other words, all three models have the appropriate complexity to fit the first dataset, but only the second and third have sufficient complexity to model the second dataset well, and only the third model can fit the final dataset appropriately. The fourth model was a vanilla XGBoost algorithm, a highly flexible tree-based ML algorithm (Section T.5.3). All four models were used to make predictions on a holdout set of the data.[28]

The results show that all four models fit the first dataset well — as should be expected. For the second dataset including the squared term, the first functional form strongly underperforms the alternatives since it cannot model the full complexity

[26]The matter-of-fact comment by Efron and Hastie regarding the use of Random Forests — a flexible ML technique — in their 2016 handbook 'Computer Age Statistical Inference' is instructive here: 'if the Random Forest does much better [than a traditional parametrized model], you probably have some work to do' (p. 347).

[27]The models were estimated using a synthetic dataset of 50,000 observations based on the observed age, work experience, and schooling in the 2018 General Social Survey. For the construction of the synthetic sample and exact functional forms underlying the three datasets, see the discussion in Chapter 4 from which this example has been taken.

[28]To generate predictions, a holdout set was partitioned off equal to 20% of the total dataset.

in the data. The same holds for the first two functional forms when fitted to the final dataset, which included the step function. Importantly, the flexible algorithm converges on the 'true' model's performance in each of these cases. It thus identifies the need for additional model complexity without requiring the researcher to formulate a pre-specified functional form. Prediction is the key benchmarking metric that allows this comparison of fit across modeling domains.[29,30]

Virtue III: Prediction can help generate insight into the behavior of complicated models

Perhaps the most important reason why prediction is traditionally underutilized in the social sciences is its supposed lack of explanatory ability. Often, predictions are merely assumed to be useful at the descriptive level at best. However, prediction can actually be used as a method to improve the interpretability of models. First, prediction can be used to make coefficient estimates in non-linear models as interpretable as those from standard linear models by intervening on observed variables and comparing the effects of such an intervention on the predictions. This is especially relevant when dealing with categorical outcome models (Breen et al., 2018; Hanmer and Ozan Kalkan, 2013). Second, a predictive analysis is amenable to more substantial interventional reasoning. For example, to assess the impact of changing a coefficient — as mentioned above — but also to compare the effect of estimating separate models on subsets of the data. Assessing these types of differences by comparing in-sample coefficients alone is practically unfeasible.[31]

To illustrate the first point, consider the mortgage application introduced in Figure 2.2 (Panel C). Logistic regression models were estimated, making the

[29]Further understanding the type of complexity that is missing from the functional forms is a topic that is explored further in Chapter 4.

[30]Akin to benchmarking, prediction provides a common metric for researchers to align their research efforts (Hofman et al., 2017; Cranmer and Desmarais, 2017). The astonishing improvements in ML methods can in part be attributed to such an alignment on a common goal: improving the predictive accuracy on benchmark sets.

[31]Note that prediction is understood here as a tool to help interpret estimated model coefficients. Prediction does not make model results causally interpretable. Causal interpretation is encoded in the research design and model estimation, whereas prediction is a tool to assess the model post-estimation (Pearl, 2009).

interpretation of the coefficient estimates less straightforward than a standard linear model, which would simply reflect the increase in the value of the outcome when increasing the covariate by one point. This ease-of-interpretation is not available for the logistic regression model. However, prediction provides a way to obtain a similarly intuitive effect size. This is illustrated for the Race coefficient in the mortgage example by comparing the predicted probabilities of success when intervening on the observed Race variable. What this means is simply changing the observed Race from White to Non-white or vice-versa and comparing predictions (Figure 2.4, Panel A).[32] As can be seen, the average probability of success decreases by 8.4 pp for Whites and increases by 11.2 pp for Non-whites when changing the Race variable. This type of do-style reasoning is easy to implement when making predictions, and improves on the common practice of simulating predictions by setting other observables to their mean or median values. The predictive approach here uses the actual data, which is arguably more informative.

Instead of using predicted probabilities, it is also possible to assess the effect of a variable in terms of the actual outcome. An increase in the probability of success does not automatically reflect a similar increase in the expected number of successful applications. Probabilities will yield predictions anywhere between 0 and 1, whereas predicted outcomes will always consist of either 0 (failure) or 1 (success). The outcome-focused equivalent of the mortgage example — where the predicted probabilities are rounded — is given in Figure 2.4, Panel B.[33] The results illustrate how intervening on the Race variable would increase the share of predicted successes from 84% to 92.8% for Non-whites, while decreasing the number of predicted successes from 98% to 94.8% for Whites.[34] The differences in the number of actual successes reflect that most Whites already have a high

[32] A single 5-fold cross validation run was used to generate a full set of predictions.

[33] The same repeated cross validation routine from Figure 2.2 (Panel C) was used for the predicted outcomes.

[34] The confidence intervals show that, depending on the fold used for evaluation, the observed percentage of successes in the fold can be higher or lower than the overall mean. This follows from the fact that the data size was limited — especially for Non-white applicants — leading to more variability in the baseline (Munnell et al., 1996). Note that the intervals should not be assessed to reflect the statistical significance of the race coefficient, but rather variability in the observed pre-intervention probability of the White or Non-white applicants in a specific fold.

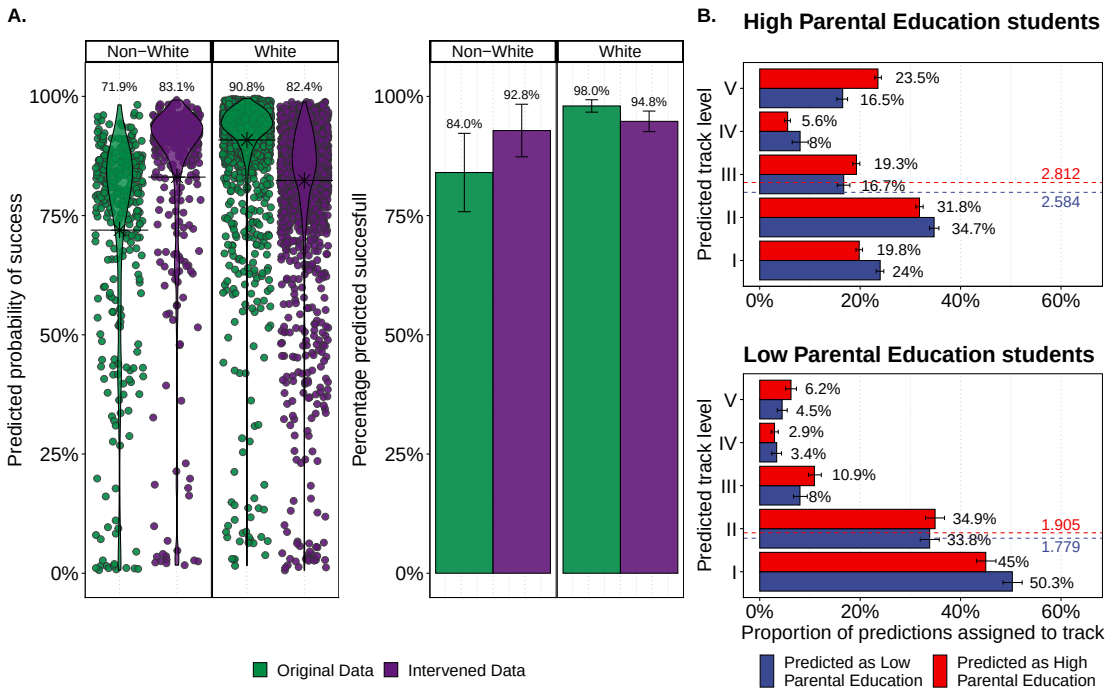


Figure 2.4: Panel A shows predicted probabilities (left) and outcomes (right) of successful mortgage applications for Whites and Non-whites in the data. Green values indicate predictions on the original data, whereas purple values show predictions when intervening on the Race variable: i.e. changing the observed Race from Non-white to White and vice-versa. The average predicted probability of success increases by 11.2 pp for Non-whites when changing the observed Race, and decreases by 8.4 pp for Whites when doing so. The predicted number of successes increases from 84% to 92.8% for Non-whites and decreases from 98% to 94.8% for Whites. Panel B shows the predicted proportion of students in each of the five track levels for students of High Parental Education (top) and Low Parental Education (bottom) when using a model fit to students of High Parental Education (red) or Low Parental Education (blue). The average predicted track level is depicted using the dashed red line when predicting using the High Parental Education model, and by the dashed blue line when predicting using the Low Parental Education model. Confidence intervals, where present, reflect 95% bounds based on variation in the predictive accuracy or number of predicted classes across the evaluated folds.

predicted probability of success prior to intervening on the Race variable. Whether to use predicted outcomes or the underlying predicted probabilities will typically depend on the particular use case.

To illustrate the use of prediction to ask more complex interventional questions, I return to the study introduced earlier concerning teacher bias in tracking (Figure 2.2, Panel D). Pre-existing work chose to model the outcome — track levels — as a continuous variable, allowing for a straightforward interpretation of the estimated coefficients. In Chapter 3, from where this example is taken, I instead estimate a hierarchical ordered probit model leading to difficult to interpret coefficient

934 estimates. However, by using the same interventional reasoning as outlined above,
935 an intuitive assessment of the impact of changing demographic characteristics like
936 parental education could easily be generated. This reasoning was then taken a step
937 further by assessing whether the effect of parental education could reasonably be
938 captured by the coefficient on a dummy-coded variable, or might manifest itself
939 through the entire model — i.e. whether students of Low parental education are
940 assessed differently on observables.

941 To this effect, separate models were estimated for students with low and high
942 parental education. This led to one model fitted to students of low parental
943 education, which could be used to make predictions for observations 'as if they were
944 low parental education students' and another model that could do the equivalent
945 but then for high parental education students. By making predictions using both
946 models — i.e. predicting outcomes twice — for both the low and high parental
947 education subsets in the data, the implied difference between the two models could
948 be assessed (Figure 2.4, Panel B).[35] As can be seen, high parental education students
949 on average obtain a track level of 2.81 when assessed as high parental education
950 students. However, when assessed as if they were low parental education students,
951 this average track level drops by about 0.25. Conversely, low parental education
952 students gain about 0.12 track levels when assessed as high parental education
953 students. Determining these differences by comparing the two estimated models
954 would not have been trivial, as multiple coefficients would have to be taken into
955 account jointly including random effects and cutoff points that are estimated as
956 part of the ordered probit model (Section T.3).

957 2.4 Taking Stock

958 This chapter set out to change the underutilization of prediction in the social
959 sciences, where prediction barely features in empirical work. This underutilization

[35] This approach is similar in intuition to decomposition methods that decompose overall group differences in some outcome into compositional differences and effect differences (Fortin et al., 2011). However, composition methods are often used within the typical linear additive framework, which can be restrictive. Exploiting prediction allows a wider variety of modeling approaches to be applied.

occurs for the wrong reasons. Many social scientists confuse the general concept of prediction with more narrow applications, like forecasting, or using predictive accuracy as an optimization metric. Yet, prediction is a much broader and simpler analytical perspective of evaluating models in terms of their ability to accurately fit the outcome of interest. Viewed in this manner, prediction becomes a logical complement to and enrichment of the methods we have grown accustomed to using throughout the social sciences. Importantly, there is absolutely no need to sacrifice traditional modeling approaches when including prediction in empirical work — contrary to the sometimes dogmatic nature of the philosophical discussion concerning prediction and explanation. Both explanatory and predictive perspectives to analysis can and should be combined.

The benefits that prediction can bring when incorporated into the typical empirical workflow of social scientists are plenty and this chapter illustrates but a few. For instance: how basic predictive consciousness can spur important debate in a research field (Figure 2.2, Panel A), but also how predictions allow for a more detailed assessment of model fit. For example by assessing the fit of individual predictions (Figure 2.2, Panel B), comparing predictive performance by subsets of the data (Figure 2.2, Panel C), using different models (Figure 2.2, Panel D), or testing our models on completely new data (Figure 2.2, Panel E). Prediction also provides a measure allowing social scientists to compare the fit of wildly varying methodological approaches (Figure 2.3, Panels A-B). This includes models from different paradigms — like ML models — which opens the way to benchmarking our models against flexible alternatives. Such benchmarking provides social scientists with a way to assess whether the models we estimate have the appropriate complexity to fit the data well (Figure 2.3, Panel C). Finally, prediction is amenable to do-style reasoning and allows us to obtain intuitive associations between variables in non-linear models (Figure 2.4, Panel A), but also to take this interventional reasoning a step further and compare how models estimated to subsets of the data differ in modeling the outcome of interest (Figure 2.4, Panel B).

In practice, the way in which prediction is applied fundamentally depends on the case at hand. There will be empirical settings where interest in the ability to predict an outcome is less natural than for example in the case of the FFC.[36] Generally speaking, three broad virtues were identified: i) using prediction as a descriptive tool to improve our understanding into the fit of a model, ii) using prediction to normatively compare different models, and iii) to help generate understanding into (complex) model behavior by interventional reasoning. These benefits help address criticisms plaguing the social sciences, like a lack of appreciation for the real-world relevance of research findings, and the use of overly simplistic models to study social life. At the same time, the sceptic might question what we exactly gain from adding prediction to empirical work. Can't we identify predictive ability by measures like the R^2? Or use fit metrics like the AIC or BIC to compare models? Aren't there specification tests to diagnose serious misspecification, and can't we identify associations in non-linear models, if we really tried?

The answer to all of these questions is: yes, although there are various nuances that make prediction preferred. For example, existing fit metrics are in-sample and can suffer from overfitting. The R^2 measure does not work well in every design and gives no insights into heterogeneity in a model's fit. Conventional non-parametric regression techniques are relatively complex to estimate and data intensive, and without a way to illustrate their necessity, most empirical work will remain wedded to the simple linear additive model. There are many more subtle nuances. The broader point is that predictions are disarmingly simple to understand and generate, and can serve multiple goals at the same time, as the examples in this chapter illustrate. Predictions also improve transparency by inviting a more rigorous assessment of a model's ability to fit the data than most aggregated in-sample metrics. Requiring researchers to make predictions is a much better way to diagnose model limitations

[36]Examples could include stylized lab experiments, or research designs like difference-in-differences, regression discontinuity or matching setups, as either the outcome is less intuitive or interest is fundamentally into an estimated coefficient. Nonetheless, many of the key benefits of prediction that have been illustrated in this chapter will be equally relevant to understand the extent to which the model fits the data, irrespective of whether the outcome itself is of central substantive interest or this is less the case.

than allowing researchers to pick their own robustness checks or descriptives, which might lead to over-engineered findings. Finally, prediction paves the way for exciting methods from other domains, like that of ML, into the workflow of social scientists. Methods that should become complementary to the social sciences.

Hopefully, this chapter can help social scientists decouple prediction from some of the field's most intriguing and sometimes heated discussions. For example whether explanation should imply prediction or what the role of ML should be in the social sciences. Although interesting, they are ultimately a distraction from what prediction as an analytical tool has to offer the social sciences. In sum, prediction's simplicity, transparency, intuitive nature, and flexibility to build on the methods we have used for decades — rather than forcing researchers into adopting new techniques — are all substantial assets, that come at virtually no price to include into our work. In other words, prediction is truly one of those illustrious free lunch buffets that social scientists continue to ignore at their own peril.

Chapter 3

To Predict and Explain: A Predictive Approach to Studying Teacher Bias

Chapter abstract: The value of prediction has only been recognized in specific empirical settings in sociology, but can be much more broadly relevant to empirical work. As argued in the previous chapter of this thesis, a predictive perspective provides a host of unique advantages and can improve understanding into the models we estimate. Using an estimated model $\hat{f}(\cdot)$ to make predictions provides intuitive and detailed insights into model fit, but can also help to generate understanding into the inner workings of the estimated model. This latter point is particularly relevant when models are estimated that defy straightforward interpretation. In this chapter, I illustrate how prediction can be easily incorporated into standard sociological work, whilst showcasing the benefits prediction brings to model interpretation. I do this by including a predictive perspective into the study of teacher bias in educational tracking in The Netherlands. I show how prediction provides a number of important nuances to the existing knowledge on the subject, which include identifying underestimation of teacher bias towards disadvantaged students in prior work, as well as revealing the importance of the institutional level in determining differences in assigned track levels.

3.1 Introduction

As has been discussed in detail in Chapter 2, prediction is an under-utilized tool in sociology. Although the use of prediction is increasing amongst sociologists applying computationally advanced methods (Molina and Garip, 2019) or those interested in estimating the predictive limits of social theories (Salganik et al., 2020), I argue that prediction is useful to quantitative sociology in a much broader way. In this chapter, I stake this claim by illustrating a number of empirical

benefits that prediction offers when complementing conventional empirical analysis. I do so by incorporating a predictive perspective to the analysis of inequality in educational opportunity through bias in teachers' evaluations of students based on their demographic characteristics. In doing so, I will showcase how prediction leads to improved understanding and richer empirical findings into the topic, in particular by making complex but more appropriate models easier to interpret. Throughout this chapter, I also show how implementing a predictive perspective is virtually costless, and seamlessly complements more conventional empirical analyses.

Emphasis in quantitative sociological work has typically been on interpreting the coefficients of a model $\hat{f}(\cdot)$, which has been estimated to some data. Using that very same model $\hat{f}(\cdot)$ to make predictions is much less common, although there are some exceptions. For example, sociologists applying computational methods like those from the ML domain tend to use prediction as a way to assess model fit and ensure their models are not overfit to the data. A more common use for prediction is to impute missing data, or as an auxiliary part of a regression analysis like two-stage least-squares. More substantively, prediction is used as a tool to generate insights into categorical outcome models, as such models yield coefficient estimates which can be difficult to interpret in an intuitive manner due to their intrinsic non-linearity (Breen et al., 2018). Finally, a more recent trend can be observed, where sociologists are becoming substantively interested in the predictive potential of their models (Watts, 2014) and finding the predictive limits in data (Salganik et al., 2020). Overall, though, prediction is not a mainstream analytical tool and its application remains restricted to relatively niche empirical settings.

As discussed at length in the previous chapter of this thesis, a predictive perspective can provide various types of empirical insights relative to more conventional analyses of estimated models. Three of these are relevant to briefly re-iterate here. First, that predictions provide a natural metric of model fit, which can easily be compared across models. For example when including different explanatory variables, functional forms or modeling assumptions that might complicate comparisons using in-sample methods. Second, making predictions allows for model fit to be assessed

at any level of disaggregation. For example, it allows a fit metric to be calculated for a specific subset of the data rather than for the entire sample, like the R^2 would do. Third, creatively applying the model $\hat{f}(\cdot)$ on permutations of the dataset can generate interpretable insights into model effects. This is particularly relevant when models become complex to the degree that coefficients become difficult to interpret — the approach of using (in-sample) prediction when interpreting categorical outcome models is a case in point.[1]

So on the one hand, prediction provides sociologists with tools to better understand how well different models actually approximate the real-world and allow the researcher to do so at arbitrary levels of disaggregation. A re-appreciation of model fit which should improve alignment between research and practice (Athey and Imbens, 2019; Hofman et al., 2017; Watts, 2017), but for instance also provides researchers with a metric that could be used to compare findings with, and align research efforts towards (Forster, 2002). Importantly, this re-appreciation should not be viewed as a call for blind fit-hunting, but rather an acknowledgement that the ability of our models to approximate the real-world matters, and can be useful to report (Watts, 2014). On the other hand, using prediction as a model interpretation tool should allow sociologists to estimate and interpret more appropriate models to study the social world. As is often the case with observational data, various complexities in model estimation arise. Examples include (unobserved) heterogeneity or non-continuous outcomes like ordinal or categorical variables. Accounting for such complexities in modeling tends to come at a price: the loss of straightforward model interpretation. The fact that categorical outcome models often lack intuitive interpretation is a case in point, which is a topic I return to in this chapter's empirical part. At the same time, it is well-known that ignoring such complexities can lead to incorrect inference due to misspecification bias (Section T.1). In practice, sociologists put a considerable premium on the straightforward interpretation of models as opposed to increasing model complexity. A preference that is evidenced

[1] As has been outlined in the previous chapter, prediction can be made in-sample as well as out-of-sample. In case of the former, prediction simply reflects the analysis of the fitted values of the model.

more broadly by the continuing dominance of the standard linear additive model throughout the social sciences (Schrodt, 2014; Hindman, 2015). Prediction provides a way to transcend this trade-off between complicated, but possibly more appropriate models and interpretability.

To illustrate the varied benefits that prediction can bring to empirical research, I apply a predictive perspective to the topic of inequity in educational opportunity, a classic sociological field of inquiry (Bourdieu and Passeron, 1979; Breen and Jonsson, 2005). I do so by studying teacher bias towards certain demographic groups in educational tracking in The Netherlands. A teacher's subjective perception of a student's ability plays an important role in educational systems, both in day-to-day teaching — by driving the amount and type of didactic material offered to the student — but especially when it comes to more comprehensive decision-making, like (informal) tracking. Therefore, when teacher evaluations are biased — e.g., by preferring girls over boys or advantaged students over disadvantaged ones net of ability — educational equity can come under pressure. I analyze teacher bias in tracking using data on primary school students in the Netherlands, who are tracked into one of five ability levels. Tracking in The Netherlands presents an important moment in the life course, as it occurs comparably early — at the age of 12 — and dictates access to the level of secondary education and tertiary education a student receives. Beyond its substantive importance, the case study of tracking is also highly amenable to a predictive perspective. The study of teacher bias in tracking faces a number of modeling complexities, which are encountered more frequently throughout sociology. These include the need for hierarchical modeling, due to possible school-level heterogeneity, as well as an ordinal outcome variable. This has made the complexity versus interpretability trade-off relevant to the study of teacher bias, as will be discussed in more detail below. In addition, the drivers of variation in the outcome — i.e., what determines track levels — is of substantive interest, which means that model fit is in fact an important estimand.

The rest of this chapter is structured as follows. In Section 3.2, I outline how a predictive perspective can complement traditional model analyses and in what

way it can be implemented into empirical research. Then, I introduce the case study of tracking in The Netherlands (Section 3.3) and the available data (Section 3.4). In Section 3.5, I present results from a conventional statistical analysis and a complementary predictive analysis. In Section 3.6, I provide concluding remarks and discuss this chapter's substantive findings as well as the potential of prediction for quantitative sociologists more broadly.

3.2 The Case for Out-of-sample Prediction

Most empirical work in sociology consists of relating some outcome y with explanatory variables X. Typically, this is done using a linear additive functional form to relate outcome and explanatory variables: $y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$. The model is estimated to the entirety of the available data $\mathcal{D}$ and yields a straightforward, in-sample assessment of the associations between explanatory variables and the outcome, captured for each variable x_k by the estimated parameter $\hat{\beta}_k$. Typically, some mechanism of interest is encoded in the model and the sign and significance of some $\hat{\beta}_k$ is used to evaluate the empirical evidence in support of some hypothesized mechanism.

This approach to empirical analysis has come under some criticism in recent decades. First, the model-building process is often cited as opaque, allowing for so-called researcher degrees of freedom in modeling assumptions and the phenomenon of 'p-hacking' (Young and Holsteen, 2017). Relatedly, the performance of research findings in structuring that which we are interested in is rarely assessed, limiting a comparative assessment of different research findings in how well they fit the data, as well as their practical relevance. In other words, sociologists are more $\hat{\beta}$- than $\hat{y}$-oriented, even though it might seem appropriate that we should also care about the ability of our theories to actually explain what we are interested in as well (Breiman, 2001; Watts, 2014). Second, many of the models typically estimated by sociologists are often criticized to be too simplistic to capture the complexities of social life. There are a number of reasons that may affect the appropriateness of the basic linear model. Data complexities can arise, e.g. when the outcome is

categorical or a count variable, or in case the observed data is selected or censored relative to the population of interest. There may also be (unobserved) sources of heterogeneity requiring effects to differ for subsets of the sample, or the mechanism under study may be more complex than the basic model allows due to non-linearities or interactions. Correctly addressing these types of complexities is paramount to valid statistic inference, but often harm the straightforward interpretation of the basic model described earlier.[2]

Although the importance of appropriate model specification is crucial to valid inference, a preference for interpretability pervades quantitative sociology (as this chapter's empirical section will showcase). This phenomenon is illustrated by the appeal of linear probability models for categorical outcomes which retain the basic model's ease of interpretation (Horrace and Oaxaca, 2006), the often encountered practice of assuming ordered categorical outcomes to be continuous (Timmermans et al., 2015; Meisenberg, 2015; Liddell and Kruschke, 2018; Domingue et al., 2021) and the continuing dominance of the linear additive model throughout sociology (Schrodt, 2014; Hindman, 2015). Put differently, there is a clear trade-off between model-interpretability, on the one hand, and model-complexity, on the other. More often than not, weight tends to lie disproportionately on the former.

Both in terms of re-appreciating a model's substantive fit, as well as providing solutions to the complexity-interpretability trade-off, prediction can play a key role for quantitative sociology. Importantly, without requiring the researcher to change much to their standard empirical workflow. As I will discuss below, a predictive perspective is a complementary analytical tool that can be implemented in most of the models routinely estimated throughout quantitative sociology.

[2]The most often encountered example of complexity affecting model interpretability is the simple case of a binary instead of continuous outcome. A logit or probit approach can address the binary nature of the outcome, but no longer retains the straightforward interpretability of coefficient estimates which the continuous outcome model had (Breen et al., 2018; Mood, 2010). Another common source of complexity is a nested or hierarchical structure in the data, which can warrant the need for varying coefficients at different levels of clustering in the data. Again, the ease of interpretation of the standard linear additive model is lost since conclusions become conditional on group membership (Gelman and Hill, 2006).

A predictive approach to model inference

As described in Chapter 2, incorporating a predictive perspective to an empirical analysis is disarmingly simple. Briefly, the process simply consists of splitting the available data $\mathcal{D}$ into an estimation set, $\mathcal{D}^{\text{estimation}}$, and an evaluation set $\mathcal{D}^{\text{evaluation}}$ using a split index $\mathcal{K}$ assigning each observation in the data to either of the two sets.[3] Then, a model $\mathcal{M}$ is estimated to $\mathcal{D}^{\text{estimation}}$ and evaluated by making predictions of the outcome using $\mathcal{D}^{\text{evaluation}}$. An example of such a predictive routine is illustrated in Algorithm 1. The reader is referred to the discussion in Chapter 2 for a more in-depth discussion of different strategies to choose evaluation and estimation sets.

Algorithm 1: Predictive Routine

for *i in 1:K* **do**
  Partition data into an `estimation set` and an `evaluation set`;
  Estimate model $\mathcal{M}$ using the `estimation set`;
  Make predictions using the `evaluation set` and estimated model $\hat{\mathcal{M}}$;
  Evaluate predictions (against observed outcome or amongst predictions);
  Store the estimated parameters of $\hat{\mathcal{M}}$;
end

As mentioned above and discussed in detail in Chapter 2, implementing a predictive routine like that in Algorithm 1 can provide standard inference with a number of useful benefits. Firstly, the predictions $\hat{y}_{\mathcal{M}}$ provide a natural fit metric of the model $\mathcal{M}$ to assess its ability to structure the outcome of interest.[4] Secondly, using predictions as a fit metric allows the comparison of widely differing models and enables the researcher to diagnose if and how models differ from one another.[5] Third, by varying $\mathcal{D}^{\text{estimation}}$ and $\mathcal{D}^{\text{evaluation}}$ creatively, intuitive model effects can be obtained for any model $\mathcal{M}$. For example, assume we are interested in the effect of sex

[3] Note that the 'train'- and 'test'-set nomenclature is more common in the fields of ML and pattern recognition. I use estimation- and evaluation-set to retain similarity to the terminology more often encountered in the social sciences.

[4] Note that in some cases the outcome itself might be difficult to interpret, for instance when they are on a log scale. In these cases, it is possible to perform similar transformations on both the predicted as well as the observed outcomes.

[5] A similar intuition is applied in ensemble methods that combine various models that may be better able to fit specific parts of the dataset than others (Baćak and Kennedy, 2019; Van der Laan et al., 2007). Predictive model fit as discussed in this chapter provides the key holistic model evaluation metric to compare different models.

on some outcome of interest — for example whether student sex has an association with a teacher's evaluation of the ability of that student — but our model does not allow for the straightforward interpretation of the coefficient on a dummy-coded sex variable. This could for example be because we estimated an ordinal outcome model. By fitting the model using $\mathcal{D}^{\text{estimation}}$ and comparing predicted outcomes for $\mathcal{D}^{\text{evaluation}}$ when changing the observed sex, an intuitive assessment of the impact of sex is obtained as encoded in the model, expressed in terms of the outcome variable. This interpretation is similar to coefficient estimates from standard linear models. In other words, predictive analysis is amenable to *do*-style interventional analysis, but implemented with a particular focus on generating understanding into the associations within a model (Pearl, 2009).[6]

Using $\mathcal{D}^{\text{estimation}}$ and $\mathcal{D}^{\text{evaluation}}$ is amenable to studying more complex effects as well. For example, we might suspect the effect of sex to manifest itself throughout the coefficients in the functional form rather than through a level difference, as would be implied by a dummy-coded variable — this is also called a 'fully moderated model' (Blackwell and Olson, 2021). In such models it can be especially difficult to identify the impact of the interacting variable on the outcome through in-sample analysis of the estimated coefficients. From a predictive perspective, we would simply estimate the model to subsets of the estimation set, and compare the predicted outcomes for the two subsets in the evaluation set using both models.[7] In the example of teacher bias based on student sex, this would boil down to estimating a model using data on male students, and a separate model using data on female

[6] *Do*-style intervention is usually framed within a causal context. In this case, it is designed as a way to obtain an indication of the association between a variable and the outcome. Whether the association has a causal interpretation depends on the causal reasoning encoded in the model $\mathcal{M}$. Whether the estimated model is interpreted through a predictive or an in-sample lens does not affect the appropriateness of a causal interpretation.

[7] The perspective of varying coefficients by group membership has been popularized within the linear additive framework through decomposition methods like the Oaxaca-Blinder decomposition (Fortin et al., 2011). The method proposed here closely resembles the same intuition: that group differences in an outcome can occur due to compositional differences — e.g., different distributions of independent variables by group — as well as effect differences — different effects of independent variables by group. Prediction using models fit to specific groups is a more general approach applying this same intuition.

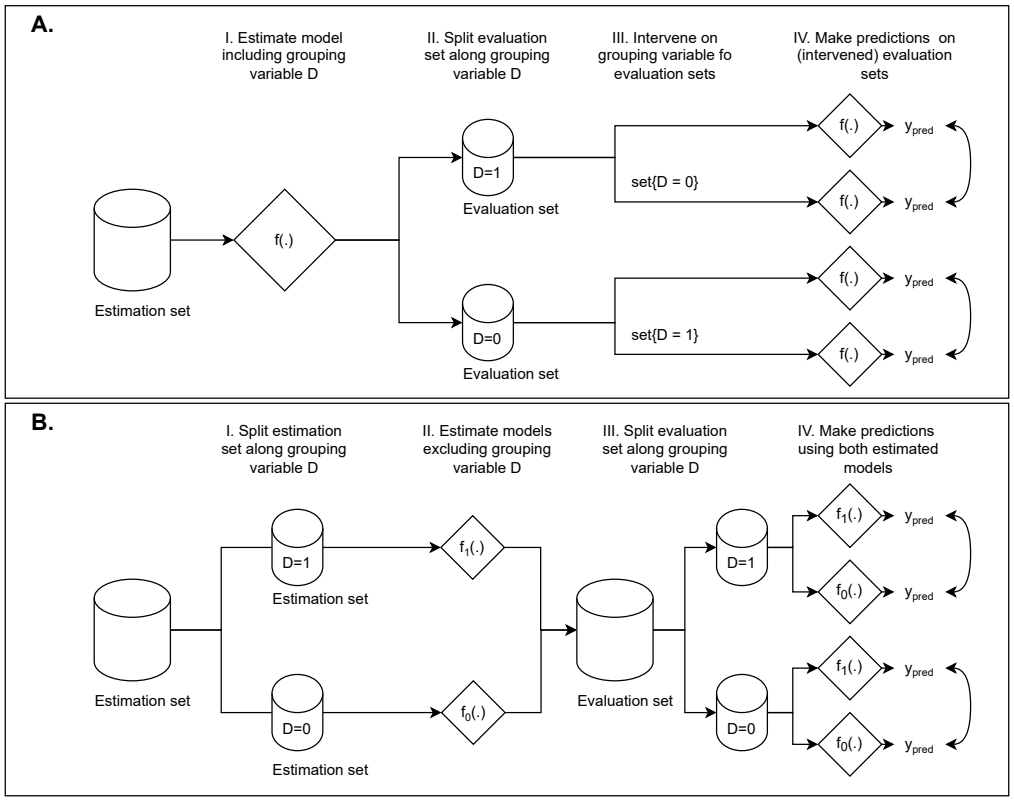


Figure 3.1: Schematic overview of varying the estimation and evaluation sets to generate insights into effect sizes. Top schema illustrates inferring a level difference from the model by varying the grouping variable of interest. Lower schema illustrates inferring group differences throughout the functional form by estimating separate models on partitions of the data along the grouping variable of interest.

students whilst excluding the sex variable from the data.[8] We could subsequently apply both models to the males and the females in the evaluation set, effectively generating predicted outcomes for some observation in the evaluation set as if they were assessed as a male students or as if they were assessed as a female student. Both the interventional approach, as well as the approach where separate models are estimated for subsets of the data are schematically visualized in Figure 3.1.[9]

In practice, Algorithm 1 can be used to append the original dataset with columns containing $\hat{y}$'s for various models. These can then be compared to i) the actual

[8]The sex variable would no longer exhibit any variation after subsetting, and be identical to an intercept.

[9]Whether such an interventional analysis is meaningful depends on the data at hand. For example, it may be that there is separation in other covariates along the grouping variable of interest, which means the observed effects may be correlated with other effects. Of course, this issue is not restricted to a predictive interpretation but rather to the general model that is estimated.

outcome y, ii) the predictions when applying different models, or iii) analyzed with respect to other observables in the data. The price of these additional insights is non-existent in a day and age of computational riches. We simply estimate the exact same model we would otherwise, but to a subset of the data and make predictions on the remaining data or permutations thereof. Crucially, the in-sample results which we would otherwise have analyzed in isolation remain completely unchanged. A further benefit of Algorithm 1 is that the predictive routine can be repeated K times. As a consequence, we obtain resampled estimates of each model's coefficient vector $\hat{\beta}$, which effectively provides insights into the sampling variability of the coefficient estimates, akin to estimating bootstrapped standard errors (Section T.4). In doing so, the process of repeated estimation to (sub-)samples provides general insights into a model's robustness as well.

As mentioned in the introduction, prediction as an analytical tool is by no means novel in empirical research. Although given limited overall attention, there has been an increased interest in assessing how well our theories are able to predict actual outcomes of interest. A notable example in sociology is the FFC (Salganik et al., 2020), discussed in detail in Chapter 2. Prediction is applied more frequently in sociology as a tool to interpret certain types of complex models, in particular categorical outcome models. The typical example is to calculate predicted probabilities, where the impact of changing some interest variable k is evaluated, setting the rest of the variables at representative values, or summarising the impact over the entire sample (Breen et al., 2018). This is akin to the interventional approach outlined in Figure 3.1.[10] The proposed approach generalizes this intuition and applies it on an out-of-sample slice of the data. In this way, the observations used to make predictions are representative of the actual data, and further provide the researcher with the various other benefits of prediction, which have been discussed above.[11] The novelty in complementing empirical work with a predictive perspective

[10]The popularity of the `margins` routine in Stata reflects the general appeal of using prediction as an interpretative tool to understand changes in conditional means in an estimated model.

[11]The standard process of setting all observables to representative values needn't lead to a representative observation, or one for which the model performs particularly well in terms of fit. Predictions don't suffer from this possible issue.

is thus not in identifying previously undiscovered benefits of prediction, but rather to propose prediction as a single, flexible tool that can provide various complementary insights to quantitative sociological work. In what follows, I will showcase this broad applicability and potential for new insights in this chapter's empirical application.

3.3 Empirical Application: Teacher Bias in Student Tracking in The Netherlands

To illustrate how prediction can complement empirical research, I use the case study of teacher bias in tracking. Teacher bias is an important topic within educational research and the life course more generally. A teacher's evaluation of a student determines the attention and type of didactic content offered to a student (Brophy and Good, 1970), and becomes even more relevant with respect to more comprehensive decision-making, like having a student repeat a year. Especially in formally tracked educational systems — where students are assigned to follow a different level of didactic content by their teachers — bias in evaluations can have a profound effect on a student's life course. At the same time, teacher bias in tracking can be equally important — and potentially less scrutinized — in informally tracked systems like that in the United States (Ansalone, 2010; Geven et al., 2018).

There are two reasons that make teacher bias a good use case for a predictive approach. First, studying tracking decisions leads to a number of modeling complexities. These include a hierarchical modeling approach — as individual teachers are nested in schools — and an ordinal outcome variable — there are multiple track levels and these follow a natural ordering. Appropriately accounting for these complexities will complicate model interpretation, which provides scope to apply a predictive approach. Notably, in most empirical work on the subject various simplifying assumptions are made with respect to these complexities, further emphasizing the potential of a predictive approach to improve alignment of the model with the data. Second, accounting for the sources leading to differences in observed tracking decisions across students is of substantive interest: in other

words, model fit is an important estimand in research on teacher bias. This is another element for which prediction is a useful tool.

Extant work on teacher bias

Overall, teachers are found to approximate the skill of students well and bias relating to observable student characteristics like gender and race tend to be minor (Jussim and Harber, 2005; Südkamp et al., 2012). This is the case in experimental designs where two groups with an equal level of ability differ on other observables (Ready and Wright, 2011), as well as when studying observational data to identify systemic biases in teacher evaluations (Sorhagen, 2013; Tach and Farkas, 2006; Watson et al., 2019), although an increasing number of cases has emerged where minority groups in specific regional contexts receive unfavourable bias (Rubie-Davies et al., 2006; Rubie-Davies, 2017; Geven et al., 2018).

The typical methodological approach throughout the literature is to use an objective measure of ability — often a testing scheme — and compare this measure to teachers' subjective reports pre-dating the test. The typical functional form is as follows:

$$y_i = \beta_0 + \beta_1 x_{\text{test},i} + \beta_2 x_{\text{bias},i} + \epsilon_i. \tag{3.1}$$

Our interest lies in the coefficient β_2 indicating whether students with certain characteristics are evaluated more or less positively net of ability. Note that interpretation is straightforward and the coefficient of interest indicates how a unit increase in the bias factor[12] — often a dummy — affects the ability measure used — often some point scale. There are two further extensions appropriate to this model when analyzing teacher bias in tracking decisions.

First, the functional form in (3.1) pools teachers together and evaluates teacher bias from the perspective of the entire educational system. However, low systemic bias does not mean that some (or all) teachers aren't biased: merely that they

[12]Throughout this chapter I refer to 'bias factors' as non-ability related characteristics towards which teachers might exhibit prejudices in assigning ability. 'Teacher bias' reflects the phenomenon of a teacher showing prejudice towards non-ability related characteristics.

aren't biased in a consistent way (Alexander et al., 1987; McKown and Weinstein, 2008; van den Bergh et al., 2010; Timmermans et al., 2015; Glock et al., 2015). Research has shown that institutional characteristics of schools can foster particular perspectives amongst teachers on the role and optimal functioning of an educational system. These factors can then manifest themselves in differences in bias at the institutional level, which may even out across the entire population and thus not show up in an analysis like (3.1) (Marcenaro-Gutierrez and Vignoles, 2015; Dronkers et al., 1998; Kauppinen, 2008). As a result, the school is a more appropriate level to study teacher bias, since interactions amongst teachers within a school are frequent, but often limited towards teachers from other schools (Geven et al., 2018). In other words, without explicit feedback loops there is a risk of group think and drift in evaluation patterns.[13] In modeling terms, this complexity warrants that some or all of the coefficients in (3.1) are allowed to differ at either the school or even teacher level:

$$y_i^{[j]} = \beta_0^{[j]} + \beta_1^{[j]} x_{\text{test},i} + \beta_2^{[j]} x_{\text{bias},i} + \epsilon_i. \tag{3.2}$$

Second, tracking is a comprehensive evaluation rather than the typical point grade often used in research on teacher bias. It has been shown that biases become more pronounced in comprehensive evaluations like deciding a student's track level or retaining a student for an extra year of education (Burkam et al., 2007; Tach and Farkas, 2006; Glock et al., 2013; Glock and Krolak-Schwerdt, 2013; Lucas, 2001). This literature suggests that teachers incorporate a broader set of characteristics — e.g., home environment, behavior and perceived motivation — as higher stakes decisions are made (Becker, 2013; Klapproth, 2018; Böhmer et al., 2017; Boone et al., 2018). From a modeling perspective, these types of comprehensive assessment introduce the need for categorical rather than continuous outcome models. Some have even questioned whether test scores are actually appropriate to be modeled

[13] Note that such drift in evaluations can be assumed to be particularly relevant in countries where residential segregation is strong, and student compositions differ strongly, as teachers are less exposed to the full ability distribution of students.

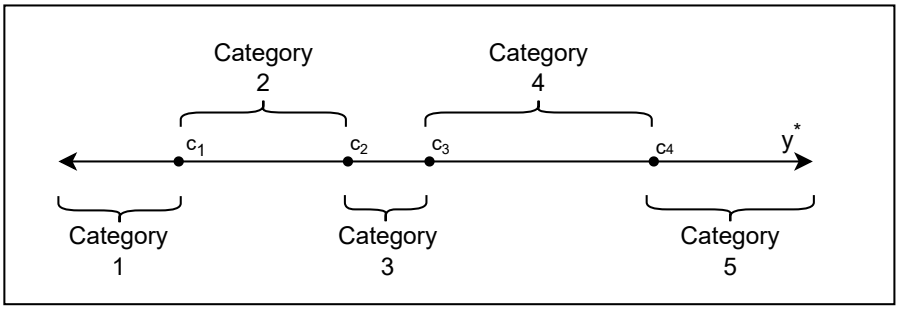


Figure 3.2: An example of the latent variable approach mapping continuous outcome y^* to five categories using cut-off points c_k, $k = 1, \ldots, 4$.

as continuous outcomes, and whether they should not be modeled as ordinal outcomes as well (Domingue, 2014).

Modeling the categorical nature of the outcome variable is straightforward, for instance using an ordered probit model (Maddala, 1986; Agresti, 2010) (Section T.3):

$$\begin{aligned} y^* &= X\beta + \epsilon \\ y &= \begin{cases} 1 & \text{if} \quad y^* <= \mu_1 \\ 2 & \text{if} \quad \mu_1 < y^* <= \mu_2 \\ \vdots & \\ K & \text{if} \quad y^* > \mu_{K-1}. \end{cases} \end{aligned} \tag{3.3}$$

Here, the standard linear additive functional form is preserved into a continuous latent variable y^* which is mapped to the categorical outcome space by defining domains of y^* corresponding to each categorical outcome based on $K - 1$ cut-offs (a visual representation is provided in Figure 3.2). A hierarchical structure like that in (3.2) can be added to this latent functional form in an analogous manner to the standard linear additive model (Section T.3).

As mentioned before, rather than incorporating the above elements a number of simplifying assumptions are typically made throughout prior work. These assumptions tend to avoid complications in the interpretation of model estimates. For example, when introducing heterogeneity through a multilevel structure, the heterogeneity of $\hat{\beta}$ across schools would have to be assessed. In case multiple coefficients are allowed to vary, these have to be assessed in tandem with one another (Gelman and Hill, 2006; Rüttenauer and Ludwig, 2020). The categorical outcome model — arguably even more so than a hierarchical model — further

complicates interpretation of the coefficients $\hat{\beta}$ (Breen et al., 2018). The effect of a coefficient depends crucially on a student's other characteristics because of the potentially unequal spacing of each category's interval and can have a different substantive impact for two observations that differ on observables. As a consequence, almost all empirical work on bias in tracking assumes the outcome variable to be continuous (Timmermans et al., 2015; Lek, 2020; Driessen et al., 2008). This approach is encountered more frequently in the social sciences, for example when studying Likert-type outcome scales in surveys (Meisenberg, 2015; Domingue et al., 2021; Liddell and Kruschke, 2018) even though it is well-established that it can lead to biased inference whenever the equal-interval assumption between observed tracks does not hold (Agresti, 2010; Domingue et al., 2021).[14] Additionally, heterogeneity across schools is usually ignored and, if modeled at all, restricted to adding a random intercept only (cf.: Timmermans et al., 2015). As will be discussed next, these same simplifying assumptions extend to the research on teacher bias in The Netherlands.

The Dutch tracking system

Tracking decisions in The Netherlands are extremely high stakes and the risk teacher bias poses could not have been conveyed any better than in the Dutch Education Inspectorate's 2016 report. In it, it is stated that the Dutch system was of a high level, "[but] at the same time [the Inspectorate] finds that not all students have the chance to obtain the level of education their ability prescribes" (Onderwijsinspectie, 2016, p.). The report goes on to mention that these differences in "chances" — the ability to reach one's potential — have always been unequal across demographic groups, but that this inequality was now on the rise.

As a case in point, Figure 3.3 shows how students with parents who enjoyed high educational attainment increasingly get tracked into higher ability levels than their objective performance would prescribe. The Inspectorate goes on to provide the telling example of two hypothetical students, both at the end of primary school and of similar overall ability. One gets tracked into a slightly higher level than the other:

[14] Note that the equal-interval assumption may also be reformulated as a lack of linearity in the functional form. From either perspective, there is a risk of misspecification bias (Section T.1).

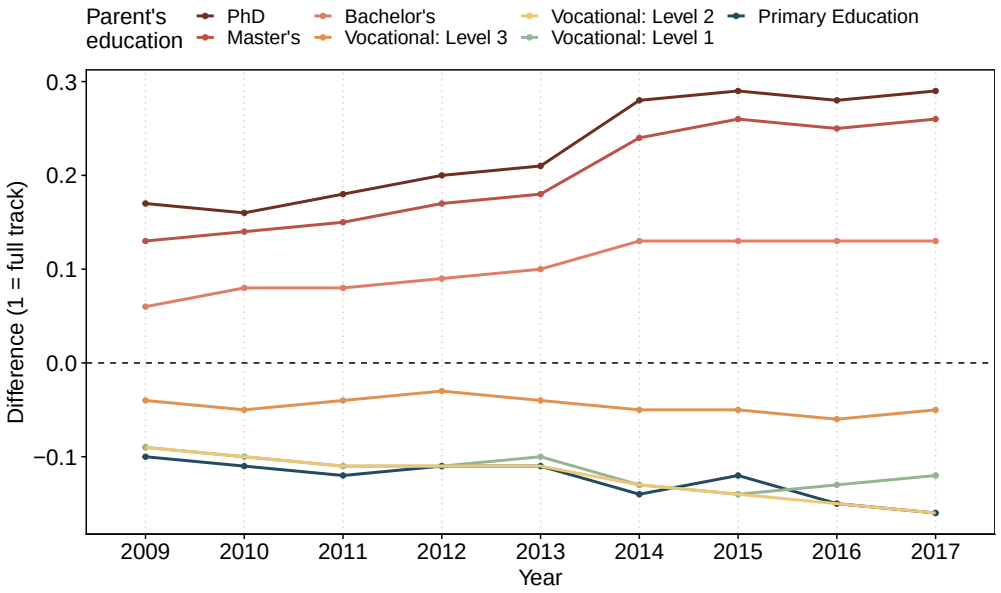


Figure 3.3: Vertical axis indicates the difference between the definitive track level assigned to a student and the track level as suggested by the student's objective performance on standardized tests for the period 2009-2017. Colors indicate the level of the parent's education. Source: CBS Microdata (calculations by Onderwijsinspectie).

a relatively small difference at face value, but which means that both students will enter different schools with different incentive structures, didactic environments and expectations. When the two hypothetical students leave secondary education, both have ended up in different educational strata, with different skill-sets and access to educational opportunities (Onderwijsinspectie, 2016). This self-reinforcing effect of small but early educational selection has been documented more generally (Lüdemann and Schwerdt, 2013; Checchi and Flabbi, 2006; Matthewes, 2021), and in The Netherlands specifically (Tolsma and Wolbers, 2010; Van de Werfhorst and Mijs, 2010; Bol et al., 2014).[15]

Importantly, there are a number of structural features of the Dutch system that exacerbate the downstream risks of tracking, leading international bodies like the OECD to criticize equity in the Dutch tracking system (Field et al., 2007). First, there are five educational tracks and seven educational certifications to be obtained, ranging from practical / vocational education to preparatory academic education. This is more than most countries with formal tracking systems. The

[15] Although students are allowed to switch tracks during secondary education, evidence shows that students tend to stay within the tracks they are assigned to (Tieben and Wolbers, 2010).

moment of tracking is also unusually early in The Netherlands: around the age of 12. Only Germany, Austria and Switzerland have similarly early tracking moments (Prokic-Breuer and Dronkers, 2012).[16] Second, the Dutch system is characterized by rigorous, between-school tracking rather than within-school tracking. This means that students in The Netherlands follow a specific didactic program and have limited interaction across tracks which further pronounce the negative consequences of tracking (Betts, 2011; Houtte and Stevens, 2009; Chmielewski, 2014; Bol et al., 2014).[17] Third, the role of individual teachers in tracking is considerable. Until 2015, track assignment was primarily governed by a central testing scheme, the school leaver's test, but teachers were allowed to propose an alternative assignment in case they deemed necessary. This already allowed for some interference by a teacher in tracking (Timmermans et al., 2015), but as of 2015 the importance of these two criteria has been reversed. The teacher's evaluation of a student now has precedence, with the test acting as a second opinion. The date of the central test was also pushed back until after the start of the official enrolment process into secondary education, making it even less consequential (Smeets et al., 2014) (see Figure 3.4).[18] Finally, the Dutch system permits considerable liberty for schools to follow their own didactic philosophy (Patrinos, 2011; Scheerens et al., 2012; Ritzen et al., 1997). Coupled with the fact that teachers tend to discuss tracking decisions amongst one another further exacerbates the risk of 'group think', where schools might collectively differ in terms of their evaluation of students due to drift over time and a lack of comparison with tracking decisions at other schools (Geven et al., 2018).

Unsurprisingly, there is a wealth of empirical work on the Dutch tracking system and teacher bias in particular (Luyten and Bosker, 2004; Timmermans et al., 2015;

[16]France, England and Italy do not track until the ages 14-16 (Hillmert and Jacob, 2010) and most Scandinavian countries and the United States do not have formalized tracking schemes at all (van der Werfhorst et al., 2015) — early tracking is even prohibited by law in Sweden (Rudolphi and Erikson, 2016).

[17]Schools are allowed to offer multiple tracks, but students follow their own set program within each track. In some cases, schools may offer multi-track classes for the first one or two years of secondary education but this is at their own discretion. In general, the number of schools offering multiple tracks has been steadily declining (Onderwijsinspectie, 2019).

[18]In 2019, the Secretary of Education announced plans to return the school leaver's test to an earlier date to coincide with the timing of a teacher's advice (Note to parliament, June 21st 2019 — reference number: 7850864).

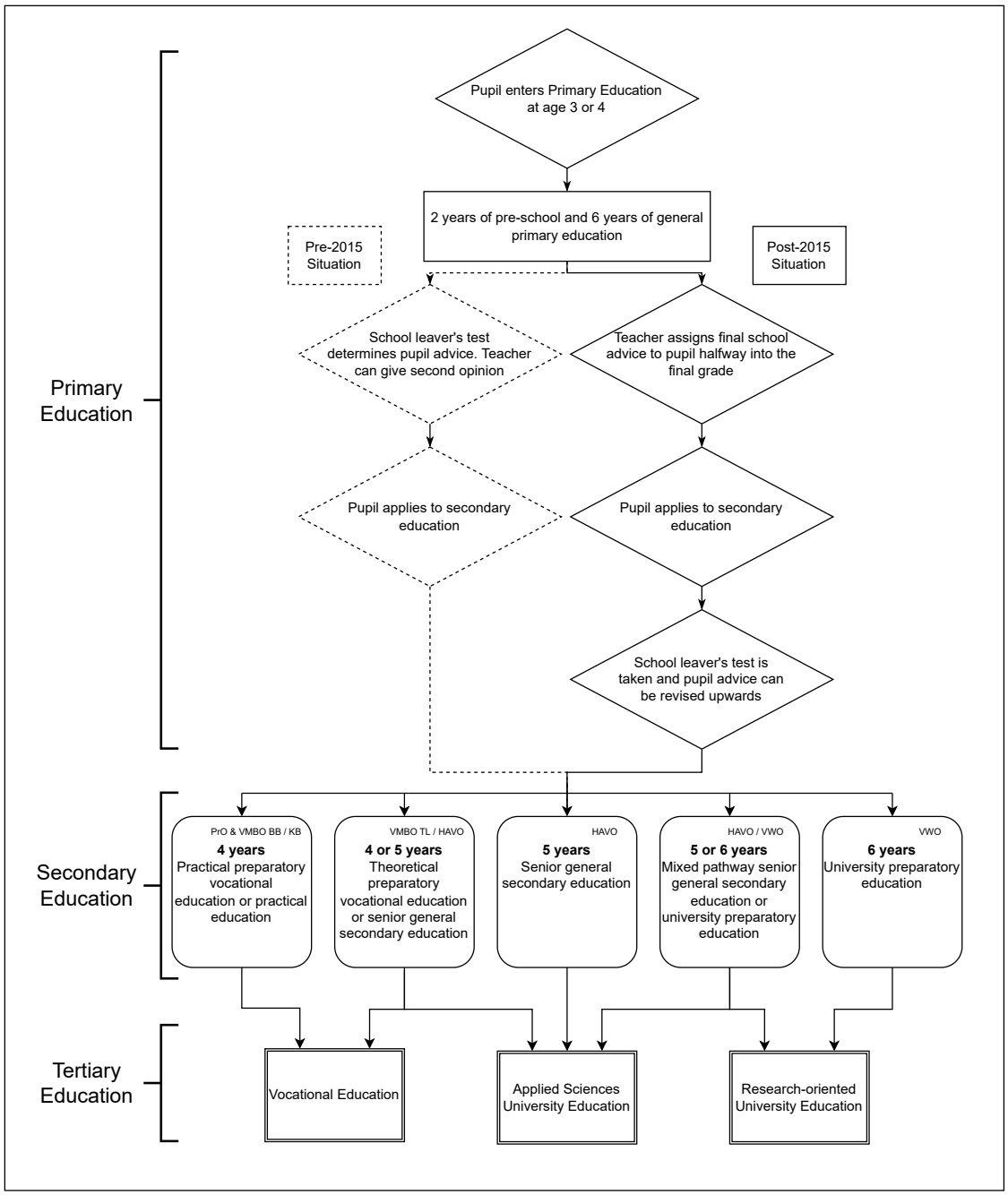


Figure 3.4: Educational structure and tracking moment in The Netherlands. The pre-2015 situation is illustrated with dashed lines, the post-2015 law change is illustrated with solid lines. The five track levels are depicted together with the formal qualifications on offer at the end of secondary education (top right per track level). Subsequent access to tertiary education is illustrated using the double-edged squares at the bottom.

Lek, 2020; van der Werfhorst et al., 2015). Typically, demographic bias factors on the student level like sex, socioeconomic status (SES) and ethnicity have been studied. Consistent evidence is found of positive biases towards girls relative to boys in

assigned track levels net of ability (Lek, 2020; Timmermans et al., 2015).[19] General bias against ethnic minorities tends to be small but has developed from positive bias to negative bias over the course of the early 2000s and 2010s (Jungbluth, 2003; de Boer et al., 2010; Timmermans et al., 2013; Roeleveld et al., 2011; Geven et al., 2018; Driessen et al., 2007). There is considerable evidence that high-SES students are tracked into higher track levels net of ability and it is hypothesized that parents from high-SES families tend to put more pressure on teachers (Dronkers et al., 1998; Onderwijsinspectie, 2016), while low-SES families are less likely to attempt to influence the tracking process (Hillmert and Jacob, 2010). Some suggest that an increase in average track level over time can be largely attributed to increased parental involvement (Timmermans et al., 2018).

As noted above, virtually all existing work on The Netherlands ignores the categorical nature of the outcome variable and assumes assigned track levels to be an interval variable or collapses the outcomes into binary variables (Timmermans et al., 2015; Lek, 2020; Driessen et al., 2008, 2007; Dronkers et al., 1998; Brandsma and Doolaard, 1999).[20] Most work further focuses on teacher bias at the system level, and might under-estimate the full extent of teacher bias when evaluated at a lower level of aggregation, like that of the school (cf.: Timmermans et al., 2015). Throughout the coming empirical analysis, I relax these assumptions whilst using prediction to aid in interpretation and analysis.

3.4 Data & Descriptive Statistics

In order to study teacher bias in The Netherlands, I use data on 746 primary schools covering nearly 50,000 students who finished primary school in either 2018

[19] Sex bias has been suggested to be driven by similar-sex teachers although evidence dismisses this hypothesis (Neugebauer et al., 2011; Coenen and Van Klaveren, 2016).

[20] To the best of my knowledge, the only case where the ordinal outcome of track decisions has been modeled is a study of the Amsterdam area where neighborhood effects are evaluated (Kuyvenhoven and Boterman, 2020). Indicative of the prevalence of assuming the outcome to be continuous are the following excerpts: 'researchers into Dutch school careers have always treated educational recommendation as an interval variable' (Driessen et al., 2008, p. 532) and '[f]ollowing Driessen, Sleegers and Smit (2008), teacher recommendation was treated as an interval variable' (Lek, 2020, p. 83).

or 2019. All student-level information comes from the CITO LVS, which is a nationally implemented Pupil Tracking System (PTS). The dataset has been fully anonymized and was developed for research purposes in partnership with the Mirror Foundation, a research institute in The Netherlands. The data concerns about 15% of the total population and representativity of the data is discussed in more detail below. The Dutch educational data infrastructure has been celebrated for its consistency and comparability by relying on similar tests for students in schools across the country (van de Werfhorst, 2021).

Assigned track level and ability test scores

Each student in the dataset is assigned one of five track levels (Figure 3.4) which is the outcome of interest. As a measure of observed ability of the student, I use standardized test scores from the PTS. These tests are meant to evaluate both new and past material across three didactic areas: reading, maths, and language. The reading test evaluates students' ability to understand written texts (Feenstra et al., 2010). The math test evaluates students' ability to solve problems concerning addition, subtraction, multiplication and division (Janssen et al., 2010). The language test evaluates students' ability to correctly spell words (de Wijs et al., 2010). Tests are taken twice yearly and are uploaded to the PTS by the schools, as is mandated by law.[21]

I average test scores for each didactic domain during the final two grades prior to the tracking moment to use as proxies for ability. Usually, three tests are used per student per didactic area since the tracking moment in the final grade happens half-way into the year. The distribution of the three averaged test scores is given in the left panel of Figure A.3.1, and the association between average test

[21] All tests are developed and psychometrically evaluated by CITO. Each test score is represented by a raw score, reflecting the student's performance on the test, which is transformed to a percentile score by CITO based on representative samples of the Dutch population. The norming by CITO ensures accurate assignment of test results to relatively broad percentile bins — specifically the 0%-10%, 10%-25%, 25%-50%, 50%-75%, and 75%-100% bins (de Wijs et al., 2010; Feenstra et al., 2010; Janssen et al., 2010). Since the dataset used for this chapter is many times larger than the sample used by CITO, I further norm the test scores within each bin such that a uniform distribution is obtained.

scores and assigned track levels is given in the right panel. As can be seen, there is a strong positive relationship between higher test scores and higher assigned track levels, as should be expected.

Student characteristics

As bias factors of interest, I use a student's sex and level of parental education. The former is readily available from the PTS, whereas schools are obliged to collect the latter as part of the government's funding scheme for primary education. So-called 'student weights' indicate whether a student might have an expected learning disadvantage when starting education and the Ministry of Education allocates additional funding to schools with more weighted students (Ladd and Fiske, 2011). All other funding is otherwise allocated on a per-student basis and there is no meaningful private / public distinction in The Netherlands. At the student level, there are three weight types: a 'high' student weight indicates no expected disadvantage and reflects those students who have at least one parent with a degree above lower secondary education; a 'low' student weight indicates neither parent has a degree above lower secondary education but both parents do have a degree above primary education; a 'lowest' student weight reflects the case where neither parent has enjoyed lower secondary education. I code this variable into two levels, reflecting high and low parental education.[22] I also add a dummy-coded control variable indicating whether the student belongs to the 2018 or 2019 cohort.

School characteristics

Finally, I include various school level characteristics to the dataset. First, a school-level version of the weight indicator that has been used for school funding.[23] Second, the denomination of the school, coded as 'Public', 'Christian' (consisting of 'Roman-Catholic', 'Protestant', 'Reformist') and 'Other' (e.g., Dalton, Montessori, Jena).

[22] Robustness checks including the variable as a three level categorical does not lead to meaningful differences in the substantive results.

[23] This measure was introduced in 2019 to replace the student level measure and is a school-level composite of parents' education, country of origin, and the duration of the mother's residence in the country, and whether parents were being assisted by the government for debt restructuring (Posthumus et al., 2016, 2019).

Variable	N	No. Schools	Mean	Std. Dev.	Min	Max
Outcome						
Track: Vocat. Practical	48,946	-	21.7%	-	-	-
Track: Vocat. Theoretical	48,946	-	29.5%	-	-	-
Track: Vocat. / Applied Science	48,946	-	17.3%	-		-
Track: Applied Science / Academic	48,946	-	10.9%	-	-	
Track: Academic	48,946	-	20.5%	-	-	-
Ability measures						
Reading	48,946	-	0.525	0.24	0	1
Maths	48,946	-	0.520	0.28	0	1
Language	48,946	-	0.521	0.27	0	1
Pupil characteristics						
Parental Education: Low	48,946	-	9.6%	-	0	1
Sex: Girl	48,946	-	50.7%	-	0	1
Cohort: 2017-2018	48,946	-	49.1%	-	0	1
School characteristics						
Denomination: Christian	48,946	461	61.2%	-	-	-
Denomination: Public	48,946	234	31.4%	-	-	-
Denomination: Other	48,946	51	7.4%	-	-	-
Urbanity: Very Low	48,946	48	4.7%	-	-	-
Urbanity: Low	48,946	115	14.3%	-	-	-
Urbanity: Medium	48,946	153	21.7%	-	-	-
Urbanity: High	48,946	236	32.4%	-	-	-
Urbanity: Highest	48,946	194	27.0%	-	-	-
Disadvantage: Low	48,946	219	34.3%	-	-	-
Disadvantage: Medium	48,946	272	35.0%	-	-	-
Disadvantage: High	48,946	255	30.6%	-	-	-

Table 3.1: Descriptive statistics of the data on tracking decisions.

Finally, I include a measure of the urbanicity of the neighborhood of the school.[24] Descriptive statistics for the dataset can be found in Table 3.1.

Representativity

The sample is constructed by the Mirror Foundation, a Dutch non-profit research institute and has been used for nationally representative research before (Engzell et al., 2021). As the main goal of the nationally implemented PTS is to allow schools to track the development of their students, there are various tools accessible to schools to analyze student test scores, set goals and make learning plans for individual students. Access to the sample at hand is mediated through schools' adoption of

[24]Most students go to a primary school in their direct neighborhood — the average distance between a student's home and school is 300 meters. Therefore, it is reasonable to use the school's postal code to identify urbanicity.

such a visualization tool. There is thus possibility for non-random selection into the sample. There are various reasons why such selection might be limited.

All schools are publicly funded with relatively more disadvantaged schools getting additional funding - private funds play no role in the Dutch system. There is thus little reason to expect selectivity in terms of resource differences between schools. Furthermore, the upload and management process of the test data is mandated by law and there is minimal upfront investment on the part of the school to use such tools. Importantly, Figure A.3.2 shows that the sample of schools aligns closely with the total population on most observables. In particular, the distribution of low parental education students is similar to the population, as is the indicator for school-level disadvantage. There are more schools containing 101-200 students and less smaller schools, as well as an under-representation of very large schools (500+ students), but the distribution is similar in terms of urbanicity and denomination. The distribution of track outcomes also closely follows that of the total population.

Analytical Strategy

To study teacher bias in The Netherlands whilst addressing possible heterogeneity at the school level as well as the ordered nature of the outcome, I estimate the following model:

$$\begin{aligned} y_i^* = &\beta_0^{[j]} + \beta_1^{[j]} x_{\text{Maths},i} + \beta_2^{[j]} x_{\text{Reading},i} + \beta_3^{[j]} x_{\text{Language},i} + \\ &\beta_4^{[j]} x_{\text{Cohort},i} + \beta_5^{[j]} x_{\text{Bias},i} + \epsilon_i, \end{aligned} \tag{3.4}$$

where y_i^* is a latent continuous variable mapped to 5 distinct levels, reflecting the five available track levels. Ability is controlled for through the three ability scores, and cohort differences through a dummy variable. The x_{Bias} variable reflects either the dummy indicator for low parental education or a student's sex, or school-level information. Possible between-school differences are added through the superscript $[j]$. The innovation of (3.4) relative to existing work is to explicitly address the ordered nature of the outcome, and allow multiple coefficients to differ by school. For illustrative purposes, I will estimate (3.4) for

various levels of complexity, i.e. by assuming the outcome to be continuous as well as ordinal and varying the number of random effects in the model. In doing so, I can compare (3.4) with the models typically estimated under the simplifying assumptions usually encountered in prior work.

After estimating (3.4) for various levels of complexity, I will include a predictive analysis to i) assess how well the various models are able to fit the data, ii) identify what factors most strongly predict bias, and iii) generate insights into associations in case a less interpretable model is required to accurately fit the data.

3.5 Results

Regression results when estimating model (3.4) including both sex and parental education are presented in Table 3.2. I estimate the model for four different levels of complexity: the first column shows coefficient estimates for a standard linear additive model where the outcome is assumed to be an interval variable and without school-level heterogeneity. The results show that performance in maths has the strongest association with a student's assigned track level, followed by reading and then language. All coefficients for the test scores have the expected sign. Additionally, girls get assigned about 0.15 track levels higher than boys, assuming the interval model is correctly specified. Conversely, low parental education students get about 0.17 track levels lower than their peers. The second column estimates the same model but with random intercepts added at the school level. Most coefficients remain unchanged although the effect associated with a student's parental education almost halves. This implies that some of the signal found in the model without random intercepts was due to school-level differences and that low parental education students cluster in schools where students tend to get assigned to lower track levels net of observed ability. The proportion of variance explained at the school level is about 16%. Typically in research on bias in tracking and in line with the existing research on The Netherlands, focus would now switch to the mechanisms underlying these results. Instead, I will estimate the ordered probit version of (3.4) and use prediction to further complement these analyses.

	Interval	Interval	Categorical	Categorical
Intercept	-0.07^{***}	-0.08^{***}	-	-
	(0.01)	(0.01)	-	-
Cohort: 2018-2019	0.02^{*}	0.02^{***}	0.06^{*}	0.05^{***}
	(0.01)	(0.01)	(0.03)	(0.01)
Ability: Reading	1.74^{***}	1.67^{***}	5.70^{***}	3.61^{***}
	(0.03)	(0.02)	(0.11)	(0.04)
Ability: Maths	3.12^{***}	3.08^{***}	9.88^{***}	6.35^{***}
	(0.03)	(0.01)	(0.13)	(0.04)
Ability: Language	0.48^{***}	0.57^{***}	1.63^{***}	1.27^{***}
	(0.03)	(0.02)	(0.08)	(0.03)
Student: Girl	0.15^{***}	0.16^{***}	0.54^{***}	0.36^{***}
	(0.01)	(0.01)	(0.02)	(0.01)
Student: Low Parental Educ.	-0.17^{***}	-0.09^{***}	-0.94^{***}	-0.36^{***}
	(0.02)	(0.01)	(0.05)	(0.02)
Cutoff point: 1|2			5.32^{***}	3.52^{***}
			(0.08)	(0.03)
Cutoff point: 2|3			9.44^{***}	6.22^{***}
			(0.12)	(0.04)
Cutoff point: 3|4			11.73^{***}	7.71^{***}
			(0.14)	(0.05)
Cutoff point: 4|5			13.27^{***}	8.73^{***}
			(0.16)	(0.05)
Random intercepts (school)	No	Yes	No	Yes
Clustered standard errors	Yes	-	Yes	-
Adj. R^2	0.78			
Num. obs.	48,946	48,946	48,946	48946
N Schools	746	746	746	746
AIC		93,779.46	79,117.62	71,616.79
Var: School (Intercept)		0.07		0.39
Var: Residual		0.38		1.00
OOS: Accuracy	57.1%	60.7%	65.1%	68.5%

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

Table 3.2: Main regression results assuming the outcome to be continuous (columns 1 and 2) or ordinal (columns 3 and 4) with random effects at the school level (columns 2 and 4) and without (columns 1 and 3). The multilevel interval model is estimated using the `lme4` package and the multilevel ordinal model using the `ordinal` package in `R`.

Columns three and four show results when addressing the categorical nature of the outcome through the ordered probit model. Significance, sign and relative magnitude of most coefficients are similar to the interval-model, although the effect of parental education seems considerably larger in the categorical models, even when including random intercepts (column 4). As mentioned, there is added complexity in interpretation, since the cut-off points determining the mapping of latent variable to categorical outcome are of a different magnitude both across the two models as well as within each individual model (cut-offs are also shown in Table 3.2). For example, the width of the second interval — mapping the outcome to the second track level — is almost twice as large as the next interval, and three times as large

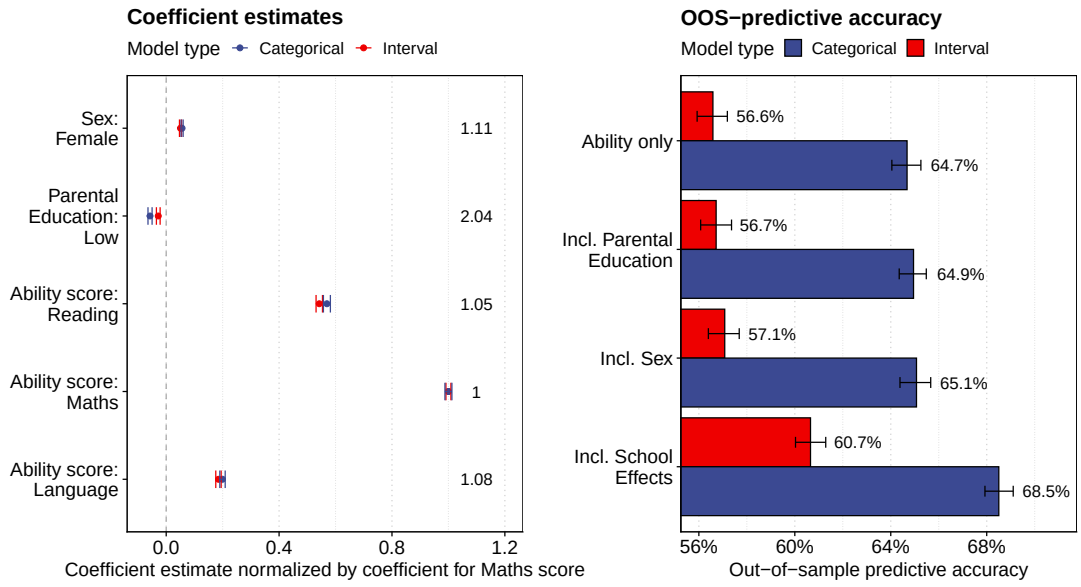


Figure 3.5: Left panel shows coefficient estimates when estimating track levels as an interval variable (red) and as a categorical variable (blue) in a model including random effects on the school level, ability scores and student demographics. All coefficients are normalized per model by the coefficient on maths for comparability. The ratio between the size of the coefficient of the categorical and interval model is depicted on the right. The right panel indicates the out-of-sample predictive performance of various models estimating track levels as an interval outcome (red) or a categorical outcome (blue). Four models are estimated to differing sets of explanatory variables. The first row includes ability scores only. The second includes ability and parental education. The third includes ability and sex. The fourth includes ability and random intercepts. Prediction is done using 250 random splits of the data into a 75% train and 25% test split, stratified at the school level.

as the fourth. This means a fixed increase in the latent variable y^* has a different effect dependent on the distributional location of an observed student. Moreover, none of the estimated coefficients have an intuitive interpretation. Note that by normalizing coefficient estimates within models — in this case by the coefficient on the maths test score — a sense of relative sizes of coefficients can be obtained. Based on these results, it indeed seems to be the case that the low parental education effect strongly increases when estimating the ordinal model (Figure 3.5, left panel).

A predictive perspective

To complement the above analysis and provide insights into the ordinal model, I perform a predictive analysis next. I start by estimating the out-of-sample predictive performance of both the interval- as well as the categorical-model when

including various sets of explanatory variables (Figure 3.5, right panel). For the interval model, continuous outcomes are rounded to the nearest integer to generate concrete track level predictions for each student when using the interval-model. This allows for the actual accuracy of predicting the correct track level to be assessed for the interval-model as well. Overall accuracy of the interval models ranges from 56-61% dependent on which covariates are included into the model, whereas accuracy is about 8 pp better for the categorical model across specifications, implying a considerably better model fit.

The lack of predictive improvement when including student level bias features to the model is striking. Accuracy improves by a mere 0.1 pp when adding parental education and 0.4 pp when including sex compared to an ability-only model. Conversely, addressing for differences in tracking between schools improves the predictive power by more than 3.5 pp. This is noteworthy because much of the existing research on teacher bias focuses on student-level characteristics, even though they add little in terms of making more accurate predictions and thus improving model fit. That the impact of schools has received less attention is possibly a consequence of the fact that school variability is usually reported in terms of the estimated variance coefficient of school-level random effects, which are not typically considered in terms of their substantive effect size and contribution to fit as is quite natural in predictive analysis.[25]

A benefit of OOS-predictions is that local model fit, as well as discrepancies between models can be easily diagnosed by comparing predicted outcomes with observed outcomes. This is illustrated in Figure 3.6 where predictions are visualized for both the interval- and categorical-model using only ability scores as explanatory variables. These figures immediately show why the interval model is misspecified and the clear non-linearity apparent in the data. Forcing an equidistant effect of the explanatory variables onto the outcome leads to poor model fit, in particular for

[25]Typically the Intra-Class Correlation is reported, indicating the relative size of the school-level variance with respect to the residual variance. This metric is dependent on the signal-to-noise ratio in the rest of the independent variables, which is fixed within categorical outcome models, like the logit or probit model. As a consequence, comparisons across models is complicated (Breen et al., 2018; Breen and Ermisch, 2021).

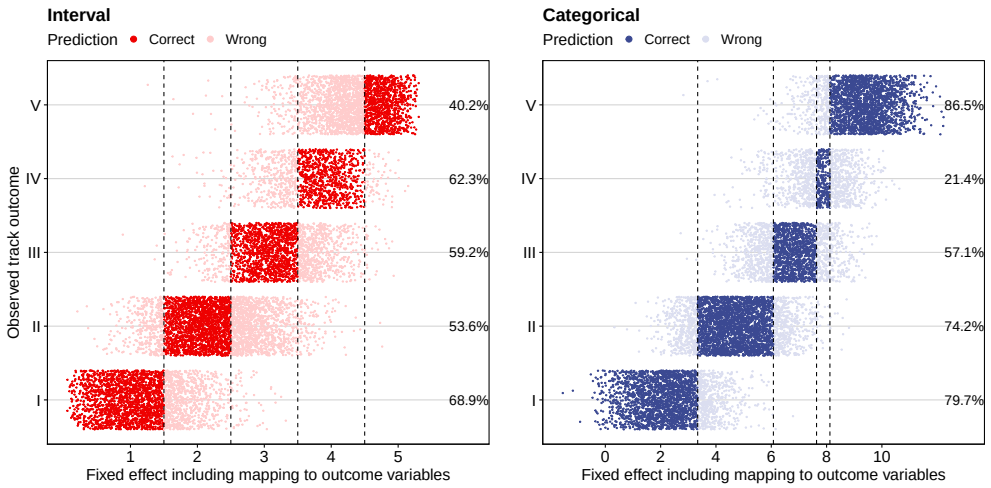


Figure 3.6: Left panel shows predictions of a model estimated when assuming track levels to be an interval variable, including ability scores as explanatory variables. Right panel shows predictions of a model estimated when assuming track levels to be a categorical variable, again using ability scores as explanatory variables. Model predictions are represented by the horizontal axis, while actual track levels are represented by the vertical axis. Vertical dashed lines indicate implied cut-off points mapping numerical predictions to actual track levels. To improve readability, predictions are scattered vertically. Predictions below 0.5 in the interval-model have been assigned the lowest track.

the top track. These graphs further indicate that the lowest track levels tend to be more consistently modeled with students often correctly assigned to the lowest two tracks. The top three tracks have considerably more variability and similar students — in terms of $\hat{y}^*$ — are often assigned to either of the top three tracks. This holds for both the interval- as well as the categorical-model. Based on these non-linearities, the strong increase in the coefficient on low parental education students in the ordinal models is probably due to the fact that these students tend to have lower test scores and estimation will thus be more strongly affected for this group due to the non-linearity in the data (see Table A.3.1 for a full correlation matrix). Pupil sex, conversely, is quite equally distributed across ability.

School level heterogeneity

A key finding from the predictive analysis is the considerable improvement in model fit when including random intercepts on the school level. When including school-level characteristics instead, these variables poorly account for the increase in predictive accuracy (Figure 3.7, left-hand panel) although most coefficients are statistically

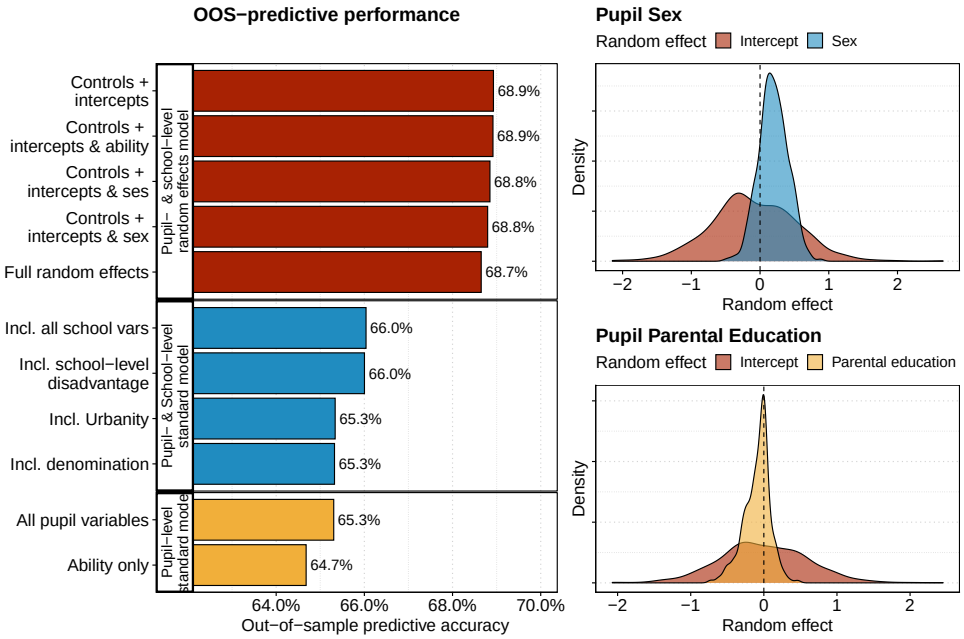


Figure 3.7: Left panel shows the predictive performance of various categorical outcome models. Bottom two models include student level characteristics (yellow). Middle four models include student and selected school level characteristics (blue). Top five models include random effects ranging from intercepts only, to random intercepts and selected random coefficients, to a full random effects model (red). Right panel shows estimated random intercepts (in red) and slopes (in blue for sex and yellow for parental education) of a model including all student level demographics and school-level random effects in the sex coefficient (top) and the parental education coefficient (bottom).

significant (Table 3.3). The school-level disadvantage indicator diminishes the student-level association with low parental education (column 2), in line with the earlier finding that part of the parental education effect resides at the school level. In terms of predictive accuracy, school-level disadvantage only improves on the baseline model by 0.7 pp. The school's urbanicity and denomination do not improve predictions at all. At the same time, adding even more heterogeneity to the model — by allowing all coefficients in the model to vary at the school-level — does not appreciably improve predictive performance either. This implies that most school heterogeneity comes from structural over- or under-evaluation of students rather than differential treatment along specific student characteristics. The right hand panel of Figure 3.7 further confirms that teacher bias towards girls and students of low parental education is unimodal, whereas there is considerable dispersion of the random intercepts which seem bimodal. This finding strongly suggests that

	Categorical	Categorical	Categorical	Categorical
Cohort: 2018-2019	0.06**	0.06**	0.06**	0.06**
	(0.02)	(0.02)	(0.02)	(0.02)
Ability: Reading	5.68***	5.68***	5.70***	5.66***
	(0.07)	(0.07)	(0.07)	(0.07)
Ability: Maths	9.91***	9.96***	9.90***	10.02***
	(0.07)	(0.07)	(0.07)	(0.07)
Ability: Language	1.68***	1.80***	1.64***	1.85***
	(0.05)	(0.05)	(0.05)	(0.05)
Student: Girl	0.55***	0.56***	0.54***	0.56***
	(0.02)	(0.02)	(0.02)	(0.02)
Student: Low Parental Educ.	−0.95***	−0.64***	−0.99***	−0.68***
	(0.04)	(0.04)	(0.04)	(0.04)
School: Other denomination	0.40***			0.24***
	(0.04)			(0.04)
School: Public denomination	0.35***			0.28***
	(0.02)			(0.02)
School: Medium disadvantage		−0.71***		−0.64***
		(0.02)		(0.02)
School: High disadvantage		−1.13***		−1.16***
		(0.03)		(0.03)
School: Low urbanity			−0.19***	−0.14***
			(0.03)	(0.03)
School: Medium urbanity			−0.18***	−0.28***
			(0.03)	(0.03)
School: High urbanity			−0.13***	−0.22***
			(0.03)	(0.03)
School: Very high urbanity			−0.56***	−0.55***
			(0.05)	(0.05)
Cutoff point: 1|2	5.48***	4.76***	5.19***	4.74***
	(0.04)	(0.04)	(0.04)	(0.05)
Cutoff point: 2|3	9.62***	9.02***	9.32***	9.03***
	(0.06)	(0.06)	(0.06)	(0.06)
Cutoff point: 3|4	11.91***	11.38***	11.61***	11.40***
	(0.07)	(0.07)	(0.07)	(0.07)
Cutoff point: 4|5	13.47***	12.97***	13.15***	13.01***
	(0.08)	(0.08)	(0.08)	(0.08)
AIC	78817.21	77138.01	78977.94	76785.04
Log Likelihood	−39396.60	−38557.01	−39474.97	−38374.52
Num. obs.	48946	48946	48946	48946
OOS: accuracy	65.3%	66.0%	65.3%	66.0%

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$
All coeficients are normalized by the coefficient for maths. Reference group in the full model are schools of Christian denomination in neighborhoods of very low urbanity with low disadvantage.

Table 3.3: Regression results of ordered probit models including various school-level characteristics.

institutional differences are a key driver of differences in tracking behavior across schools. This is in line with concerns that a relative freedom for schools to follow their own didactic philosophy and lack of feedback mechanisms can lead to group think.

Bias factors

Given that the less interpretable ordered probit models led to significantly improved fit, prediction can be used to generate understanding into the substantive effects in the estimated ordinal models. To do so, I follow the two strategies outlined in Figure

3.1. I start by estimating separate models for the various bias factors and using these various models to generate predictions for the same student. This approach in practice allows for a full interaction of the variables in the model with the bias factor. Then, I perform a simpler analysis where I estimate the full model and make predictions where I simply change the observed bias factor and compare predictions.

For the question of sex bias, I estimate two models after partitioning the estimation set into boys and girls. One model is estimated using only data on girls, $\hat{f}_{\text{girls}}(\cdot)$, and another using data on boys, $\hat{f}_{\text{boys}}(\cdot)$. The evaluation set is similarly partitioned into a set only including girls and another only including boys. By making predictions for either evaluation set using both estimated models, I effectively obtain predictions for boys when evaluated as if they were boys — i.e., when using $\hat{f}_{\text{boys}}(\cdot)$ as the prediction model — and as if they were evaluated as girls — i.e., when using $\hat{f}_{\text{girls}}(\cdot)$. The same approach can be done for the girls in the evaluation set. Results of this approach are presented in Figure 3.8 for both student sex as well as parental education for which I apply the same routine but using students' parental education as the partitioning variable.

In the top left panel, track levels are predicted for boys using a model estimated to boys (dark blue) and using a model estimated to girls (dark red). The average difference between the two approaches is about 0.15, implying that if boys get tracked as if they were girls they would on average be assigned 0.15 track levels higher. The top right panel shows a similar reversed result. Note that the implied teacher bias towards girls of 0.15 is about the same as was found in the interval model earlier, which might be expected since the distribution of girls and boys is very similar across observables which means that the coefficient is less affected by non-linearities in the process of mapping the latent variable to the outcome space. Note that these results succinctly summarize a full interaction of the covariates with a sex dummy in terms of the impact of sex on the outcome. This is in spite of the fact that none of the coefficients in the ordered probit model are straightforward to interpret.

The bottom two panels of Figure 3.8 show the same approach for parental education. When students of high parental education get tracked as if they were

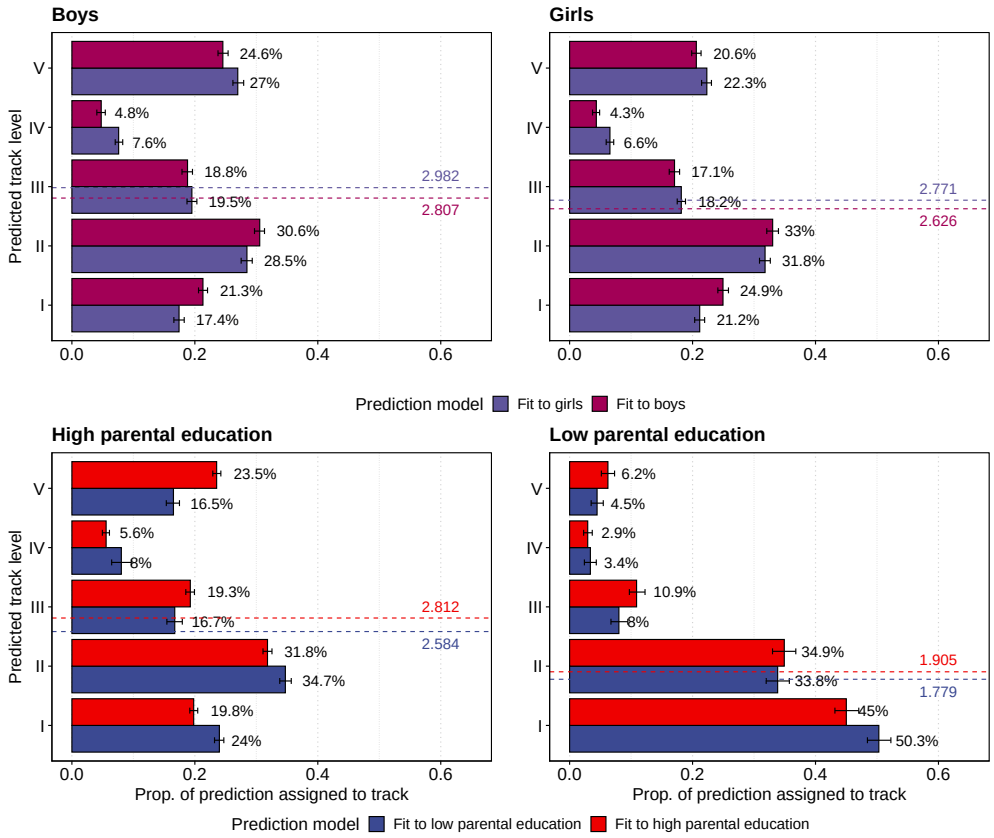


Figure 3.8: Top two panels show predicted track levels for boys (top left) and girls (top right) when using a model fit to boys (dark red) and a model fit to girls (dark blue). Bottom two panels show predicted track levels for students with high parental education (bottom left) and low parental education (bottom right) when using a model fit to high parental students (red) or low parental students (blue). Horizontal dashed lines indicate the average track level amongst each group. Prediction is done using 250 random splits of the data into a 75% train and 25% test split, stratified at the school level.

of low parental education (bottom left), they get assigned about 0.23 track levels lower than they would otherwise. This is considerably more than the interval model suggested. Conversely, low parental education students would gain about 0.13 track levels when treated as high parental education students (bottom right). The difference in effects stem from the fact that most low parental education students are clustered in the lower two track levels, for which the mapping intervals were considerably larger. When the set of students of low parental education is resampled such that it resembles the set of high parental education students in terms of their observed ability, the positive effect for low parental education becomes of a similar size to the reverse effect described earlier.

For both sources of bias, a more simple routine can be estimated by simply intervening on the observed bias feature and comparing predictions — i.e., changing the observed sex for males to female and vice-versa (top schema, Figure 3.1). This gives a direct interpretation of the coefficients from the full model in column 4 in Table 3.2. For sex bias, this routine yields almost identical results as the full interaction model, with boys obtaining 0.16 track levels higher when changing their sex to girls, whereas for girls the reverse effect is found. For parental education, intervening on high parental education students leads to a reduction of 0.17 track levels. Intervening on low parental education students leads to an increase of 0.13 track levels. These estimates are slightly lower than from the full interaction models described above. There might thus be more evidence for different evaluation practices between students of varying parental education than varying sex. Across both routines, however, bias towards low parental education students is considerably larger than those found in the interval model which were estimated to be 0.09 track levels lower; an almost halving of the estimated effect.

Differential evaluation processes across schools

The same predictive approach can be applied to school effects as well. By estimating two models on the estimation set — one with and one without school effects — and making predictions on the evaluation set using both models, the impact of school heterogeneity can be estimated. In other words: track predictions are made using a model estimated without any school heterogeneity — the 'system' model, $\hat{f}_{\text{system}}(\cdot)$ — and compared to predictions by a model estimated with school heterogeneity — the 'school-varying' model, $\hat{f}_{\text{school}}(\cdot)$.

The results are shown in Figure 3.9. Comparing predictions using the population model and the school-varying model indicates the influence of school effects on average track levels. Some schools assign students an entire track lower or higher than the average school would — as proxied for by the model without school effects. Based on frequent resampling, confidence intervals for these effects can be visualized. 30% of the schools consistently deviate from the population model (in blue) with a

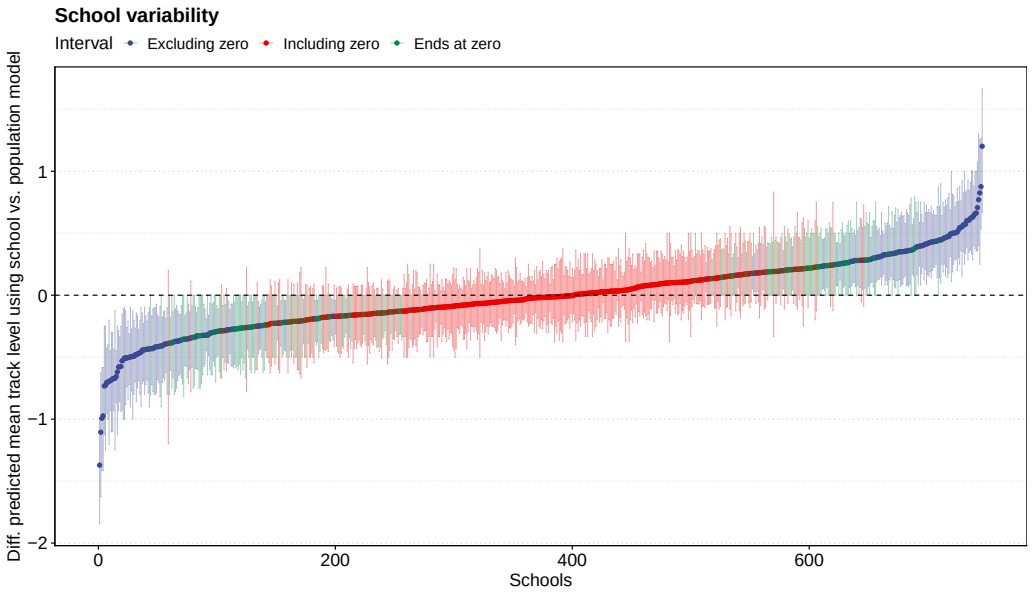


Figure 3.9: Figure shows the difference between the average predicted track levels for a school when using a full random effects model and a population model without any variation on the school-level. Horizontal line indicates no difference between school- and population-model. Prediction is done using 250 random splits of the data into a 75% train and 25% test split, stratified at the school level. Blue dots indicate schools where at least 95% of estimates are above or below zero (30% of schools). Green dots indicate school where at least 95% of estimates are above / below or end at zero (21% of schools). Red dots indicate schools where less than 95% of estimates are above or below zero (49% of schools).

further 21% of schools' confidence intervals ending at zero (in green). In the most extreme case, the most over-biasing and under-biasing schools differ by almost two entire track levels, implying considerable differences in evaluation practices across different schools. Again, using prediction as an analytical tool leads to intuitive effect sizes of the random effects in the ordered probit model whereas these would not be easily interpretable based on the estimates themselves.

In sum, the predictive perspective pursued in this work leads to a number of nuances in the analysis of teacher bias in tracking in The Netherlands. First, it illustrates the importance of addressing the ordinal nature of the outcome for model fit, and the reason for the lack in model fit of the interval-model. Second, it provides interpretable insights into teacher bias towards students' demographic characteristics. Especially notable is the finding that teacher bias towards students with low parental education has likely been under-estimated in prior work. Third, it identifies the importance of schools in generating differences in tracking decisions

net of ability. Many of these findings have been missed throughout prior work, which resorted to more conventional model analyses.

3.6 Discussion & Concluding Remarks

This chapter set out to illustrate the use of prediction as an analytical tool that is broadly applicable for quantitative sociologists. Predictions provide a natural descriptive tool to diagnose model fit at any level of aggregation, and to compare different models. Additionally, predictions can generate insights into complicated functional forms which yield estimates that are difficult to interpret. Usually, such complexity in interpretation leads researchers to opt for more simplistic models which retain a level of direct interpretability, which I have referred to as the complexity-interpretability trade-off. Unfortunately, this ease of interpretation can invalidate inference if the functional form is misspecified. A case in point is the difference in teacher bias estimates towards low parental education students that was found when ignoring the categorical nature of the outcome in this chapter's empirical application. Comparing individual predictions across models illustrated the nature of the misspecification, an understanding into model fit which is not usually accessible in traditional metrics of model fit.

Substantively, this chapter investigated whether teachers in The Netherlands are biased in their track assignment process towards students based on their sex or parental education, as well as the role of the institutional level. This chapter is the first to answer these questions using data after a 2015 law change that made teachers more fundamental to the tracking process. Although broadly in line with work prior to 2015, I find that the size of bias towards students of low parental education will likely have been structurally under-estimated in the past due to a lack of appreciation for the categorical nature of the outcome variable. The predictive approach championed in this chapter allowed for intuitive insights into models addressing the ordinal outcome variable, even though they are known to be notoriously difficult to interpret. Furthermore, this chapter's predictive approach showed that bias towards student-level observables is present, but is less

consequential than the variation in tracking processes between schools, corroborating calls to better study institutional differences.

The predictive perspective pursued in this chapter's empirical application leads to two main substantive takeaways. First, that teacher bias towards pupil characteristics that tend to be irregularly associated with track outcomes — e.g., clustering with high or low ability levels — can suffer from bias when ignoring the ordinality of the outcome variable. This is due to a violation of the equal-spacing assumption implicit in using interval models. As a consequence, past work should be re-analysed to ensure this bias has not affected substantive conclusions and categorical models should be used in future analyses of tracking. Second, institutional differences in tracking are considerable and confirm fears that a mixture of freedom for schools to pursue didactic philosophies of their choice, teacher-based tracking, and a lack of feedback mechanisms have led to deviations or drift in tracking practices amongst schools. More guidance in calibrating tracking decisions across schools seems appropriate. In particular, stimulating interactions between teachers at different schools — which is currently limited in The Netherlands — could be a fruitful avenue to ensure tracking is more consistently applied throughout the system.

There are also limitations to this study's empirical application. Past work has shown that migrant background is another potential source of bias, although findings have been mixed. This source of teacher bias could not be evaluated in this chapter due to a lack of data on a student's migrant background. However, these students tend to score lower on ability scores and likely suffer from similar under-estimation as was found for students of low parental education. Furthermore, due to the 2015 changes to the tracking mechanism, this chapter had to use a different metric of objective ability than most prior work on The Netherlands: I used a student's grades in the final two years, whereas most prior work uses the school leaver's test. This means that direct comparisons across studies is difficult.

More generally, this chapter champions prediction as a method that can be used to broadly complement existing empirical approaches in sociology. Although a key tool in certain empirical settings, a broad appreciation and inclusion of

prediction in the standard empirical work of sociologists is missing. In this chapter, I have argued that a predictive perspective can complement virtually any empirical model estimated by sociologists. In a time of computational riches, there are ever fewer reasons to remain wedded to simplified models. Prediction should be a key companion for sociologists as more appropriate methods to study the complexities of social life are adapted. Unfortunately, prediction as an analytical tool is still underutilized throughout sociology, often for the wrong reasons. Prediction is too often conflated with the use of convoluted models and the intense philosophical debate on whether prediction and explanation are conceptually the same has generated an unnecessary dichotomy between the two. When viewing prediction as a simple tool to improve understanding into an estimated model, a host of applications becomes available, none of which requires a dogmatic choice between prediction or explanation. The empirical use case in this chapter provides a clear illustration of how prediction is both simple to implement and completely complementary to the typical empirical workflow of sociologists. More importantly, it has led to nuanced insights into the existing sociological knowledge on teacher bias in tracking that is representative for the many benefits a predictive perspective can offer quantitative sociologists.

Chapter 4

Incorporating Machine Learning into Sociological Model-building

Chapter abstract: Machine Learning (ML) models have shown an impressive ability to find patterns in data, but have mostly been restricted to predictive questions within sociology. However, ML's ability to identify intricate relationships amongst variables can be equally useful for explanatory work. In this chapter, I propose a framework that can improve sociological model-building by using methods from the ML domain to provide an indication of the fit potential in a dataset. This potential can be used to identify a lack of fit in a researcher's own model, and advances in the field of Explainable AI (X-AI), like Shapley values, can be used to generate understanding into missing patterns when large differentials in fit are found. The framework crucially relies on the use of prediction as a holistic benchmarking measure — a benefit of prediction that has been discussed in the previous chapters — and should improve alignment of our models with the data and guard against misspecification bias. It also improves transparency into the research process by providing insights that are solely dependent on the available data, rather than the data *and* the researcher. The framework is also aligned with calls to use ML to build better interpretable models, rather than replace existing methods. I illustrate the approach through a range of simulation and real-world examples.

4.1 Introduction

ML methods have shown an impressive ability to find intricate patterns in data. As a result, they are increasingly being adopted in industry and research (Rahal et al., 2022; Athey and Imbens, 2019) to assist in finding structure relating explanatory variables, X, with an outcome of interest, y. The key difference with standard empirical work in sociology is that ML methods exploit the exponential increase in computational resources available in this day and age to let the data dictate the relationship amongst variables, rather than relying on the researcher to hypothesize

the functional relationship between the two. Often, this flexibility leads to better fitting models that improve on researcher-hypothesized models in fitting the data (Baćak and Kennedy, 2019). However, ML methods are almost exclusively applied within a predictive context (Mullainathan and Spiess, 2017; Shmueli, 2010) and its inclusion into sociological inquiry — which has an overwhelmingly explanatory focus — remains limited.[1] This is unfortunate, as the fact that ML methods can find relevant associations amongst variables clearly holds value for most explanatory work as well.

In this chapter, I propose a framework to incorporate ML methods into the empirical workflow of sociologists, specifically during model-building. ML methods' ability to find intricate patterns in data can be used by sociologists to provide an indication of the fit potential amongst variables in a dataset. This potential can be used as a benchmark of a researcher's own hypothesized model. In the case of large differences in fit between a researcher's model and the ML model, methods from the X-AI domain can be used to unpack the ML model and provide insights into why the ML alternative leads to improved fit. A popular explanation method which will be introduced in this chapter is the Shapley values approach. More generally, the rapidly evolving field of X-AI provides a range of methods aimed at enabling researchers to understand what patterns are picked up by ML models that are often referred to as 'black boxes', making them more applicable to help answer explanatory questions.

The proposed framework uses ML to address two related and important problems facing quantitative sociologists. First, to help diagnose misspecification and better align our functional forms with the empirical data. Second, to provide increased transparency in a generally opaque empirical research process. It also provides a welcome realignment of empirical methods with the computational reality of this day and age (Efron and Hastie, 2016; Muñoz and Young, 2018, p. 4). Typically, sociologists reduce the social world into tractable mechanisms which can be modelled

[1] Indicative of the predictive focus of ML is the discussion by Mullainathan and Spiess (2017) who explicitly call for the application of ML in predictive contexts or even reformulating explanatory questions into predictive ones. Similar assessments can be found by Molina and Garip (2019) and Shmueli (2010).

and analysed. As data is often observational, we tend to specify functional forms relating an outcome of interest with some variable(s) reflecting our hypothesized mechanism and include possible confounders — the *conditioning-on-observables* approach (Morgan and Winship, 2015). For example, we might hypothesize a model of the following form:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_k x_{ki} + \epsilon_i. \tag{4.1}$$

As has been discussed at various points in this thesis, valid inference crucially depends on a correctly specified relationship between the outcome of interest, y, and the explanatory variables, X (Buja et al., 2019a; Cameron and Trivedi, 2005; Long and Trivedi, 1992) (Section T.1). But how do we know whether the model in (4.1) is correct? The problem is that we typically never know the true functional form. This is a considerable issue given the reliance of inference on correct specification. In practice, though, much more attention is typically paid to identifying relevant variables to include into the model than making sure the functional relationship amongst these variables is correct. This is reflected by the fact that variables are often simply specified to affect the outcome in a linear and additive way (like above). At the same time, there are various examples — and likely many more that have gone unnoticed — where more complicated functional relationships, which could for instance include non-linearities or interactions, have led to reversed findings (Heckman et al., 2018; McClintock, 2017; Freedman, 2009; Christensen and Christensen, 2014; Dougherty et al., 2015) (see also the example in the next section).

Criticism of the ability of linear additive models like the above to capture the complexity of social life has been around for many decades (Abbott, 1988), but computational limitations have historically made such simplifications a necessity.[2] Fortunately, in an age of computational riches, such simplifications are no longer required. ML's ability to find patterns in data through exploiting computational power is a valuable asset available to researchers in this day and age to develop

[2]Linear additive functional forms plugged into exponential family distributions have convenient properties that make estimation considerably more straightforward than non-parametric or even Bayesian approaches to inference (Efron and Hastie, 2016, Chapter 4).

better fitting models and help guard against misspecification. The potential of ML-based model fit to guide model-building has already been illustrated through a number of examples in psychology and behavioral work (Agrawal et al., 2020; Kleinberg et al., 2017; Peterson et al., 2021; Peysakhovich and Naecker, 2017).

Another consequence of a relative lack of emphasis on functional form specification is that it allows researchers to effectively evaluate a large number of possible functional forms and choose which one to report (Sala-i Martin, 1997; Muñoz and Young, 2018; Simonsohn et al., 2020). For example, whether to include a possible confounder in a linear, logged or squared form. Such 'researcher degrees of freedom' are to blame for a large number of important empirical findings turning out to be unreproducible and a more general 'crisis in science' where results are often dependent on over-engineered models (Gelman and Loken, 2013; Young, 2018). Allowing researchers to cherrypick specifications — knowingly or unknowingly (Epley and Gilovich, 2016; Gerber and Malhotra, 2008) — can lead to positive findings far above the reported certainty thresholds (Ioannidis, 2005). Introducing ML methods into the empirical workflow does not solve the problem of 'p-hacking' entirely, but it does introduce additional transparency into the research process in two ways. First, comparing a researcher's carefully tailored model to the fit of a ML method estimated to the same data can reveal a severe lack of model fit and provides a more general insight into the potential fit in the data. Second, ML methods put considerable emphasis on re-estimating models on many sampled subsets, a practice which should help protect further against p-hacking. Beyond these traits, ML is in line with many of the benefits of prediction which have been discussed throughout the previous chapters in this thesis.

Concretely, I propose introducing ML methods into the typical empirical workflow in the following way. Assume a researcher develops some hypothesized functional form, $\tilde{f}(\cdot)$, for which it is unclear whether it correctly represents the underlying relationship amongst variables in the data. First, the hypothesized functional form is estimated and, in parallel, a non-parametric ML model is estimated to the same dataset. The flexible model thus uses the same set of variables as identified by the

researcher in their own hypothesized model, and follows the same inferential logic.[3] However, the functional form in which variables relate to one another is left to be decided by the data. Second, the fit of both the hypothesized model and the flexible ML model are evaluated using cross validation, and a possible lack of fit in $\tilde{f}(\cdot)$ is identified by comparing the fit of $\tilde{f}(\cdot)$ and the ML method. Third, in case a considerable lack of appropriate specification is found, the flexible model is unpacked to better understand the lack of specification in $\tilde{f}(\cdot)$ such that it can be improved.

The framework crucially relies on a number of concepts from ML or closely adjacent fields. First, that cross validation can be used to provide a comparable indication of a model's fit across model domains (Stone, 1977; Rose, 2013; Chapter 2). Second, that ML methods can be used as efficient function approximators (Van der Laan et al., 2007). Third, that methods from the field of X-AI allow researchers to distill understanding from ML methods, which can be used to improve our own models (Samek et al., 2019). These elements, in particular the last, allows for an iterative process where ML is used in a supporting role in model-building. The overarching goal of the framework is still to generate interpretable models, but by using ML methods as a guide rather than a replacement for classical methods (Rudin, 2019; Agrawal et al., 2020).[4]

The rest of this chapter is structured as follows. In Section 4.2, I introduce a toy example to illustrate the risk of model misspecification and researcher degrees of freedom and provide a brief discussion on the topic. Then, in Section 4.3, I introduce each of the three steps of the proposed framework and illustrate its potential by applying the framework to the same toy example from the previous section. In Section 4.4, the empirical part of this chapter, I apply the framework to three cases; one simulated and two real-world datasets. In doing so, I illustrate how

[3]By following a set inferential logic (namely that of the original model), complexity in the proposed framework is introduced solely through the functional relationship amongst variables and distinguishes itself from 'model complexity' as is often understood as the problem of selecting amongst a (large) set of potential covariates (Bucca and Urbina, 2019).

[4]The iterative and complementary role envisioned for ML is in line with an increasing body of literature that recommends using ML in a supportive role throughout the analytical pipeline whilst warning against completely replacing interpretable methods for ML equivalents (Agrawal et al., 2020; Rudin et al., 2010; Rudin, 2019, p. 3).

the framework identifies i) missing non-linearities and interactions in a simulated dataset on the association between schooling and earnings, ii) non-linearities and spatial heterogeneity in a large dataset on London house sales, and iii) various intricate patterns in voting preferences amongst US voters in the GSS. In Section 4.5, I provide a discussion and concluding remarks.

4.2 Incorrect Models and Risks to Inference

Consider the following toy example. We are interested in whether having grandchildren affects elderly individuals' worries about climate change. To this effect, we analyze a survey ($N = 900$) where respondents were asked their level of concern with climate change on a 7-point scale. In addition, we have information on the respondent's age, sex, whether they completed high school and, our interest variable, whether they have grandchildren or not. We might propose the following model ('Model 1'):

$$y_i = \beta_0 + \beta_1 x_{\text{age},i} + \beta_2 x_{\text{sex, i}} + \beta_3 x_{\text{high_school},i} + \beta_4 x_{\text{grandchildren, i}} + \epsilon_i, \tag{4.2}$$

which simply plugs all the relevant controls and our interest variable into a linear additive model. We might also consider including a non-linearity in the effect of age by adding a square ('Model 2'):

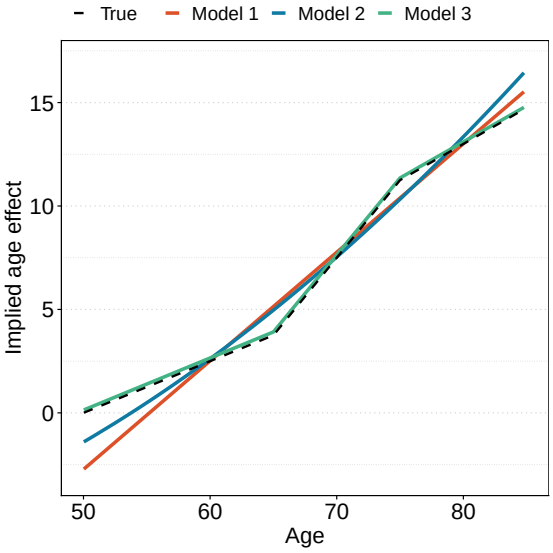
$$y_i = \beta_0 + \beta_1 x_{\text{age},i} + \beta_2 x_{\text{sex, i}} + \beta_3 x_{\text{high_school},i} + \beta_4 x_{\text{grandchildren},i} + \beta_5 x^2_{\text{age},i} + \epsilon_i. \tag{4.3}$$

Results for estimating both models on a simulated dataset are presented in the first two columns in the left-hand panel of Table 4.1. We find a substantial difference between the two models in terms of the estimated association between having grandchildren and concerns with climate change. The effect is statistically insignificant for Model 1 with a point estimate of -0.02, whereas it is negative and significant ($p < 0.01$) for Model 2 with a point estimate of -0.68. A Likelihood Ratio (LR) test strongly prefers the second model over the first ($p < 0.001$). We might feel comfortable to conclude our empirical analysis at this stage.

Table 4.1: Table on the left shows regression results when estimating a functional form assuming linear age, quadratic age, and a step function for age. Plot on the right shows implied effect of age for the three model specifications.

	Model 1	Model 2	Model 3
(Intercept)	−19.02***	−2.16	9.97***
	(0.51)	(3.42)	(0.23)
Age : linear	0.52***	0.03	
	(0.01)	(0.10)	
Grandchildren	−0.02	−0.68**	0.57*
	(0.21)	(0.25)	(0.23)
Sex: female	−0.97***	−0.99***	−0.95***
	(0.10)	(0.10)	(0.08)
High school	0.41***	0.38***	0.36***
	(0.10)	(0.10)	(0.08)
Age: squared		0.00***	
		(0.00)	
Age: 50+			0.26***
			(0.02)
Age: 65+			0.49***
			(0.03)
Age: 75+			−0.41***
			(0.04)
Age: 85+			−0.17
			(0.13)
R^2	0.88	0.89	0.91
Adj. R^2	0.88	0.89	0.91
Num. obs.	900	900	900
RMSE	1.44	1.42	1.23

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



Imagine, however, that the true effect of age is indeed non-linear, but piece-wise linear rather than a second-degree polynomial. Specifically: there is a strong increase in associated concerns amongst those aged 65-75 with more modest increases at other ages. If we had estimated a functional form in line with this non-linearity, we would find results as reported in the third column of the left panel of Table 4.1, which shows that the effect of having grandchildren is in fact associated with a higher level of concern. The estimated age effect per functional form is illustrated in the accompanying figure to the right of Table 4.1. The problem is that both the linear, as well as the quadratic function of age overshoots the effect as the respondent's age increases. Coupled with the fact that grandchildren are more prevalent under the older part of the population, this effect leads to incorrect inference.[5]

To generalize the above, the central issue with model misspecification is that the assumption of exogeneity, i.e., $E(\epsilon|X) = 0$, is violated. Consider four bivariate

[5]Code to reproduce the dataset is available in this thesis' GitHub repository.

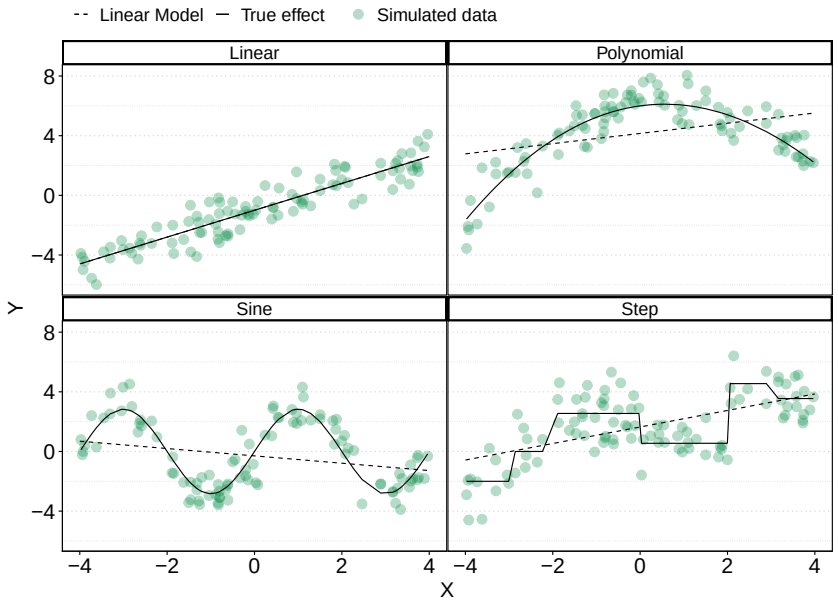


Figure 4.1: Implied association from a linear model (dashed line) relative to the true effect (solid line). The linear models are estimated based on 80 uniformly distributed observations for the X variable, plugged into one of four functional forms (see Section A.4.1) with an additive normally distributed error. Data points are shown in green.

relationships plotted in Figure 4.1.[6] Three of the four models include a non-linear relationship, and although a linear coefficient estimated to the data ($N = 80$) is statistically significant at the conventional 5% level, exogeneity is violated in each of the three cases. This is most clearly visible in the quadratic specification where squared residuals are much larger at the ends and the middle of the X distribution and are clearly correlated with the covariate. This violation of exogeneity leads to a form of omitted variable bias that renders estimated coefficients inconsistent (Long and Trivedi, 1992; Cameron and Trivedi, 2005, Chapter 8) (Section T.1).

To guard against misspecification, various statistical tests have been proposed in the literature. Most popular are White's test for functional misspecification (White, 1980, 1981) and Ramsey's RESET test (Ramsey, 1969). White's test statistic is based on a comparison of the estimated coefficients from a hypothesized model $\hat{\beta}$ with those from a weighted regression $\hat{\beta}_{\text{WLS}}$, whereas the Ramsey test statistic compares the hypothesized model with a model which includes higher order versions of the explanatory variables and evaluates their statistical significance. Both

[6] Choice of functional forms for this example have been inspired by Polley et al. (2011).

provide computationally efficient statistics to evaluate basic misspecification of the functional form, although both also assume that the true complexity of the model lies within a specific class of functions $\mathcal{M}$ with possibly restrictive assumptions (Golden et al., 2016).[7] Since the introduction of the White and Ramsey tests, more computationally intensive (semi-)parametric methods have been developed to estimate parameters of interest in case of non-linear effects in control variables which have been used as the basis for a different set of misspecification tests (discussed in more detail below) (Robinson, 1988a,b; Yatchew, 1997). At the same time, there has been substantial follow-up research building on the principles of the White and Ramsey tests (Golden et al., 2016). However, there are various well-known limitations to these types of specification tests (Buja et al., 2019a).

First, they are generally under-powered and struggle to identify misspecification in multivariate settings (Buja et al., 2019c, p. 616).[8] Second, they provide limited insight into what to do next, whenever a test is rejected. Third, and perhaps most importantly, the application of misspecification tests differs substantially throughout published empirical work (Long and Trivedi, 1992; Open Science Collaboration, 2015), and continues to this day. It is safe to say that the econometric frontier has left the practical frontier far behind and the wealth of follow-up research on specification tests has barely penetrated into typical empirical analysis. As a consequence, researchers enjoy relative freedom in specifying their functional form, as well as the robustness and specification tests they use to verify its appropriateness. In other words, the crisis in reproducibility and the impact of researcher degrees of freedom on empirical results extend to the (lack of) specification checks chosen by researchers to validate their model assumptions (Simonsohn et al., 2020; Young, 2018).

Even without malicious intent, many functional forms are from an age of computational limitations that are no longer present (Muñoz and Young, 2018;

[7] For example, neither test will pick up an interaction of some continuous variable x_k and a binary variable D if the (linear) slopes are mirrored for both groups.

[8] To illustrate, misspecification tests are generally able to identify the bivariate non-linearities in Figure 4.1, although a standard RESET test assessing a possible second degree polynomial fails to reject the step function and requires a third degree polynomial. For the toy example, White's test fails to reject misspecification of the quadratic model.

Efron and Hastie, 2016; Buja et al., 2019c, p. 615). The linear additive functional forms plugged into exponential family probability distributions were historically preferred due to their ease of estimation, not their de-facto appropriateness to study social life. This pragmatic preference for parsimony has led to a prevalence of fairly simplistic functional forms which remain in use to this day, even though the wealth of qualitative research in the social sciences consistently implies that social life is in fact highly complex and probably not appropriately modelled using such simple functional forms (Abbott, 1988). As the toy example above illustrates, it is likely that under-specification of functional forms is a prevalent problem throughout sociology that can easily go unnoticed. Failure to consistently apply the tools to diagnose such a lack of model specification will maintain this status quo and further increase the number of false and irreproducible findings throughout empirical work.

4.3 A Computational Framework to Improve Model-building

Instead of (selective) application of specification tests to hypothesized functional forms, I propose a computational framework that exploits ML methods to obtain an indication of the potential model fit in a dataset. This potential can then be compared to the fit of the hypothesized model, and thus used to diagnose a lack of specification in the latter. The proposed framework has a number of benefits. First, it is simple to apply and provides the researcher with an intuitive assessment of how well the data *could* be modelled versus how well the data *is* modelled by their hypothesized model. Second, as long as a broad suite of flexible form models is applied to generate the fit potential, there is no room for researcher degrees of freedom to affect findings. Third, using ML models to benchmark model fit allows for an explicit feedback loop to understand how to improve an underspecified hypothesized model. In case a lack of fit is found, the ML model outperforming the hypothesized model can be unpacked to understand what patterns are picked up and possibly missed by the hypothesized model.

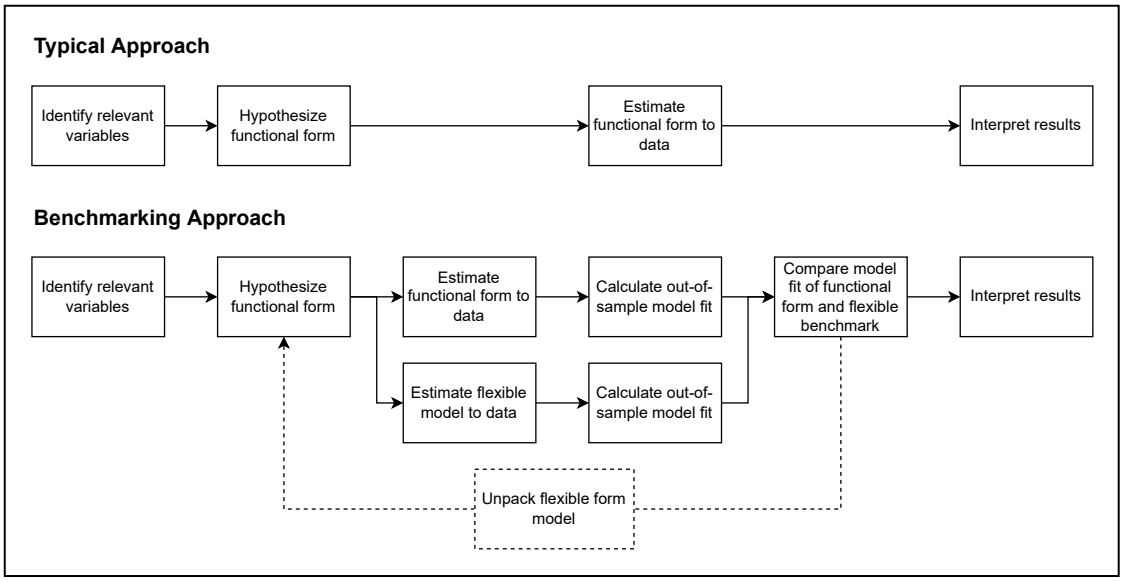


Figure 4.2: Schematic illustration of the proposed benchmarking framework.

Given a dataset $\mathcal{D}$ and a researcher-hypothesized functional form $\tilde{f}(\cdot)$, relating outcome of interest y with independent variables X, the proposed framework consists of the following three simple steps:

1. Estimate hypothesized model $\tilde{f}(\cdot)$ to the data, as well as a ML model $\bar{f}(\cdot)$;
2. Evaluate the model fit of both $\tilde{f}(\cdot)$ and $\bar{f}(\cdot)$ and assess a possible lack of fit in $\tilde{f}(\cdot)$;
3. Diagnose why $\bar{f}(\cdot)$ improves on the hypothesized model and improve $\tilde{f}(\cdot)$ accordingly.

The proposed benchmarking framework including each of the three steps is illustrated schematically in Figure 4.2. Note that benchmarking uses the same variables as originally hypothesized by the researcher. The framework is not designed for data-mining where a large number of possible explanatory variables are included into the model, which risks including variables that might improve model fit but could harm our understanding of the underlying processes — typical examples would be to (accidentally) include post-treatment variables or colliders.[9]

[9]As the flexible model uses the exact same inferential reasoning in terms of which explanatory variables are included into the model, much of the typical criticism regarding model fit is not

Rather, the approach here is to evaluate the functional form assumptions given the exact same inferential curation of variables and logic as present in the researcher's originally hypothesized model.[10]

The framework relies crucially on three components. First, a flexible form *benchmarking model* to estimate the potential fit in the data. Second, a *fit metric* to compare the flexible model's performance to that of the hypothesized model and identify a possible lack of model fit. Third — in case a considerable lack of fit is found — an *unpacking method* is needed to generate understanding into why the hypothesized functional form underperformed the benchmark. In what follows, I will describe each of the three components of the proposed framework in-depth and illustrate them using the toy example introduced earlier.

Step I: The benchmarking model

The first component of the proposed framework is a flexible form model $\bar{f}(\cdot)$ to serve as a benchmark of the possible model fit in the data. This model should include the same variables and thus follow the same inferential logic as the hypothesized model $\tilde{f}(\cdot)$, but evaluate patterns not necessarily considered by the researcher's functional form. A number of ML methods are highly appropriate for this task. In general, ML methods do not rely on pre-specified functional forms, but distill the functional form from the observed data (Mullainathan and Spiess, 2017; Molina and Garip, 2019; Baćak and Kennedy, 2019). They also retain more structure than non-parametric approaches like local regression and suffer less from the curse of dimensionality (Bishop, 2006, Chapter 1). ML methods are still relatively costly from

applicable here. Often, 'fit-hunting' is criticized based on examples where all causal logic is effectively abandoned in favor of model fit. A case in point is the example brought up by Young (2019) of a study into the gender wage gap done at Google, which included job promotions as an explanatory variable. Clearly, including promotions would improve fit, but invalidates any inference as it is a result (i.e., post-treatment) of the very process of interest. Although this example showcases accurately that a singular focus on fit is problematic, it mostly emphasizes that a focus on model fit still requires causal logic, just as in typical model-building.

[10]Note that ML methods *are* often used to identify relevant variables amongst a (large) set of possible covariates (Bucca and Urbina, 2019). This is another fruitful applications of ML into sociological model-building, but outside of the scope of this chapter.

a data perspective, but are feasible to estimate when analyzing datasets commonly encountered in the social sciences (Hastie et al., 2009; Athey and Imbens, 2019).

In particular tree-based models — a subset of ML methods — lend themselves exceptionally well to the envisioned benchmarking approach (Section T.5). First, tree-based methods can deal with various types of data often encountered in the social sciences — e.g., categorical as well as numerical variables — and handle missing data well. Second, tree-based methods do not require exponential increases in sample size as the feature space increases, like other popular methods from the ML domain like neural networks. Third, the most popular tree-based methods like the Random Forest (RF) or Gradient Boosting (GB) model incorporate regularization and variable selection by design, requiring less tuning on the part of the researcher and allowing the method to be applied out-of-the-box or with limited effort to most datasets.[11] Finally, tree-based methods have been shown to perform well amongst ML methods across various applications whenever data is tabular, as is almost always the case in the social sciences (Lundberg et al., 2020; Chen and Guestrin, 2016).

Take the two most common tree-based methods, the RF and the GB model. Both approaches share the flexibility of a decision tree, but address risks of overfitting to the sample at hand by averaging many decision trees estimated to slightly altered (for RF) or weighted (for GB) versions of the data (Breiman, 1996; Freund and Schapire, 1996) (see Section T.5.2 for a more detailed discussion of tree-based models). This allows for the estimation of highly non-linear patterns without requiring any type of functional form specification on the part of the researcher. As a case in point, consider the non-linear associations in Figure 4.3, which had been introduced earlier. As before, the black lines illustrate true associations and the green dots a hypothetical dataset generated by the true association and some white noise. The blue and red dots are out-of-sample predicted values based on a

[11] For some ML methods the optimal solution is dependent on the scale of the explanatory variables. For example, least absolute shrinkage and selection operator regression introduces an additional term to the squared error loss of a standard OLS regression which reflects the absolute size of all coefficients $\hat{\beta}$. This effectively shrinks large coefficients which provide limited improvement to fit towards zero. Clearly, dividing or multiplying an explanatory variable x_k by 100 would focus or divert attention of the algorithm on $\hat{\beta}_k$. Decision trees are scale invariant as single binary splits are placed on individual explanatory variables.

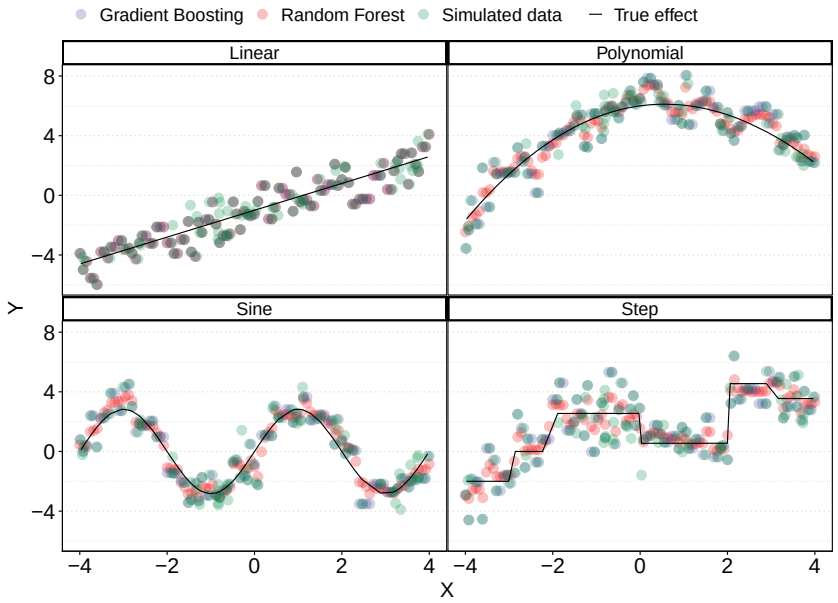


Figure 4.3: Predicted outcome using a GB (blue) and RF (red) model fit to noisy data (green). True effect is depicted with the black line. Each model is estimated based on 80 uniformly distributed observations for the X variable, plugged into the specified functional forms (see Section A.4.1). Predictions are made using another 100 uniformly distributed observations. The default parameter values for the RF and GB (with `nrounds` set to 200) Models are used from the `randomForest` and `xgboost` packages in `R`.

GB and a RF model trained to 80 observed data points and fed with 100 uniformly distributed X values to generate $\hat{y}$-predictions for. As can be seen, both models learn the non-linearity in the data well and do so without requiring any sort of structure or pre-hypothesized functional relationship.

Besides a particular tree-based model, other ML or non-parametric methods could be used as well. More generally, limiting the benchmarking step to a single researcher-curated ML model to serve as $\bar{f}(\cdot)$ would clearly invite some of the very risks this framework is designed to address in terms of researcher degrees of freedom in model-building. Although ML methods tend to fit tabular data well, tuning parameters can easily be calibrated in a way to diminish model fit and thus imply appropriate fit of $\tilde{f}(\cdot)$ when this is not the case.[12] Relatedly, some methods fit certain patterns in data better than others with often limited a-priori guidance (Rose, 2013; Baćak and Kennedy, 2019; Berk and Bleich, 2013). The benchmarking model

[12]For example by forcing trees to be unreasonably shallow or deep, possibly leading to limited fit or severe overfitting with poor performance out-of-sample.

can be cherrypicked just as well as a functional form. As a consequence, a principled approach would be to estimate not a single but an ensemble of flexible form models during the benchmarking phase. Such an approach has a number of benefits.

An ensemble or 'Super Learner' can consist of multiple models, as well as the same model with different parameter settings. By evaluating the ensemble, the most appropriate model to estimate the patterns in a specific dataset is identified amongst the included models using cross validation. The oracle result by Van Der Laan and Dudoit (2003) shows that a Super Learner is indeed optimal to identify the best fitting model to approximate an unknown DGP amongst the evaluated models, and the price of including large numbers of models into the Super Learner is minor in terms of performance (Van der Vaart et al., 2006; Van der Laan et al., 2007; Baćak and Kennedy, 2019) (see Section T.6 for a detailed discussion). Ensemble ML methods are implemented more generally as universal function approximators when functional forms are unknown (Hornik et al., 1989; Agrawal et al., 2020). Using a Super Learner also improves transparency as it can easily be specified prior to model estimation or even pre-registered (Open Science Collaboration, 2015). Naturally, the full set of models and the resulting fit of each should be part of the research output (as is illustrated in the examples below).[13]

I put the first step of the framework to the test by estimating a Super Learner including a number of GB and RF models with various parameter estimates to the toy example introduced earlier. In addition, I include the two researcher-hypothesized models to the Super Learner as well, specifically 'SL.GLM_Linear' and 'SL.GLM_Quadratic' reflecting the model including linear age and quadratic age, respectively. For illustrative purposes, I also add the true model, 'SL.GLM_PieceLinear', even though I did not in fact hypothesize this model originally. A truncated output from the Super Learner can be found in Table 4.2 and shows the root mean squared error (RMSE) of various models included in the Super Learner, estimated based on 10 folds using cross-validation (for the full output see Table A.4.1). As can be seen, the

[13]Throughout the empirical examples in this chapter, I will use the Super Learner approach as implemented in the `R` package `SuperLearner` although various alternatives exist in different programming languages (Polley et al., 2021).

Model	Ave	SD	Min	Max
SL.GLM_PieceLinear	1.218	0.060	1.046	1.320
SL.GB_200_1_0.1	1.229	0.059	1.107	1.327
SL.GB_500_1_0.1	1.231	0.059	1.104	1.323
SL.GB_100_1_0.1	1.241	0.059	1.129	1.351
RF_200_2	1.383	0.064	1.229	1.544
SL.GLM_Quadratic	1.384	0.063	1.174	1.536
SL.GLM_Linear	1.413	0.065	1.321	1.569
SL.GB_200_5_0.01	2.698	0.088	2.484	2.926
SL.GB_200_6_0.01	2.698	0.088	2.486	2.927

Super Learner performance based on RMSE & 10-fold CV.

GB parameters: `n_rounds_max_depth_eta`, RF parameters: `n_trees_mtry`.

Output is truncated for brevity, see Table A.4.1 in the Appendix for the full output.

Table 4.2: SuperLearner performance of the Toy Example

best performing model is — unsurprisingly — the true model, but its performance is closely tracked by a GB model which strongly outperforms both the Linear as well as the Quadratic specifications. What is more, including the GB model with different parameters (in this case a slow learning rate combined with high or low tree depth) leads to overfitting and a strong underperformance in terms of fit. This further illustrates the need to include various models during benchmarking. The results from the Super Learner suggests, correctly, that a linear or quadratic specification misses important patterns in the data which are picked up by the GB model.

Step II: Estimating and comparing model fit

The second component of the proposed framework is a metric to compare the benchmarking model $\bar{f}(\cdot)$ directly with the hypothesized model $\tilde{f}(\cdot)$ in terms of its ability to fit the data. Comparative statistics like the AIC or BIC, or LR tests (in case of nested models) are typically used to compare two or more explicitly hypothesized models. These statistics are not easily transferred to ML methods, mainly because they rely on in-sample diagnostics and specified functional forms or require estimation of degrees of freedom known to be challenging for ML models (Janson et al., 2015). More generally, ML methods should not be evaluated on their in-sample performance due to the risk of over-fitting and are typically assessed via

Variable	N	Mean	SD	Min	Pctl. 5	Pctl. 95	Max
RMSE							
GB	1000	1.241	0.06	1.062	1.145	1.34	1.414
Linear	1000	1.417	0.064	1.195	1.314	1.522	1.643
Quadratic	1000	1.391	0.065	1.181	1.284	1.495	1.602
PieceLinear	1000	1.222	0.06	1.049	1.124	1.316	1.414
GB vs. Linear	1000	-0.176	0.049	-0.315	-0.254	-0.097	-0.024
GB vs. Quadratic	1000	-0.149	0.046	-0.284	-0.222	-0.073	0.008
GB vs. PieceLinear	1000	0.019	0.02	-0.042	-0.013	0.051	0.109
R^2							
GB	1000	0.656	0.005	0.638	0.647	0.664	0.671
Linear	1000	0.626	0.005	0.609	0.617	0.635	0.64
Quadratic	1000	0.63	0.005	0.613	0.621	0.639	0.645
PieceLinea	1000	0.656	0.005	0.639	0.648	0.665	0.672
GB vs. Linear	1000	0.03	0.002	0.024	0.027	0.033	0.037
GB vs. Quadratic	1000	0.025	0.002	0.019	0.022	0.028	0.032
GB vs. PieceLinear	1000	-0.001	0	-0.002	-0.001	0	0

Table 4.3: Three linear additive models are estimated on a train-set consisting of 80% of the total dataset. The three linear additive models assume the effect of age to be respectively linear, quadratic or piece-wise linear. A GB model is also fit using the `xgboost` package in `R` with the following parameters `depth=1, nrounds=200, eta=0.1` which were shown to be optimal in the Super Learner routine (Table 4.2). RMSE and OOS R^2 is calculated for each model based on the remaining 20% of the data. The last three rows of the RMSE and R^2 parts of the Table show bootstrapped results of the difference between the optimal GB model and each of the linear additive models. 1,000 splits of the total dataset into train- and test-set are evaluated.

cross-validation (Shmueli, 2010).[14] The universal comparability of models' out-of-sample prediction discussed earlier in Chapter 2 makes them an appropriate fit metric for the framework. The reader is referred there for a more detailed discussion of the various benefits of prediction and approaches to operationalize them. Throughout the rest of this chapter, I will use K-fold cross validation and Monte Carlo cross validation to assess out-of-sample model fit in terms of the RMSE and R^2.

Putting the second step in practice, I compare the out-of-sample performance of the GB model identified in the previous step with the two hypothesized alternatives for the toy example. As mentioned above, I include two fit metrics: the out-of-sample

[14]For example, it is non-trivial to find in-sample equivalents of model fit when including certain types of ML methods relative to parametric models. For example, for tree-based method it can be challenging to define an analogue to the degrees of freedom available in the parametric context. Out-of-sample evaluation provides a simple solution to the problem of comparing fit between parametric and ML methods (Hastie et al., 2009, Chapter 7).

R^2 and the RMSE, and evaluate both using 1,000 splits of the total dataset into an estimation and evaluation set. The results are summarized in Table 4.3. The top half of the table reflects the RMSE, whereas the lower half reflects the R^2. For both metrics, the flexible model identified by the Super Learner, and the two hypothesized models are evaluated. Again for illustration's sake, the true piece-wise model is included as well. The bottom three rows for both metrics show the difference in the fit metric between the flexible model and the hypothesized models. As can be seen, the flexible model strongly outperforms the linear and quadratic models, but does not improve on the true model as should be expected. This leads us to conclude there is substantial fit improvement possible for the two hypothesized models.[15]

Step III: Unpacking the ML model

The third and final component of the proposed framework is a method to generate understanding into the ML model when it outperforms the hypothesized model in a way that warrants re-evaluation of the originally hypothesized functional form $\tilde{f}(\cdot)$.[16] Increasingly, researchers and practitioners have called for better understandable ML models given their broad application in everyday life and the risks a lack of understanding might bring when deploying ML in practice.[17] These developments have led to the emergence of X-AI, which focuses on understanding *why* ML models generate the predictions they do (Ribeiro et al., 2016; Samek et al., 2019). Two

[15] As mentioned, the 1,000 repeated estimations of the hypothesized models has the added benefit of providing bootstrapped standard errors of the coefficient estimates which are presented in Figure A.4.1 together with the estimated coefficient on the full set. As can be seen, in this case the reported coefficient lies at the extremes of the bootstrapped distribution of β's for the two misspecified models.

[16] Note that a lack of model fit need not be problematic in and of itself given a specific research question. For example, an optimal model from a fit perspective does not necessarily equate to the most appropriate model to answer a causal question (Young, 2019). In other words, the decision to unpack the flexible form model depends on the researcher's assessment of whether the lack of fit is potentially problematic and whether or not they feel comfortable defending an apparent lack of fit in the proposed model. However, as the flexible model includes the exact same variables as the originally hypothesized model, clear improvements in model fit will typically reflect improved estimation of the relationship amongst variables which can be expected to impact inference.

[17] Common examples include unnoticed ethnic profiling, or models fit to aspects that do not generalize outside of the sample used (Subrahmanian and Kumar, 2017; Lundberg and Lee, 2017). As a case in point, recent legislation in the European Union now increasingly requires companies that use algorithms to support decision-making to be explainable, the so-called 'Right to Explanation' (Goodman and Flaxman, 2017).

approaches can be identified: global and local explainability. The former attempts to describe the mechanics of a model in general terms, i.e., which variables tend to be important for the model's overall performance. Variable importance measures are an example of this approach (Baćak and Kennedy, 2019; Brand et al., 2021). Local explainability takes a more detailed approach and attempts to explain the drivers of individual predictions made by a model. As global explainability only gives a high-level understanding of which variables are important to model fit rather than an understanding into the functional relationship or direction of associations amongst variables, local explanation methods are more appropriate for the framework.

There are many local explanation methods available, all of which attempt to provide understanding into why a ML model makes the predictions it does. Which method is more appropriate fundamentally depends on the substantive question being asked of the model (Doshi-Velez and Kim, 2017) and, more generally, the definition of interpretability tends to differ across substantive and expert domains (Zhou et al., 2021; Lipton, 2018).[18] For example, most explainability research is focused on providing additional context to practitioners making (high-stakes) decisions based on a ML method in active deployment. In such cases, the ability to quickly calculate and extract information from an explanation could be more relevant than emphasizing fidelity. Conversely, an ethical reviewer of a system-in-action might require more fine-grained and detailed explanations, whereas in low-stakes cases with well-verified performance this may again be different. Within the proposed framework, the onus is on understanding the underlying patterns found by the ML model rather than practical concerns around the deployment of the model. This motivates the use of Shapley values which have a number of unique and attractive properties with extensive theoretical grounding discussed in

[18]The variety of perspectives in explainability have led some to call for a more rigorous reporting of the envisioned goals of generating explanations (Doshi-Velez and Kim, 2017). Following the taxonomy of Doshi-Velez and Kim (2017), explanation in this framework is meant to generate understanding into the unknown patterns in the data, leading to the choice for local explainability at the model level. Temporal efficiency and alignment with operational processes, which sometimes constrains explainability when deployed as a decision support system, is thus irrelevant.

more detail below.[19] Recent advances have made their estimation computationally more feasible and rekindled interest into their use more broadly (Aas et al., 2021; Samek et al., 2019; Heskes et al., 2020). More importantly, they are shown to align closely with human intuitions of model interpretation making them appropriate for the iterative process envisioned in the framework (Doshi-Velez and Kim, 2017; Lundberg and Lee, 2017; Rudin, 2019). Shapley values have been applied in various fields, notably medicine (Lundberg et al., 2020; Tang et al., 2021), but have also found applications in sociological research, for instance as a way to understand complex network dynamics (van der Laan et al., 2022) as well as to resolve path dependencies in decomposition metrics like segregation indices (Elbers, 2021).

Shapley values for model explanation

Shapley values — originally developed within Cooperative Game Theory (Shapley, 1953) — decompose each prediction made by a model into a sum of K Shapley values, one for each variable in the model:

$$\hat{y}_i = \phi_0 + \sum_{k=1}^{M} \phi_k x_{k,i}, \tag{4.4}$$

where $\hat{y}_i$ is a prediction made by a model, and ϕ_0 resembles the overall mean across predictions. The Shapley values thus decompose the deviance of $\hat{y}_i$ with respect to the mean prediction in an additive way, leading to the typical waterfall plots associated with Shapley values illustrated in Figure 4.4. As every prediction is decomposed into K Shapley values, the approach effectively leads to a matrix of $N \times K$ values, one for every variable x_{ki} in the model. Thus, every observed value of x_{ki} has an associated Shapley value ϕ_{ki} reflecting that variable's impact on a specific outcome $\hat{y}_i$. In a linear additive model without an intercept, the Shapley value for variable k would simply be the coefficient β_k times the observed

[19] Another consequence of the wide variety in perspectives on what matters in explanation is that there are few generally accepted quantitative metrics to assess explanation quality and different methodological approaches can provide differing explanations. As explanation becomes more critical to ML research, such metrics and comparative approaches are increasingly being introduced into the field (Krishna et al., 2022). The uniqueness of Shapley values in providing a number of theoretical guarantees makes them especially attractive for generating model understanding as understood in this chapter.

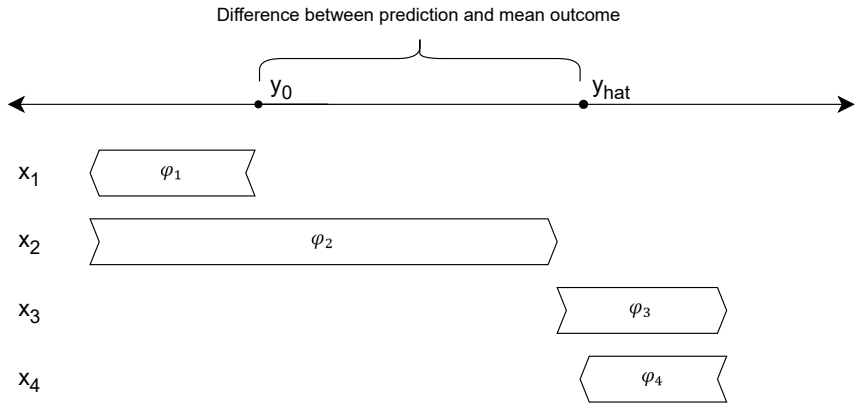


Figure 4.4: Prediction $\hat{y}$ and mean value y_0 based on a model with four explanatory variables. The difference between the prediction and overall mean is decomposed along the Shapley values of each of the four variables.

value x_{ki}: $\phi_{ki} := \beta_k x_{ki}$ (Aas et al., 2021). Plotting pairs of Shapley values and the values of their associated explanatory variable $[\phi_{ki}, x_{ki}]$ thus provides a visual representation of the global functional relationship between x_k and the outcome which can be used to generate understanding of the patterns encoded in the ML model (Lundberg et al., 2020, p. 59).

Calculating Shapley values ϕ_k for variable k given some prediction $\hat{y}_i$ relies on a perturbation-based approach where the impact of omitting information on variable k on the model's prediction is assessed.[20] This is done by defining all possible information sets — which is simply a set of $s \leq K$ variables for which information is available to the model — including variable k, $\mathcal{M}^k$, and its complement where each information set excludes information on variable k, $\mathcal{M}^k/k$. For each of the information sets in $\mathcal{M}^k$, a prediction is made using that set as well as its equivalent in $\mathcal{M}^k/k$ leading to two predictions — one including information on k and one without — which are then differenced.[21] The Shapley value ϕ_k is defined as the weighted

[20] The approach has many resemblances to the *do*-operator popularized by Pearl (2009), as it effectively concerns intervention on the covariate space (Lundberg and Lee, 2017; Heskes et al., 2020).

[21] Consider calculating the Shapley value for variable k in a model $\hat{f}(\cdot)$ which is estimated to four explanatory variables. First, all 2^{K-1} possible information sets containing combinations of the remaining $K-1$ variables have to be considered. When $K = 4$ we consider 8 unique information sets of the three other variables, ranging from including none of the variables — $\mathcal{I} = [0, 0, 0]$ — to including all three — $\mathcal{I} = [1, 1, 1]$ — into the prediction. For each information set $\mathcal{I}$, the difference in prediction is calculated when adding information on the variable k — i.e., the observed x_k — to the prediction: $\bar{f}(\mathcal{I} \cup k) - \bar{f}(\mathcal{I})$ where $\bar{f}(\cdot)$ is the flexible model.

mean over all these differences in predictions for each information set in $\mathcal{M}^k$:

$$\phi_k = \overbrace{\sum_{\mathcal{I} \in \mathcal{M}^k}}^{\text{Sum over all information sets}} \underbrace{\frac{|\mathcal{I}|!(|\mathcal{M}| - |\mathcal{I}| - 1)!}{\mathcal{M}!}}_{\text{Weighting function}} \overbrace{(\bar{f}(\mathcal{I} \cup k) - \bar{f}(\mathcal{I}))}^{\text{Difference in prediction}}. \tag{4.5}$$

This process is then repeated for every observation in the dataset leading to the $N \times k$ matrix of Shapley values.

Shapley values have a number of unique theoretical properties that make them attractive as a method to generate understanding into models. First, the sum of all Shapley values ϕ_k and the mean prediction, ϕ_0, match the actually observed prediction $\hat{y}_i$. Second, whenever the inclusion of variable k into the information set has no impact on the model's prediction — $\bar{f}(\mathcal{I}) = \bar{f}(\mathcal{I} \cup k)$ for all $\mathcal{I}$ — its Shapley value is zero. Third, whenever two features j and k contribute equally to every prediction — $\bar{f}(\mathcal{I} \cup j) - \bar{f}(\mathcal{I}) = \bar{f}(\mathcal{I} \cup k) - \bar{f}(\mathcal{I}) \quad \forall \mathcal{I} \in \mathcal{M}^k$ — their Shapley values are the same: $\phi_j = \phi_k$. Fourth, they are consistent with respect to addition and multiplication across models. Shapley values have been shown to be the only local explanation method to possess these properties (see Shapley (1953); Young (1985); Lundberg and Lee (2017); Lundberg et al. (2020) for detailed discussion and proofs) and as a consequence have been shown to align closely with a human intuition of model understanding (Pearl, 2009; Sloman, 2005; Lundberg and Lee, 2017; Lundberg et al., 2020).

There are a number of computational challenges in calculating Shapley values that have limited their applicability in the past. This computational burden stems from three elements. First, the number of possible information sets is 2^K and scales exponentially with the number of features, quickly becoming intractable. Second, calculating the prediction $\bar{f}(\mathcal{I})$ when the information set $\mathcal{I}$ consists of a subset $\mathcal{S} \subset \mathcal{K}$ requires an expectation $\mathrm{E}[x|x^\star = x_s]$ to be evaluated for the variables $\bar{\mathcal{S}}$ not in $\mathcal{I}^{\mathcal{S}}$. Typically, the empirical density of the missing variables is used but this assumes independence of the feature space. Joint Gaussian and copula-based methods have been proposed as an alternative, although these further increase computation time

(Aas et al., 2021; Heskes et al., 2020). Finally, many predictions have to be explained to assure many tuples $[\phi_k, x_k]$ to infer the underlying functional form, thus further scaling the computational requirements linearly. However, computationally efficient methods have been developed to calculate Shapley values, notably the Kernel SHAP approach due to the seminal paper by Lundberg and Lee (2017) that relies on sampling from the universe of information sets to approximate the exact Shapley values (see Section T.7.1 for a more detailed discussion).

Throughout this chapter, I will use the estimation method optimized for tree-based models that provide a number of benefits to estimation. First, by exploiting the structure of the estimated tree to determine the relevant information sets, estimation becomes efficient to the point that exact Shapley values can be calculated, rather than having to rely on numerical approximations (Lundberg et al., 2020, p. 64-65). Furthermore, this efficiency allows pairwise Shapley values to be calculated in reasonable time, which enables interaction effects to be explicitly evaluated (Lundberg et al., 2020; Lapuschkin et al., 2019). In the standard setup, interactions would be divided amongst the relevant interacting variables in the standard additive decomposition in (4.4).[22] Shapley interaction values assess the joint omission of *two* variables from the model and the subsequent impact on the outcome (see Section T.7.3 for a more detailed discussion). Second, the independence assumption of variables that is implicit when using the empirical density for those variables which are excluded from the information set is relaxed. This is again a result from exploiting the estimated tree when making predictions using an incomplete information set (Section T.7.2).

To illustrate how Shapley values can be used to generate understanding into a ML model, I calculate the Shapley values for the age variable in the GB model identified in the toy example above. This leads to N Shapley values associated with the N values of x_{age}. A joint scatterplot is presented in Figure 4.5. These show the tuples $[\phi_{ki}, x_{ki}]$ for all N observations in the data, where ϕ_{ki} reflects the contribution

[22] Interactions can still be identified from visualizing Shapley values for a variable x_k separately, conditional on values of another variable x_j (as is illustrated in the first empirical example). However, they do not explicate interactions.

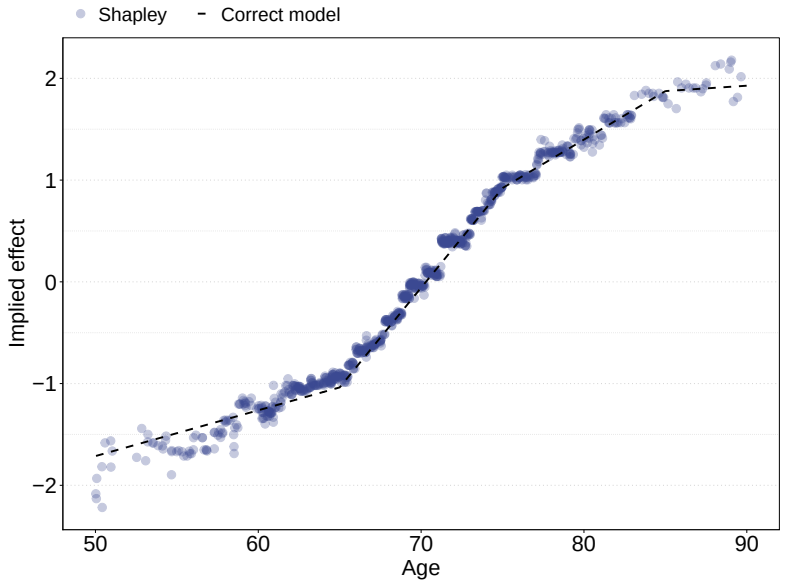


Figure 4.5: Implied effect size of age through Shapley values versus the correctly specified model. Note that both implied effects have been normalized to allow for comparisons with the true functional form.

of variable x_{ki} to the prediction $\hat{y}_i$ relative to ϕ_0. As the underlying model is linear additive, these Shapley values should represent the underlying coefficient β_k times the observed value of the covariate x_{ki} for each observation. Note that I normalize the Shapley values as well as the true underlying effect such that they can be visualized jointly.[23] As can be seen, the Shapley values accurately recover the piecewise linear association in the underlying data, and would provide us with concrete insights into the way in which our hypothesized functional forms could be improved from the linear and quadratic specifications of age. Again, no functional form specification was required to generate these insights.

Resemblances to other computational frameworks

Before applying the framework to three empirical use cases, I briefly discuss a number of methods and frameworks that share a similar intuition and computational appetite to the proposed framework. In spirit, the first two steps of the framework are indebted to modern types of specification tests that rely on comparing residuals from non-parametric regression to those provided by a researcher's own hypothesized

[23] The Shapley value for variable k reflects the impact of x_k into the information set relative to ϕ_0. In linear regression including an intercept, interpretation is relative to the overall mean.

model. A fully non-parametric method can identify a broader range of non-linear functional forms than those for example considered by the RESET test, although formal tests of improved fit can be challenging to compute due to the dependence of local estimates to the data at hand. More generally, the amount of data required to compute local estimates scales exponentially with the number of independent variables (Bishop, 2006; Yatchew, 2003).[24] Although very similar in spirit, the ML methods in the framework retain considerably more structure and are more efficient from a data perspective than non-parametric tests, in addition to better post-estimation analysis of the fitted model.[25]

The proposed framework also shares similar goals to the field of model robustness. The emphasis in model robustness lies on exposing the potential role of researcher degrees of freedom during the model-building process (Sala-i Martin, 1997; Simonsohn et al., 2020; Young and Holsteen, 2017). Similarly to the proposed framework, a number of models slightly different to the hypothesized model, $\tilde{f}(\cdot)$, are estimated to guard against researchers cherrypicking a specification. The key difference in model robustness is that variation in $\tilde{f}(\cdot)$ is typically induced by excluding variables in the model rather than assessing different functional relationships amongst variables (Simonsohn et al., 2020; Young and Holsteen, 2017). In general, model robustness emphasizes inference into a single parameter, e.g. β_k, and its consistency and substantive size across the exclusion of (sets of) variables. The functional complexity as well as the fit of the models is of limited importance (Slez, 2019; Young, 2019). The proposed framework instead focuses on the functional relationship amongst a set number of variables rather than variable

[24]Typically in non-parametric regression a general function $y_i = m_h(x_i)$ is estimated as a simple weighted average of observations close to the observed x_i in euclidean space. Inclusion of data points is generally governed by a bandwidth parameter h defining a region $[x_i - h, x_i + h]$. Different weighting functions or kernels can be used to determine the importance of the data points within the interval, for example by weighing points closer to x_i more than those further away (Yatchew, 2003). As the dimensionality of X increases, the number of data points required increases exponentially, limiting the practical applicability of the approach — the so-called 'curse of dimensionality' (Bishop, 2006, Chapter 1).

[25]Note that some ML methods are similarly affected by the 'curse of dimensionality'. Notably, methods like neural networks require considerable amounts of data to estimate (Bishop, 2006). This is the main reason why tree-based or shrinkage methods are usually applied within the social sciences, as such methods are even applicable for very small N cases (Liao et al., 2017; Athey and Imbens, 2019).

inclusion: the emphasis is on specification. Furthermore, model robustness relies on the exclusion criteria of variables to generate 'plausible' specifications (Young and Holsteen, 2017, p. 3). However, this might include omitting a known and important confounding variable (Slez, 2019). This raises the question whether the various alternative specifications are equally plausible in practice.[26] Importantly, repeated estimation of the researcher-hypothesized model $\tilde{f}(\cdot)$ to samples of the full dataset, which is done in Step II of the framework, automatically generates a large number of $\hat{\beta}_k$ estimates which should already guard against severe p-hacking. Heavily over-engineered estimates of a coefficient $\hat{\beta}_k$ should become visible across repeated estimation on many resampled datasets. Despite their differences, both frameworks are similarly indebted to the growth in computational power to enable researchers to become more critical of modeling assumptions and to improve transparency into the research process (Muñoz and Young, 2018, p. 4).[27]

Closest in spirit to the proposed framework is an emerging literature in behavioral economics and psychology using ML models to find the predictive limit of behavioral theories (Kleinberg et al., 2017; Agrawal et al., 2020; Peterson et al., 2021). The typical approach is to compare the performance of some behavioral theory to predict the outcomes of a stylized experiment — e.g., the fairness perception of two vignettes (Agrawal et al., 2020) — with a flexible ML model trained to the same data. In case the ML method leads to better predictions than a behavioral theory, those cases where the behavioral theory underperformed relative to the ML benchmark are studied. This research focuses solely on large-scale experimental data and has not been extended to empirical work as broadly as this chapter proposes. More generally, the onus in this literature is on theory completeness

[26] For example, if an important confounder is omitted in 50% of the alternative 'reasonable' specifications, we shouldn't be surprised to see high variation in the estimates of the coefficient of interest.

[27] Model robustness has recently been expanded to 'Multiverse' analysis by Young and Stewart (2021) who, in addition to varying the number of variables in the model, further include a number of reasonable functional forms to assess the degree to which different findings can be supported by data. Like model robustness, fit is of secondary importance and the framework relies on an a-priori specified set of reasonable functional forms. However, in spirit it is more closely aligned to the proposed framework, even though it remains fundamentally reliant on the researcher to determine what constitutes 'plausible' specifications, rather than letting computational power dictate the patterns in the data.

rather than correct functional form specification. Post-estimation interpretation of the ML method — the third step in the proposed framework — is mostly ad-hoc, if present. However, the central premise of using ML to find patterns in data not necessarily hypothesized by a researcher is a key resemblance amongst the two approaches. An increasing interest in finding the predictive potential of social theories — and using ML to do so — can be observed more generally throughout the social sciences (Hofman et al., 2017) and was for instance central in the FFC discussed in earlier chapters (Salganik et al., 2020).

A final strand of research similar to the proposed framework is the field of 'autometrics', which also aims to find better fitting models than a researcher's own hypothesizing might yield. The autometrics approach is to include a large number of base transformations of the explanatory variables into a linear additive functional form. This 'stacked' model is then iteratively trimmed to reach an optimal model using in-sample fit metrics and specification tests (Doornik and Hendry, 2015). Autometrics is a more classical approach to flexible model-building compared to the ML methods proposed in this chapter. Specifically, the autometrics setup can be subsumed by the set of ML methods called 'generalized additive models' and spline-based methods which can be folded into Step I of the proposed framework (Hastie and Tibshirani, 1987; Hastie et al., 2009).[28] Related, some advocate a similarly stacked model with interactions but instead apply variable selection through least absolute shrinkage and selection operator (LASSO) regression (Blackwell and Olson, 2021; Beiser-McGrath and Beiser-McGrath, 2020). The proposed framework shares the same broad intuition of the above approaches, but uses a broader set of methods such that a wider range of patterns can be identified. The proposed framework further emphasizes out-of-sample evaluation over in-sample evaluation, reducing the risk of overfitting. An advantage of autometrics is that the flexible model is interpretable by design: the functional form is known, whereas we need to employ X-AI techniques to generate understanding into flexible models from the proposed framework. Note, though, that multiple transformations of a variable

[28]The MARS algorithm takes a similar approach to finding better specified models from a stacked specification by iteratively trimming variables from a full specification (Friedman, 1991).

x_k in a linear additive model can quickly become difficult to interpret as well in which case X-AI methods like Shapley values could also be necessary to generate intuitive understanding into the final model.

4.4 Applying the Framework in Practice

I apply the proposed framework to three case studies. The first is a simulation based on the Mincerian wage equation, a classic field of research investigating the returns of additional years of schooling on a person's income (Lemieux, 2006a). The second is a hedonic regression of house prices applied to a large dataset of transactions in the London retail market (Malpezzi, 2003). The third studies the demographic determinants of voting preferences in the United States using the General Social Survey (GSS) (Davis and Smith, 1991). The Mincerian Wage application is relevant because it provides an example of a functional form which has been actively innovated over the past decades and provides an illustration of how the proposed framework can speed up model-building. The hedonic regression and voting examples are chosen because both enjoy considerable academic interest (Heckman et al., 2018; Lemieux, 2006a; Can, 1992) and typical models include a number of standard control variables with a (novel) interest variable. Since the latter is often correlated with the control variables, a correct functional relationship between outcome and control variables is crucial. For these latter two case studies I will evaluate whether the typical complexity in which control variables feature in the functional form is sufficient, and improve them where necessary.

Each of these use cases is meant to illustrate the potential of the proposed framework to identify more intricate patterns in the data. In the first, both a non-linearity and interaction effect are identified using the framework. In the second, the framework is used to identify considerable non-linearities in a number of explanatory variables, as well as that spatial heterogeneity is poorly controlled for using neighborhood variables. In the third, the potential of pairwise Shapley values is used to identify a wide range of interactions amongst variables. Note that none of the

examples are designed with a specific substantive research question in mind, which might require particular control variables to address specific sources of confounding.

Application I: Mincerian Wage simulation

The Mincerian wage equation is a classic economic tool used to estimate the effect of an additional year of education on an individual's wages — the so-called 'return to education' (Lemieux, 2006a). As mentioned above, this example is appropriate to showcase the framework since the functional form used to estimate the Mincerian wage equation has been innovated over the past decades. In other words: model specification has improved. The original functional form simply related log yearly wages with years of education and work experience:

$$\ln(\text{wages}_i) = \beta_0 + \beta_1 x_{\text{educ},i} + \beta_2 x_{\text{exp},i} + \epsilon_i. \tag{4.6}$$

Subsequently, a square term was added to allow for a level of non-linearity in the effect of work experience:

$$\ln(\text{wages}_i) = \beta_0 + \beta_1 x_{\text{educ},i} + \beta_2 x_{\text{exp},i} + \beta_3 x^2_{\text{exp},i} + \epsilon_i. \tag{4.7}$$

And more recently, a step function was added in the effect of education to allow for different linear effects by level of education:

$$\begin{aligned}\ln(\text{wages}_i) =& \beta_0 + \beta_1 x_{\text{educ_0_8},i} + \beta_2 x_{\text{educ_9_10},i} + \beta_3 x_{\text{educ_11_12},i} + \\ & \beta_4 x_{\text{educ_13_14},i} + \beta_5 x_{\text{educ_15+},i} + \beta_6 x_{\text{exp},i} + \beta_7 x^2_{\text{exp},i} + \epsilon_i.\end{aligned} \tag{4.8}$$

For illustrative purposes, I also study a fourth, hypothetical functional form where each coefficient slightly differs by sex:

$$\begin{aligned}\ln(\text{wages}_i) =& \text{I}(x_{\text{sex},i} = \text{Female})[\beta^*_0 + \beta^*_1 x_{\text{educ_0_8},i} + \beta^*_2 x_{\text{educ_9_10},i} + \beta^*_3 x_{\text{educ_11_12},i} + \\ & \beta^*_4 x_{\text{educ_13_14},i} + \beta^*_5 x_{\text{educ_15+},i} + \beta^*_6 x_{\text{exp},i} + \beta^*_7 x^2_{\text{exp},i} + \epsilon_i] + \\ & \text{I}(x_{\text{sex},i} = \text{Male})[\beta_0 + \beta_1 x_{\text{educ_0_8},i} + \beta_2 x_{\text{educ_9_10},i} + \beta_3 x_{\text{educ_11_12},i} + \\ & \beta_4 x_{\text{educ_13_14},i} + \beta_5 x_{\text{educ_15+},i} + \beta_6 x_{\text{exp},i} + \beta_7 x^2_{\text{exp},i} + \epsilon_i].\end{aligned} \tag{4.9}$$

To illustrate the proposed framework I simulate data using the four specifications above as DGPs, plugging coefficient estimates as found in the recent empirical

Variable	N	Mean	Std. Dev.	Min	Pctl. 25	Pctl. 75	Max
Age	50,000	41.09	13.70	18	29	53	65
Schooling (years)	50,000	13.79	2.97	0	12	16	20
Work experience (years)	50,000	23.01	13.7	0	11	35	51
Sex: Female	50,000	50.1%	-	-	-	-	-
Sex: Male	50,000	49.9%	-	-	-	-	-
Log wages (Linear-I)	50,000	8.29	1.25	5.24	7.238	9.38	10.96
Log wages (Linear-II)	50,000	7.58	0.69	5.20	7.066	8.09	9.17
Log wages (Linear-III)	50,000	6.74	0.64	4.62	6.254	7.30	7.99
Log wages (Linear-IV)	50,000	6.70	0.80	3.73	6.325	7.16	8.56

Table 4.4: Descriptive statistics of the simulated wage data

literature into each specification. I vary the coefficients of sex to fall within a standard error of the coefficients found in the literature to generate the fourth DGP (Section A.4.1) (Lemieux, 2006a; Heckman et al., 2018).

Based on these four DGPs and a synthetic sample of 50,000 individuals' age, years of education, years of work experience and sex I generate four outcomes: one using each DGP. The synthetic sample used is based on the GSS 2018, such that the distribution of the explanatory variables is representative for an actual working population. For illustrative purposes, I calibrate the error term in each functional form such that the proportion of explainable variance is constant in every dataset irrespective of the DGP used to generate the outcome variable. Descriptives for the dataset can be found in Table 4.4. Linear-I refers to the first functional form above provided by (4.6), Linear-II refers to (4.7) and so on.[29]

As a result, there are four datasets consisting of a vector of outcomes y and the same matrix of explanatory variables X. Each outcome vector comes from a different underlying functional relationship of X and y. The first is based on a world where both education and work experience affect the outcome linearly, the second where work experience follows a second degree polynomial and so on. I next propose four hypothetical functional forms, $\tilde{f}(\cdot)$, to estimate to each of the four datasets. The first three functional forms are equal to those defined in (4.6)-(4.8), as well as a fourth functional form which is equal to (4.8) but includes

[29]The same functional forms (4.6)-(4.8) have been used for Figure 2.3, Panel C, in Chapter 2.

a dummy for sex. This means that none of the four hypothesized models aligns with the true DGP in the fourth dataset, which follows the DGP provided in (4.9). I include this fourth form for illustrative purposes, as it is common practice to 'control' for group differences by simply including a dummy variable whereas this might be simplifying the true pattern in the data.

Estimating these four functional forms means that for the first dataset, all four hypothesized models should have the appropriate flexibility to model the data. For the second dataset, only the second, third and fourth functional form should and for the third dataset, only the third and fourth functional forms should be able to fit the data well. None of the proposed functional forms have sufficient flexibility to estimate the underlying DGP in the fourth dataset. In addition to the four hypothesized functional forms, I estimate a Super Learner including various tree-based ML methods (Tables A.4.2 and A.4.3 in Section A.4, selected models in boldface) — the first step of the framework. As can be seen, the GB model performs best across all four datasets. For each dataset, the best performing model is selected as the flexible model (indicated in boldface). Next, when comparing the out-of-sample model fit of the flexible model with the four hypothesized models — Step II in the framework — the flexible model is able to match the true functional form's performance in the first three datasets and strongly outperforms the most flexible functional form in the fourth dataset (Figure 4.6). The framework thus identifies the underlying model without requiring the researcher to hypothesize a functional form in each of the four datasets.

Finally, the flexible model can be unpacked to infer why it improved on the hypothesized functional forms — the third step in the framework. We know the true DGP in this case, making it obvious why the flexible model outperformed the under-specified functional forms, but even if we didn't; the Shapley values easily identify the correct association of the independent variables with the outcome (Figure 4.7). As can be seen, the Shapley values capture the linearity of both explanatory variables in the first dataset, the non-linearity in years of work experience in the second dataset, as well as the step-wise function in the third dataset. What is more, the

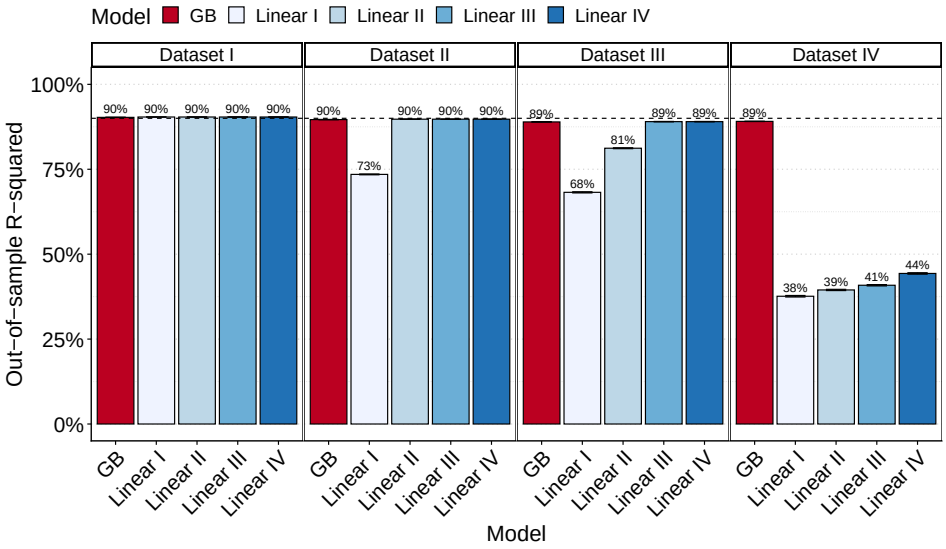


Figure 4.6: Out-of-sample R^2 for the four datasets with varying DGPs, using four functional forms and the GB model. As can be seen, the GB model converges on the correct functional form's model fit in each of the datasets, whereas the pre-specified models only reach the maximum explanatory power when the functional form's complexity equals or outperforms that of the DGP. 100 splits into an estimation set of 80% and evaluation set of 20% are used.

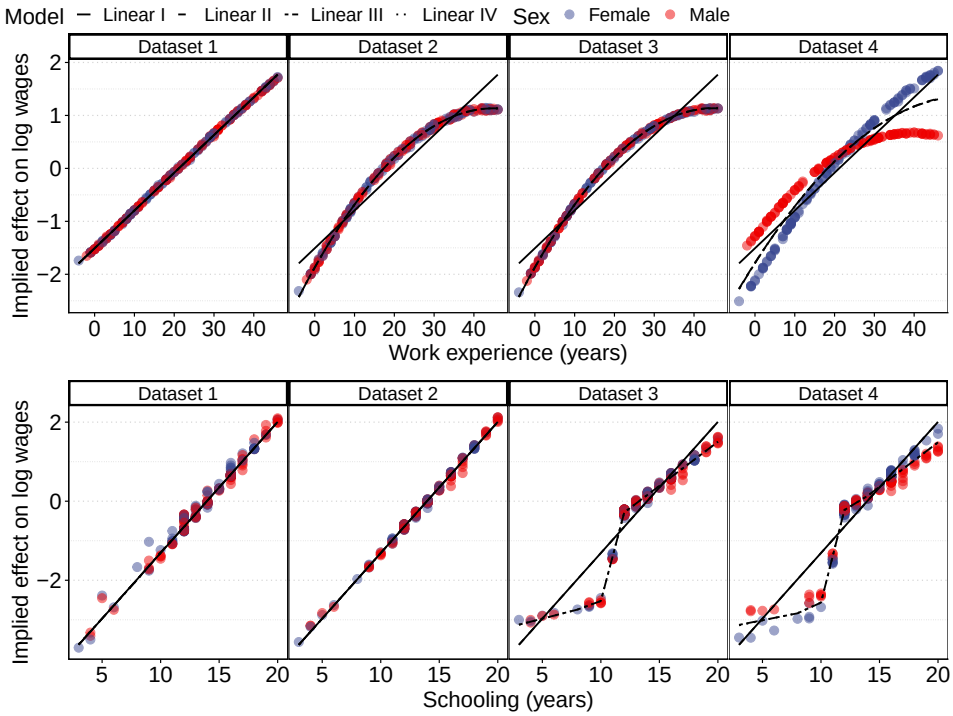


Figure 4.7: Effect of work experience (red) and schooling (blue) as predicted by the four estimated functional forms, and as implied by Shapley values. As can be seen, the correct functional form is approximated well by the Shapley values for each of the datasets and clearly shows an interaction effect between sex and both effects in the fourth dataset. Shapley values and implied effects from the estimated models have been scaled such that they can be easily compared visually.

Shapley values pick up the interaction with sex in the fourth dataset, as is illustrated by varying the colors of the Shapley values by sex. Again, without requiring the researcher to formulate any functional relationship between variables a-priori.

Application II: London house prices

As a second case study, I use a large dataset on transactions in the London housing market. The typical approach to modeling house prices is through a hedonic pricing model that assumes each house to be composed of various traits for which buyers have certain preferences. Buyers effectively combine the value of individual traits to determine the (monetary) value of an entire house. In hedonic regression, interest typically lies in the price elasticity of certain traits, e.g. how much the house price increases with an additional square meter of living space or the inclusion of a garden. In practice, linear additive models are typically estimated relating observed house characteristics with log prices although many authors have suggested that the assumptions of additive linearity that is implicit in this functional form might be unreasonable (Malpezzi, 2003; Fan et al., 2006).

In this application I use typical house characteristics, like the house size, number of rooms, and property type as explanatory variables. I also include a number of neighborhood characteristics. In many applications, researchers attempt to address spatial heterogeneity by including neighborhood-level observables like crime indices, travel distances to local centres or deprivation scores. Again, in the typical model these variables are added in a linear additive framework although it is often argued that spatial heterogeneity is considerably more complex (Elhorst, 2010). To estimate the hedonic regression, I use a dataset of nearly 630,000 house sales containing basic house characteristics and a neighborhood identifier. The transaction data were collected and merged by the Reshare project hosted at the UK Data Service (Chi et al., 2021). I include neighborhood level data from the Department of Transport and Communities and Ministry of Housing, Communities & Local Government.[30]

[30] I include these indices at the Lower Super Output Area level, for which the average population in London is around 1700.

Variable	N	Mean	Std. Dev.	Min	Pctl. 25	Pctl. 75	Max
Log House Price	629,669	12.92	0.63	9.21	12.49	13.25	17.44
House size (square meter)	629,669	88.89	46.45	10	60	104	959.74
Number of rooms	629,669	4.17	1.71	1	3	5	20
Distance to local centre (minutes)	629,669	11.52	4.79	5	8.87	13.28	60.74
Crime and disorder index	629,669	0.35	0.57	-2.01	-0.05	0.75	2.17
Deprivation index	629,669	21.72	11.85	1.7	12.17	29.75	66.21
Property type: Detached	629,669	0.06	-	-	-	-	-
Property type: Flats/Maisonettes	629,669	0.44	-	-	-	-	-
Property type: Semi-Detached	629,669	0.18	-	-	-	-	-
Property type: Terraced	629,669	0.33	-	-	-	-	-
Transaction type: Marketed sale	629,669	0.76	-	-	-	-	-
Transaction type: Private rental	629,669	0.17	-	-	-	-	-
Transaction type: Non-marketed sale	629,669	0.01	-	-	-	-	-
Transaction type: Other	629,669	0.05	-	-	-	-	-
New property	629,669	0.01	-	-	-	-	-
Year	629,669	2015	2.40	2011	2013	2017	2019
Month	629,669	6.61	3.36	1	4	9	12

Table 4.5: Descriptive statistics of the London house price data

I also include the year and month of the transaction. Descriptive statistics of the dataset can be found in Table 4.5.

As the hypothesized model, I estimate a typical linear additive functional form relating log house prices to the variables depicted in Table 4.5. As is customary, I assume linear trends for all continuous variables, as well as for the temporal variables. The exact functional form is as follows:

$$\begin{aligned}\ln(\text{House price}_i) =& \beta_0 + \beta_1 x_{\text{area},i} + \beta_2 x_{\text{rooms},i} + \beta_{3,\ldots,5} x_{\text{property type},i} + \beta_6 x_{\text{new},i} + \\ & \beta_7 x_{\text{travel_time},i} + \beta_8 x_{\text{crime},i} + \beta_9 x_{\text{deprivation},i} + \beta_{10} x_{\text{year},i} + \\ & \beta_{11} x_{\text{month},i} + \beta_{12,\ldots,16} x_{\text{transaction_type},i} + \epsilon_i.\end{aligned} \tag{4.10}$$

Following the framework, I start by estimating both the hypothesized model and a Super Learner to the data. The best fitting model is again a GB approach, which will be used as the flexible model (see Table A.4.4 in Section A.4, selected model in boldface). In the next step, the flexible model is compared in terms of its out-of-sample fit with the hypothesized model. The results are summarized in Figure 4.8 and are striking. Already amongst the simplest housing characteristics like size and property type the ML model improves on the linear additive functional form by about 5.5 pp in terms of the OOS R^2. However, the largest improvement by far is observed when adding neighborhood variables like 'travel time to hub'

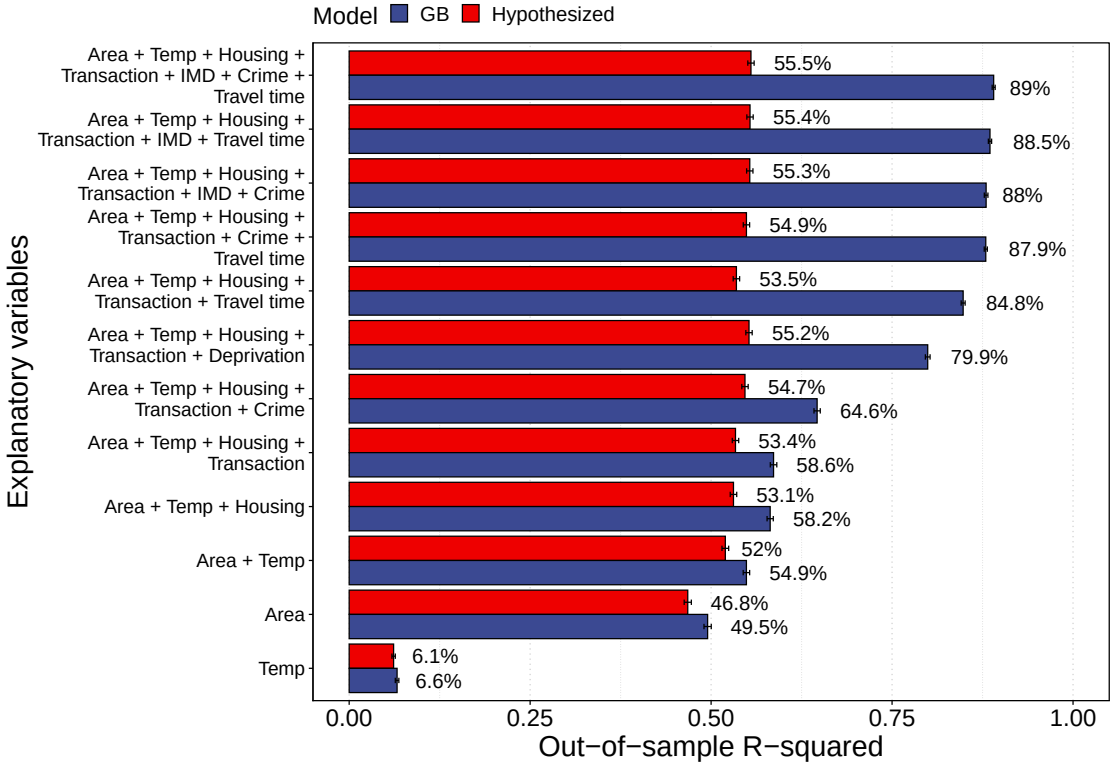


Figure 4.8: Out-of-sample R^2 for the hypothesized model and the flexible model. 'Temp' contains yearly and monthly time trends. 'Area' contains the house size and number of rooms. 'Housing' includes the property type, and whether the house was newly built. 'Transaction' reflects the type of transaction. 'Crime', 'Deprivation', and 'Travel time' reflect the crime index, the deprivation index, and the travel time to a local hub of the neighborhood, respectively. 100 splits into an estimation set of 80% and evaluation set of 20% are used.

to the model, with the difference in R^2 more than 30 pp. This strongly implies that spatial heterogeneity is poorly addressed by the functional form in (4.10), a point to which I will return below.

To evaluate why the flexible model outperforms the linear additive framework, I calculate Shapley values for the flexible model next. These are visualized for six explanatory variables: the size of the house, the travel time to a local hub, the crime index, number of rooms in the house, and the two temporal variables (year and month of sale). The results can be found in Figure 4.9. As can be seen, non-linearities are observed for most of these variables ranging from a slightly decreasing elasticity for the size of the house to piece-wise linearity in the effect for travel time and the crime index. It is also clear that the number of rooms, year, and month variables should all be modelled in a non-linear way. When including

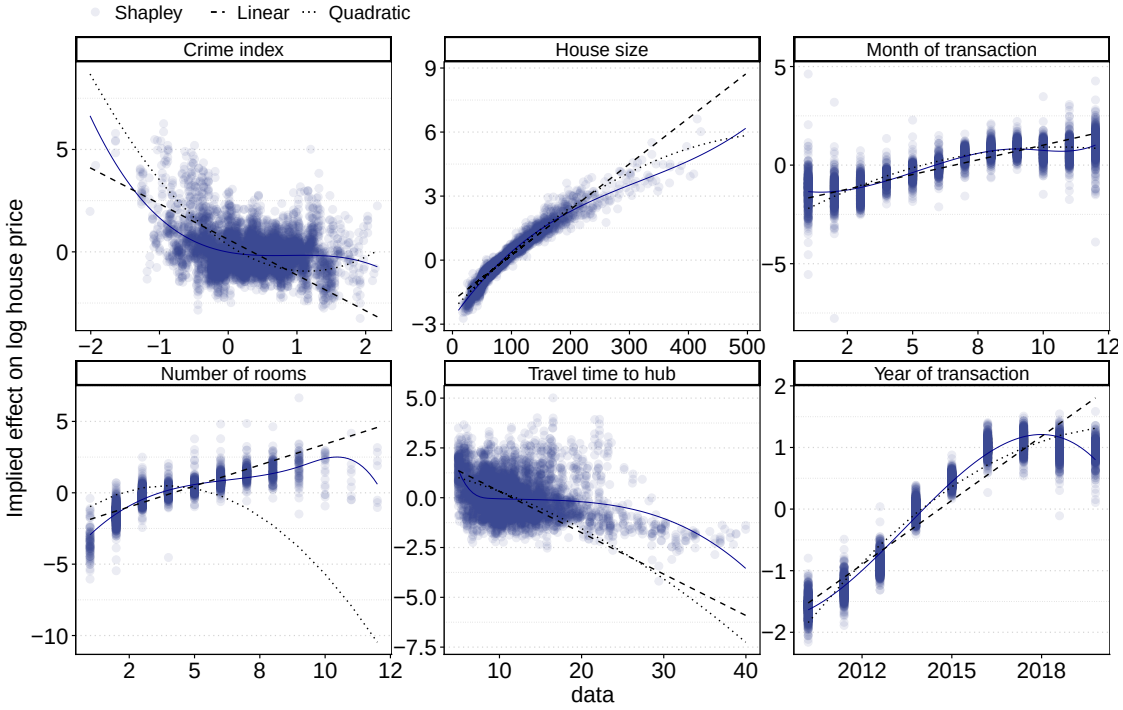


Figure 4.9: Plots show implied effect from a typical linear additive model (dashed line), when adding a squared term (dotted) and as implied by Shapley values (scatter). A LOESS fit is included into the Shapley plots to indicate the association between each variable and the outcome. Implied effects are scaled to allow for visual comparisons.

these non-linearities into the linear additive model, the fit improves and is strictly preferred according to a LR test[31] although a remaining gap in model fit still points at further interactions and non-linearities amongst variables, possibly across time. Specifically, the considerable noise in the Shapley values of the neighborhood-level variables implies that these variables do not follow a very precise pattern, and including non-linearities may not suffice to capture the underlying patterns well.

Given the stark increase in model fit when adding neighborhood-level variables, I include random intercepts on the neighborhood level to a simple model including the house size and temporal explanatory variables. This effectively provides a fully non-linear association for each neighborhood's characteristics and the outcome variable, which seems reasonable based on the large increases in model fit when adding any of the neighborhood characteristics to the ML model, combined with

[31] Adding squared versions of the neighborhood indicators, area, number of rooms and temporal variables improves the adjusted R^2 by 2.5 pp. A LR test accordingly prefers the updated model ($P < 0.001$).

the variation in their Shapley values. Based on this modification of the functional form, the out-of-sample R^2 of the model improves to 84%, which is much closer to that of the ML model. In other words, the heterogeneity amongst neighborhoods could not be captured by the three variables when included in the model in either a linear or polynomial form — although this strategy is often encountered in the literature (Malpezzi, 2003) — and requires more intricate modeling. In this case, it seems like the ML model's ability to estimate flexible patterns has led to the identification of individual neighborhoods thus mimicking a random intercept approach. The framework thus illustrates that spatial heterogeneity is highly predictive, but poorly accounted for by a functional association of the available neighborhood characteristics (Elhorst, 2010).

Application III: Party identification in the United States

As a third and final case study, I evaluate the demographic determinants of party identification in the United States using the GSS (Davis and Smith, 1991). Party identification has been of substantive interest in the social sciences (Freeden et al., 2013), and the GSS is an often used resource for this purpose (Joslyn et al., 2017; Hooghe and Oser, 2017). Throughout this literature, a number of demographic variables are used as typical controls. Usual suspects include the respondent's age, sex, race, educational attainment and income. Typically, another variable of substantive interest is included into the analysis, e.g. cognitive ability (Meisenberg, 2015) or social class (Morgan and Lee, 2017). These interest variables naturally tend to correlate with the control variables, making correct specification critical. The GSS also encompasses many waves and spans multiple decades. Temporal changes might thus be relevant as well.

Table 4.6 shows descriptive statistics of the GSS containing information on voting preferences and demographic characteristics between 1974 and 2018. The outcome of interest is a 7-point scale indicating whether the respondent identifies strongly with the Democratic party (value of 1) or the Republican party (value of 7). The GSS also includes the respondent's age, years of schooling, income level across

Table 4.6: Descriptive statistics of the GSS

Variable	N	Mean	Std. Dev.	Min	Pctl. 25	Pctl. 75	Max
Party Identification	26,011	3.67	1.99	1	2	6	7
Age	26,011	45.31	17.14	18	31	58	89
Schooling (Years)	26,011	13.13	3.00	0	12	15	20
Income	26,011	10.19	2.74	1	9	12	12
Sex: Female	26,011	0.56	-	-	-	-	-
Sex: Male	26,011	0.44	-	-	-	-	-
Race: Black	26,011	0.14	-	-	-	-	-
Race: Other	26,011	0.05	-	-	-	-	-
Race: White	26,011	0.81	-	-	-	-	-
Survey Year	26,011	1996.00	13.46	1974	1987	2010	2018

12 brackets, sex, and race. In addition, the year of the survey wave is available. As hypothesized model, I use a standard linear additive model as often encountered in the literature (Freeden et al., 2013). Most functional forms include a linear time trend for the year of the survey, and add most demographic variables as a linear determinant or dummy. The hypothesized functional form is as follows:

$$\begin{aligned} y = &\beta_0 + \beta_1 x_{\text{age}} + \beta_2 x_{\text{female}} + \beta_3 x_{\text{black}} + \beta_4 x_{\text{other}} + \\ &\beta_7 x_{\text{education}} + \beta_8 x_{\text{income}} + \beta_9 x_{\text{survey_year}} + \epsilon_i \end{aligned} \tag{4.11}$$

and resembles that found in e.g. Meisenberg (2015).

I start by estimating (4.11) as well as a Super Learner containing various RF and GB models. The best performing model is again the GB model, although the RF also performs well (see Table A.4.5 in Section A.4, selected model in boldface). I use the GB approach as the flexible model. The out-of-sample R^2 when including different sets of explanatory variables to the hypothesized and flexible model is visualized in Figure 4.10 and show that the explanatory power of the flexible model is almost double that of the hypothesized model indicating a considerable lack of appropriate specification. The flexible model improves most considerably when race is added to the model, although the hypothesized model already underperforms when simply including temporal, age, and sex variables. Note also that overall fit is comparatively low.

To understand why the GB model outperforms the linear additive model, I illustrate both Shapley values and Shapley interaction values for the age, schooling,

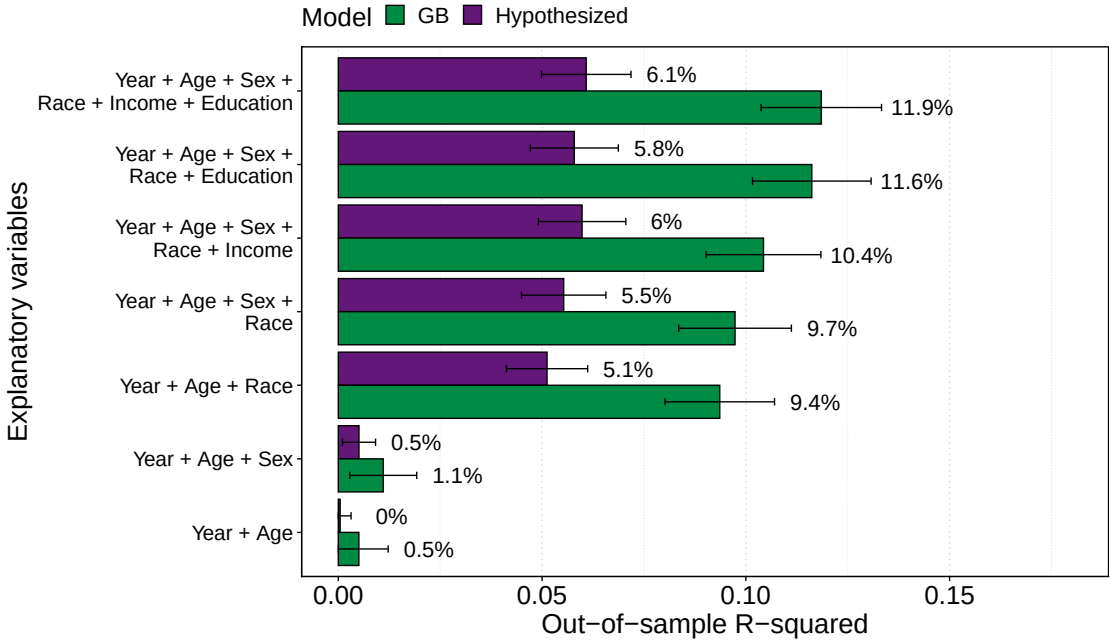


Figure 4.10: Out-of-sample R^2 for the hypothesized model as compared to the flexible model. The 'Year' set reflects the survey wave. The 'Age' set reflects the respondent's age, 'Sex' their sex, 'Race' their race, 'Income' their income, and 'Education' their years of education. The optimal model from the Super Learner routine has been estimated to 80% of the data, and the OOS R^2 is calculated on the remaining 20%. The data has been split 1,000 times into estimation- and evaluation-sets.

and income variables in Figure 4.11. These Shapley values show a number of interesting dynamics. First of all, there are clear non-linearities in the associations of age, income, and years of schooling with the outcome. Assuming these associations to be linear does not fit the implied effect well, and adding square terms improves the model fit considerably (Figure 4.11, panel A).[32]

These direct effects can be further decomposed into a number of interactions by estimating pairwise Shapley values (Section T.7.3). These interactions allow Shapley values for a single variable to be decomposed into a direct and indirect effect. The indirect effects indicate how Shapley values for certain observed characteristics deviate from the overall effect of the variable conditional on the height of the second variable. The estimates in panels B-E all show indirect effects and illustrate that younger respondents associated less strongly with the Democrats in earlier waves

[32] Adding square terms to all three variables and separate intercepts for each survey year improves the R^2 from 6.1% to 7.0%. Unsurprisingly, a LR test comparing both models prefers the updated model ($P < 0.001$).

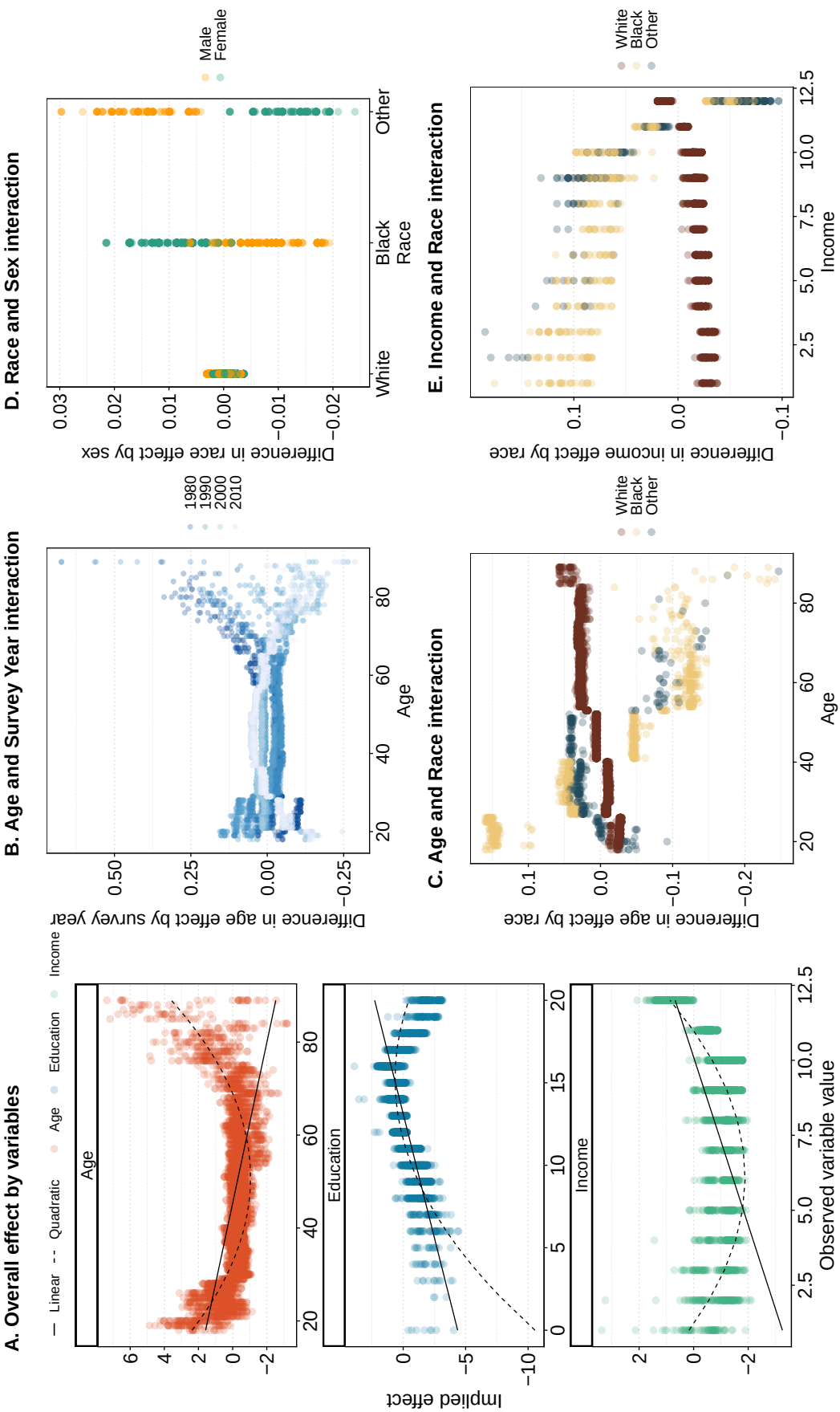


Figure 4.11: Overall implied effect as estimated by Shapley values for the age, education, and income variables including effects as implied by a linear (solid) and quadratic (dashed) functional form (Panel A). Differences as implied by pairwise Shapley values in the effect per age relative to the overall effect by survey year (Panel B), and by race (Panel C). Differences as implied by pairwise Shapley values in the effect per race for females and males (Panel D). Differences as implied by pairwise Shapley values in the effect per race for income (Panel E).

than in later waves (Figure 4.11, Panel B), but also that Whites have a stronger positive age effect than Non-whites (Figure 4.11, Panel C). Similarly interesting dynamics can be observed when interacting sex and race, which show that the sex effect is more pronounced for Blacks but is less pronounced for Others (Figure 4.11, Panel D). The effect of income shows similarly complicated dynamics where higher income is associated more strongly with a Republican identification for Whites than for Non-whites (Figure 4.11, Panel E). Adding race interactions for sex, income, and age as implied by the Shapley values improves the fit of the hypothesized model using conventional in-sample fit statistics.[33]

4.5 Discussion & Concluding Remarks

This chapter set out to incorporate ML methods into the empirical workflow of sociologists to address a key problem: that we often don't have a clear indication what the appropriate functional form might be to model a dataset. Coupled with a historic preference for parsimonious and easy-to-estimate models, this has lead to the prevalence of relatively simplistic functional forms which run the serious risk of lacking the appropriate flexibility to study the social world. In addition, this lack of emphasis on developing specifications that best fit the data prolongs the risk of researcher degrees of freedom muddying empirical findings. That many functional forms are often overly simplistic is well-known, and criticism of the linear additive framework has been consistent, but remain without a clear pathway to improve. Misspecification tests are known to be underpowered in multivariate settings and typically limited to specific types of misspecification. They also lack guidance on how to improve misspecification going forward. Perhaps most importantly, their application in empirical work varies strongly and can be prone to cherrypicking.

I propose a framework which emphasizes the use of ML methods to generate a data-driven estimate of the fit potential in a dataset. This fit potential indicates

[33] Adding age by race and sex by race interactions further improves the adjusted R^2 to 7.2%. A LR test prefers the updated model again ($P < 0.001$). Finally, adding income by race improves the adjusted R^2 by another 0.1 pp with a LR test also preferring the added flexibility to the functional form ($P < 0.001$).

how well an outcome could be modelled when the functional relationship amongst variables is dictated by the data rather than a researcher's hypothesizing. Such an estimate provides an indication whether a researcher's own functional form might miss important nuances like interactions or non-linearities. Crucially, the fit potential is a feature of the data and researcher degrees of freedom do not play a role in its estimation. Furthermore, the framework incorporates valuable ML concepts like out-of-sample model evaluation and cross-validation to guard against over-fitting and, as a bonus, provides bootstrapped confidence intervals into the coefficient estimates of our own hypothesized model which further guards against p-hacking. Most importantly, in case a ML method finds more intricate patterns in the data than our own models do, we can unpack the former to provide guidance on how to improve the latter. This is contrary to popular belief that ML models are fundamentally black boxes which cannot convey intuition into the patterns they identify.

The approach presented in this chapter is not a panacea, and has a number of limitations. First, uncovering intricate patterns from the data will still depend on the available data and might provide limited guidance in low N settings, although many of the methods proposed here can be applied to datasets typically encountered by sociologists. Second, although considerable progress has been made in recent years, unpacking flexible ML methods remains challenging. This is in part because patterns can become complex to the degree that even local explanation methods like Shapley values will not yield easily digestible insights into the underlying functional form, especially when multiple interactions may be at play. More generally, active debate regarding the practical as well as philosophical aspects of explanation methods' accuracy to reflect how ML methods operate remains ongoing. However, the constant developments in the X-AI field are both reassuring and exciting. Finally, when causal questions rather than optimal fit are of interest, there is limited guidance whether a difference in model fit between a hypothesized model and ML alternative is actually problematic. That having been said, the framework is not a blind search for fit across a host of variables but follows the same inferential logic as a researcher's own model. Therefore, any improvement in fit should reflect missed functional

relationships amongst variables that have been purposefully curated by a researcher and their functional relationship should thus be expected to affect inference.

Already in its current form, I have shown how the benchmarking framework can be fruitfully applied in practice. Using the canonical example of the Mincerian wage equation, I show how benchmarking could have sped up the historic process of model-development considerably. The framework effortlessly identified underspecification, and subsequent analysis of the ML models identified the necessary improvements to the functional form. The other two empirical examples, a hedonic regression of house prices and a model explaining party identification, both illustrate how often encountered modeling strategies lead to considerably lower fit than a flexible ML model. In the case of house prices in London, a number of simple non-linearities were identified. Most importantly, the fundamental inability to address spatial heterogeneity through the inclusion of standard neighborhood characteristics became clear through the immense differences in fit between the hypothesized and ML method when including neighborhood data. For party identification, a lack of complexity was similarly evident from applying the framework, and unpacking the flexible models showed important interactions and non-linearities between the explanatory variables and outcome, which are not commonly implemented in the functional forms used to study party identification.

At the root of almost every empirical sociological finding lies a functional form which is assumed to be correctly specified. There is limited evaluation whether this functional form is appropriate for the data, leading to a number of serious risks. The model might not accurately reflect the patterns in the data, affecting the validity of statistical inference. Simple reporting of a single or curated number of functional forms further exposes sociology to p-hacking. These practices stem from a state of the world where limitations on computational power necessitated parsimony. These constraints are no longer applicable, yet the models we estimate retain a simplicity that belies the intricacies of the social mechanisms we are interested in. This is evidenced in the work of our qualitative colleagues, as well as the ever increasing number of empirical examples where ML models outperform the linear additive

models usually estimated by sociologists. This chapter proposes a framework to address this issue by exploiting the benefits of ML methods — to find intricate patterns in data — to address this key issue throughout empirical work. A symbiosis of quantitative sociological work and the computational riches of this day and age that is long overdue, and will only bear more fruit as the goals of the ML community increasingly align with the explanatory focus of sociologists.

Chapter 5

Using Tree-based Machine Learning and Registry Data to Study the Equitability of the Dutch Decentralization of the Social Domain

Chapter abstract: Decentralization is frequently applied in social care systems, but key questions regarding its merits and risks remain. Chief among them is the impact on equitability when responsibility is transferred from central to local entities. Although equity is important, existing research makes use of aggregate measures that do not provide insights into how an intervention might have affected certain groups in the population differently. Furthermore, it is often unclear a-priori along what groups heterogeneity might manifest itself. The problem of not knowing the underlying patterns in the data can be approached using methods from the machine learning (ML) domain, as has been the central topic of the previous chapter. I show how ML allows for robust identification of demographic groups reacting differently to a policy intervention, without the need to specify these groups a-priori. To illustrate, I study the decentralization of the social domain in The Netherlands, using rich administrative data. By applying causal trees and forests, I find considerable heterogeneity in the policy's effect on social assistance use, both across geographic regions as well as demographic groups.

5.1 Introduction

Territorial decentralization of care — the process of delegating the responsibility of some care provision to the regional rather than national level — has been an important lever with which policymakers have experimented over the past decades. In particular within the context of rising costs of care for aging populations (Rechel

et al., 2013), decentralization is often touted as a panacea that not only improves care provision, but supposedly does so at lower costs (Sumah et al., 2016; Lobao et al., 2009; Pike et al., 2012). As a consequence, decentralization can be observed to varying extents in care systems across the world (Kazepov, 2010). However, lively debates on the merits of (de)centralization remain (Crook and Sverrisson, 2003; Mosca, 2006). Central amongst these debates is the question what the impact of decentralization is on the equitability of care. In particular whether decentralization affects access and use of care differently for certain demographic groups in the population, and as a consequence might exacerbate inequality in the system.[1]

The central problem this chapter addresses is that empirical assessment of the impact of decentralization — or any large-scale policy intervention — on equity is challenging. Particularly when there aren't strong a-priori hypotheses as to which groups in the population might be differently affected by a policy change. The identification of such groups becomes problematic from both a data as well as a methodological perspective. As a consequence, existing empirical work relies on aggregate measures of equity like the Gini coefficient, which reflect the distribution of care use across a secondary dimension, for example income (Van Doorslaer et al., 2006). Evaluating changes in such a measure before and after a decentralization can shine some light on aggregate equitability, but will hide more subtle nuances in how a policy change might have affected more specific groups in the underlying population. However, aggregate measures often lack the ability to accurately reflect what happened and to whom as a consequence of a policy intervention and cannot provide policymakers with the required information to evaluate the policy's complete implications on equity.

In an ideal world, policymakers would know how a policy intervention like decentralization affected i) the overall use of some care provision — typically not included in aggregate measures — as well as ii) whether this effect differed for groups in the population — a necessary condition for any change in overall

[1]Throughout this chapter, the terms 'access' and 'use' are used interchangeably to refer to the combined process of an individual becoming i) conscious of their needs and the possibility to obtain care, and ii) to actually obtain this care through the care system in place.

equity. Such an account of the overall impact of a policy, as well as groups for which this effect differs, *would* provide policymakers with the information they need to appropriately assess the impact of a policy in terms of equity and guide necessary changes. To reach this goal, a simple Heterogeneous Policy Effect (HPE) approach is proposed where some overall policy effect τ is estimated which is then allowed to differ as a function of observables: $\tau_i = \tau(x_i)$. For example, allowing the effect of the policy to be different for high income groups relative to low income groups: $\tau(\text{low_income}) \neq \tau(\text{high_income})$. Obviously, these effects could always be aggregated back up to aggregate measures of equity.

There are two reasons why a HPE approach is not typically implemented in empirical work studying the impact of decentralization on equity. The first is that it requires individual and temporal data of considerable size. Traditionally, such an analysis would require the deployment of large-scale surveys consisting of multiple waves which would be costly and often still lack the size to explore unexpected sources of heterogeneity not incorporated in the survey design a-priori. The second reason is that we often do not know how $\tau(x_i)$ should be operationalized. Specifically, for what demographic groups we should allow heterogeneity. It can quickly become intractable to assess all possible groups in a dataset as these scale exponentially with the number of observables, leading to risks of spurious findings (Athey and Imbens, 2016) and p-hacking (Gelman and Loken, 2013). In particular for a policy intervention like decentralization, dimensionality is problematic as each region might exhibit its own heterogeneous patterns. Effectively, the problem is that we do not know the appropriate functional form underlying the data; the same problem as has been the central topic of the previous chapter.

Fortunately, two recent developments have made the implementation of a HPE approach to policy interventions increasingly realistic. First, the rise in large-scale temporal data in the form of administrative records that are increasingly available for research can address the considerable data demands of a HPE perspective. Second, and more importantly, methods have been developed within the ML domain that can help identify heterogeneous groups in terms of effect sizes in a robust and data-driven

manner. Especially relevant are causal trees and causal forests (Athey and Imbens, 2016; Wager and Athey, 2018), both of which exploit the spectacular increases in data and computational resources that are revolutionizing the social sciences more broadly (Salganik, 2019; Rahal et al., 2022; Watts, 2017). In this chapter, I showcase how the combination of administrative data and ML methods can be used to describe a policy intervention in a detailed way that provides policymakers with the insights they need to evaluate a policy in a more complete manner than existing approaches.

As a substantive case study, I assess the 2015 decentralization of social assistance in The Netherlands. Social assistance concerns various types of services provided to people in need, like help with the groceries and other daily chores or transport assistance. Access to social assistance used to be organized centrally. However, to improve alignment between local needs and providers and in order to cut bureaucratic overhead, it was decided to delegate the responsibility of administering social assistance to the municipality level (Vermeulen, 2015, Chapter 5). Effectively, this means that each municipality now determines themselves when to provide citizens with social assistance, a policy shift dubbed the 'decentralization of the social domain' (Vonk and Schoukens, 2019, p. 152). Many criticized the supposed cost efficiencies which a shift to decentralization should bring to be based on wishful thinking (Allers, 2013; Maarse and Jeurissen, 2016), as well as labeling the effective delegation of responsibility away from the central government as undemocratic (Duyvendak and Tonkens, 2017). Most problematically, though, critics predicted that the policy change could lead to heterogeneity between regions in how access to social assistance was mediated and could exacerbate inequalities in access (Vonk and Tollenaar, 2012). Studying the validity of this latter point of critique is the key focus of this chapter.

To inform ongoing debate on the Dutch decentralization and showcase the HPE approach using methods from the ML domain, I use causal trees and forests to empirically evaluate the following two research questions:

RQ1: Has social assistance use expanded or contracted as a consequence of the 2015 decentralization of the social domain in The Netherlands?

RQ2: Has the overall effect on social assistance use been different for demographic groups and / or regions due to the 2015 decentralization of the social domain in The Netherlands?

Substantively, I aim to inform the discussion of decentralization in The Netherlands. More generally, I aim to illustrate how the HPE approach and the use of ML methods in particular can be used to effectively monitor policy interventions going forward. This is particularly relevant as similar substantive questions like the ones asked above are being asked regarding decentralizations across the world (Jiménez-Rubio et al., 2008; Costa-Font and Gil, 2009; Mosca, 2006; Sumah et al., 2016). This chapter can hopefully provide a blueprint to answering them in a robust manner, making use of the computational riches available to researchers today.

The rest of this chapter is structured as follows. In Section 5.2, I provide some background into the decentralization debate and a discussion of existing empirical research into the impact of decentralization on equity. In Section 5.3, I introduce tree-based methods from the ML domain which allow heterogeneous effects to be estimated. Then, I introduce the Dutch case study in Section 5.4. The data and research design is presented in Section 5.5 and the results in Section 5.6. Finally, in Section 5.7, I summarize this chapter's substantive findings and discuss the potential of the HPE approach, and ML methods in particular, within policy evaluation more generally.

5.2 Background

Many countries have enacted laws holding the state responsible for the provision of social welfare to its citizens. In its most narrow sense, social welfare is thought of as 'providing a minimum income to all members of society (universal) or to select groups (categorical) such as the elderly' (Bahle et al., 2010, p. 1). In practice, social welfare has become much broader and escapes a clear functional definition, often also including further needs like education and healthcare, as well as specific ones like home care or other forms of social assistance (Vonk and Schoukens, 2019). This

has led to what is now known as the modern welfare state which can be observed to different extents across the world. The scope, ambition and implementation of social welfare has diverged considerably over time, with some countries still relying on kinship ties to cover certain needs whereas others have expanded the central government's responsibility in providing care further (Leitner, 2003).

The start of the modern welfare state can be traced to the latter parts of the 19th century.[2] Prior to the state-governed social welfare that emerged in the late 19th century, caring for the needy had predominantly been a local affair (Kuhnle and Sander, 2010). For example, the *Landrecht* enacted in 1794 in Prussia delegated care for the poor to local communities, and many countries relied on local religious chapters to take care of the poor (Dorwart, 2013). This meant that the developments of the late 19th century constituted both a formalization as well as a centralization of social welfare. Interestingly, a reverse trend can be observed from the late 20th century onwards as states have started to delegate responsibility back to local entities (Kazepov, 2010; Andreotti et al., 2012).[3] This is especially the case for in-kind services like social assistance — e.g., job programs or help in the household — whereas welfare in the form of monetary assistance has remained mostly centralized. The process of decentralization typically entails delegating the provision of some welfare service from a central function — often a national government — to a smaller entity.[4] In this process, different functional elements of the welfare provision might be decentralized. Typically, the administrative function is delegated first with the financial, legal and in particular the normative terms still being set centrally (Vonk and Schoukens, 2019).

[2]At the time, a commonly shared view emerged that there should be some form of social security for all, and this was a matter to be handled by the state (Flora and Heidenheimer, 1981; Kuhnle and Sander, 2010). Prior to this period, views towards the poor differed considerably. For example, the Elizabethan *Act for the Relief of the Poor* enacted in 1601 already aimed to provide an institutionalized form of relief to individuals in need. At the same time, in 19th century France 'the feeling was still that the poor had to be threatened with the possibility of starvation to be kept industrious' (Rimlinger, 1971, p. 46).

[3]Note also that some formerly decentralized healthcare systems have started to centralize again (Saltman, 2008).

[4]There is a distinction between functional decentralization, which delegates to a specific organization that still operates nationally, and territorial decentralization, which delegates to a sub-national region. Throughout this chapter, decentralization exclusively refers to territorial decentralization.

The key rationale behind decentralization is that it should lead to a better provision of social services whilst simultaneously reducing costs. However, many of the key arguments in favour of decentralization remain untested. The first argument assumes that local regions should be better able to identify the needs of individuals, and provide appropriate solutions to these needs (Hayek, 1945; Musgrave, 1959; Ostrom et al., 1994). This so-called 'customization' argument has been the key driver for the many decentralizations in Europe, including The Netherlands (Vonk and Schoukens, 2019, p. 158). Secondly, as the scope of welfare has gradually expanded over time, it has been argued that bureaucratic overhead has increased excessively as a consequence. It is argued that decentralization should reduce red tape and therefore lead to a more cost-efficient provision of welfare (Tiebout, 1956; Oates et al., 1972; Sumah et al., 2016; Maarse and Jeurissen, 2016). These theoretical arguments have been shown to depend at least in part on other structural features of the system in which decentralization is implemented, and theoretical counter-arguments like economies of scale could be raised against decentralization's efficiency argument as well (Sumah et al., 2016).

At the same time, there are various sources of criticism against decentralization. For example, decentralization effectively reduces the central government's responsibilities by transferring them to local policymakers. Decentralization thus undermines the broader responsibility and democratic principles through which social welfare is typically governed (Blair, 2000).[5] Most critical, however, is the risk that localization of welfare provisions might lead to inequities in access and quality of care dependent on one's location. Individuals might strategically migrate to regions with higher quality provision, leading to differential pressure on the care system across regions which can be especially problematic when resources are unequally distributed across these regions. In addition, there is the considerable risk of arbitrariness in how access to welfare is decided when there is a lack of central oversight. The latter includes problems like (unconscious) discrimination of marginalized groups, but

[5]Note that at the same time others have argued that localization in fact empowers local citizens (Blair, 2000).

also conscious differences due to capture of the administrative process by pressure groups. Differences may also arise simply due to drift over time.[6]

The risk of inequity in access should not be underestimated, especially as barriers to access can be broadly observed even in settings without decentralization (Van Groenou et al., 2006). Research consistently shows that demographic groups differ in their use of social assistance, even when controlling for objective need (Floridi et al., 2021). This inequity can stem from differences in the ability of a central system to reach relevant social groups, in combination with differences in the ability of these groups to navigate the administrative hurdles of a care system and other knowledge asymmetries. The combination of different theoretical mechanisms that could be at work, the structural features of a care system, and the possibility of baseline differences in access make it highly complex to formulate any theoretical predictions about the effect of a decentralization. This means that careful empirical monitoring is crucial to ensuring the equity of decentralization.

Standard approaches to studying equity

Existing research on the impact of decentralization on equity overwhelmingly relies on aggregate measures of equity, necessitated by limited data availability (Sumah et al., 2016). Examples of commonly used aggregate measures are the Gini index — relating the distribution of the use of some care provision along some other demographic dimension like income — or concentration indices comparing expected use — usually based on age and sex — with actual use, again along some other dimension capturing possible inequity (Wagstaff et al., 1991; Jiménez-Rubio et al., 2008).

There are four reasons why this approach is limited. First, wholesale changes in the underlying use of some care provision can occur without any actual changes to an aggregate equity measure. Second, and related, equity measures do not provide insights into the overall expansion or contraction of care use and might thus hide important overall effects. Third, small groups in the population might

[6]Many of the centralizations that occurred in the 20th century were in fact explicitly motivated to reduce inconsistencies across regions that developed both purposefully as well as due to drift.

experience serious effects due to a policy intervention but will have only a limited effect on aggregate measures which are based on the entire population. Fourth, most aggregate measures only allow equity to be assessed along a single dimension — e.g., income — whereas inequity might be more complicated and concern combinations of demographic variables (for example, low income individuals with a migrant background). Additionally, much empirical work studying aggregate measures of equity as a consequence of decentralization take a cross-sectional rather than temporal approach — i.e., comparing aggregate equity statistics across regions with differing levels of decentralization (Hölsch and Kraus, 2004; Costa-Font and Gil, 2009; Lopez-Casasnovas et al., 2005). This is problematic if there were inequities prior to decentralization. Especially if such differences might have led to the choice to decentralize in the first place, which would lead to reverse causality (Sumah et al., 2016; Abimbola et al., 2019). Not knowing which groups were affected by a policy change, and in what way, severely limits the relevance of insights from conventional approaches to equity for policy action.

Conversely, changes in an aggregate measure of equity must be a consequence of differences in how groups in the population have reacted to a policy intervention — e.g., low income groups expanded their use more or less than other social groups. Studying such changes at the group level would represent a richer picture of changes to equity, which is the HPE approach taken in this chapter. Instead of assessing differences in the aggregate, the HPE approach studies the overall effect of decentralization on individuals' use, as well as whether groups in the population exhibited heterogeneous effects. Effectively, it allows equity to be studied one step prior — namely the group level — to aggregating these effects into aggregate measures of equity. The two analytical approaches are illustrated in Figure 5.1.

The reason why a HPE approach is less established in the study of decentralization is likely due to data limitations. This is particularly problematic in decentralizations. Consider the simple case of the decentralization of a care provision from a national level to 20 administrative regions. Even considering only a handful of demographic variables like sex (binary), migrant background (in three categories),

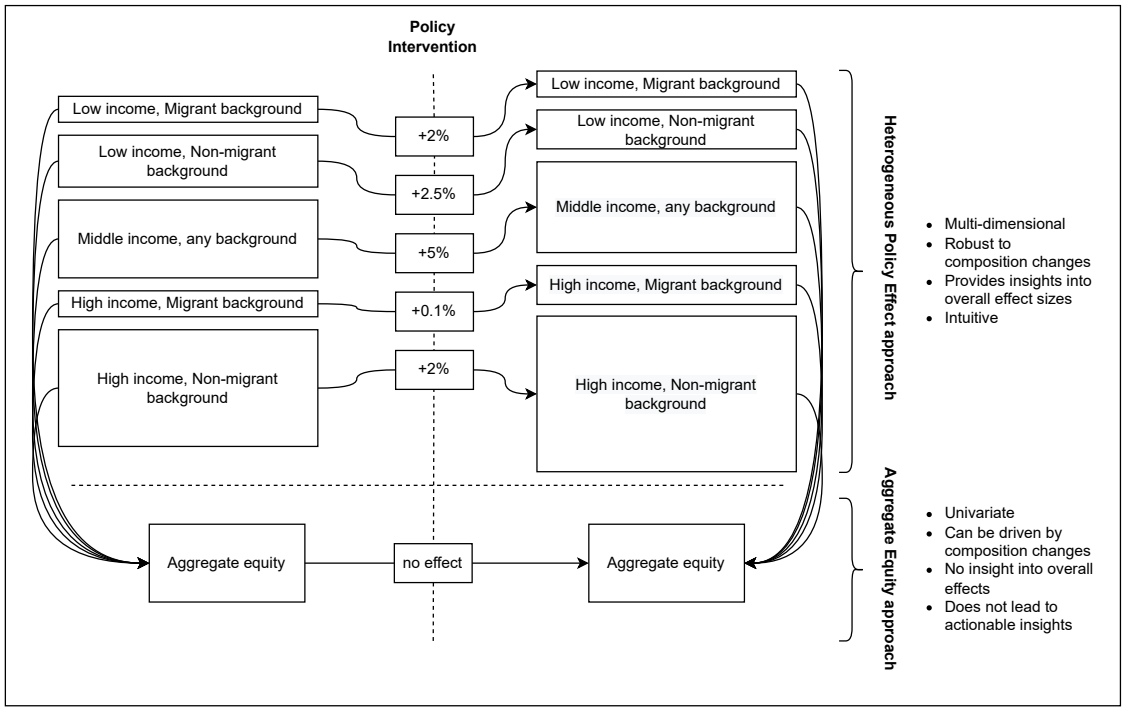


Figure 5.1: Schematic illustration of the Heterogeneous Policy Effect approach and the standard 'Aggregated' approach to studying equity. Overall equity in the system changed due to both compositional differences in the population as well as heterogeneous effects for the five depicted groups in the population.

and income (in five categories), the potential number of heterogeneous effects becomes in excess of 500 as every group might experience a different effect in each region. Assessing all these possibly heterogeneous effects quickly becomes infeasible to do manually, prone to leading to false positives, and requires considerable amounts of data to retain sufficient power for inference (Athey and Imbens, 2016; Wager and Athey, 2018). Fortunately, increases in data availability and the development of new methods that exploit computational power to robustly identify complex patterns in data make a HPE approach increasingly feasible and flexible.

5.3 A Heterogeneous Policy Effect Approach to Equity

A HPE approach is based on the standard treatment effect framework. Following the potential outcomes framework of Rubin (Rubin, 1974; Imbens and Wooldridge, 2009), we are interested in the difference between the outcome when an individual

is exposed to a policy, $Y^{(1)}$, versus the outcome when that individual is not exposed, $Y^{(0)}$. The fundamental problem of causal inference is that we only ever observe one of the two potential outcomes. The typical approach to making progress in this situation is to assume that conditional on observables, X_i, the potential outcomes $\{Y^{(1)}, Y^{(0)}\}_i$ are independent of whether individual i received the treatment or not, the so-called 'unconfoundedness' assumption:

$$\{Y^{(1)}, Y^{(0)}\}_i \perp\!\!\!\perp W_i | X_i. \tag{5.1}$$

Unconfoundedness means that conditional on observables, we can simply study the difference in outcomes between treated and untreated. This condition would be violated if there were further unobserved characteristics that might increase the probability to be part of the treatment, but also have an effect on the potential outcomes — e.g., self-selection into treatment (Imbens and Wooldridge, 2009; Heckman, 1979). Random assignment of observations to treatment would be one way to guarantee that (5.1) holds. The idea that is implicit in unconfoundedness is that conditional on observables, treatment assignment can be assumed as random (Imbens and Rubin, 2015). Popular approaches to estimating causal effects are balancing methods, where observations are weighted in such a way that the treatment and control group are similar in terms of observables (Imbens and Rubin, 2015).[7,8] However, the most common approach is to explicitly model the outcome as a function of the observables X_i and treatment W_i. Typically, a linear additive model is used for ease of interpretability and estimation:

$$y_i = X_i\beta + \tau W_i + \epsilon_i, \tag{5.2}$$

where W_i is an indicator whether observation i was exposed to treatment or not. Whether opting for covariate balancing or matching / weighting procedures, they

[7] Another popular approach is to weight observations along the propensity of receiving treatment: $e(x_i) = \Pr[W_i = 0 | X = x_i]$, which can be more efficient when the covariate space is large (Hirano et al., 2003).

[8] In policy interventions, an example of a variable confounded with both the treatment and the outcome could be age, as both treatment propensity and age are correlated with time. Matching the treatment and control groups on age could address this specific source of confounding.

all lead to an estimated effect $\hat{\tau}$ reflecting the overall impact of belonging to the treatment group on the outcome. In this chapter, $\hat{\tau}$ would reflect how being exposed to decentralization affected the overall use of some care service. However, interest further lies in *heterogeneity* in the overall effect, i.e., whether the treatment τ might differ based on observables:

$$\tau_i = f(X_i). \tag{5.3}$$

Understanding (5.3) is a key estimand when studying inequity, as any change in equity due to an intervention has to result from differences in τ_i across groups.

Studying heterogeneity in effects

The conventional approach to assessing (5.3) is to either interact the treatment indicator with other observables or to re-estimate effects to sub-groups of the population — effectively a fully interacted model. The question is what interactions to assess or subsets to evaluate. One approach would be to try all possible interactions and subsets, but this approach quickly becomes intractable as the number of possibly heterogeneous groups scales exponentially with the observables leading to the risk of spurious findings and susceptibility to p-hacking (Athey and Imbens, 2016; Gelman and Loken, 2013). When studying decentralization this problem becomes particularly pronounced due to the regional dimension as a key source of additional heterogeneity. A more commonly observed approach is to assess a smaller, theoretically motivated set of observables along which to study heterogeneity which are pre-registered by the researcher (Wooldridge, 2010; Athey and Imbens, 2016; Brand et al., 2021). The problem with this 'theory-driven' approach is that we might miss important subgroups which the researchers did not explicitly hypothesize a-priori.[9] In other words, the functional relationship given

[9]Missed heterogeneity could be due to an important dimension of heterogeneity being ignored, or limited complexity in forming groups — e.g., if heterogeneity in fact lives in a detailed sub group like 'highly educated single individuals from a Western migrant background' rather than along single dimensions.

in (5.3) might be more complex than expected (a problem discussed throughout previous chapters in this thesis).

Increasingly, ML methods are being used to identify heterogeneous effects in data. ML methods excel in finding patterns in data when the functional form is unknown and possibly complex (Hastie et al., 2009; Athey and Imbens, 2019; Chapter 4). However, they are typically applied where an outcome y_i is modelled as a function of a set of observables X_i. That is, where both the outcome and explanatory variables are observed. This allows overfitting of the model to be assessed by evaluating how well the model predicts the outcome out-of-sample. Clearly, this is not possible for estimands we don't observe in practice like treatment effects. However, tree-based ML methods have recently been adapted to allow for the study of various estimands including those we do not observe explicitly, like treatment effects (Athey and Imbens, 2016; Wager and Athey, 2018) and have led to successful applications studying the heterogeneous effects of college (Brand et al., 2021), summer school programs (Davis and Heller, 2020), state coercion on voting (Yin et al., 2021), and classroom interventions (Athey and Wager, 2019).[10]

Compared to other ML approaches to effect heterogeneity, tree-based approaches have a number of benefits. First, they go beyond the linear framework inherent in shrinkage and LASSO-based approaches (Imai and Ratkovic, 2013; Tian et al., 2014), which is particularly relevant when interactions might be present. Second, tree-based methods deal well with high dimensional data, as variable selection is incorporated into the model by design, thus avoiding the curse of dimensionality which is more prevalent in nearest neighbor approaches (see Section T.5). Third, the statistical properties of tree-based methods improve over approaches which do not provide similar statistical guarantees, like Bayesian Trees (Wager and Athey, 2018, p. 1229).

Tree-based methods partition a dataset into a number of distinct groups — so-called leaves. Leaf membership depends on iteratively partitioning the data by picking an explanatory variable and splitting the data based on a single value of that variable. For example, a split might occur using the variable 'age' where all

[10]To the best of my knowledge, this is the first application of heterogeneous effect estimation in the effects of decentralization using tree-based ML.

individuals aged 60 and over are placed in one partition, and all those under 60 in another. A process that is then repeated for each subsequent partition. The splitting variable and value are chosen in such a way that the observations in each leaf are maximally similar in terms of the outcome *within* leaves, and maximally dissimilar *across* leaves. Tree-based methods can best be thought of as adaptive nearest neighbor methods where instead of using a general distance metric to group observations together like a kernel, grouping is done along a sequence of decision rules along which variation in the outcome is maximal (Hastie et al., 2009, Chapter 9).

To adapt trees to estimating causal effects, splits are no longer based on the outcomes y in a leaf l, but rather the difference between the treated and untreated within a leaf: $\mathrm{E}[y_i|w_i = 1, i \in l] - \mathrm{E}[y_i|w_i = 0, i \in l]$. The challenges are i) how to address overfitting of the data, and ii) the consistency of the estimates.[11] A key idea is 'honest' estimation, which uses a separate partition of the data to estimate the structure of the tree — e.g., which splits to make — and uses the remaining partition to estimate the substantive estimands in each leaf. As shown by Athey and Imbens (2016), honest estimation leads to consistent estimates of heterogeneous treatment effects. A full description of the causal tree algorithm is provided in the Appendix (Section T.8.2), but it is informative to briefly consider the loss function used to build the structure of the causal tree, as opposed to the simple residual sum of squares normally minimized for standard decision trees:

$$-\mathrm{E\hat{M}SE}_\tau(\mathcal{S}^{\mathrm{tr}}, N^{\mathrm{est}}, \Pi) = \overbrace{\frac{1}{N^{\mathrm{tr}}} \sum_{i \in \mathcal{S}^{\mathrm{tr}}} \hat{\tau}^2(X_i; \mathcal{S}^{\mathrm{tr}}, \Pi)}^{\text{Variation in estimated treatment effects}} - \left(\frac{1}{N^{\mathrm{tr}}} + \frac{1}{N^{\mathrm{est}}}\right) \cdot \underbrace{\sum_{l \in \Pi} \left(\frac{S^2_{\mathcal{S}^{\mathrm{tr}}_{\mathrm{treat}}}(l)}{p_l} + \frac{S^2_{\mathcal{S}^{\mathrm{tr}}_{\mathrm{control}}}(l)}{1 - p_l}\right)}_{\text{Variation in } Y^{(1)} \text{ and } Y^{(0)} \text{ per leaf}}. \tag{5.4}$$

Here, $\mathcal{S}^{\mathrm{tr}}$ reflects the partition of the data used to estimate the structure of the tree and N^{tr} the number of observations in this set. Π reflects the tree structure.

[11] For example, naively placing splits by maximizing treatment effects in both leaves will lead to over-estimation of these effects as both estimated effects and splits are based on the same data (Athey and Imbens, 2016, p. 7335).

$S^2_{\mathcal{S}}(l)$ reflects the variance of the outcome variable in leaf l using data in set $\mathcal{S}$ and p_l the proportion of treated individuals in leaf l. Finally, $\tau(X_i; \mathcal{S}^{\text{tr}}, \Pi)$ reflects the estimated treatment effect as a function of X_i which assigns observations to an individual leaf l according to the tree structure Π.

This loss function illustrates a number of key elements in the estimation of the causal tree structure under honest estimation. The first term reflects that variation of the treatment effects estimated across the leaves in the tree is maximized. The second term reflects a trade-off between variability in τ — the first term — and the variance in the outcomes of both the treated, $Y^{(1)}$, and untreated, $Y^{(0)}$, in each leaf l. In other words, leaves with large variation in the outcome amongst treated and untreated are penalized and there has to be sufficient precision in estimating the treatment effect for a split to improve the objective function. This condition implies that a split may be made that does not increase variation in τ but does make the estimates in a leaf more precise. During the tree-building process, it is customary to use a subset of $\mathcal{S}^{\text{tr}}$ that has not been used in estimation to evaluate the criterion above when considering a new split, a process called 'honest cross-validation' (Athey and Imbens, 2016, p. 7356).[12] When the full tree structure Π has been built according to (5.4), the actual effects τ are estimated using the estimation partition not used during tree-building.

Often, a question that remains is whether observations are sufficiently similar within the leaves of a causal tree to allow for the unconfoundedness assumption (see e.g. the discussion in Brand et al. (2021)). As a solution, balancing methods can be applied within leaves. For the study of large-scale policy interventions like decentralization, such questions are less relevant as outside of immigration there is little scope for selectivity into treatment. However, the regional dimension does increase the number of variables in the model considerably. This can inflate risks that multiple trees are equally able to identify heterogeneous effects or that there might be path dependence in estimating a single causal tree to the data, a common

[12]Formula (5.4) is consistent for the criterion to be optimized, but becomes biased when repeatedly using the same sample whilst building the tree. Therefore, using a held-out set of data to evaluate the criterion is recommended (Athey and Imbens, 2016, p. 7356).

issue when using decision trees in isolation (Hastie et al., 2009, Chapter 15). The standard solution is to combine a large amount of trees into a forest, where each tree is estimated on a slightly permutated sample of the data. By averaging across all the trees, smoother partition rules are obtained which tend to have better fit and reduce the risk of locally optimal solutions.

The same rationale has been incorporated for treatment effects through causal forests (Wager and Athey, 2018). These forests have attractive properties which include estimation of confidence intervals for estimated treatment effects and an omnibus test for treatment heterogeneity in the data.[13] The downside is that they are less interpretable as compared to a causal tree. Instead of visually inspecting the tree itself, post-estimation inference has to be done via variable importance metrics that summarize the relevance of a certain variable across the trees — usually based on the number of times a variable features in the partitioning rules. However, as will be shown below, causal forests and trees can be combined in an effective manner where the former is used as a first pass to identify relevant variables, which can then be investigated further using causal trees for interpretability. In what follows, the HPE approach based on tree-based ML will be implemented to identify heterogeneous groups in the decentralization of the social domain in The Netherlands.

5.4 Empirical Case Study: The Dutch Decentralization of the Social Domain

In 1956, social security became a right in The Netherlands with the enactment of the ABW (General Assistance Act) which tasked the government with providing a minimum level of subsistence to all inhabitants (Vonk and Schoukens, 2019).[14] The original ABW heavily relied on local diversification with little central control.

[13] Wager and Athey (2018) develop a general asymptotic theory for random forests which extend to the use of causal forests. Their approach relies on incorporating auxiliary noise into the estimation process combined with repeated sub-sampling, which is similar in spirit to bootstrap methods.

[14] Social security became a fundamental social right in 1983 through enactment of Article 20 of the Constitution stating that 'the means of subsistence of the population is a concern of the authorities' `https://www.denederlandsegrondwet.nl/id/vkugbqvdutxn/artikel_20_bestaanszeker_heid_welvaart`.

In practice, this led to instances where similar individuals in need were treated differently across localities (Vonk and Schoukens, 2019, p. 157). This was considered undesirable, leading to revisions in 1972 with more extensive national control being imposed.[15] However, by the early nineties, the centralized social welfare system had become increasingly complex as multiple schemes were added in the interim period. As a result, the system was perceived to be cumbersome and inefficient (Schut and van den Berg, 2009). Simultaneously, a re-appreciation for the originally local character of social welfare emerged (Maarse and Jeurissen, 2016), leading to a push for increased (re-)delegation of responsibilities to the municipality level.

Further revisions in 1992 and 2003 tasked municipalities with enacting their own local legislation, including increased discretion on deciding how to spend social welfare finances, although overall financial control remained at the central level. The biggest overhaul occurred when a major law change, named the 'decentralization of the social domain', was enacted in 2015. This act completely delegated the administrative functions of a number of care domains to the municipal level. Municipalities received extensive freedom in identifying needs and administering care, and were explicitly encouraged to innovate on standard practices as they deemed appropriate. These tasks did not include income-related welfare policies like unemployment benefits, but rather in-kind services provided to citizens to help with specific care needs — examples include extra help in the household, day activities or home improvements like a stairlift (Mot et al., 2010).[16] This is relevant as social security in the financial sense allows for more determinism in its implementation, e.g., by setting subsistence thresholds. The delegated activities were considerably more variable, and means-testing was explicitly forbidden. In this chapter, I focus on what is known as 'social assistance' (WMO) which, after the law change, can broadly be classified as i) help in the household, ii) services like transport or home

[15]The main mechanism of control was the financial functioning of the social care system, with the government now able to decline covering incurred costs if the provided care did not follow set rules. This allowed the government to frame the general terms through which social welfare was provided (Kamerstukken II 1991/92, 22 545, nr. 3).

[16]See also *Memorie van Toelichting Wet Maatschappelijke Ondersteuning 2015.* Kamerstuk; January 14, 2014.

adjustments, iii) home assistance including care as well as other activities, and iv) short-term care at an external facility.[17] Social assistance should be juxtaposed with intensive, long-term care (LTC) which falls under the WLZ (Long-term Care Act), a separate care arm that is not studied here.

Social assistance before and after 2015

The PW (Social Care Act) was enacted on January 1st 2015 and split what was formerly known as the AWBZ (Special Care Act) into three distinct elements. A key element of the policy change was to assign parts of the social care elements into a separate function called the WMO, covering social assistance. The administration of the WMO was then fully decentralized to the municipality level. The PW was part of a broader transition towards a so-called 'participation society', understood as a system that was 'geared more to the individual's "own strengths and real needs"' and as a consequence required a significant 'customization' of social assistance (Vonk and Schoukens, 2019, p. 158). By delegating administrative responsibility to the municipality, it was deemed to give 'more opportunities to provide people with integrated customized support' and 'creat[ing] a system at [the] local level that was less dependent on central rules and more responsive to the needs of the individual citizens' (Vonk and Schoukens, 2019, p. 152). This view is aptly summarized by the 'one household, one plan, one provider' mantra that was promoted by the government as part of the law change. Clearly, the customization argument for decentralization had been heavily invoked in the 2015 law change (Vermeulen, 2015), and its introduction was accompanied by considerable budget cuts (Vonk and Schoukens, 2019, p. 158).[18]

In practice, each municipality was mandated by law to propose a concrete policy plan describing their planned provision of social assistance. This included how to identify those in need and the way in which assistance was to be provided.

[17] Across the decentralization, care types are consistent and broadly speaking the same regulatory processes remained in place to ensure quality of care.

[18] The objective set out by the central government was to cut administrative costs by 20% as part of the decentralization (Kamerstuk, Rapport brede heroverwegingen, 18. Openbaar bestuur, Tweede Kamer der Staten-Generaal, Den Haag, 1 April 2010).

This plan was then to be formalized through a bye-law.[19] Broad guidelines were provided centrally, for example stating that so-called 'kitchen table talks' and 'multi-disciplinary neighborhood teams' should play a role in identifying care needs, but concrete implementation was left to the municipality.[20,21] Kitchen table talks were meant to reflect the ambition that unspecified social care professionals would develop a custom plan in consultation with and tailor-made for a single household. Neighborhood teams, on the other hand, were supposed to help identify those in need at the local level (van der Tier et al., 2021). Qualitatively it has already been found that municipalities differed substantially in their interpretation and implementation of these functions (Claessen et al., 2019; Jansen et al., 2021).[22]

The 2015 decentralization of the social domain can be expected to have had a far-reaching impact on access to and subsequent use of social assistance. First, the intervention was explicitly motivated to make social assistance more 'customized' to the local population, making regional heterogeneity a central goal. Second, the way in which the policy change was enacted invites considerable subjectivity into the decision-making process. The multitude of pathways in which assistance use might have been affected — e.g., through heterogeneity in policymakers' interpretation of guidelines and social assistance all the way to differences in operationalization on the ground — make it difficult to hypothesize what might have happened to inhabitants' use of social assistance after the policy change. In what follows, I apply a HPE approach to improve understanding into the impact of the decentralization of the social domain in terms of overall expansion or contraction in social assistance use, as well as its heterogeneity across demographic groups and geographic regions.

[19] Article 2.1.3 WMO.

[20] For example, municipalities were not allowed to abolish social care elements like public home care (Vonk and Schoukens, 2019).

[21] See Article 8 and 8a of the PW. A system where decentralization occurred through a national legislative framework that provides municipalities with competences to set their own standards and procedures in providing actual social care is also known as a sophisticated form of "regulated decentralization" (Vonk and Schoukens, 2019).

[22] For example, some municipalities employed actual care providers in kitchen table talks, whereas other municipalities instead employed individuals who were more knowledgeable about the administrative side of the care system (Claessen et al., 2019).

5.5 Data and Research Design

To study the impact of the 2015 decentralization of the social domain on the use of social assistance in The Netherlands, I use registry data collected by Statistics Netherlands (CBS).[23] The period of interest are the years directly following the policy change in 2015 up until 2019. The registry contains information on every individual registered to be living in the Netherlands at the start of each year based on municipal population registers.[24] I use a range of demographic characteristics including sex, age, and household income — based on tax data — expressed as a percentage of the target income set by the government as the 'social minimum'.[25,26,27] I also include migrant background — based on the country of birth of an individual's mother — classified into Dutch (no migrant background), Non-western, and Western.[28] Finally, I include household type, specifically whether an individual is part of i) a single-person household, ii) a couple without children, iii) a couple with children, iv)

[23] All results presented here are based on own calculations based on non-public registry data from CBS, accessed through the Remote Access environment. CBS was not involved in the calculation of any of the results presented here.

[24] See `https://www.cbs.nl/en-gb/our-services/customised-services-microdata` for a full description of the dataset.

[25] Whenever household income is unavailable, individuals are either living in an institutional household or income is unknown to the tax authorities. The inclusion of an indicator for institutional households allows the former to be distinguished from the latter. When either of these cases occur, income is coded as a negative value (-2 for unknown incomes, and -1 for institutional incomes).

[26] Expressing income as a percentage of the social minimum addresses the need to discount income, for example due to inflation. More generally, it reflects the way in which the government perceives poverty. The household-level social minimum further depends on the household composition, and thus more accurately reflects subsistence thresholds than actual real income. For example, the social minimum in 2016 for an individual above 22 years of age was set at 8.87 EUR per hour which amounts to approximately 16,000 EUR yearly (Kamerstukken II, 2007–2008, 31 200 – XV, nr. 84, p. 2.).

[27] I include education as an alternative operationalization of socio-economic status as a robustness check. The reason for preferring income data is twofold. First, educational data is missing for parts of the population. The main reason for this gap is that educational credentials are not always accurately registered for the elderly who obtained their credentials in previous incarnations of the educational system, which cannot be accurately mapped to current credential classifications. The same holds for degrees obtained in different countries. Second, income and educational attainment are highly correlated in The Netherlands and typically lead to similar substantive results (De Graaf and Ganzeboom, 1993).

[28] Western is classified in line with the definition of CBS and includes all countries in Europe (excluding Turkey), North America, Oceania, Indonesia and Japan. All remaining countries are classified as Non-western. Individuals born abroad but for whom both parents were born in the Netherlands are classified as Dutch.

a single parent with children, v) some other multi-person household structure, or vi) an individual with an unconventional income, which includes both individuals living in 'institutional' households, as well as those without a registered income during the year of observation.[29]

As outcome of interest, I define social assistance use as a binary variable indicating that an individual has made use of any form of social assistance during a calendar year. Prior to 2015, information on social assistance use was collected and submitted to the registry by the CAK (Central Administrative Office AWBZ). This information concerned all individuals who received funding for social assistance services *and* had paid a personal monthly contribution into the general fund from which social assistance was financed. This latter detail is important, as most individuals were required to pay a means-tested personal contribution, but certain individuals — e.g., those in the lowest income brackets — could receive exemptions.[30] As a consequence, there had never been complete data on social assistance use at any geographic aggregation level when data was still submitted to the registry by the CAK.[31]

From 2015 onwards individual-level use has been submitted directly to the registry by individual municipalities. This collection is no longer conditional on a personal contribution, allowing for complete information on social assistance use for those municipalities that submitted data to the registry. Submitting data to the registry is voluntary but strongly recommended, and there are various incentives to do so — for example, it allows municipalities to use the so-called Social Assistance Monitor which the government developed based on the submitted data to provide municipalities with comparative insights.[32] The majority of municipalities have

[29]Household structure is based on tax records. About half of the households in the unconventional income category are living in so-called institutional households. These households consist of individuals living in elderly care homes, but also other government-run institutions like military bases or jails. Throughout the causal tree and forest analyses, institutional households can be further distinguished through the income indicator.

[30]`https://www.hetcak.nl/regelingen/wet-maatschappelijke-ondersteuning/wmo-in-2022/eigen-bijdrage`.

[31]`https://www.cbs.nl/nl-nl/nieuws/2020/19/ruim-1-1-miljoen-mensen-met-wmo-maatwerk-in-2019`.

[32]Conditional on uploading municipal data, a report can be generated at `https://www.waarstaatjegemeente.nl/`.

	Total population		Submitted to registry		Analysis sample	
Year	Munics.	N	Munics.	N	Munics.	N
2015	394	16,978,307	179	7,956,420	-	-
2016	391	17,081,138	243	11,993,491	208	11,107,561
2017	388	17,181,921	359	14,466,471	-	-
2018	381	17,283,020	314	15,593,417	-	-
2019	355	17,408,280	320	16,164,810	208	11,347,048

Table 5.1: Number of municipalities and the size of the population living in these municipalities across time for i) the total population in The Netherlands, ii) those municipalities submitting data to the registry, and iii) those municipalities submitting data in both 2016 and 2019.

indeed submitted data from the very first year after the policy change, although some municipalities only started submitting data in the latter years. For example, coverage of the data already exceeded 65% of the population in 2016 (see Table 5.1).

In order to study the use of social assistance across time I subset the total population in three ways. First, I limit the analysis to comparing social assistance use in the years 2016 and 2019. This means using a time point after the actual law change as the 'pre-policy' period. The reason for choosing this starting point is that the first two years after the policy change were deemed 'transitional' with minimal customization across municipalities. The period following this initial phase was deemed to be 'transformational' (van der Ham et al., 2018, Chapter 3). Especially the first year following the decentralization was dominated by the administrative transfer from the central to the municipal level, and the aim during this period was to provide 'a seamless transition' rather than customization across municipalities (van der Ham et al., 2018, p. 40).[33] In other words, the substantive effects of the decentralization will not have kicked in until the latter years. I confirm this in the robustness section. From a pragmatic perspective, 2016 was the first year where more than half of the population is covered thus providing a more complete picture of the population as well. Second, I only focus on individuals aged 18 and over as younger individuals make use of a separate policy arm (the Jeugdwet). Third, I only include municipalities which have submitted data to the registry in both 2016

[33]See also *Voortgang en ambitie Wmo, volwaardig meedoen.* Brief van de staatssecretaris van Volksgezondheid, Welzijn en Sport. Tweede Kamer, vergaderjaar 2015/2016, 29538, nr. 214.

and 2019. Some municipalities merged over time, meaning that some municipalities who submitted in 2016 drop out of the analysis if they merged with municipalities that did not. This only affects a tiny fraction of the 2016 population.[34]

Basic descriptives of the sample

Descriptives for the population under study are provided in Table 5.2 for both 2016 and 2019. To assess selectivity due to including only municipalities that submitted data to the registry, I provide descriptives for the full population in 2016 and 2019 as well. As can be seen, demographics across the three-year period are fairly consistent. There was a net population growth and most of this increase was due to in-migration, which can also be seen from the relative increase in the number of people with either a Western or Non-western migrant background living in the Netherlands (Mooij et al., 2020). The population also slightly aged during this period. From the full population columns it follows that selectivity due to only including municipalities that submitted data does not lead to meaningful differences in the sample under study.

We observe an expansion of social assistance use of about 0.8 pp between 2016 and 2019. Figure 5.2 further shows considerable baseline differences across municipalities, as well as differences in the change of social assistance use. Note that a disaggregated analysis of each type of social assistance in isolation is outside the scope of this chapter, although some insights are provided in the results section.[35]

Demographic determinants of social assistance use

To get a better sense of the determinants of social assistance use, Table 5.3 shows descriptives statistics of social assistance users and the rest of the population for the years 2016 and 2019. As can be seen, there are considerable differences with

[34]Mergers usually occur for logistical / policy convenience. The total number of municipalities decreased from 391 in 2016 to 355 in 2019. In some cases, a municipality was partially merged with multiple other municipalities, specifically, this was the case for the municipalities Littenseradeel and Winsum.

[35]A further extension when focusing on a single type of care could be to investigate an enumeration of the number of times a type of assistance is used rather than a binary outcome.

	Analysis Sample				Total Population			
	2016		2019		2016		2019	
Variable	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.
Age	48.690	18.479	48.938	18.745	49.164	18.510	49.449	18.796
Income: % of social minimum	281.680	177.788	298.875	188.807	285.499	176.484	302.872	187.335
Sex: Male	0.492	-	0.493	-	0.492	-	0.493	-
Sex: Female	0.508	-	0.507	-	0.508	-	0.507	-
Migrant Background: Native Dutch	0.752	-	0.733	-	0.774	-	0.758	-
Migrant Background: Non-Western	0.138	-	0.150	-	0.127	-	0.137	-
Migrant Background: Western	0.110	-	0.117	-	0.105	-	0.105	-
Household type: Single parent	0.065	-	0.067	-	0.062	-	0.065	-
Household Type: Single person	0.220	-	0.217	-	0.210	-	0.208	-
Household Type: Unconventional	0.030	-	0.032	-	0.029	-	0.030	-
Household Type: Other multi-person	0.032	-	0.033	-	0.030	-	0.032	-
Household Type: Couple with children	0.354	-	0.350	-	0.361	-	0.357	-
Household Type: Couple without children	0.300	-	0.301	-	0.308	-	0.309	-
Highest Education: Unknown	0.414	-	0.379	-	0.426	-	0.388	-
Highest Education: Low	0.148	-	0.142	-	0.143	-	0.137	-
Highest Education: Middle	0.250	-	0.259	-	0.252	-	0.262	-
Highest Education: High	0.187	-	0.221	-	0.180	-	0.213	-
Social Assistance Use: Yes	0.069	-	0.077	-	-	-	-	-
n	8,894,433	-	9,170,345	-	13,677,385	-	14,069,936	-

Table 5.2: Descriptive statistics for 2019 and 2016 of all 208 municipalities who submitted social assistance data from 2016 onwards to the Dutch registry, and total population descriptive statistics for the years 2016 and 2019 (final columns).

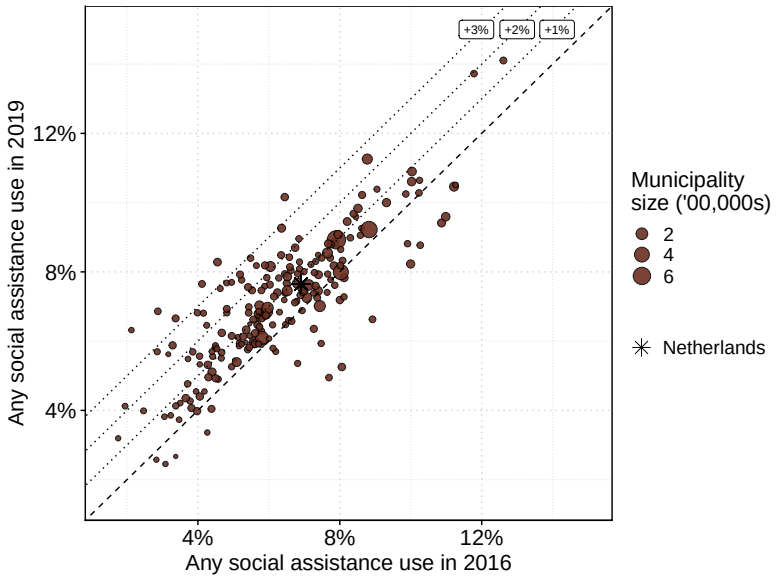


Figure 5.2: Development of social assistance usage between 2016 (horizontal axis) and 2019 (vertical axis) for all 208 municipalities that submitted social assistance data to the registry from 2016 onwards.

	2016		2019	
Variable	Non-Users	Users	Non-Users	Users
Age	47.322	67.138	47.508	66.180
Income: % of social minimum	291.556	148.493	310.796	155.102
Sex: Female	0.498	0.639	0.497	0.626
Migrant Background: Native Dutch	0.750	0.774	0.731	0.756
Migrant Background: Non-Western	0.139	0.131	0.150	0.149
Migrant Background: Western	0.112	0.095	0.119	0.094
Household Type: Single parent	0.065	0.058	0.068	0.063
Household Type: Single person	0.198	0.512	0.193	0.500
Household Type: Unconventional	0.025	0.096	0.027	0.085
Household Type: Other multi-person	0.033	0.021	0.034	0.021
Household Type: Couple with children	0.374	0.077	0.372	0.082
Household Type: Couple without children	0.305	0.235	0.305	0.249
Highest education: Unknown	0.395	0.672	0.359	0.609
Highest education: Low	0.144	0.206	0.135	0.224
Highest education: Middle	0.262	0.095	0.270	0.125
Highest education: High	0.199	0.027	0.236	0.042
N	8,280,459	613,974	8,468,206	702,139

Table 5.3: Descriptive statistics of the analysis sample by social assistance use and year across all municipalities that submitted data on social assistance use to the registry in both 2016 and 2019.

the elderly and less affluent making more use of social assistance than the rest of the population. Importantly, there also seems to have been changes to the composition of social assistance users across time. For example, the average age of social assistance users has declined across the period under study. Additionally, the average income of users has increased, albeit to a lesser extent than the overall population (see Table 5.2).

To get a multivariate sense of the determinants of social assistance use, I estimate a simple linear probability model with social assistance use as outcome and the descriptive characteristics included in Table 5.2 as covariates to 2016 data. The results are shown in Figure 5.3 and indicate a strong age effect in social assistance use with a negative overall effect up until the age of about 55 and a strong positive association at older age. Women have an increased probability of using social assistance (+1.8 pp), as do single person households (+4.9 pp) and individuals in the rest category of unconventional households (+17 pp). Likely, use of social assistance is driven by almost half of these individuals who live in institutional households. Couples without children, who can likely rely on one another for assistance, show a considerably lower use (-5 pp). There is also a strong income gradient with richer individuals considerably less likely to use social assistance. Individuals with a migrant background are also less likely to use social assistance, which might indicate barriers to access.

Demographic determinants go some way to explaining baseline differences in use across municipalities. I illustrate this by making predictions for the expected number of social care users based on the above model and plot this prediction against the observed prevalences.[36] I visualize the results in Figure 5.4. Compositional differences make up some of the variability across municipalities, but not all. This further speaks to the importance of assessing the effects of decentralization in a heterogeneous way that includes possible interactions between municipalities and demographic characteristics.

[36]Note that I exclude the municipality for which the prediction is made from the estimation routine to ensure the prediction is made out of sample and accurately reflects the model's ability to predict the outcome (Chapter 2).

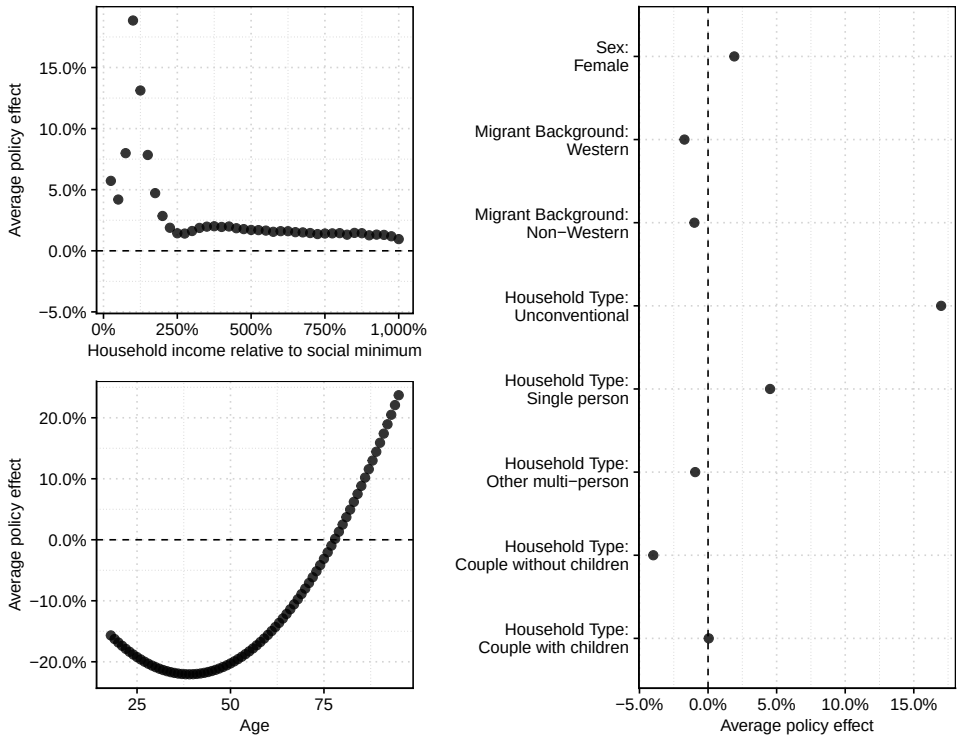


Figure 5.3: Coefficient estimates of a standard linear probability model with social care use as outcome.

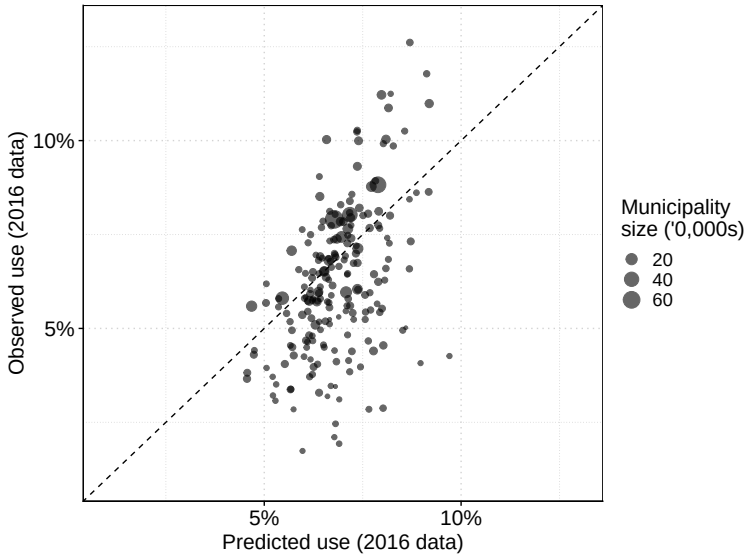


Figure 5.4: Predicted versus observed social assistance usage at the municipality level based on a linear probability model estimated to 2016 data.

Analytical strategy

To assess how social assistance use has developed after the policy change, I pool all individuals in the two interest years — 2016 and 2019 — including a treatment indicator set to one for individuals in 2019 and zero for those in 2016. In line with the research questions set out earlier, I start off by analyzing the overall expansion in social assistance use controlling for observables through various approaches. First, I estimate a linear probability model including all covariates — the 'covariate balancing' approach. Second, I apply weighting methods like propensity score matching and entropy balancing (Morgan and Winship, 2015). Third, I use the causal forest described in Section 5.3 which can also be used to assess the overall policy effect by calculating a simple weighted average of the effects estimated in each leaf.

I then assess heterogeneity in effects using a two-step approach. I start by estimating a causal forest to a one third random sample of the dataset. Sampling is done for computational reasons[37] and includes individuals across both control and treatment groups, all relevant demographics as well as a dummy-encoded matrix of indicator variables indicating the municipality an individual resides in. I use variable importance metrics to identify which variables the causal forest deems meaningful for heterogeneity of the policy effect. In particular, I use this step to identify municipalities that exhibit heterogeneous patterns, which I call the set of 'alternate municipalities' that follow a different heterogeneity pattern from the remaining set of municipalities — the 'consistent municipalities.' I then subset the total sample into separate sets for each alternate municipality and a pooled set with all the consistent municipalities. Based on these sets, I estimate causal trees to samples of 500,000 observations for each, unless an alternate municipality is smaller in which case I just use the entire sample available for that municipality.[38] This analytical approach is visualized in Figure 5.5.

[37]Estimating the causal forest based on a one third random sample and minimum node size of 2,500 on the heavy computing servers of the CBS takes about 24 hours. Estimation time when increasing the sample to 50% lasted at least ten days after aborting the routine prematurely.

[38]Estimating a single causal tree to 500,000 observations using a minimum node size of 2,500 on the heavy computing servers of the CBS takes between 2 and 8 hours.

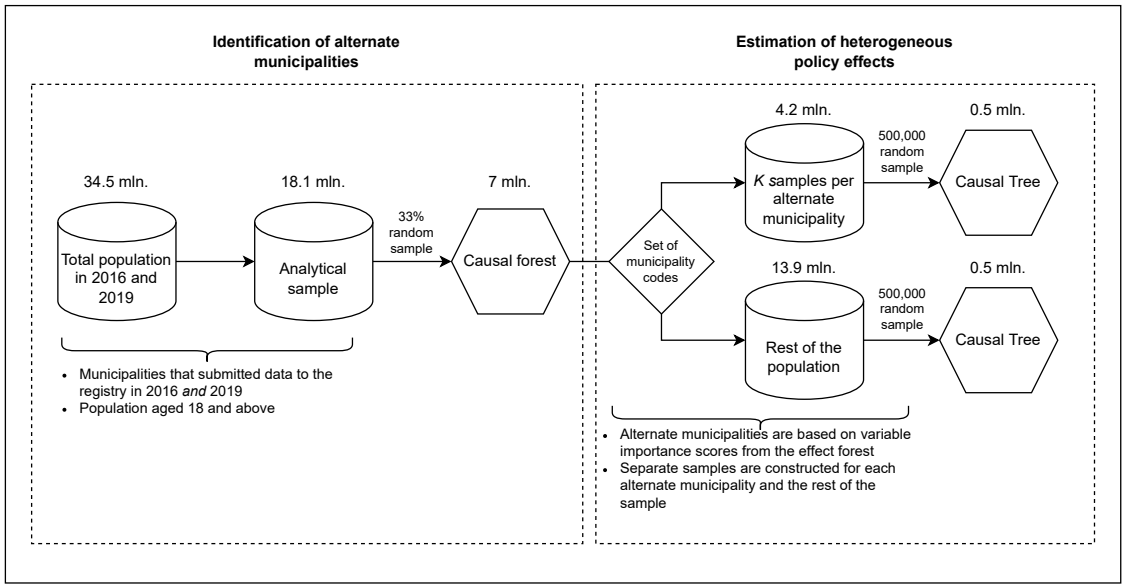


Figure 5.5: Analytical approach to estimating effect heterogeneity. Note that sample size of the set of alternate municipalities has already been reported in this figure (see Section 5.6).

The two-step process described above exploits the improved robustness of causal forests in order to identify relevant determinants of overall heterogeneity in the data, in particular municipalities, but retains the interpretability of individual causal trees to analyze the subsequent samples. As the dataset is so sizeable, the typical dependence of causal trees on sample variability can be expected to be minimal, meaning that the loss in efficiency due to retaining a causal tree approach should be minor. I set the minimum size of the leaves in the causal trees and forest to 2,500 which should allow detailed groups to be identified without overburdening the computational demands of estimating both approaches at lower levels of detail. Post-estimation, I limit the depth of the tree by inspecting the out-of-sample error. To verify that the leaves found by the tree indeed reflect authentic heterogeneity, I further assess whether the heterogeneous policy effects that were found in the sampled data hold in the entire population data. This can be seen as an out-of-sample test of the found heterogeneity, by evaluating the policy effects on data not used in estimating the heterogeneous effects.

Method	Estimate	Std. Error	N
Covariate balancing	0.00971	0.00011	18,064,778
Propensity matching	0.00970	0.00012	18,064,778
Entropy balancing	0.00973	0.00012	18,064,778
Causal forest	0.00942	0.00019	5,961,377

Table 5.4: Estimated ATEs using various estimation methods.

5.6 Results

I start by presenting results for RQ1 regarding the overall change in social assistance use between 2019 and 2016 controlling for demographic characteristics (Table 5.4). As can be seen, the period under study can be characterized by an expansion of around 0.97 pp in social assistance use when applying conventional methods like entropy balancing, propensity score weighting, and covariate balancing. This is higher than the descriptive expansion in the period under study even though the population aged, and likely is a result of the population becoming relatively wealthier in the intermittent period, which would otherwise be associated with a decline in use. Results across methods are highly similar, both in terms of point estimates as well as standard errors. This will be in part because balance on observables across the 2019 and 2016 population was already fairly good prior to estimation, which is unsurprising given that selection into treatment was impossible other than via immigration (see Figure A.5.3).[39]

In addition to the overall estimates from conventional methods, I include effect estimates from the causal forest as well. The causal forest might match similar groups of individuals together in a potentially more complex way than conventional methods, possibly incorporating interacted effects and non-linearities. Nonetheless, estimating a causal forest to the data leads to an average treatment effect (ATE) which is broadly in line with those found from the entropy and propensity matching approaches, albeit slightly lower.[40]

[39]Most variables were already well within conventional balance bounds of 0.1 deviance prior to matching. Note that the overall propensity score — a combined metric of balance — did indicate the need for further balancing.

[40]Note that a one third sample is used for the causal forest approach. Re-estimating the

Variable	Importance
Income: % of social minimum	0.19929
Household Type: Single person	0.15122
Household Type: Couple with children	0.09813
Age	0.09474
Migrant Background: Western	0.07771
Municipality: The Hague	0.06446
Migrant Background: Non-Western	0.05665
Household Type: Single parent	0.05055
Sex: Female	0.04728
Municipality: Amsterdam	0.04483
Household Type: Couple without children	0.03898
Migrant Background: Native Dutch	0.03759
Municipality: Rotterdam	0.02987
Household Type: Unconventional	0.00711
Household Type: Other multi-person	0.00106
Municipality: Utrecht	0.00052
Municipality: Almere	0.00002

Table 5.5: Variable importance scores for the causal forest estimated to one third of the total population including municipality codes. Municipality codes are highlighted in boldface.

Identifying alternate municipalities

In the next step, I again estimate a causal forest to the data which includes dummy variables indicating each individual's municipality of residence. Of key interest in this step are variable importance metrics which indicate the variables the forest used to estimate the heterogeneity in estimated policy effects. These are presented in Table 5.5. As can be seen, five out of a total of 183 dummy-encoded municipalities feature in the partitioning process and are thus labeled as alternate municipalities going forward.[41] This implies that heterogeneity in the policy effect across municipalities is limited to a select number, and heterogeneity is otherwise stable throughout most of the population without varying considerably from one municipality to the next. Interestingly, these alternate municipalities include some of the larger municipalities in The Netherlands like Amsterdam and Rotterdam, which tend to have considerably more civil servants in both absolute as well as relative terms, including dedicated

conventional methods to the same sample yields similar findings (results not shown here), implying that the difference between the forest and conventional weighting methods are not driven by the sample used to estimate the overall policy effect.

[41]Note that I grouped together all municipalities with less than 25,000 inhabitants for computational efficiency, which combines 26 of the 208 municipalities into a single dummy and relates to about 2.4% of the sample.

innovation teams (Panteia and E'til, 2020).[42] In terms of the substantive size of the importance metrics, municipalities Utrecht and especially Almere feature less frequently than the others. This could suggest that they differ less markedly from the rest of the population. Furthermore, all demographic characteristics seem to play a role in estimating heterogeneity and an omnibus test confirms heterogeneity to be present across the policy's effect (see Table A.5.3).

Causal tree analysis: Consistent municipalities

I start by estimating a causal tree to the set of consistent municipalities, illustrated in Figure 5.6. Descriptively, the difference between 2019 and 2016 — shown at the top of the tree — was 0.79 pp descriptively, and 1.06 pp when applying conventional balancing methods (see Table A.5.2).[43] This shows there was stronger overall expansion in the set of consistent municipalities relative to the entire population. Note that a tree depth parameter has to be chosen to decide the complexity of the final tree, i.e., the number of leaves to show. To pick this parameter, I assess the decrease in out-of-sample prediction error as a function of tree depth and set the complexity parameter where out-of-sample improvement plateaus (see Figure A.5.4).

Following the structure of the tree, the first partition occurs at a fairly high cutoff of income — a household income of 292% of the social minimum — and partitions 46% of the sample into a terminal node with an estimated policy effect of 0.27 pp for individuals with incomes above the cut-off point. The effect is almost eight times as large in the other partition, estimated at 2.1 pp. In other words, most if not all of the observed expansion is located at the lower end of the income distribution. Assessing the less wealthy partition further, the next split is made on being a Dutch native versus having some form of migrant background. Those with a migrant background recorded a lower policy effect of 1.0 pp whereas the

[42] Amsterdam, Rotterdam, The Hague and Utrecht are all part of the so-called 'G4' group in The Netherlands containing the four large urban centres. The group was formalized in 1994 to align on major policy topics, although it has been expanded to the G40 in latter years to include additional municipalities (45 since July 2022).

[43] The nomenclature used here to describe a tree is 'partition' for every group in the tree unless it is a terminal partition — without further splits — in which case it is called a 'leaf'.

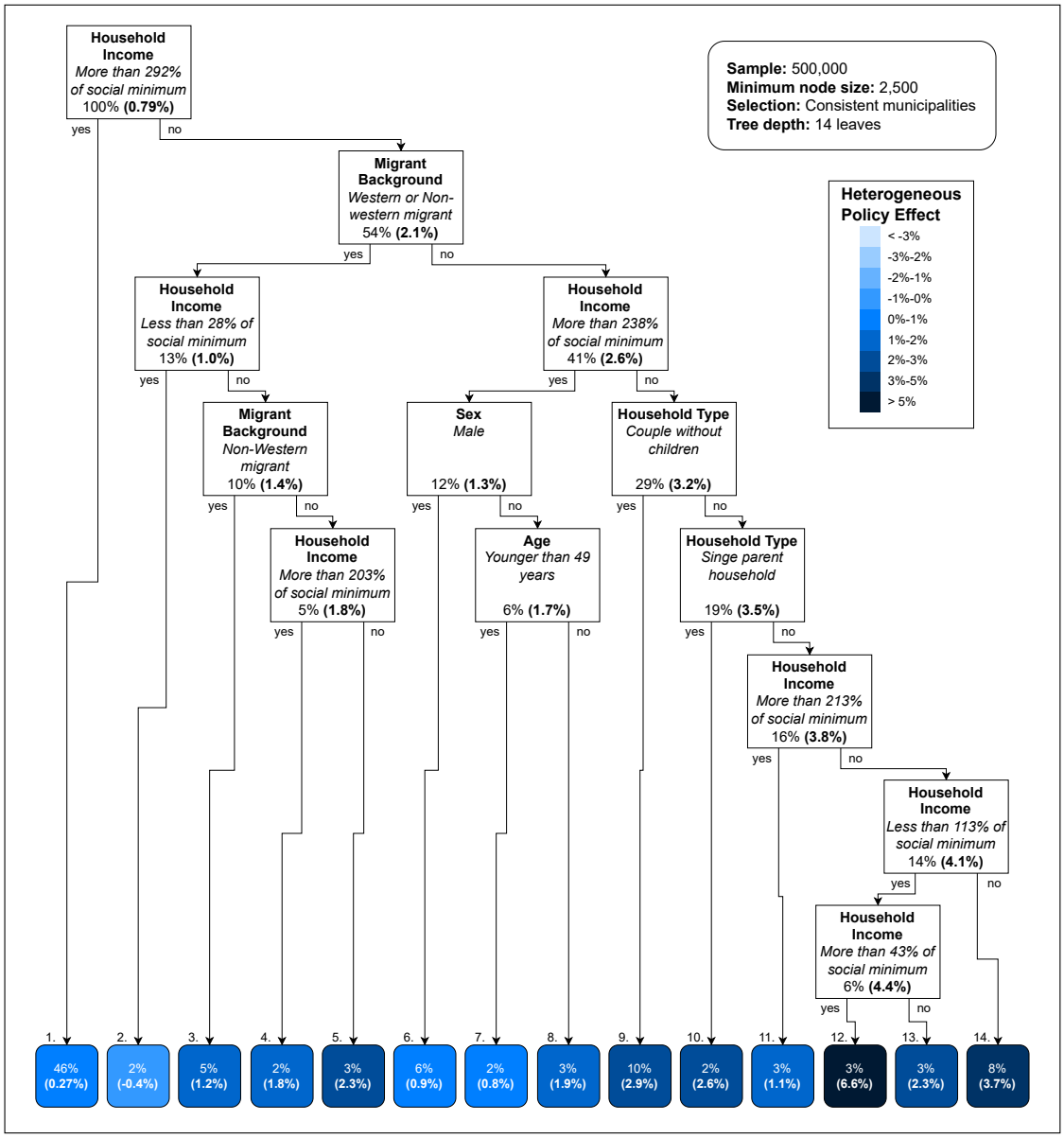


Figure 5.6: Causal tree results based on 500,000 individuals from the set of consistent municipalities across the years 2019 (treatment) and 2016 (control). Causal tree has been pruned at a complexity level of 14 nodes based on visual inspection of the out-of-sample error. For each node, the percentage of the total sample reaching that node is represented as well as the estimated policy effect in boldface parentheses.

remaining set of natives recorded an increase of 2.6 pp. Within the partition of relatively less affluent natives, a further partition is made on income, this time at 238% of the social minimum. Those above this point show effects ranging from 0.9 pp to 1.9 pp with older females having a stronger effect than younger females or males of any age. In the partition of natives with incomes below 238% of the social

minimum, we observe that couples without children and single-parent household both recorded effect sizes of around 2-3 pp whereas the rest of the household types show a strong income gradient with effects ranging from 1 to 6.6 pp dependent on income, with lower incomes generally recording larger effects.

A couple of remarkable features appear from the causal tree analysis. Firstly, there is a strong income gradient visible, where poorer individuals exhibited stronger increases in the use of social assistance. This gradient is considerably stronger under the native Dutch population than those with any migrant background. This limited expansion for individuals with migrant backgrounds could reflect difficulties in reaching these groups. It is noteworthy that most immigrants in the Netherlands live in the major cities like Amsterdam, Rotterdam and The Hague, which follow a different heterogeneity pattern and aren't covered here (see Table A.5.8). This could point at a degree of local capture, where individuals with a migrant background who live in municipalities that consist of a (large) majority of natives are effectively left behind. Another noteworthy finding is that amongst Dutch natives with incomes below 43%, a relative decrease in effect is observed relative to those just above this threshold. This is likely due to the fact that this group includes those with unconventional incomes — which have incomes coded as -1 or -2. A finding that is further confirmed when assessing the descriptive statistics of each group (see Table A.5.4). This group is characterized by a degree of unconventional participation in society, containing those without any labour-related income. The reason for a lower policy effect amongst this group could be because these individuals already made quite extensive use of assistance services prior to the policy intervention (Table A.5.1).

In sum, in the consistent set of municipalities there appears to have been a considerable expansion of social assistance use at the lower end of the income distribution which would suggest improved equity. However, this uptake is lower under those with a migrant background as compared to native Dutch people. This finding shows that amongst this group, barriers to access remain and may have increased further since the policy change, at least at set levels of income.

Causal Tree Analysis: Individual Alternate Municipalities

The Hague area: The results from estimating a causal tree to the The Hague municipality are presented on the left of Figure 5.7. Straightaway, we observe a descriptive contraction of 0.13 pp instead of the overall expansion of social assistance use observed in the population. Estimating this effect using conventional methods indeed shows that expansion in the area was minimal and around 0.18 pp (Table A.5.2). Interestingly, the initial partition of the tree shows that declines in use are mainly driven by those over 72, with the first partition placed at age rather than income. Those over 72 exhibited a contraction of 3.6 pp whereas the rest of the population expanded somewhat by 0.11 pp. In particular the 72+ without a migrant background showed strong negative effects (-4.3 pp) as compared to those with a migrant background (-1.9 pp). Amongst the younger partition, the data is further split on the group of households without conventional incomes, which also show a strong decline in use (-3 pp). These findings indicate that groups which tended to make more use of social assistance initially — the elderly and those in unconventional households — contracted significantly following the policy change in The Hague.

Beyond these strong age and household effects, we further observe that the richer part of the sample aged below 73 — with incomes over 207% of the social minimum — exhibit a null effect overall, whilst those below this threshold recorded an expansion (+1.2 pp). This income gradient is consistent with earlier findings. As was the case at high age, Dutch natives exhibit smaller effects overall when income is between 123% and 208% of the social minimum (+0.98 pp versus +1.2 pp). Finally, beyond a continuing strong income gradient, where lower incomes show stronger effect sizes, we observe increased effects amongst poor individuals living in multi-person households (+3.3 pp) relative to living alone (+1.3 pp) and a considerable difference between younger Non-western individuals (-0.67 pp) and older (+1.5 pp).

Compared to the findings across the consistent municipalities the differences are quite remarkable. In particular the strong negative age effect observed in The Hague is noteworthy, as is the contraction amongst households without a conventional income. Finally, there has been a relative increase in use amongst Non-western

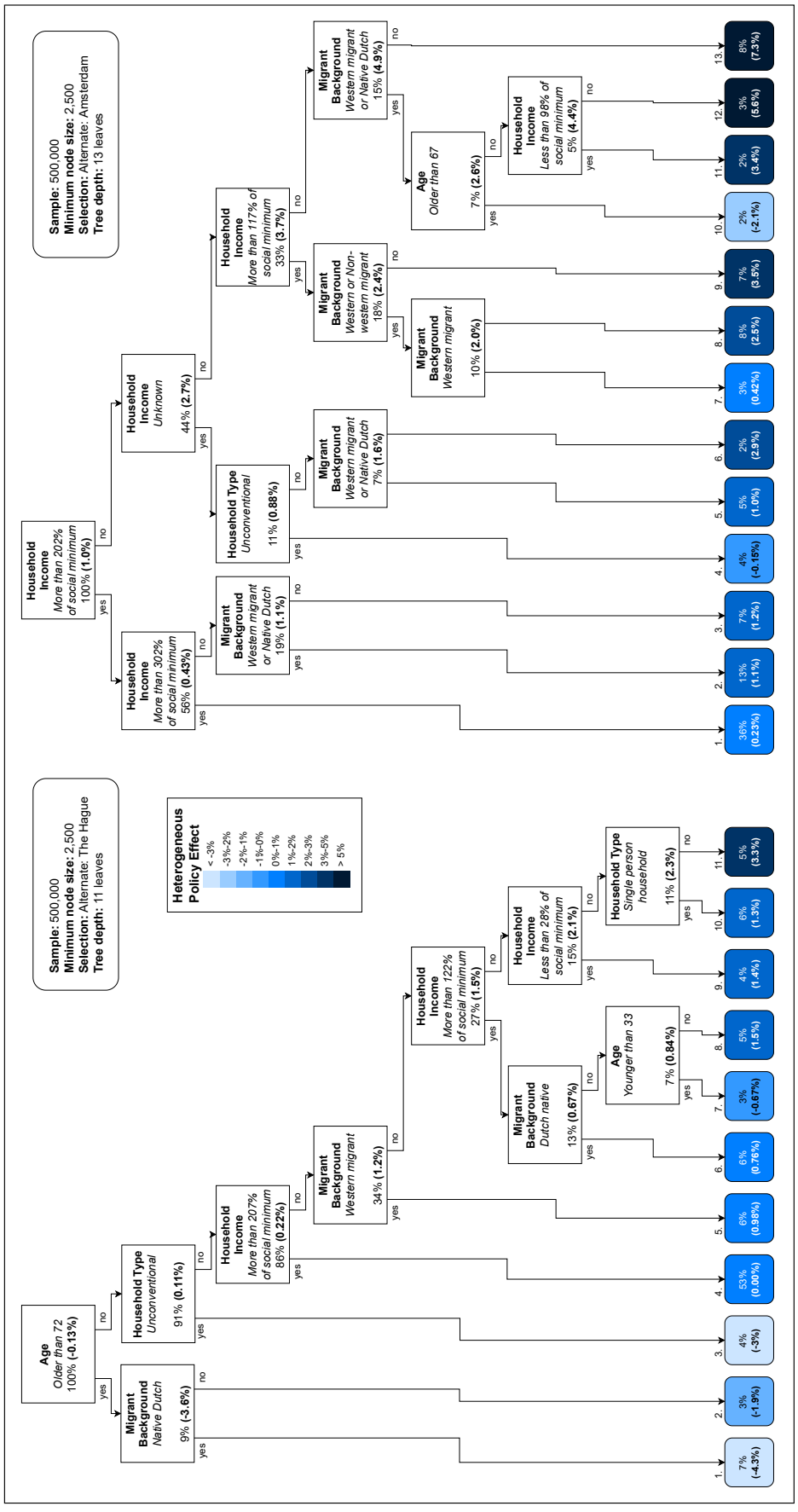


Figure 5.7: Causal tree results based on 500,000 individuals from the The Hague area (left) and the Amsterdam area (right). Causal tree has been pruned at a complexity level of 11 leaves for The Hague and 13 leaves for Amsterdam based on visual inspection of the out-of-sample error. For each node, the percentage of the total sample reaching that node is represented as well as the estimated policy effect in boldface parentheses.

individuals as compared to their native Dutch and Western counterparts. Again, this is contrary to the findings presented earlier. These differences across migrant background might not be surprising, as The Hague is characterized by a much more mixed population than the rest of The Netherlands with more than half of inhabitants having some migrant backgrounds (Table A.5.8). This finding supports the hypothesis that local capture may have played a role in how decentralization has affected different groups in the population.

Amsterdam area: Like the overall sample but in contrast to The Hague, social assistance use in the Amsterdam area has shown an overall expansion. Descriptively, an increase of 1.0 pp is observed with an estimated policy effect of 1.25 pp which is considerably higher than any of the other subsets studied (Table A.5.2). As was the case for the consistent municipalities, the initial splits in the tree follow a strong income effect (right hand side of Figure 5.7). Starting with a split at household incomes of 202% with a further split occurring at incomes of 302%. As was observed in the overall population, effect sizes decline for the richest (+0.23 pp) with more pronounced effects for those between income levels of 202% and 303% (+1.1 pp) and even higher for those below 202% (+2.7 pp). Similar to The Hague, further splits are made on households without conventional income and, as in The Hague, individuals within this group exhibited contractions in use instead of expansions (-0.15 pp).

By far the strongest positive effect sizes occur for the least wealthy partitions of the sample, with household incomes below 202% expanding their use of social assistance by 3.7 pp when excluding the unconventional households — an expansion which further increases when partitioning the sample at an income level of at most 117% of the social minimum (+4.9 pp). Amongst the group of individuals with incomes between 118% and 202% there is a clear gradient by migrant background apparent: the native Dutch have the strongest expansion (+3.5 pp) whereas those with a Non-western migrant background show slightly lower effects (+2.5 pp) and those with a Western migrant background exhibit almost no expansion (+0.42 pp). A reverse trend can be observed in the partition of the population living at or below

the social minimum: those with a Non-western background strongly increased their use of social assistance (+7.3 pp) whereas those with a Western background and natives exhibited a less strong effect (+2.6 pp). Striking is the decline amongst the group of native Dutch individuals living around the social minimum when over the age of 66 (-2.1 pp) as compared with the effect for those below 66 (+4.4 pp). This indicates, again, a reversed age trend relative to the consistent municipalities, albeit for a much more specific partition of the sample than compared to The Hague.

Relative to the findings from The Hague, the Amsterdam area shows slightly more similar heterogeneity with the consistent set, with income the main determinant of effect sizes. Consistent with The Hague is the contraction observed in unconventional households, as well as the stronger expansion of those with a Non-western migrant background and low income, whereas it were mainly the native Dutch low income individuals who expanded their use in the rest of The Netherlands. This could again be a result of the fact that the population in the Amsterdam area is much more mixed than The Netherlands overall in terms of migrant backgrounds (Table A.5.8).

Rotterdam area: In Rotterdam, the descriptive expansion in social assistance use was 0.56 pp with the estimated expansion around 0.77 pp (Table A.5.2). As was the case in Amsterdam and the consistent municipalities, income represents the first partitioning variable (Figure 5.8). However, the initial cutoff point is considerably lower than in any of the other municipalities or the rest of The Netherlands: those over 107% of the social minimum are partitioned into a group that exhibits no meaningful expansion in social assistance use, versus an expansion of 2.9 pp for those at or below the social minimum. Those above 107% are further partitioned into those over 72, which exhibited a contraction of 2.6 pp. For the younger partition of these relatively high income individuals, a second income cutoff is placed at 257% of the social minimum, reflecting a clear income gradient where the wealthier partition shows close to a null effect (+0.14 pp) and those with a lower income exhibiting an expansion of about 1.0 pp, conditional on younger ages.

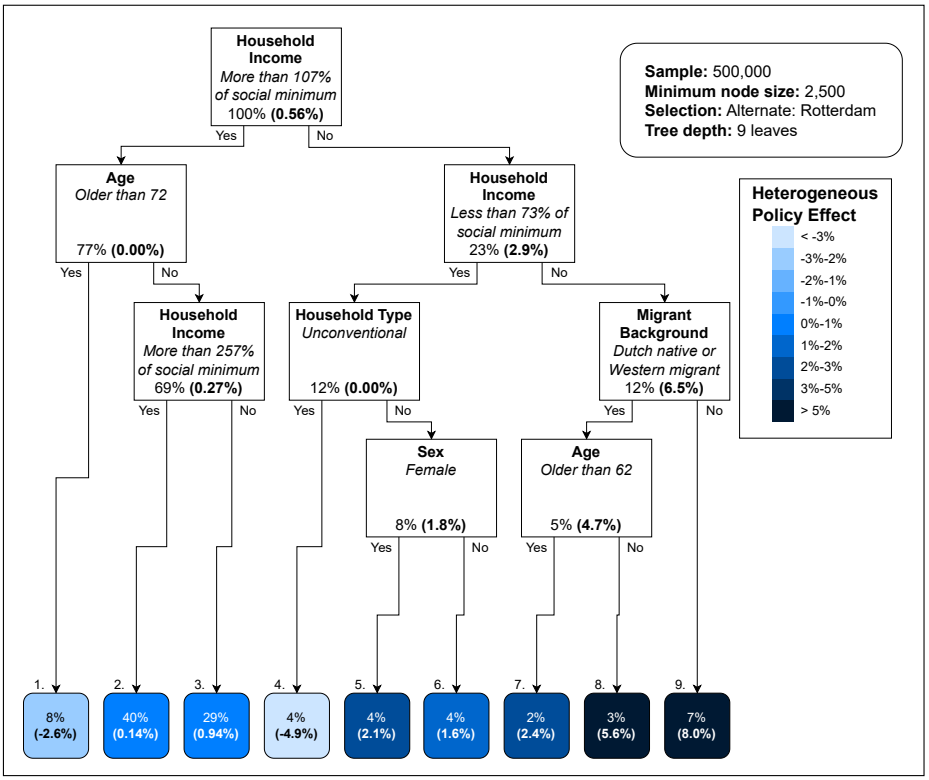


Figure 5.8: Causal tree results estimated to a sample of 500,000 individuals from Rotterdam area. Causal tree has been pruned at a complexity level of 9 leaves based on visual inspection of the out-of-sample error. For each node, the percentage of the total sample reaching that node is represented as well as the estimated policy effect in boldface parentheses.

In general, the causal tree pertaining to Rotterdam is fairly small with only nine leaves identified, although these reflect quite substantial heterogeneity. This is a result of the fairly low initial income cut-off, and could reflect a focused approach of targeting those individuals below the social minimum. Amongst those below the social minimum, unconventional household types are again partitioned off into a group showing a considerable contraction of use (-4.9 pp). A similar finding as was discovered in both The Hague and Amsterdam. In addition, there is another clear stepped income gradient, although interestingly this effect seems to be inverted. Those with income levels between 73% and 108% show an overall effect of 6.5 pp, whereas those below 73% (excluding unconventional households) exhibit an effect of 1.8 pp — an increase in use which is more pronounced for females, at 2.1 pp, than males, at 1.6 pp. Note further that the set of individuals with income levels between 73% and 108% are split into those with a Non-western migrant background

which show increases in social assistance use of 8 pp. Those with a Western migrant background, and native Dutch within this income bound show an age gradient in estimated effects, where those above the age of 62 show markedly less increases in use (+2.4 pp) than those below this age cutoff (+5.6 pp).

As has been the case with all alternate municipalities discussed up until now, a key difference in the Rotterdam area is the strong contraction in use amongst unconventional households, as well as an inverted age gradient. Rotterdam also shows stronger expansions amongst those with a Non-western migrant background at lower income levels than the native Dutch. Notably, most of the expansion in use occurs at incomes around the social minimum rather than at the very lowest parts of the income distribution, and null effects already kick in above relatively low levels of household income.

Almere and Utrecht areas: Both the Almere and Utrecht areas show some similar patterns to the other alternate municipalities (Figure 5.9). Overall effects are lower than the rest of The Netherlands, and there is a clear income gradient present. The Utrecht area is characterized by a contraction amongst older, richer individuals (-1.7 pp) which was also observed in Rotterdam and The Hague. For both areas, unconventional households again exhibit contractions. However, both regions differ from the other alternate municipalities in that there is a less marked increase in social assistance use for those with a Non-western background: they exhibit similar expansions in Utrecht (+2.7 pp for low income Dutch natives versus +3.0 pp for those with a Non-western background), whereas in Almere the strongest effects occur amongst Dutch natives (+3.2 pp) as compared to those with a migrant background (+0.95 pp). Almere is thus more in line with the consistent municipalities in terms of differences between natives and those with a migrant background. This is surprising, as most of the alternate municipalities showing strong increases amongst this group are characterized by a large proportion of inhabitants from a Non-western background. This is also the case in Almere and Utrecht (albeit both to a lesser extent, see Table A.5.8), but this does not

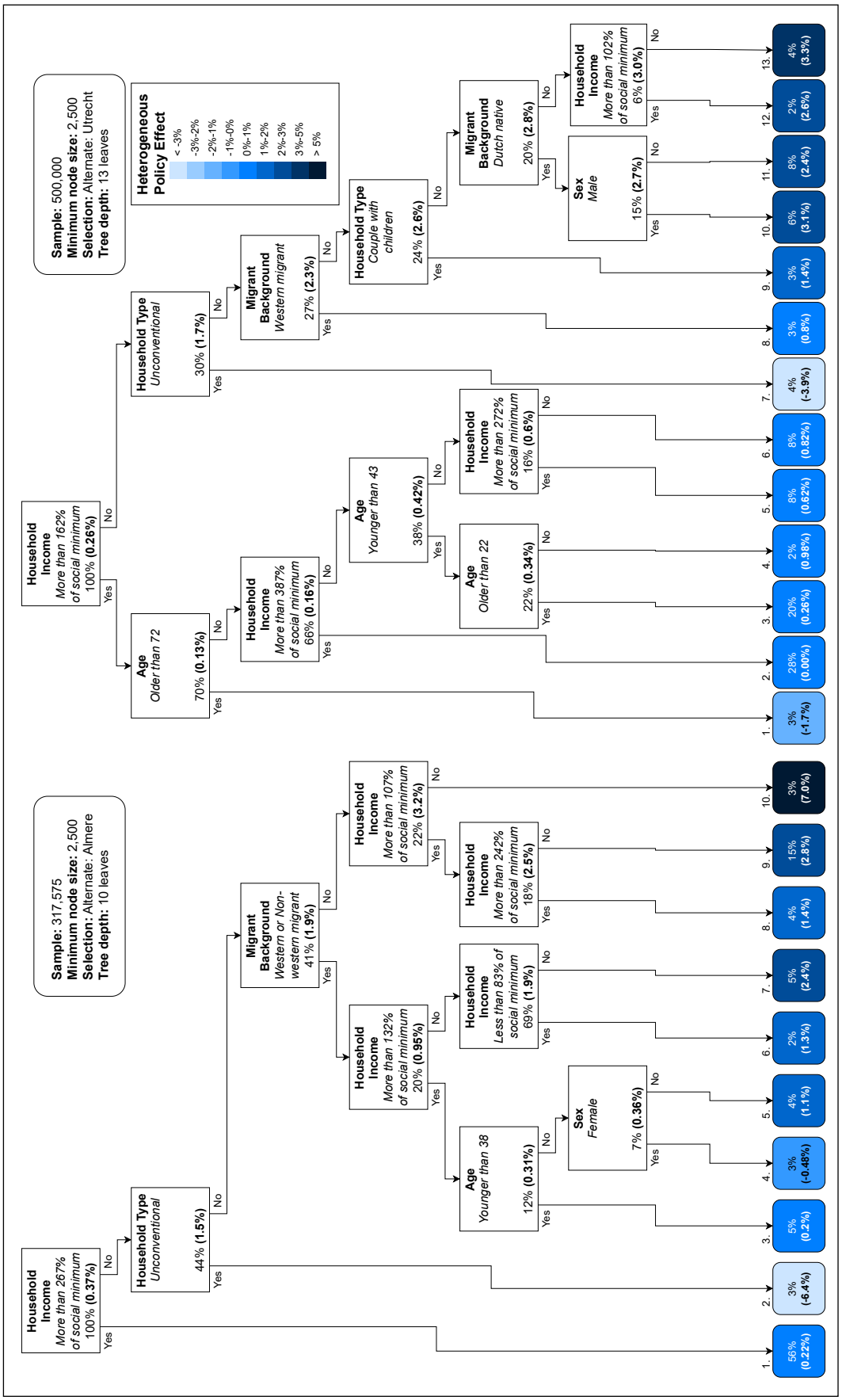


Figure 5.9: Causal tree results based on 317,575 individuals from the Almere area (left) and 500,000 individuals from the Utrecht area (right). Causal tree has been pruned at a complexity level of 10 leaves for Almere and 13 leaves for Utrecht based on visual inspection of the out-of-sample error. For each node, the percentage of the total sample reaching that node is represented as well as the estimated policy effect in boldface parentheses.

result in stronger expansion in social assistance use amongst this group. A possible explanation could be that Almere is a young municipality, having been built and populated in the late seventies and early eighties of the last century. Thus both the migration patterns, as well as the institutional setup of the municipality are different as compared to the other alternate municipalities which are multiple centuries old. Almere is also characterized by a relatively right-wing and anti-immigrant local government compared to the rest of the alternate municipalities.[44]

Comparative insights of heterogeneity across samples

To provide a comparative assessment of the differences in heterogeneity across the six samples under study, I assign each individual their group-level heterogeneous effect and decompose these effects as a function of demographics. I model the effects as a linear additive function which includes a squared term for age and dummies for every 10% increase in household income to allow for non-linearities in income — this seems appropriate given the complex nature of the income effect observed. The full regression results are presented in Table A.5.7, and these are summarized in Figure 5.10.

These results illustrate a couple of things: first of all, as was already apparent from the causal tree analyses, unconventional households have experienced an expansion of social care use in the set of consistent municipalities, whereas a (strong) decline can be observed in alternate municipalities. This has been most pronounced in Almere, Rotterdam and Utrecht and least pronounced in Amsterdam. The reason for this difference might be due to different approaches to those living in institutional households, which are contained within this larger group. For example, urban centres have shifted towards a policy of 'ageing in place', through which municipalities attempt to keep the elderly living in their own homes longer in contrast to living in a care home. This would be consistent with less social

[44] In the 2018 municipal elections, the centre-right People's Party for Freedom and Democracy — the VVD — was re-elected as the largest political party with the far-right, anti-immigrant Freedom Party — the PVV — coming in as the third largest. In Amsterdam, Utrecht, and The Hague, centre-left parties received the most votes. In contrast, Rotterdam has had a right-wing local party, Leefbaar Rotterdam, which has been largest since 2002.

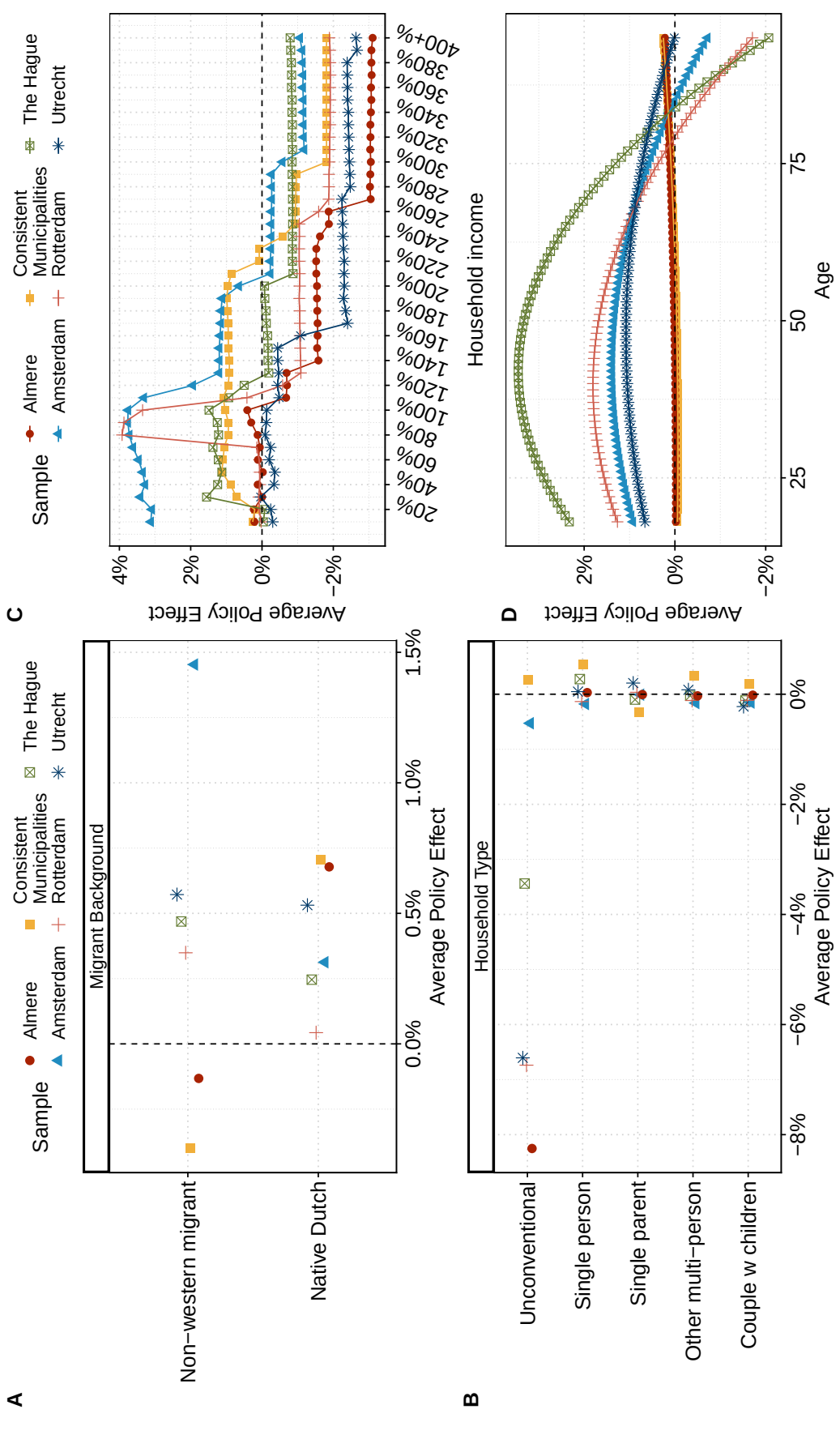


Figure 5.10: Coefficient estimates from linear models relating heterogeneous effects to demographic characteristics for the set of consistent municipalities and the five alternate municipalities. Panel A shows the association with migrant background. Panel B shows the association with household type. Panel C shows the association with income. Panel D shows the association with age (full results can be found in Table A.5.7).

assistance being targeted towards institutionalized individuals which frequently live in care homes. At the same time, institutional households consist of only 50% of the broader group of unconventional households. Additional research will be necessary to better understand the mechanisms underlying these differences in policy effects concerning the set of unconventional households.

Secondly, those with a Non-western migrant background registered lower effect sizes than native Dutch individuals in the set of consistent municipalities. The same pattern can be observed in Almere. For each of the other alternate municipalities we observe stronger expansions amongst those with a Non-western background, in particular in Amsterdam (+1.4 pp versus +0.3 pp for natives as compared to those with a Western background). The effects are about equal for the Utrecht area and show slightly higher effect sizes amongst those with a Non-western background in Rotterdam and The Hague. As mentioned before, it might not be surprising that the alternate municipalities are better able to improve use amongst these groups, as those with a migrant background are represented much more strongly across their populations (Table A.5.8). This finding is consistent with the idea that decentralization should improve connection with local needs, as an overall barrier to access for non-natives was observed prior to the policy intervention. However, it suggests that a considerable proportion of individuals of a certain demographic group are necessary for access to improve. This indicates there might be a degree of local capture apparent. Noteworthy within this reasoning, though, is the converse finding for Almere where those with a Non-western background exhibit lower effects even though the area is equally characterized by considerable proportions of individuals with a Non-western background.

Thirdly, in terms of age we see strongly contrasting effects across the samples. Both Almere and the set of consistent municipalities show a fairly flat age curve, with a slight increase towards older ages. In comparison, all the other municipalities instead show a marked uptake amongst younger individuals with a sharp decline towards older ages. As has been the case with individuals from a migrant background, Almere is more similar to the rest of The Netherlands — a finding which is not

surprising given that the Almere area had the weakest importance scores from the causal forest (Table 5.5). The differences between alternative municipalities and the rest of The Netherlands may be a result of the former actively seeking out needy individuals earlier in the life course, as experimenting with these types of interventions is considerably more common in the larger urban centres.[45]

Finally, we observe a similar stepped decline in effect sizes for higher income groups as compared to lower income groups. However, the inflection points differ across municipalities as well as the relative size of each step. The set of consistent municipalities shows a smaller stepped decrease across the lower parts of the income distribution whereas most other municipalities have stronger differences in effects for the lower end of the distribution. The remarkable increase amongst individuals around the social minimum in Rotterdam discussed above is also apparent from these graphs. Overall, the Amsterdam area seems to have recorded the strongest expansions amongst low income individuals throughout.

Diving deeper into possible reasons behind the observed policy effects, I provide two further sets of analyses. The first provides insights into whether the expansion is driven by changes in the type of care being provided. As can be seen from Figures A.5.1 and A.5.2, there isn't clear evidence for substitution amongst the four types of care that constitute social assistance. At the same time, a number of municipalities strongly decreased care in one type and increased care in another. This could reflect a level of functional specialization and/or difference in approach to social assistance post-intervention. It is outside the scope of this research to study such differences in-depth, but these insights provide interesting avenues for future work in the form of comparative case studies. Evidence for consistent substitution as a possible explanation for the overall increase in use is however limited.

The second set of further analyses is to assess how policy effect sizes relate to the pre-policy use of social assistance. In particular, whether expansions were particularly large in groups with low or high initial use. In Figure A.5.5 it can

[45]For example, the Amsterdam area has multiple 'early signaling teams' used to identify early risk of criminality, debt, poverty, as well as need for care and social assistance (`https://algoritmeregister.amsterdam.nl/vroeg-eropaf/`).

be seen that evidence seems mixed. There is a positive correlation between policy effects and baseline prevalence in the set of consistent municipalities, indicating that those who already made considerable use of social assistance expanded this further. This pattern is much less clear in the alternate municipalities. There, groups with high original prevalence tend to have decreased the strongest although there is little overall consistency apparent. This echoes findings from the causal tree analyses regarding the elderly and unconventional households, which tended to make considerable use of social assistance pre-policy. More broadly, these findings provide further evidence that the more urban and populous alternate municipalities may have approached the decentralization in a more idiosyncratic and perhaps assertative way than the rest of the country. In the latter, it seems like the size of policy effects closely tracks pre-policy prevalences. In the former, more interventional effects can be found which contrast with pre-policy use.

Robustness checks

Even though the causal trees have been estimated using 'honest estimation', meaning that effect sizes have been estimated using data not used to construct the tree, I further assess the robustness of the found heterogeneity for each sample. Specifically, for each causal tree I further perform subset analyses for each of the found groups in the total population instead of the sample used to build the tree. I then compare the estimated effect size in the full population with that found using the sample used to estimate the original effects. The results are shown in Figure A.5.6 and broadly reflect that the found heterogeneity is robust. Almost all the heterogeneous effects that were found via subset analysis on the entire population correlate strongly with those estimated in the causal trees and deviate in a similar way from the overall effect in the population.

In addition, I provide a number of robustness checks regarding some of the assumptions made in this analysis. First of all, the 2016 year was used to reflect social assistance use pre-policy even though the policy was enacted in 2015. This reasoning was based on various reports indicating that the first two years after the

policy change were transitional and not yet characterized by innovations in the way in which social assistance has been granted at the municipality level. To verify this assumption, I estimate the ATE in use between 2015 and 2016. Results are presented in Table A.5.9. As can be seen, there indeed seems to have been barely any empirical expansion between 2015 and 2016 with the ATE estimated to be 0.04 pp for the year, whereas the 2016-2019 period was characterized by an expansion that was almost 25 times as large. It is also noteworthy that the effect is only statistically significant at the 1% level, even though the number of observations used exceeds 10 million. These findings indicate that it indeed seems to be the case that it took time for the 2015 policy to lead to substantive differences in the administration of social assistance, as had also been widely reported.

In a similar vein, I assess pre-2015 trends in social assistance use to evaluate whether the expansion observed in the 2016-2019 period is substantial compared to a similar time window in previous years. Note that full information on social assistance use is not available, as only individuals who paid a personal contribution into the general fund from which social assistance was financed feature in the registry pre-2015 (see Section 5.4). Nonetheless, any expansion or contraction amongst the rest of the population can still be evaluated. The prevalence of social assistance use in the period 2010-2014 is visualized in Figure A.5.7 and indicate consistent use up until the policy change. Use increased by 0.14 pp between 2010 and 2011 and by 0.1 pp between 2011 and 2012 before decreasing again by about 0.1 pp in 2013. Overall, there has never been a yearly expansion or contraction in excess of 0.15 pp.

I also reran the causal tree analysis including data on education. This is not the preferred approach as education data is limited for large parts of the elderly population as almost 40% of adults miss information on education, with gaps centred around those above 60 (see Table A.5.10). Including education does not meaningfully alter the found heterogeneity, with education in some cases replacing cutoffs which were otherwise placed using the income variable. This finding is broadly consistent with earlier results that income and education tend to be highly correlated in The Netherlands (De Graaf and Ganzeboom, 1993) and further

motivates the use of income, which is considerably richer than the three-level education variable available at the registry.[46]

5.7 Discussion & Concluding Remarks

As countries around the world are enacting decentralization policies to reign in escalating costs of care, important questions remain regarding the impact of these interventions on equity. However, heterogeneity in the effect of an intervention — a necessity for equity to change — can be challenging to estimate. Limitations on data and methods have led to relatively simple assessments of equity, typically through aggregate measures that do not provide the level of detail required for a policy's effect on equity to be assessed fully. A HPE approach explicitly models the effects of a policy for sub-groups of the population and is increasingly viable due to the advent of large-scale administrative data and ML methods tailor-made to identify effect heterogeneity in a robust manner.

Substantively, I study the 2015 nationwide decentralization of the social domain in The Netherlands which delegated the administrative responsibility for social assistance from the national to the municipal level. Like the majority of decentralizations, the policy presumed that care would be i) better aligned with local needs when provided by municipalities rather than through a central portal, and ii) more cost-efficient. Concerns regarding the policy's effect on equity were raised at the time, but did not deter policymakers from enacting these sweeping changes. Using registry data, I estimate both the overall expansion / contraction of social assistance use (*RQ1*) and differences in these effects by demographic groups using causal tree and causal forest approaches (*RQ2*).

I find that the period following the policy intervention has been characterized by a substantial expansion in the use of social assistance. Correcting for demographic developments, an overall expansion of about 1.0 pp in use can be observed across adults in The Netherlands. This is considerable, as the initial prevalence of social assistance use had been 6.9% and stable prior to intervention. Combined with the

[46] Causal tree results including education are omitted for brevity.

budgetary cuts accompanying the policy this seems to suggest that efficiency has indeed improved as a result of the decentralization, at least in terms of the number of people making use of social assistance.[47] Some have argued that there may have been a transfer of care from other domains, for instance away from intensive LTC and into the relatively cheaper social assistance domain. In the specific case of LTC, prevalence has been stable and slightly increasing as well, which suggests this particular mechanism might not be driving the observed expansion.[48] That having been said, a complete assessment of the efficiency argument would require possible transfers across domains to be taken into account as well.

In terms of equity, causal forests showed strong evidence for heterogeneity in the policy's effect and identified five municipalities that differed from the rest of The Netherlands in terms of heterogeneity. Notably, these included the four most populous urban centres formerly part of the 'G4', which are characterized by a larger number of civil servants and dedicated innovation teams to deal with the specific challenges arising in the municipalities. Causal trees were estimated to samples of each of these municipalities and the remaining set and found remarkable sources of heterogeneity by demographics. The overall expansion in social assistance use identified across the period under study is almost entirely centred around low income individuals. However, there are nuances across the six samples as well. The alternate municipalities show equal or more considerable expansions in social assistance use amongst those with a Non-western background with the exception of Almere. This is in contrast to the rest of the municipalities in The Netherlands where expansion of use is centred around native Dutch individuals. This is noteworthy as the alternate municipalities are characterized by high proportions of inhabitants with a Non-western background compared to the rest of the country. Furthermore,

[47] Direct comparisons are difficult between the pre- and post-intervention costs of social assistance, as care was effectively divided across three new care arms and accompanying budgets. However, the original legislation was explicitly accompanied with budgetary cuts as a result of expectations that customization would decrease bureaucratic overhead (Allers, 2013; Vonk and Schoukens, 2019, p. 158).

[48] See `https://www.cbs.nl/nl-nl/cijfers/detail/84529NED` for the development of LTC between 2015 and 2019. Descriptively, these indicate an increase in the use of LTC from 1.73% in 2016 to 1.78% in 2019.

unconventional households — either institutional or otherwise characterized by a lack of income — and the elderly recorded contractions in use in the alternate municipalities as compared to expansions in the rest of The Netherlands. More generally, heterogeneity in the latter seems to be positively associated with pre-policy prevalence as compared to more complex patterns in the alternate municipalities. Again, this might reflect more targeted innovations across the more populous and urban alternate municipalities as compared to the rest of the country.

Consistently, though, the strongest determinant of heterogeneity in policy effects has been household income. The much stronger expansion recorded amongst the lowest income levels reflects that equity, at least from an income perspective, has increased as a consequence of the decentralization. As social assistance is meant for those in need and without the necessary means to support themselves, it seems like the decentralization has had the desired effect of reaching an increasing number of individuals who at least qualify for the latter in terms of subsistence. At the same time, equity has decreased ceteris paribus in most of the country for individuals with migrant backgrounds. This is worrying, as barriers to access were already observed prior to policy changes being enacted. That municipalities with large representations of individuals with a migrant background do not exhibit these patterns further suggests that decentralization might foster local capture. National policymakers should thus wonder whether letting go of central control can lead to undesirable consequences to equity, and monitoring the further divergence in access between native Dutch and those with migrant backgrounds is crucial going forward. Further work is needed to qualitatively understand why these differences emerge, with Almere a particularly relevant case study as it exhibits both a large migrant population as well as lower expansion under this demographic group.

Taking a step back, this chapter illustrates how a HPE approach can lead to the identification of sources of heterogeneity that might have been difficult to hypothesize a-priori. By using ML to identify patterns in the data, the level of detail in which a policy's impact can be measured far exceeds that of typical approaches to studying equity. It also provides policymakers with actionable insights into the

consequences of policy interventions on-the-ground, and identifies groups reacting differently to the intervention that could otherwise have been missed. Relatedly, this chapter does not aim to settle the decentralization debate in terms of its impact on equity, although evidence has been found that some of the typical worries like local capture and divergence seem warranted. Instead, this chapter calls for an increased appreciation for the complexity of policy interventions. It might be desirable to make sweeping statements about the general impact of certain classes of policies, but in reality this impact is a result of complicated interactions between the actual policy and the wider context within which it is implemented. If anything, the details and nuances this chapter identifies reflect this complexity and the need for (continuous) empirical investigation into policy effects to guard against unwanted consequences.

Most policy interventions are designed with the best intentions in mind, but the true effects of a policy intervention in a population, when implemented in practice, can hardly ever be predicted theoretically. Both in terms of overall effects, but especially in terms of heterogeneity in these effects across demographic groups. Too often, such heterogeneous effects are glossed over, perhaps in part due to the complexity in estimating them accurately. As this chapter shows, new methods developed in domains like ML — embracing an abundance of computational resources that characterizes this day and age — can provide assistance here. Especially when combined with the increasing availability of large-scale data for research purposes, these methods should become a key tool in the analysis of policy interventions. Hopefully, beyond its substantive findings, this chapter provides a compelling blueprint for how such an approach can work in practice.

Chapter 6

Discussion & Concluding Remarks

With this thesis, I set out to bridge the gap between current empirical sociological work and exciting methods and ideas developed in the past decades that embrace the spectacular increases in computational power available to researchers today. The goal throughout this endeavour has been to use ideas and methods from computational domains like ML to assist quantitative sociologists in generating understanding of the social world. In particular, I have focused on the potential of such methods to help sociologists address three enduring practical issues within empirical work, which have formed the recurring topics to which I have returned throughout this thesis.

The first is that we often ignore how well our models predict the outcomes we are interested in. This focus on $\hat{\beta}$- rather than $\hat{y}$-questions pervades quantitative sociology and means that we might miss important insights from our models. More generally, it precludes a broader understanding of the ability of social theories to actually predict the real world. The second is that most empirical sociological work makes use of relatively simplistic and possibly incorrect functional form assumptions, which date from an era of computational limitations that should no longer constrain our model-building. The widespread prevalence of the linear additive functional form to relate explanatory variables with an outcome is a case in point. The third is that researcher degrees of freedom and p-hacking have led to what some have called a 'crisis in science' and diminished trust in research findings. I've put forward two simple ideas that build on methods and concepts from computational domains like ML, which I believe can help sociologists address these issues. The

first is that prediction is a valuable model analysis tool, which sociologists should implement more frequently in their work. The second is that ML models' ability to identify intricate patterns in data holds immense value for empirical work, especially when we don't know the underlying patterns in the data — a frequent reality in quantitative sociology. Throughout this thesis, I have attempted to illustrate how I believe these two ideas can be valuable for quantitative sociology.

In this process, I've attempted to retain the standard empirical workflow of the sociologist, and to view computational methods as a complement to — rather than a substitute for — existing methods. This is in line with a broader view that ML methods should be incorporated into existing empirical workflows that have interpretable models of the social world as their ultimate goal, from which we can subsequently extract meaningful insights about social life. This complementary mindset should be clear from the various empirical examples I present throughout this thesis. Throughout Chapters 2 and 3, I used prediction to gain additional insight into models that are routinely estimated in sociological work. Even when actively including new modeling methods — as was the case in Chapter 4 — the onus firmly remained on building interpretable models of the world, which any quantitative sociologist would feel comfortable analyzing. This was further evidenced by the fact that in Chapter 5, where a considerable part of the empirical workflow had been delegated to causal trees and causal forests, the identified patterns were still confirmed using standard subset regressions. The empirical examples in this thesis should provide evidence that computational methods can easily be folded into the sociologist's standard workflow, without the need for wholesale changes to how we do research. In other words, the often polarizing question whether 'to explain or to predict' is mostly one regarding the frame of mind of the researcher rather than a fundamental decision with severe implications for our empirical workflow or the type of questions we can ask.

Having established that computational ideas and methods can easily fit into our empirical workflow, the question that remains is what they have to offer to quantitative sociologists. Again, the empirical use cases discussed in this thesis

should provide compelling arguments in favor of their inclusion in empirical work. For instance, Chapters 2 and 3 illustrate the value of assessing model fit at a lower level of aggregation than standard aggregate measures of model fit. Such insights in fit improve our model-checking, as well as our theory-building and the general alignment of our work with the real world. These chapters also illustrate how assessing predictive improvements when including different sets of explanatory variables to our models can lead to substantive insights that are not obvious from standard model analyses. As was illustrated in Chapter 3, prediction can also provide solutions to the trade-off between model interpretability and complexity that pervades sociology, especially when difficult-to-interpret models are estimated. These are all important traits that will improve our understanding and confidence in the models we estimate. On a broader level, prediction re-emphasizes our model's ability to accurately predict the things we are interested in. A welcome rebalancing, away from a monolithic focus on $\hat{\beta}$ and more towards a healthy co-appreciation of $\hat{y}$.

The quality of our empirical models can be further improved when we don't use prediction as a tool to analyze our estimated models, but to benchmark them against alternatives. As shown in Chapter 4, benchmarking allows us to test our own hypothesized models against flexible models like those from the ML domain. Comparing the fit across the two can help guard against the risk of using overly simplistic models to capture social life, which can invalidate our inferences. More generally, the ability of ML methods to find patterns in data without relying on a researcher's hypothesizing can be immensely beneficial in situations where we don't know the true patterns in the data. This was illustrated in Chapter 5, where causal trees and causal forests identified various complex patterns that would have been challenging to identify otherwise.

Last but not least, the most important virtue of computational methods might very well be their ability to improve trust and transparency in the research process. The fundamental risk of computational methods to overfit the data at hand becomes a blessing in disguise, as it makes robustness a central theme throughout computational work. Frequent re-estimation of models on resampled datasets and

providing assessments of the variability of research findings are all indicative of a strong focus on robustness within the computational domain, an emphasis that is less strongly encountered within sociology. Incorporating a similar focus on the stability of findings might be the most important take-away to garner from the computational mindset.

A crucial contextual note to the above is that the three issues I outlined, which computational methods are supposed to help sociologists address, are fundamentally practical in nature. They do not arise from a lack of statistical theory but rather from what happens when we transfer this theory to real-world data. For example, researcher degrees of freedom are irrelevant when researchers fully commit to pre-registration and refrain from attempting multiple specifications on a dataset or, if they do, they choose to use a separate set to report results for the final model or control for multiple hypothesis testing. Similarly, if we know the underlying DGP there is no need to worry about a lack of complexity nor a need to benchmark our models using ML. We can rest assured our models are well-specified. This would also make the use of prediction to diagnose our model's (disaggregated) fit less relevant. However, this is the reality we encounter in stylized examples in methods textbooks. In practice, many of these assumptions are unlikely to hold, as these same textbooks will typically attest. We generally do not know the true model and should expect that it is often more complex than the linear additive models routinely estimated throughout sociology. The large amount of irreproducible research findings in empirical work further suggests that the research process suffers more from researcher degrees of freedom than journal articles convey. It is within this practical context — where we are faced with real-world data of which we know relatively little — that the methods and ideas described in this thesis will be most valuable.

This same practical context is also the very reason that it is difficult to determine theoretical guarantees whether these issues can in fact be solved. There is a reason why many statistical tests rely on correctly specified (families of) functional forms or nesting of models, as this allows for some level of tractability, asymptotic results

and accompanying statistical guarantees. These results don't necessarily transfer very well to situations where we don't know much about the underlying DGP and are unwilling or unable to make many assumptions about it. It is informative here to revisit Breiman's 'two cultures in statistical modeling' which were introduced in Chapter 1. Breiman described two distinct ways of going about statistical modeling; the first assumes that data is generated by some known 'data model' and empirical researchers proceed to estimate its parameters. The second approach is to assume that the mechanisms underlying the data are fundamentally unknown and we can use algorithms to generate understanding into these mechanisms. Breiman goes on to note that the large majority of statisticians subscribe to the former viewpoint — as is the case with most quantitative sociologists — and laments how this viewpoint may *seem* to lead to a better understanding of the world and ever more advanced (statistical) theory, but that this progress is misleading and falls apart once we concede that we typically don't know much about the underlying data model.

Although inconvenient, the second culture is likely more realistic and a state of the world where computational methods like those described in this thesis can be of great assistance. The perhaps uncomfortable prospect conjured up by Breiman's second culture is somewhat similar to that sketched by Duncan Watts in calling for additional scrutiny of the predictive ability of social theories. As he notes, making empirical work both $\hat{\beta}$- as well as $\hat{y}$-relevant will make our work as sociologists considerably harder. Nonetheless, that our theories should be able to make predictions is an inconvenient reality that is difficult to escape. Similarly, the fact that we have a fundamental lack of knowledge of the true processes underlying our data is a similarly inconvenient reality, but one we have to face up to.

In practical terms, embracing Breiman's second culture of statistical modeling means that sociologists will likely find themselves more frequently in situations without the type of clearcut answers we have become used to when assessing the p-values of our statistical tests. A telling example from this thesis is how to decide whether the difference in fit between our own hypothesized model and that of a ML model is sufficiently large to warrant a re-evaluation of the originally hypothesized

functional form (Chapter 4). As mentioned there, this will often become a subjective choice on the part of the researcher. Another such example is how to generate understanding from the ML model that outperforms our own hypothesized model when both rely on the exact same explanatory variables. Although Shapley values are suggested as a solution in this thesis, there are other X-AI methods available as well. The choice of heuristic measure when deciding the level of complexity of a causal tree is another such example, where we will often have to resort to making decisions under uncertainty.

This reliance of computational methods on heuristics and analyses that are more akin to exploratory data analysis reflect the most important limitations of the vision for sociological work presented in this thesis. As computational methods become more common in sociology, we will have to formalize how to deal with the many choices presented to sociologists in trying to understand the processes underlying our data, and in what way to report them in our research. Visualization will be a key tool to convey modeling results from computational methods, similar to the literature on specification curves which also relies heavily on visual representations. Another important avenue will be whether and how to pre-register the types of computational methods that are applied. For example, what models and parameters will be evaluated when benchmarking, as well as the sampling frames used to estimate and evaluate such methods, or even the computer seeds underlying them. These are important elements that have to be decided upon on the road to incorporating computational methods into sociology.

That the nature of empirical modeling will become more heuristic as sociology increasingly adopts computational methods seems without doubt. The bottom line is that this subjectivity stems from the fact that we start to embrace that we live in Breiman's second culture, at least most of the time. The prospect of more heuristic empirical sociological work is thus unrelated to the choice whether to use computational methods or not. Rather, computational methods become a part of the solution to this new reality. As a consequence, a new mindset will be required

to deal with a fundamental uncertainty regarding the processes underlying our data, and computational methods are designed for this exact state of the world.[1]

Realizing the vision for computer age sociology that has been set out in this thesis will require some changes to the way in which we approach sociological questions, but perhaps the most important requirement will be dedicated investments in the ways in which we train future sociologists. Although none of the methods or ideas presented throughout this work require drastic changes to how we do research, it does rely on a number of skills and training that may not yet be part of most graduate degrees in sociology. First, ML methods take a different approach to extracting insights from data than classical statistics do and will require dedicated courses to get sociologists acquainted with ML ideas and concepts. Fortunately, there are many online courses and textbooks that provide excellent introductions to these topics. However, broad adoption will require targeted efforts by sociology departments. Second, libraries and modules for computational methods are developed more commonly in open source programming languages like `R` and `Python`, which rely less on graphical user interfaces than for example `SPSS` or `STATA`.[2] Although sociologists increasingly make use of such open source languages, much empirical work is still done in software that does not incorporate many of the more recent computational advances used in this thesis. For computer age sociology to truly take off, this will have to change as well.

The most important development needed for the methods and ideas presented in this thesis to become more mainstream is an acceptance that once we leave the assumptions as set out in our typical methods textbooks behind, we are at risk of running into issues for which there aren't clearcut solutions. By relaxing ever more assumptions about how the social world operates — like viewing the world in a linear additive manner — sociologists embark on a journey which will seem

[1] An important side note to mention is that embracing computational methods does not mean that uncertainty bounds as we have come accustomed to can no longer be quantified. Computational methods like the bootstrap or other resampling based methods combined with approaches relying on data splitting like honest estimation allow for standard uncertainty estimates to be estimated, even for estimands like forest-based predictions.

[2] Besides open source programming languages typically being more flexible than proprietary software, they have the additional benefit of being considerably more transparent in terms of what happens 'under the hood.'

to be considerably less structured. However, embracing uncertainty can also lead to new types of cooperation and give birth to new approaches to social theory. A connection can be made to one of the classic success stories from the ML domain: that of highly accurate speech and image recognition models. A combination of the widespread availability of large datasets with well-defined performance metrics and a view that we need to learn the underlying structures of interest through diving deeply into the available data has been a key factor in this success story. It allowed researchers from across the world to compare and build on each other's approaches. If we view social processes as fundamentally unknown and set ourselves the common goal of gaining understanding in the structure of social life, perhaps we can realize a similar revolution in terms of progress instead of continuing to formulate $\hat{\beta}$-hypotheses in relative isolation and implementing our own models of the world. At any rate, facing up to the fact that much remains unknown — a view that is crucial to the relevance of computational methods — will require a cultural shift away from Breiman's first culture and towards his second. This is something that has to be done throughout quantitative sociology.

This is not to say that the core sociological toolbox is obsolete. Far from it. As I have been careful to emphasize in this thesis' introductory chapter, the ideas presented throughout this work are designed to help sociologists with just a small part of the broader sociological endeavour: to better formulate, estimate and interpret empirical models of the social world. Arguably, this is the easy part of sociological work. The parts where we theorize about possible social theories and regularities, define the research design required to assess them, or interpret the results of our models and make sense of them within the existing knowledge on a topic all ask more from the purely sociological mind. At the same time, even though the domain on which this thesis focuses might be comparatively small, it is absolutely paramount that it is done well and we can use all the help we can get to ensure that this is the case. The ideas outlined in this thesis, and the broader field of research they are inspired by, provide many such helping hands

and we would do well to bridge the gap that currently exist between them, and conventional tools in quantitative sociology.

Although the barriers to making computer age sociology a reality might be considerable, they are worthwhile to try and overcome. The problems which computational methods can help address are not incidental, but fundamental to the enduring relevance of sociological science. We should embrace tools that can help us combat these issues. We can no longer pretend that the enormous impact which computational methods have already had on so many facets of research provide no value to quantitative sociology. Of course, it is all too easy to find examples where computational methods lead to questionable findings when naively transferred to sociological questions, and based on such incidents we could brazenly dismiss any possible value the field might have for sociology as a result. In fact, there will undoubtedly already be cases of such problematic implementations and likely many more will come. But so will cases of p-hacking, bias due to overly simplistic functional forms and statistically significant but practically irrelevant social theories. We should instead assume computational methods can be relevant to quantitative sociology as well, and direct our efforts towards finding how and when this value is maximized. Hopefully, this thesis provides a compelling example of such an effort, and inspires others to join in.

Appendices

A Chapter appendices

A.1 Chapter 1

[This section is deliberately left blank]

A.2 Chapter 2

[This section is deliberately left blank]

A.3 Chapter 3

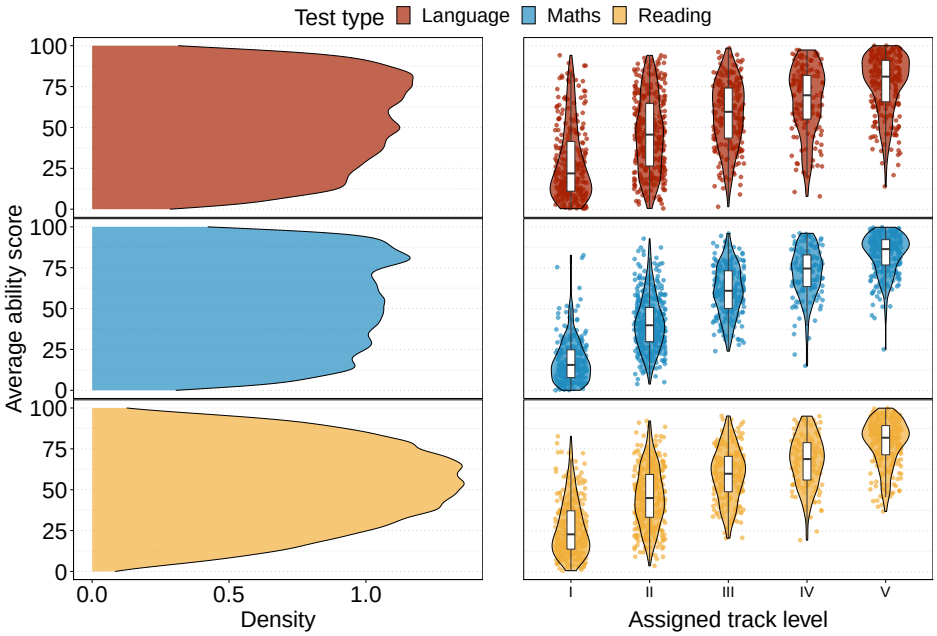


Figure A.3.1: Left panel shows density plots of test scores for each of the three didactic areas included in the sample. For every student, the average percentile test score across all tests made in the final two grades is shown. Right panel shows box-plots indicating average percentile score per didactic area for each track level. A random sample of 2.5% of the dataset is used to illustrate the relationship further through scatter-plots.

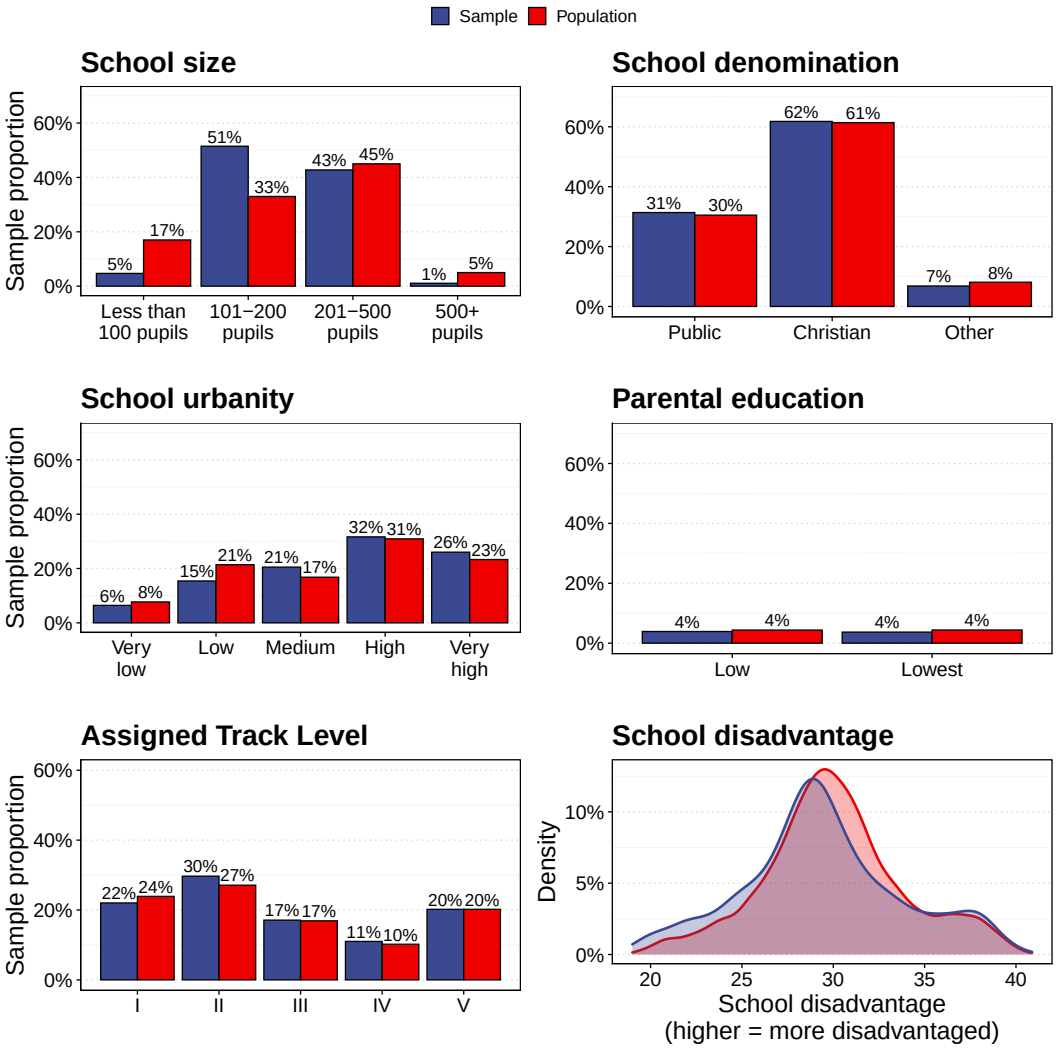


Figure A.3.2: Representativity of the sample relative to the population. Each graph compares the sample used (blue) to the total population in The Netherlands (red). Sources: CBS Statline, DUO, and Onderwijsinspectie CBS (accessed 7th December 2020); Dienst Uitvoering Onderwijs (accessed 10th December 2020); Inspectie van het Onderwijs (accessed 7th December 2020). For each of the plots concerning school-level characteristics, the number of schools in the sample and population is shown. For each of the plots concerning student-level characteristics, the number of students in the sample and population is shown.

	Reading	Maths	Language	Sex: Female	Cohort: 2019	Parent. Educ.: Low	Denom.: Chris-tian	Denom.: Other	Denom.: Public	Track: I	Track: II	Track: III	Track: IV	Track: V	School weight: Low	School weight: Middle	School weight: High
Reading	1.00	0.60	0.67	0.06	-0.02	-0.17	-0.02	0.02	0.01	-0.55	-0.21	0.12	0.20	0.53	0.16	-0.02	-0.15
Maths	0.60	1.00	0.55	-0.14	-0.01	-0.20	-0.01	0.02	0.00	-0.63	-0.25	0.15	0.25	0.59	0.19	-0.02	-0.17
Language	0.67	0.55	1.00	0.04	-0.01	-0.13	0.02	0.01	-0.02	-0.47	-0.18	0.10	0.17	0.46	0.10	-0.01	-0.09
Sex: Female	0.06	-0.14	0.04	1.00	-0.01	0.01	0.01	-0.01	-0.01	0.00	0.01	0.00	-0.00	-0.01	-0.02	0.01	0.01
Cohort: 2019	-0.02	-0.01	-0.01	-0.01	1.00	-0.00	0.01	0.00	-0.01	0.01	0.00	-0.00	0.00	-0.01	0.00	-0.00	0.00
Parent. Educ.: Low	-0.17	-0.20	-0.13	0.01	-0.00	1.00	-0.02	-0.02	0.04	0.22	0.02	-0.07	-0.06	-0.13	-0.16	-0.09	0.26
Denom.: Christian	-0.02	-0.01	0.02	0.01	0.01	-0.02	1.00	-0.35	-0.85	0.00	0.04	0.02	-0.01	-0.06	-0.10	0.10	-0.01
Denom.: Other	0.02	0.02	0.01	-0.01	0.00	-0.02	-0.35	1.00	-0.19	-0.02	-0.01	-0.01	0.02	0.03	0.06	0.01	-0.07
Denom.: Public	0.01	0.00	-0.02	-0.01	-0.01	0.04	-0.85	-0.19	1.00	0.01	-0.04	-0.01	0.01	0.04	0.07	-0.11	0.05
Track: I	-0.55	-0.63	-0.47	0.00	0.01	0.22	0.00	-0.02	0.01	1.00	-0.34	-0.24	-0.18	-0.27	-0.19	-0.00	0.20
Track: II	-0.21	-0.25	-0.18	0.01	0.00	0.02	0.04	-0.01	-0.04	-0.34	1.00	-0.30	-0.23	-0.33	-0.07	0.04	0.03
Track: III	0.12	0.15	0.10	0.00	-0.00	-0.07	0.02	-0.01	-0.01	-0.24	-0.30	1.00	-0.16	-0.23	0.03	0.02	-0.05
Track: IV	0.20	0.25	0.17	-0.00	0.00	-0.06	-0.01	0.02	0.01	-0.18	-0.23	-0.16	1.00	-0.18	0.06	-0.01	-0.05
Track: V	0.53	0.59	0.46	-0.01	-0.01	-0.13	-0.06	0.03	0.04	-0.27	-0.33	-0.23	-0.18	1.00	0.19	-0.04	-0.15
School weight: Low	0.16	0.19	0.10	-0.02	0.00	-0.16	-0.10	0.06	0.07	-0.19	-0.07	0.03	0.06	0.19	1.00	-0.53	-0.48
School weight: Middle	-0.02	-0.02	-0.01	0.01	-0.00	-0.09	0.10	0.01	-0.11	-0.00	0.04	0.02	-0.01	-0.04	-0.53	1.00	-0.49
School weight: High	-0.15	-0.17	-0.09	0.01	0.00	0.26	-0.01	-0.07	0.05	0.20	0.03	-0.05	-0.05	-0.15	-0.48	-0.49	1.00

Table A.3.1: Correlation table of all variables included in the dataset.

A.4 Chapter 4

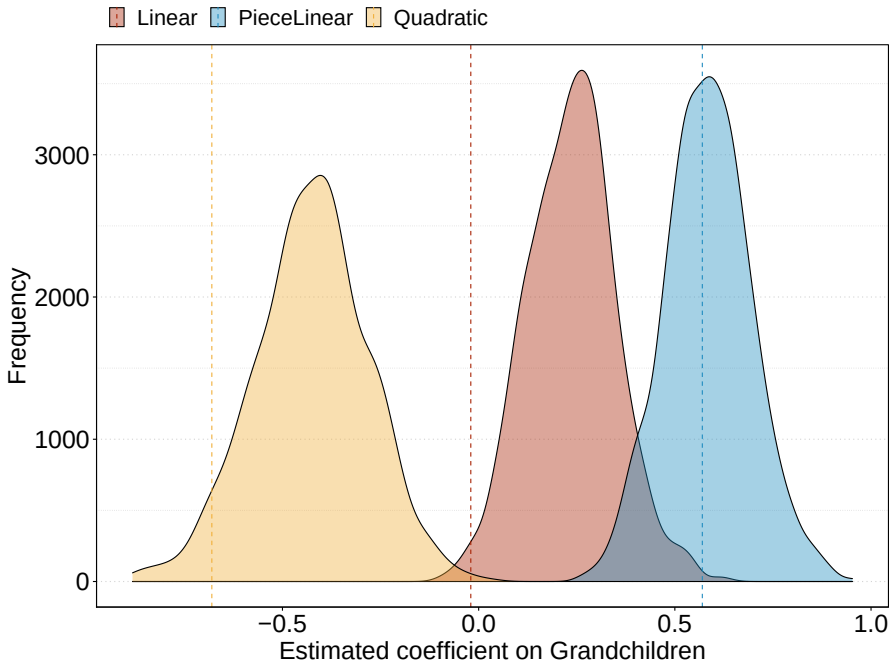


Figure A.4.1: Bootstrapped distributions of the coefficient of interest from Step II of the framework. A natural result of estimating the fit of the hypothesized model over a large amount of resampled sets is that it allows for insights into the variability of the found coefficients across samples. In this case, 80% of the data is randomly sampled 1,000 times to generate these distributions.

Model	Ave	SD	Min	Max
SL.glm_Step	1.218	0.060	1.046	1.320
SL.GB_200_1_0.1	1.229	0.059	1.107	1.327
SL.GB_500_1_0.1	1.231	0.059	1.104	1.323
SL.GB_100_1_0.1	1.241	0.059	1.129	1.351
SL.GB_100_2_0.1	1.242	0.060	1.113	1.342
SL.GB_500_3_0.01	1.247	0.060	1.122	1.357
SL.GB_100_3_0.1	1.252	0.061	1.098	1.349
SL.GB_200_2_0.1	1.253	0.061	1.107	1.354
SL.GB_500_2_0.01	1.254	0.059	1.139	1.364
SL.GB_500_4_0.01	1.259	0.060	1.120	1.364
SL.GB_500_5_0.01	1.268	0.061	1.125	1.364
SL.GB_200_3_0.1	1.268	0.062	1.101	1.351
SL.GB_100_4_0.1	1.271	0.062	1.103	1.362
SL.GB_500_6_0.01	1.275	0.061	1.116	1.376
SL.GB_100_5_0.1	1.279	0.061	1.094	1.366
SL.GB_500_2_0.1	1.284	0.063	1.120	1.367
SL.GB_100_6_0.1	1.286	0.062	1.110	1.363
SL.GB_200_4_0.1	1.292	0.063	1.107	1.381
SL.GB_200_5_0.1	1.297	0.062	1.106	1.385
SL.GB_500_3_0.1	1.302	0.063	1.116	1.393
SL.GB_200_6_0.1	1.313	0.063	1.132	1.394
SL.GB_500_4_0.1	1.330	0.064	1.139	1.419
SL.GB_500_5_0.1	1.341	0.064	1.157	1.440
SL.GB_500_6_0.1	1.357	0.065	1.198	1.443
SL.GB_500_1_0.01	1.366	0.062	1.255	1.503
SL.RF_200_2	1.383	0.064	1.229	1.544
SL.glm_Quadratic	1.384	0.063	1.174	1.536
SL.RF_200_4	1.393	0.066	1.224	1.538
SL.RF_500_4	1.394	0.066	1.220	1.545
SL.RF_500_2	1.398	0.065	1.233	1.578
SL.RF_100_4	1.398	0.066	1.229	1.552
SL.RF_100_2	1.408	0.067	1.250	1.675
SL.glm_Linear	1.413	0.065	1.321	1.569
SL.RF_200_1	2.342	0.098	2.029	2.671
SL.RF_100_1	2.394	0.100	2.106	2.815
SL.GB_200_3_0.01	2.696	0.089	2.477	2.919
SL.GB_200_4_0.01	2.697	0.088	2.483	2.925
SL.GB_200_5_0.01	2.698	0.088	2.484	2.926
SL.GB_200_6_0.01	2.698	0.088	2.486	2.927
SL.GB_200_2_0.01	2.725	0.093	2.488	2.952
SL.GB_200_1_0.01	2.911	0.114	2.567	3.188

Super Learner performance based on RMSE & 10-fold CV.
Selected model in boldface. Hypothesized models in italic.
GB parameters: `n_rounds_max_depth_eta`.
RF parameters: `n_trees_mtry`.

Table A.4.1: Super Learner performance, simulated data.

Model	Ave	SD	Min	Max
SL.Linear	0.12	0.00	0.12	0.12
SL.GB_100_3_0.1	0.12	0.00	0.12	0.13
SL.GB_200_3_0.1	0.12	0.00	0.12	0.12
SL.GB_500_3_0.1	0.12	0.00	0.12	0.12
SL.GB_100_4_0.1	0.12	0.00	0.12	0.12
SL.GB_200_4_0.1	0.12	0.00	0.12	0.12
SL.GB_500_4_0.1	0.12	0.00	0.12	0.12
SL.GB_100_5_0.1	0.12	0.00	0.12	0.12
SL.GB_200_5_0.1	0.12	0.00	0.12	0.12
SL.GB_500_5_0.1	0.12	0.00	0.12	0.12
SL.GB_500_5_0.01	0.13	0.00	0.13	0.14
SL.GB_500_4_0.01	0.13	0.00	0.13	0.14
SL.GB_500_3_0.01	0.14	0.00	0.14	0.14
SL.GB_200_4_0.01	1.07	0.00	1.07	1.08
SL.GB_200_5_0.01	1.07	0.00	1.07	1.08
SL.GB_200_3_0.01	1.08	0.00	1.08	1.09
SL.RF_3_200	0.47	0.00	0.46	0.48
SL.RF_3_500	0.47	0.00	0.46	0.49
SL.RF_1_100	0.48	0.00	0.47	0.50
SL.RF_3_100	0.48	0.00	0.46	0.51
SL.RF_1_200	0.48	0.00	0.45	0.50
SL.RF_1_500	0.48	0.00	0.45	0.49
SL.GB_100_3_0.01	2.90	0.01	2.90	2.92
SL.GB_100_4_0.01	2.90	0.01	2.89	2.92
SL.GB_100_5_0.01	2.90	0.01	2.89	2.91

Super Learner performance based on RMSE & 10-fold CV.
Selected model in boldface. True model in italic.
GB parameters: `n_rounds_max_depth_eta`.
RF parameters: `n_trees_mtry`.

Model	Ave	SD	Min	Max
SL.Linear	0.07	0.00	0.07	0.07
SL.GB_100_3_0.1	0.07	0.00	0.07	0.07
SL.GB_200_3_0.1	0.07	0.00	0.07	0.07
SL.GB_500_3_0.1	0.07	0.00	0.07	0.07
SL.GB_100_4_0.1	0.07	0.00	0.07	0.07
SL.GB_200_4_0.1	0.07	0.00	0.07	0.07
SL.GB_500_4_0.1	0.07	0.00	0.07	0.07
SL.GB_100_5_0.1	0.07	0.00	0.07	0.07
SL.GB_200_5_0.1	0.07	0.00	0.07	0.07
SL.GB_500_5_0.1	0.07	0.00	0.07	0.07
SL.GB_500_3_0.01	0.09	0.00	0.09	0.09
SL.GB_500_4_0.01	0.09	0.00	0.08	0.09
SL.GB_500_5_0.01	0.09	0.00	0.08	0.09
SL.RF_3_100	0.26	0.00	0.25	0.28
SL.RF_1_200	0.26	0.00	0.26	0.28
SL.RF_3_200	0.26	0.00	0.26	0.28
SL.RF_1_100	0.27	0.00	0.25	0.28
SL.RF_1_500	0.27	0.00	0.25	0.28
SL.RF_3_500	0.27	0.00	0.26	0.27
SL.GB_200_3_0.01	0.96	0.00	0.96	0.97
SL.GB_200_4_0.01	0.96	0.00	0.96	0.96
SL.GB_200_5_0.01	0.96	0.00	0.95	0.96
SL.GB_100_3_0.01	2.61	0.00	2.60	2.62
SL.GB_100_4_0.01	2.61	0.00	2.60	2.61
SL.GB_100_5_0.01	2.61	0.00	2.60	2.61

Super Learner performance based on RMSE & 10-fold CV.
Selected model in boldface. True model in italic.
GB parameters: `n_rounds_max_depth_eta`.
RF parameters: `n_trees_mtry`.

Table A.4.2: Super Learner performance, Mincerian wage simulation Linear-I (top) and Linear-II (bottom).

Model	Ave	SD	Min	Max
SL.Linear	0.07	0.00	0.07	0.07
SL.GB_100_3_0.1	0.07	0.00	0.07	0.07
SL.GB_200_3_0.1	0.07	0.00	0.07	0.07
SL.GB_500_3_0.1	0.07	0.00	0.07	0.07
SL.GB_100_4_0.1	0.07	0.00	0.07	0.07
SL.GB_200_4_0.1	0.07	0.00	0.07	0.07
SL.GB_500_4_0.1	0.07	0.00	0.07	0.07
SL.GB_100_5_0.1	0.07	0.00	0.07	0.07
SL.GB_200_5_0.1	0.07	0.00	0.07	0.07
SL.GB_500_5_0.1	0.07	0.00	0.07	0.07
SL.GB_500_5_0.01	0.08	0.00	0.08	0.08
SL.GB_500_4_0.01	0.08	0.00	0.08	0.08
SL.GB_500_3_0.01	0.09	0.00	0.08	0.09
SL.RF_1_100	0.24	0.00	0.23	0.26
SL.RF_3_100	0.24	0.00	0.23	0.24
SL.RF_1_200	0.24	0.00	0.22	0.26
SL.RF_3_200	0.24	0.00	0.23	0.25
SL.RF_1_500	0.24	0.00	0.23	0.25
SL.RF_3_500	0.24	0.00	0.23	0.25
SL.GB_200_3_0.01	0.85	0.00	0.85	0.85
SL.GB_200_4_0.01	0.85	0.00	0.84	0.85
SL.GB_200_5_0.01	0.85	0.00	0.84	0.85
SL.GB_100_4_0.01	2.30	0.00	2.30	2.30
SL.GB_100_3_0.01	2.30	0.00	2.30	2.31
SL.GB_100_5_0.01	2.30	0.00	2.30	2.30

Super Learner performance based on RMSE & 10-fold CV.
Selected model in boldface. True model in italic.
GB parameters: `n_rounds_max_depth_eta`.
RF parameters: `n_trees_mtry`.

Model	Ave	SD	Min	Max
SL.GB_200_3_0.1	0.09	0.00	0.09	0.09
SL.GB_500_3_0.1	0.09	0.00	0.09	0.09
SL.GB_100_4_0.1	0.09	0.00	0.09	0.09
SL.GB_200_4_0.1	0.09	0.00	0.09	0.09
SL.GB_500_4_0.1	0.09	0.00	0.09	0.09
SL.GB_100_5_0.1	0.09	0.00	0.09	0.09
SL.GB_200_5_0.1	0.09	0.00	0.09	0.09
SL.GB_500_5_0.1	0.09	0.00	0.09	0.09
SL.GB_100_3_0.1	0.10	0.00	0.10	0.10
SL.GB_500_5_0.01	0.10	0.00	0.10	0.10
SL.GB_500_4_0.01	0.11	0.00	0.11	0.11
SL.GB_500_3_0.01	0.13	0.00	0.12	0.13
SL.RF_1_100	0.30	0.00	0.29	0.33
SL.RF_3_100	0.30	0.00	0.28	0.32
SL.RF_1_200	0.31	0.00	0.29	0.32
SL.RF_3_200	0.30	0.00	0.29	0.32
SL.RF_1_500	0.30	0.00	0.28	0.33
SL.RF_3_500	0.31	0.00	0.29	0.32
SL.Linear	0.44	0.00	0.44	0.45
SL.GB_200_4_0.01	0.85	0.00	0.85	0.86
SL.GB_200_5_0.01	0.85	0.00	0.84	0.85
SL.GB_200_3_0.01	0.87	0.00	0.86	0.87
SL.GB_100_3_0.01	2.30	0.00	2.30	2.31
SL.GB_100_4_0.01	2.30	0.00	2.29	2.30
SL.GB_100_5_0.01	2.29	0.00	2.29	2.30

Super Learner performance based on RMSE & 10-fold CV.
Selected model in boldface. Hypothesized model in italic.
GB parameters: `n_rounds_max_depth_eta`.
RF parameters: `n_trees_mtry`.

Table A.4.3: Super Learner performance, Mincerian wage simulation Linear-III (top) and Linear-IV (bottom).

Model	Ave	SD	Min	Max
SL.GB_1500_7_0.3	0.227	0.001	0.226	0.228
SL.GB_1750_7_0.3	0.227	0.001	0.224	0.229
SL.GB_1250_6_0.3	0.227	0.001	0.226	0.228
SL.GB_1500_7_0.3	0.227	0.001	0.224	0.230
SL.GB_1250_7_0.3	0.227	0.001	0.227	0.228
SL.GB_1250_6_0.4	0.228	0.001	0.226	0.229
SL.GB_1250_7_0.3	0.228	0.001	0.227	0.229
SL.GB_1000_6_0.4	0.229	0.001	0.227	0.231
SL.GB_1000_7_0.3	0.229	0.001	0.228	0.231
SL.GB_1000_6_0.3	0.230	0.001	0.229	0.231
SL.GB_1250_6_0.5	0.231	0.001	0.229	0.232
SL.GB_1500_8_0.3	0.231	0.001	0.229	0.232
SL.GB_1250_8_0.3	0.231	0.001	0.229	0.232
SL.GB_1500_8_0.3	0.231	0.001	0.228	0.233
SL.GB_1250_7_0.4	0.231	0.001	0.230	0.232
SL.GB_1000_6_0.5	0.231	0.001	0.229	0.233
SL.GB_1000_7_0.4	0.231	0.001	0.230	0.233
SL.GB_1750_8_0.3	0.231	0.001	0.229	0.233
SL.GB_1250_8_0.3	0.231	0.001	0.230	0.233
SL.GB_1000_8_0.3	0.231	0.001	0.230	0.233
SL.GB_1500_7_0.4	0.232	0.001	0.229	0.233
SL.GB_1750_7_0.4	0.232	0.001	0.230	0.234
SL.GB_750_7_0.4	0.232	0.001	0.231	0.234
SL.GB_750_6_0.5	0.233	0.001	0.231	0.234
SL.GB_750_8_0.3	0.233	0.001	0.232	0.235
SL.GB_750_6_0.4	0.233	0.001	0.231	0.235
SL.GB_750_7_0.3	0.233	0.001	0.231	0.235
SL.GB_1250_9_0.3	0.235	0.001	0.233	0.236
SL.GB_1500_9_0.3	0.235	0.001	0.234	0.236
SL.GB_750_7_0.5	0.235	0.001	0.234	0.236
SL.GB_1000_7_0.5	0.236	0.001	0.234	0.237
SL.GB_1250_7_0.5	0.236	0.001	0.235	0.237
SL.GB_1000_8_0.4	0.237	0.001	0.235	0.237
SL.GB_750_8_0.4	0.237	0.001	0.235	0.238
SL.GB_1250_8_0.4	0.237	0.001	0.236	0.238
SL.GB_1500_8_0.4	0.237	0.001	0.236	0.239
SL.GB_750_6_0.3	0.237	0.001	0.236	0.238
SL.GB_1750_8_0.4	0.238	0.001	0.236	0.239
SL.GB_500_6_0.4	0.242	0.001	0.239	0.244
SL.GB_750_8_0.5	0.242	0.001	0.240	0.244
SL.GB_1000_8_0.5	0.243	0.001	0.241	0.244
SL.GB_1250_8_0.5	0.244	0.001	0.242	0.245
SL.GB_500_5_0.4	0.254	0.001	0.252	0.255
SL.RF_500	0.369	0.001	0.366	0.370
SL.GLM	0.421	0.002	0.419	0.423

Super Learner performance based on RMSE & 10-fold CV.
Selected model in boldface. Hypothesized model in italic.
GB parameters: `n_rounds_max_depth_eta`.
RF parameters: `n_trees_mtry`.

Table A.4.4: Super Learner performance, London house price data.

Model	Ave	SD	Min	Max
SL.GB_50_4_0.15	1.86	0.01	1.83	1.89
SL.GB_75_4_0.15	1.86	0.01	1.83	1.89
SL.GB_100_3_0.1	1.86	0.01	1.83	1.89
SL.GB_75_3_0.15	1.86	0.01	1.83	1.89
SL.GB_100_4_0.1	1.86	0.01	1.83	1.89
SL.GB_75_4_0.1	1.86	0.01	1.83	1.89
SL.GB_100_3_0.15	1.86	0.01	1.83	1.89
SL.GB_50_5_0.15	1.86	0.01	1.83	1.89
SL.GB_50_5_0.1	1.86	0.01	1.83	1.89
SL.GB_75_3_0.1	1.86	0.01	1.84	1.89
SL.GB_50_3_0.15	1.86	0.01	1.84	1.89
SL.GB_75_5_0.1	1.86	0.01	1.83	1.89
SL.GB_50_4_0.1	1.86	0.01	1.84	1.89
SL.GB_100_3_0.1	1.86	0.01	1.82	1.91
SL.GB_100_4_0.15	1.86	0.01	1.83	1.89
SL.GB_100_5_0.1	1.86	0.01	1.83	1.89
SL.GB_200_3_0.1	1.86	0.01	1.82	1.90
SL.GB_100_4_0.1	1.86	0.01	1.81	1.91
SL.GB_75_5_0.15	1.86	0.01	1.83	1.89
SL.GB_50_3_0.1	1.86	0.01	1.84	1.89
SL.GB_100_3_0.2	1.86	0.01	1.82	1.91
SL.GB_100_5_0.15	1.86	0.01	1.84	1.90
SL.GB_200_3_0.2	1.87	0.01	1.82	1.91
SL.GB_100_5_0.1	1.87	0.01	1.82	1.92
SL.GB_200_4_0.1	1.87	0.01	1.82	1.91
SL.GB_100_4_0.2	1.87	0.01	1.82	1.91
SL.RF_4_200	1.87	0.01	1.83	1.93
SL.RF_4_100	1.87	0.01	1.83	1.93
SL.RF_2_100	1.87	0.01	1.83	1.93
SL.RF_4_500	1.87	0.01	1.83	1.93
SL.RF_2_500	1.87	0.01	1.83	1.93
SL.RF_2_200	1.87	0.01	1.83	1.93
SL.GB_200_5_0.1	1.87	0.01	1.83	1.93
SL.GB_100_5_0.2	1.87	0.01	1.83	1.92
SL.GB_200_4_0.2	1.87	0.01	1.83	1.92
SL.GB_200_5_0.01	1.89	0.01	1.84	1.94
SL.GB_200_4_0.01	1.89	0.01	1.84	1.95
SL.GB_200_5_0.2	1.89	0.01	1.85	1.95
SL.GB_200_3_0.01	1.90	0.01	1.84	1.95
SL.GLM	1.92	0.01	1.90	1.95
SL.GB_100_5_0.01	2.04	0.02	1.98	2.10
SL.GB_100_4_0.01	2.05	0.02	1.98	2.11
SL.GB_100_3_0.01	2.05	0.02	1.98	2.11

Super Learner performance based on RMSE & 10-fold CV.
Selected model in boldface. Hypothesized model in italic.
GB parameters: `n_rounds_max_depth_eta`.
RF parameters: `n_trees_mtry`.

Table A.4.5: Super Learner performance, GSS data.

A.4.1 Simulations

Non-linear DGP's

Below the four DGPs used to generate the true data in Figures 4.1 and 4.3 are provided:

$$y_{\text{linear}} = -1 + 0.9X + \epsilon_1 \tag{A.1}$$

$$y_{\text{polynomial}} = -6 + 0.4X - 0.36X^2 + 0.005X^3 + \epsilon_2 \tag{A.2}$$

$$y_{\text{sine}} = 2.83\text{sin}(\frac{\pi}{2X}) + \epsilon_3 \tag{A.3}$$

$$y_{\text{step}} = -2\text{I}(X > 0) + 4\text{I}(X > 2) - \text{I}(X > 3) + \epsilon_4 \tag{A.4}$$

Where $X \sim \text{Unif}(-4, 4)$ and each error term normally distributed and calibrated in such a way that the total variance of each outcome is 5.

Mincerian wage equation simulation

Below the four DGPs used to generate the four Mincerian wage outcomes in Figures 2.3, 4.6 and 4.7 are provided:

Linear-I

$$\begin{aligned} \ln(\text{wages}) &= 4.5 + 0.125x_{\text{years_educ}} + 0.09x_{\text{years_exp}} + \epsilon \\ &\epsilon \sim \mathcal{N}(0, 0.07) \end{aligned} \tag{A.5}$$

Linear-II

$$\begin{aligned} \ln(\text{wages}) &= 4.5 + 0.125x_{\text{years_educ}} + 0.09x_{\text{years_exp}} - 0.001x_{\text{years_exp_squared}} + \epsilon \\ &\epsilon \sim \mathcal{N}(0, 0.07) \end{aligned} \tag{A.6}$$

Linear-III

$$\begin{aligned}\ln(\text{wages}) \quad = \quad & 4.5 + 0.02x_{\text{years_educ_0_8}} + 0.03x_{\text{years_educ_9_10}} + 0.3x_{\text{years_educ_11_12}} + \\ & 0.06x_{\text{years_educ_13_14}} + 0.06x_{\text{years_educ_15+}} + 0.09x_{\text{years_exp}} + \\ & 0.001x_{\text{years_exp_squared}} + \epsilon \\ & \epsilon \sim \mathcal{N}(0, 0.07)\end{aligned} \tag{A.7}$$

Linear-IV

$$\begin{aligned}\ln(\text{wages}) \quad = \quad & I(x_{\text{sex}} = \text{Female})[3.5 + 0.025x_{\text{years_educ_0_8}} + 0.06x_{\text{years_educ_9_10}} + \\ & 0.35x_{\text{years_educ_11_12}} + 0.06x_{\text{years_educ_13_14}} + 0.06x_{\text{years_educ_15+}} + \\ & 0.1x_{\text{years_exp}} - 0.0005x_{\text{years_exp_squared}} + \epsilon] + \\ & I(x_{\text{sex}} = \text{Male})[5.5 + 0.015x_{\text{years_educ_0_8}} + 0.02x_{\text{years_educ_9_10}} + \\ & 0.25x_{\text{years_educ_11_12}} + 0.04x_{\text{years_educ_13_14}} + 0.04x_{\text{years_educ_15+}} + \\ & 0.06x_{\text{years_exp}} - 0.001x_{\text{years_exp_squared}} + \epsilon^*] \\ & \epsilon \sim \mathcal{N}(0, 0.09) \\ & \epsilon^* \sim \mathcal{N}(0, 0.08)\end{aligned} \tag{A.8}$$

(A.9)

A.5 Chapter 5

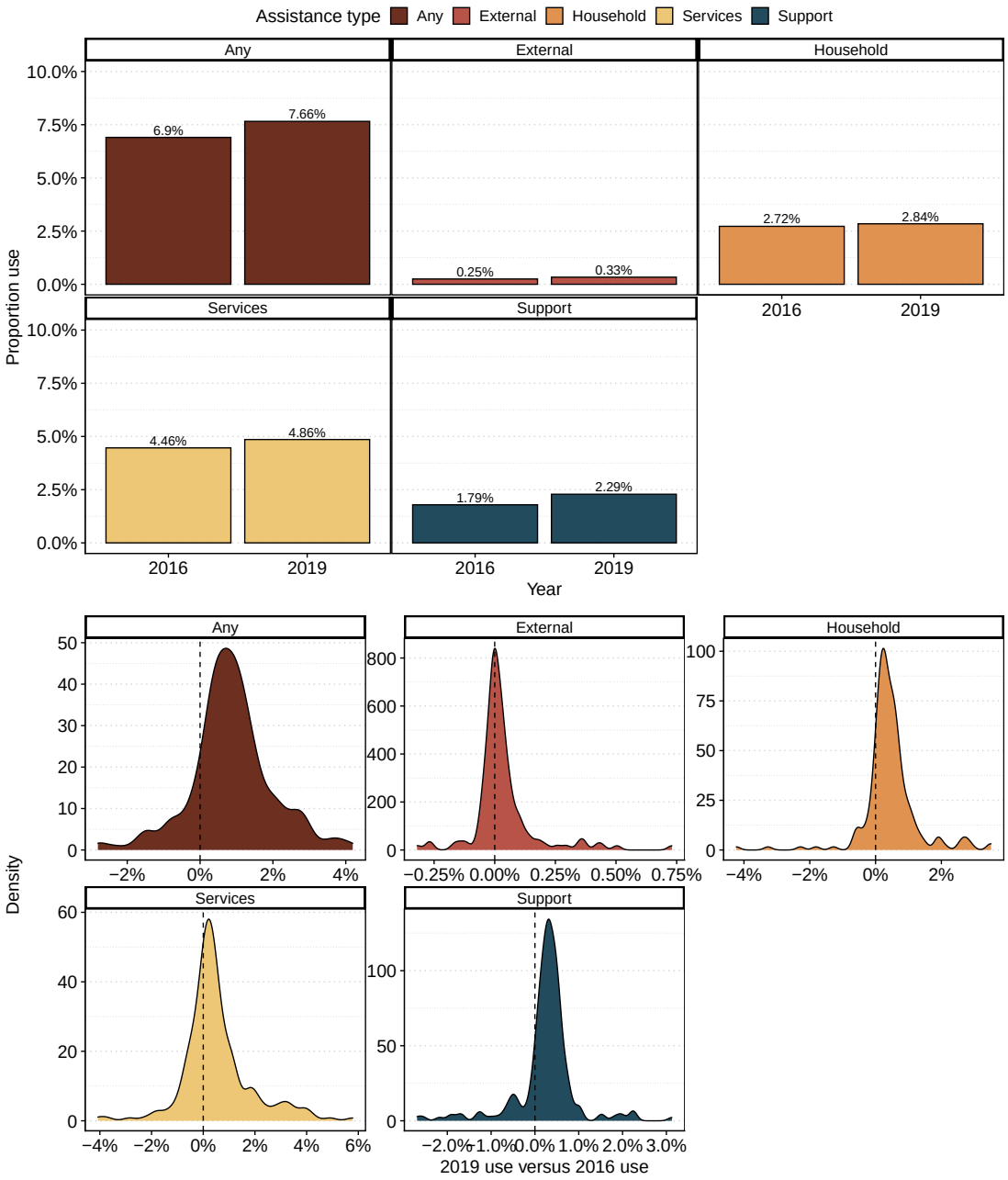


Figure A.5.1: Overall social assistance use and by type in 2019 and 2016 for The Netherlands as a whole (top) and difference by type of social assistance across municipalities (bottom).

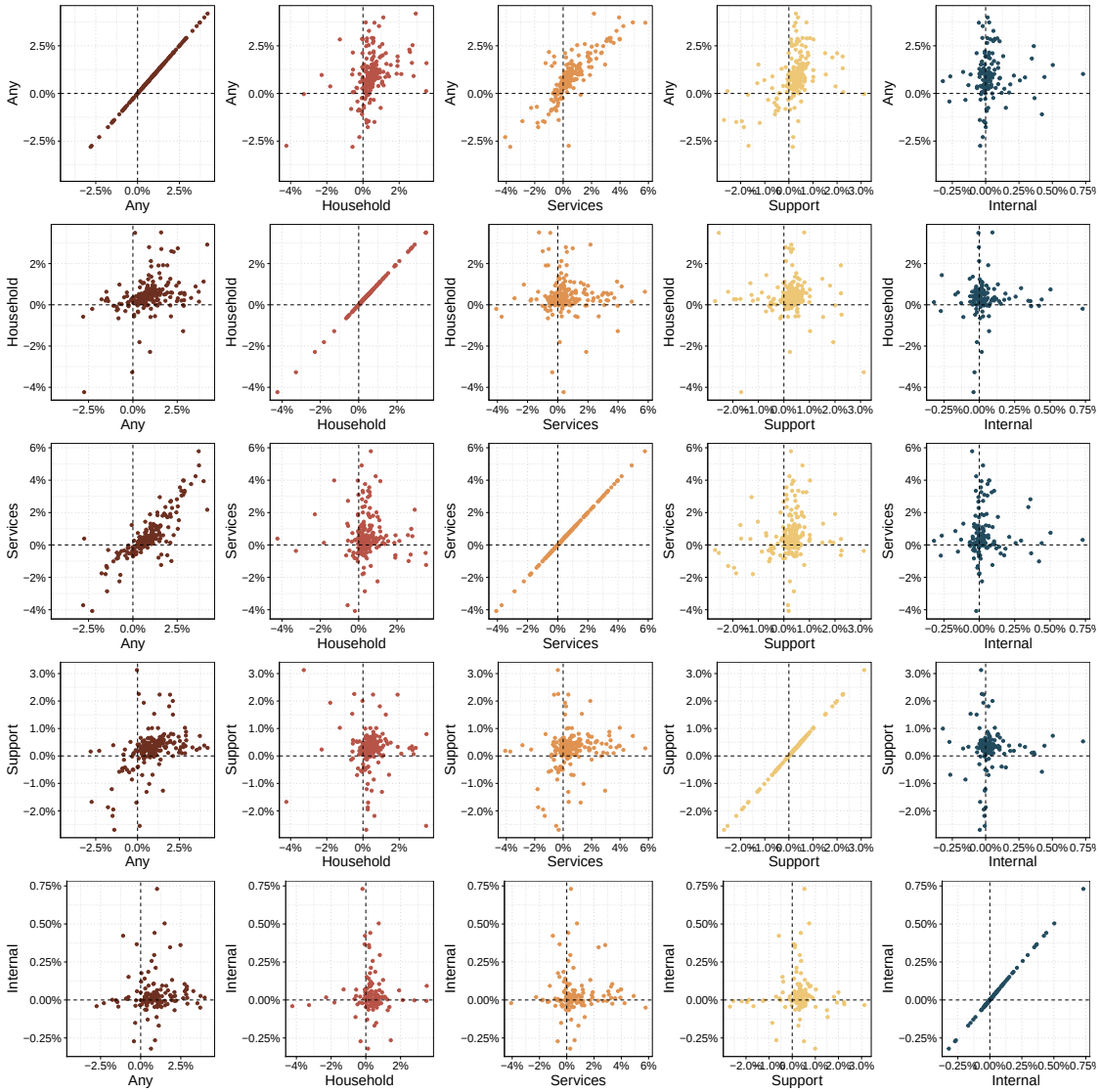


Figure A.5.2: Correlation plots for all municipalities between the difference in various types of social assistance use between 2019 and 2016.

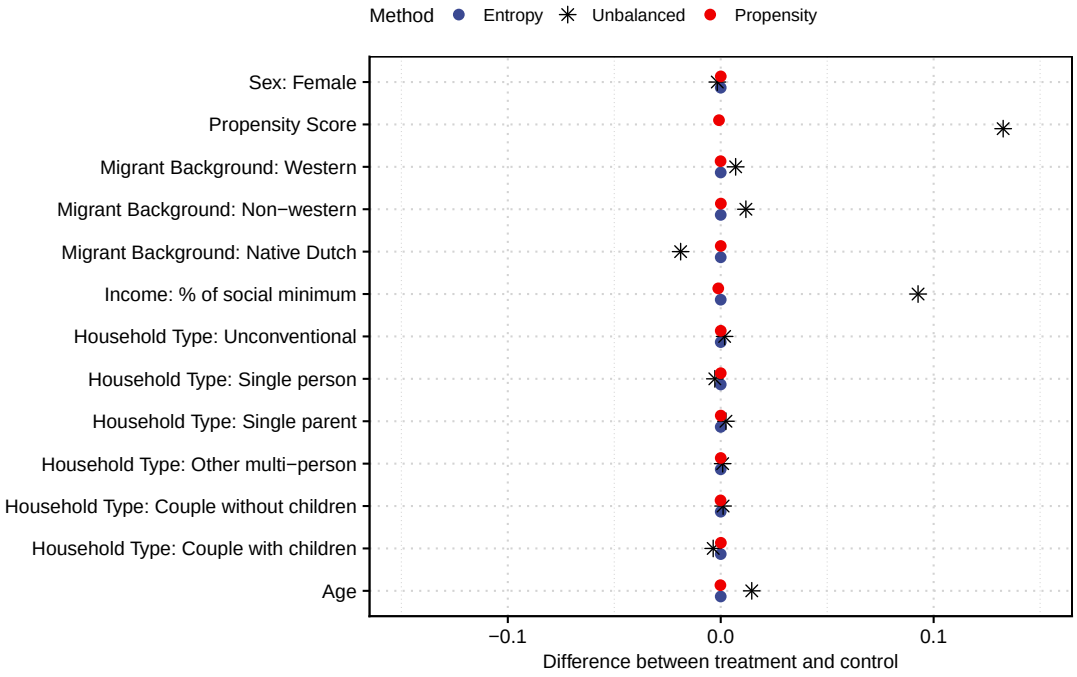


Figure A.5.3: Covariate balance before and after balancing using propensity score matching (blue) and entropy balancing (red). Unadjusted balance is shown with the black asterisks.

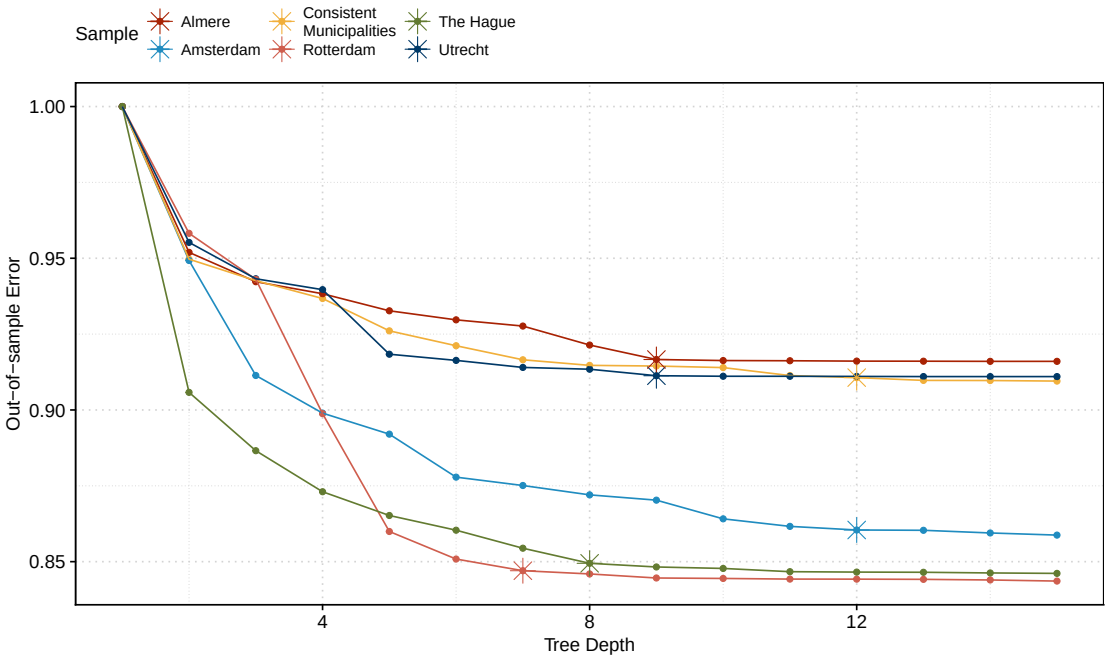


Figure A.5.4: Out-of-sample error of causal trees estimated as a function of tree depth. Asterisks indicate selected depth.

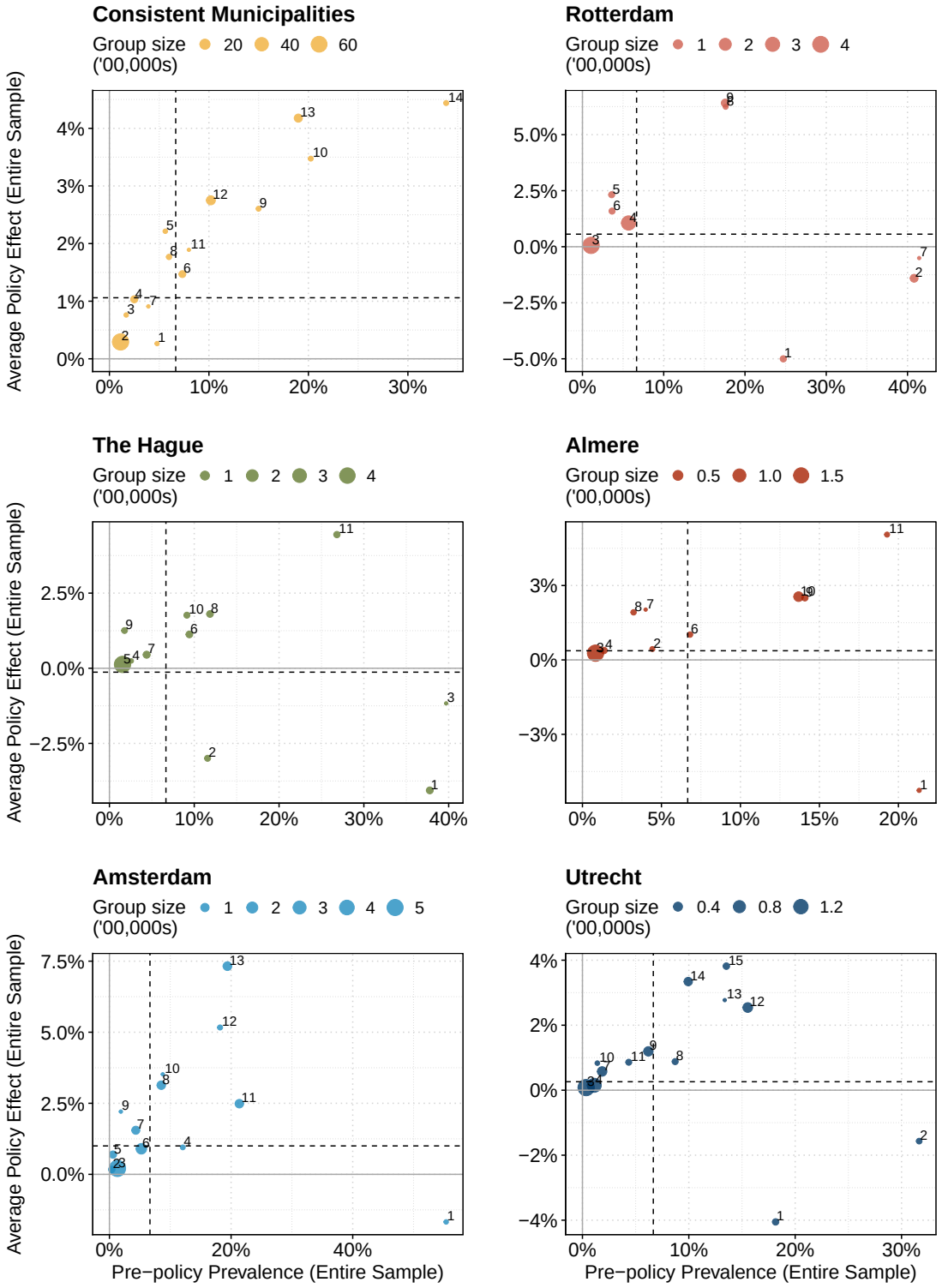


Figure A.5.5: Estimated policy effect for each of the identified heterogeneous groups in the entire sample on the y-axis, compared with the prevalence in 2016 on the x-axis. Each group is numbered in line with the tree representations presented in Chapter 5. The vertical dashed line illustrates the overall baseline prevalence and the horizontal dashed line illustrates the overall policy effect.

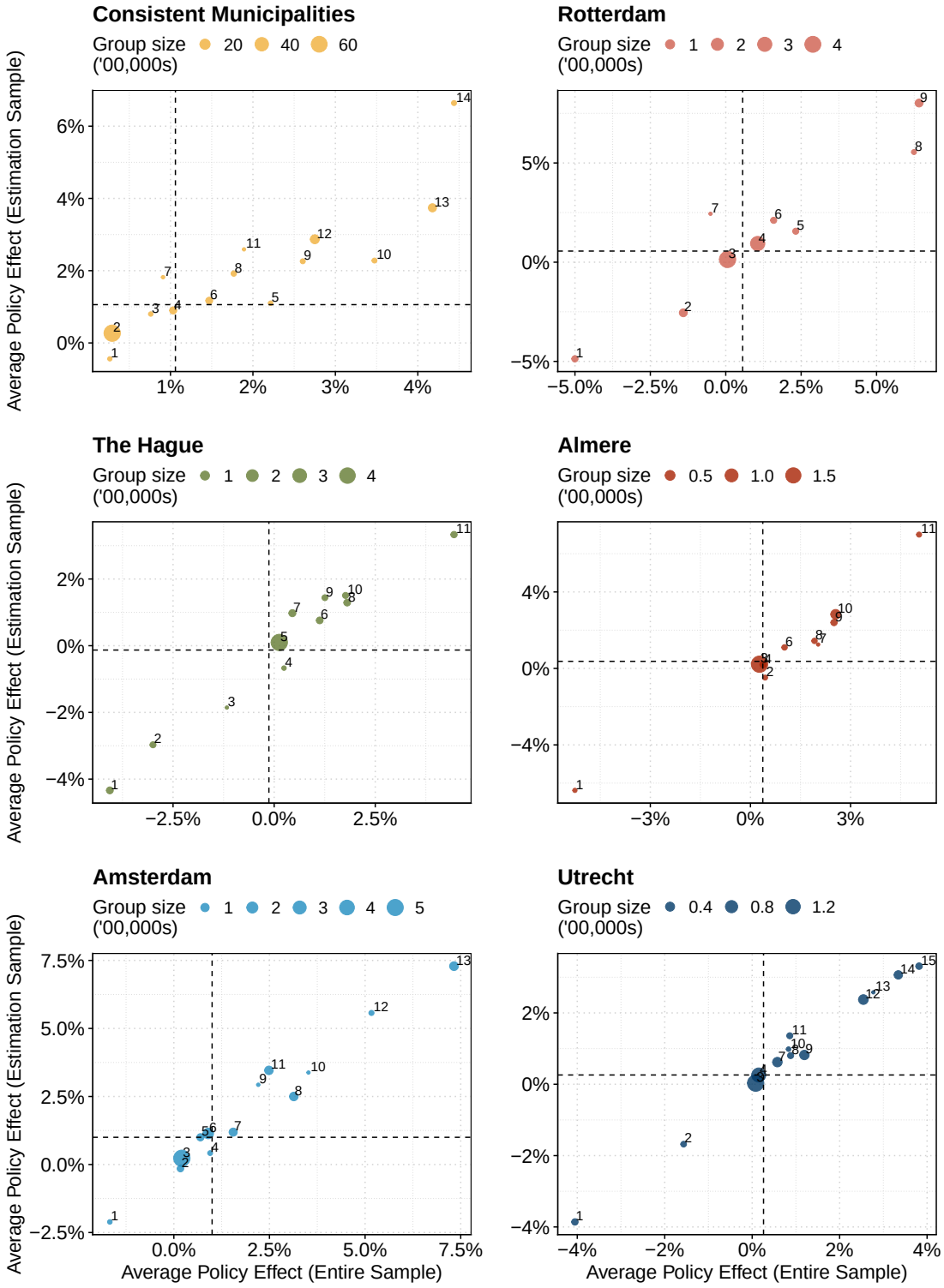


Figure A.5.6: Estimated policy effect based on the sample used to estimate the causal tree on the y-axis for each of the identified heterogeneous groups compared with the observed policy effect for that group in the entire sample on the x-axis. Each group is numbered in line with the tree representations presented in Chapter 5. The dashed lines illustrate the descriptive effect in the entire sample.

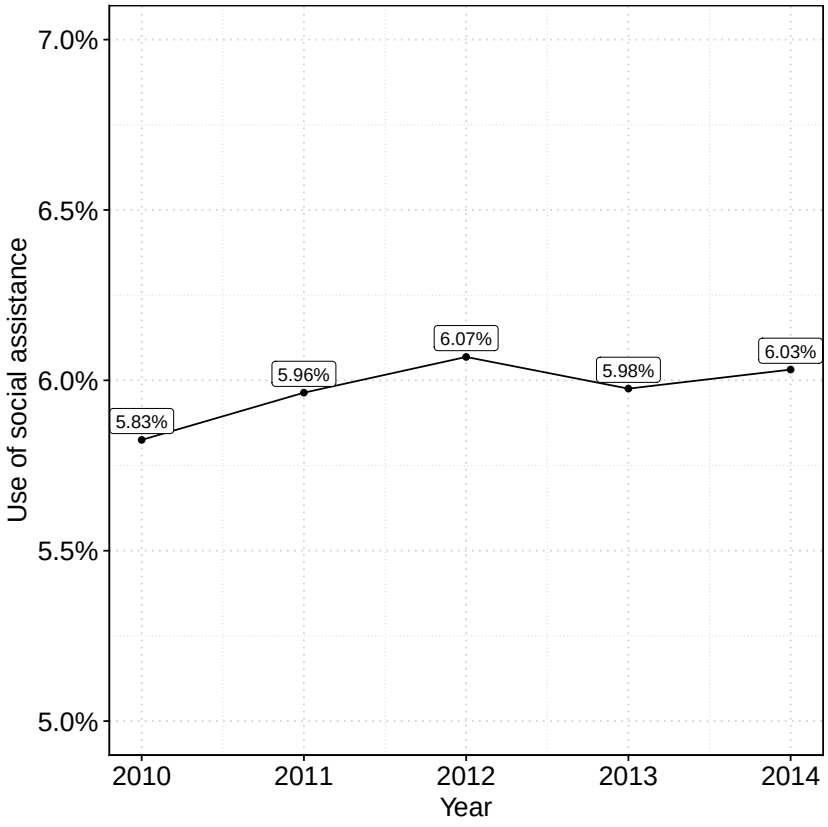


Figure A.5.7: Pre-2015 trends of social assistance use based on AWBZ data pertaining to those individuals making use of social assistance who also contributed to the general fund.

	2016		2019	
Group	N	Use	N	Use
Sex: Male	4,376,845	0.0507	4,522,820	0.0580
Sex: Female	4,517,590	0.0868	4,647,525	0.0946
Migrant background: Native Dutch	6,685,005	0.0711	6,720,975	0.0790
Migrant background: Non-western	1,227,535	0.0654	1,375,255	0.0762
Migrant background: Western	981,895	0.0593	1,074,110	0.0617
Household type: Single parent	577,040	0.0621	618,645	0.0718
Household type: Single person	1,953,185	0.1611	1,989,220	0.1765
Household type: Unconventional	265,500	0.2215	289,455	0.2057
Household type: Other multi-person	283,565	0.0459	302,610	0.0488
Household type: Couple w children	3,146,960	0.0150	3,211,205	0.0179
Household type: Couple w/o children	2,668,180	0.0541	2,759,205	0.0634
Age: 18-24	1,126,165	0.0264	1,163,130	0.0318
Age: 25-34	1,303,735	0.0236	1,387,990	0.0290
Age: 35-44	1,551,535	0.0273	1,519,865	0.0332
Age: 45-54	1,489,545	0.0354	1,456,510	0.0411
Age: 55-64	1,565,660	0.0520	1,645,560	0.0589
Age: 65-74	1,016,080	0.0880	1,086,090	0.0974
Age: 75-84	650,570	0.2751	708,145	0.2751
Age: 85-94	175,850	0.5627	185,965	0.5716
Age: 95-104	15,260	0.6212	17,060	0.6401
Household Income: -12,5% - 12,5% of soc. min.	616,495	0.1067	622,650	0.1133
Household Income: 12,5% - 37,5% of soc. min.	12,675	0.0722	14,115	0.1860
Household Income: 37,5% - 62,5% of soc. min.	24,715	0.0621	25,280	0.0864
Household Income: 62,5% - 87,5% of soc. min.	74,105	0.1105	68,155	0.1498
Household Income: 87,5% - 112,5% of soc. min.	621,220	0.2539	597,070	0.2926
Household Income: 112,5% - 137,5% of soc. min.	473,915	0.2173	449,805	0.2463
Household Income: 137,5% - 162,5% of soc. min.	472,370	0.1522	442,240	0.1773
Household Income: 162,5% - 187,5% of soc. min.	517,445	0.1100	477,040	0.1326
Household Income: 187,5% - 212,5% of soc. min.	522,670	0.0721	483,025	0.0920
Household Income: 212,5% - 237,5% of soc. min.	547,270	0.0514	511,480	0.0668
Household Income: 237,5% - 262,5% of soc. min.	554,955	0.0371	530,330	0.0492
Household Income: 262,5% - 287,5% of soc. min.	552,455	0.0289	539,815	0.0391
Household Income: 287,5% - 312,5% of soc. min.	528,290	0.0211	524,435	0.0286
Household Income: 312,5% - 337,5% of soc. min.	498,770	0.0167	505,835	0.0219
Household Income: 337,5% - 362,5% of soc. min.	456,800	0.0136	476,530	0.0174
Household Income: 362,5% - 387,5% of soc. min.	403,575	0.0114	437,535	0.0145
Household Income: 387,5% - 412,5% of soc. min.	349,140	0.0099	388,360	0.0126
Household Income: 412,5% - 437,5% of soc. min.	295,700	0.0097	338,270	0.0116
Household Income: 437,5% - 462,5% of soc. min.	244,470	0.0082	287,760	0.0105
Household Income: 462,5% - 487,5% of soc. min.	198,235	0.0078	240,940	0.0094
Household Income: 487,5% - 512,5% of soc. min.	161,205	0.0070	199,255	0.0088
Household Income: 512,5% - 537,5% of soc. min.	129,820	0.0069	163,025	0.0084
Household Income: 537,5% - 562,5% of soc. min.	103,420	0.0064	133,665	0.0083
Household Income: 562,5% - 587,5% of soc. min.	83,260	0.0061	109,235	0.0074
Household Income: 587,5% - 612,5% of soc. min.	65,855	0.0063	87,955	0.0073
Household Income: 612,5% - 637,5% of soc. min.	53,235	0.0060	71,825	0.0068
Household Income: 637,5% - 662,5% of soc. min.	43,065	0.0055	58,305	0.0063
Household Income: 662,5% - 687,5% of soc. min.	36,000	0.0053	48,125	0.0070
Household Income: 687,5% - 712,5% of soc. min.	29,285	0.0048	40,070	0.0066
Household Income: 712,5% - 737,5% of soc. min.	25,095	0.0046	33,180	0.0065
Household Income: 737,5% - 762,5% of soc. min.	20,995	0.0050	28,240	0.0060
Household Income: 762,5% - 787,5% of soc. min.	17,795	0.0048	23,795	0.0055
Household Income: 787,5% - 812,5% of soc. min.	15,300	0.0049	20,205	0.0049
Household Income: 812,5% - 837,5% of soc. min.	13,080	0.0046	17,170	0.0055
Household Income: 837,5% - 862,5% of soc. min.	11,610	0.0047	15,130	0.0056
Household Income: 862,5% - 887,5% of soc. min.	10,085	0.0055	13,330	0.0060
Household Income: 887,5% - 912,5% of soc. min.	8,770	0.0046	11,580	0.0047
Household Income: 912,5% - 937,5% of soc. min.	7,810	0.0051	10,080	0.0055
Household Income: 937,5% - 962,5% of soc. min.	6,830	0.0037	8,950	0.0056
Household Income: 962,5% - 987,5% of soc. min.	6,190	0.0040	8,035	0.0050
Household Income: 987,5% - 1012,5% of soc. min.	80,465	0.0044	108,510	0.0051

Table A.5.1: Proportion of social assistance use between 2016 and 2019 by univariate demographic characteristics.

	Consistent set		Almere		Utrecht		The Hague		Rotterdam		Amsterdam	
Method	Estimate	S.E.	Estimate	S.E.	Estimate	S.E.	Estimate	S.E.	Estimate	S.E.	Estimate	S.E.
Covariate balancing	0.01059	0.00013	0.00619	0.00078	0.00533	0.00058	0.00183	0.00055	0.00770	0.00051	0.01252	0.00042
Propensity matching	0.01057	0.00014	0.00619	0.00083	0.00533	0.00063	0.00183	0.00059	0.00768	0.00056	0.01251	0.00047
Entropy balancing	0.01060	0.00014	0.00623	0.00083	0.00533	0.00063	0.00185	0.00059	0.00770	0.00056	0.01250	0.00047
N	13,872,919		317,575		561,434		851,604		1,040,025		1,421,221	

Table A.5.2: Estimated ATEs using various estimation methods applied to the set of consistent municipalities as well as to each of the alternate municipalities.

Variable	Estimate	S.E.	T value	P-value
Mean	0.962	0.0219	43.994	$< 2.2e^{-16}$
Difference	1.937	0.0467	41.484	$< 2.2e^{-16}$

Omnibus test of heterogeneity (Wager and Athey, 2018).
Significance of the `Difference` term reflects rejection of the null of no heterogeneity.
Estimates close to 1 of the `Mean` term reflects accuracy in estimating the overall effect.

Table A.5.3: Omnibus test of heterogeneity in the overall policy effect.

Group	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Age	45.71	34.13	42.25	51.53	56.23	51.73	34.55	65.99	67.65	39.74	48.98	58.42	46.14	57.55
Sex	0.48	0.46	0.51	0.51	0.57	0.00	1.00	1.00	0.50	0.75	0.50	0.62	0.51	0.58
income: % of soc. min.	441.45	-1.66	164.44	247.19	141.14	265.38	266.20	264.21	170.26	144.10	225.17	98.97	-0.67	164.01
Migrant Background: Native Dutch	0.87	0.00	0.00	0.00	0.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Migrant Background: Non-western	0.05	0.50	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Migrant Background: Western	0.08	0.50	0.00	1.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Household Type: Single parent	0.04	0.04	0.15	0.06	0.10	0.04	0.08	0.05	0.00	1.00	0.00	0.00	0.00	0.00
Household Type: Single person	0.07	0.31	0.22	0.22	0.41	0.20	0.14	0.22	0.00	0.00	0.46	0.87	0.43	0.68
Household Type: Unconventional	0.00	0.52	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.50	0.00
Household Type: Other multi-person	0.02	0.02	0.07	0.05	0.04	0.04	0.04	0.03	0.00	0.00	0.08	0.03	0.02	0.06
Household Type: Couple w children	0.57	0.05	0.41	0.28	0.15	0.32	0.58	0.11	0.00	0.00	0.46	0.10	0.05	0.26
Household Type: Couple w/o children	0.30	0.05	0.15	0.39	0.31	0.40	0.16	0.59	1.00	0.00	0.00	0.00	0.00	0.00
Heterogeneous Policy Effect	0.0027	-0.0044	0.0117	0.0182	0.0226	0.0090	0.0080	0.0192	0.0287	0.0259	0.0110	0.0664	0.0228	0.0374
N	6,466,825	331,570	753,525	297,925	392,675	810,785	342,105	448,855	1463,270	297,040	347,640	396,805	391,025	1,132,875

Table A.5.4: Descriptive statistics of the heterogeneous groups identified for the consistent municipalities (Top) and the The Hague area (Bottom).

Group	1	2	3	4	5	6	7	8	9	10	11
Age	81.70	80.42	35.51	43.98	40.15	51.14	25.42	49.30	31.39	43.39	50.68
Sex	0.60	0.58	0.44	0.49	0.55	0.54	0.51	0.51	0.49	0.61	0.45
income: % of soc. min.	192.88	176.20	-1.80	399.26	94.06	166.98	163.52	163.22	-1.46	100.48	98.21
Migrant Background: Native Dutch	1.00	0.00	0.15	0.56	0.00	1.00	0.00	0.00	0.37	0.24	0.45
Migrant Background: Non-western	0.00	0.44	0.39	0.27	0.00	0.00	1.00	1.00	0.63	0.76	0.55
Migrant Background: Western	0.00	0.56	0.45	0.17	1.00	0.00	0.00	0.00	0.00	0.00	0.00
Household Type: Single parent	0.02	0.03	0.00	0.07	0.10	0.12	0.23	0.15	0.09	0.31	0.00
Household Type: Single person	0.48	0.47	0.00	0.17	0.49	0.40	0.18	0.21	0.66	0.00	1.00
Household Type: Unconventional	0.08	0.06	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Household Type: Other multi-person	0.01	0.05	0.00	0.04	0.04	0.04	0.12	0.09	0.05	0.06	0.00
Household Type: Couple w children	0.01	0.04	0.00	0.44	0.18	0.13	0.37	0.41	0.10	0.38	0.00
Household Type: Couple w/o children	0.39	0.35	0.00	0.27	0.19	0.31	0.10	0.13	0.10	0.25	0.00
Heterogeneous Policy Effect	-0.0434	-0.0186	-0.0297	0.0010	0.0098	0.0076	-0.0067	0.0151	0.0144	0.0129	0.0333
N	51,390	19,630	37,750	450,930	54,775	49,430	23,075	37,850	35,695	48,735	42,345

Group	1	2	3	4	5	6	7	8	9	10	11	12	13
Age	42.04	46.62	40.78	30.92	27.99	31.92	49.01	42.31	57.05	77.57	43.78	47.88	48.89
Sex	0.48	0.52	0.49	0.48	0.54	0.49	0.55	0.52	0.56	0.64	0.49	0.52	0.51
Income: % of soc. min.	489.37	252.17	249.28	-2.00	-2.00	-2.00	160.65	157.41	159.77	86.48	52.19	104.29	89.79
Migrant Background: Native Dutch	0.59	0.74	0.00	0.06	0.65	0.00	0.00	0.00	1.00	0.79	0.64	0.72	0.00
Migrant Background: Non-western	0.21	0.00	1.00	0.37	0.00	1.00	0.00	1.00	0.00	0.00	0.00	0.00	1.00
Migrant Background: Western	0.20	0.26	0.00	0.57	0.35	0.00	1.00	0.00	0.00	0.21	0.36	0.28	0.00
Household Type: Single parent	0.06	0.08	0.14	0.00	0.03	0.11	0.10	0.21	0.09	0.01	0.09	0.13	0.16
Household Type: Single person	0.19	0.45	0.21	0.00	0.79	0.67	0.48	0.23	0.53	0.68	0.53	0.72	0.41
Household Type: Unconventional	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.14	0.14	0.00	0.06
Household Type: Other multi-person	0.03	0.04	0.10	0.00	0.05	0.06	0.04	0.10	0.03	0.01	0.03	0.01	0.04
Household Type: Couple w children	0.38	0.15	0.40	0.00	0.01	0.08	0.13	0.36	0.07	0.00	0.09	0.05	0.20
Household Type: Couple w/o children	0.34	0.28	0.14	0.00	0.12	0.08	0.25	0.11	0.28	0.17	0.13	0.08	0.14
Heterogeneous Policy Effect	0.0023	0.0114	0.0119	-0.0015	0.0100	0.0293	0.0042	0.0250	0.0346	-0.0211	0.0338	0.0557	0.0729
N	516,080	179,115	94,960	59,210	74,115	29,775	36,260	108,355	107,420	33,865	29,750	38,440	113,880

Table A.5.5: Descriptive statistics of the heterogeneous groups identified for the Amsterdam area (Top) and the Rotterdam area (Bottom).

Group	1	2	3	4	5	6	7	8	9
Age	81.17	42.39	44.76	41.64	30.94	32.36	74.64	44.62	49.32
Sex	0.56	0.47	0.53	0.47	1.00	0.00	0.66	0.56	0.57
Income: % of soc. min.	192.92	417.23	182.06	-1.65	5.25	5.69	99.63	97.96	97.34
Migrant Background: Native Dutch	0.85	0.61	0.42	0.26	0.35	0.37	0.82	0.73	0.00
Migrant Background: Non-western	0.06	0.27	0.46	0.42	0.46	0.44	0.00	0.00	1.00
Migrant Background: Western	0.09	0.13	0.13	0.32	0.20	0.19	0.18	0.27	0.00
Household Type: Single parent	0.02	0.06	0.15	0.00	0.15	0.04	0.01	0.19	0.19
Household Type: Single person	0.46	0.15	0.29	0.00	0.57	0.68	0.75	0.60	0.44
Household Type: Unconventional	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
Household Type: Other multi-person	0.02	0.04	0.06	0.00	0.06	0.06	0.01	0.02	0.03
Household Type: Couple w children	0.03	0.45	0.26	0.00	0.08	0.08	0.01	0.08	0.19
Household Type: Couple w/o children	0.47	0.30	0.23	0.00	0.13	0.13	0.23	0.10	0.15
Heterogeneous Policy Effect	-0.0255	0.0014	0.0094	-0.0487	0.0211	0.0156	0.0244	0.0555	0.0802
N	70,455	422,300	306,410	41,905	39,600	39,425	21,600	27,885	70,440

Group	1	2	3	4	5	6	7	8	9	10
Age	44.21	40.85	28.16	53.66	52.97	36.53	45.50	50.16	55.88	48.44
Sex	0.48	0.44	0.53	0.00	1.00	0.51	0.60	0.49	0.55	0.61
Income: % of soc. min.	412.98	-1.59	202.88	206.35	202.99	13.86	107.06	255.15	181.64	68.94
Migrant Background: Native Dutch	0.66	0.27	0.00	0.00	0.00	0.00	0.00	1.00	1.00	1.00
Migrant Background: Non-western	0.23	0.53	0.81	0.70	0.69	0.81	0.84	0.00	0.00	0.00
Migrant Background: Western	0.11	0.19	0.19	0.30	0.31	0.19	0.16	0.00	0.00	0.00
Household Type: Single parent	0.06	0.00	0.21	0.04	0.23	0.16	0.25	0.08	0.11	0.17
Household Type: Single person	0.08	0.00	0.18	0.22	0.17	0.50	0.34	0.20	0.29	0.57
Household Type: Unconventional	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Household Type: Other multi-person	0.03	0.00	0.11	0.09	0.10	0.08	0.04	0.05	0.04	0.02
Household Type: Couple w children	0.56	0.00	0.40	0.43	0.32	0.17	0.25	0.30	0.18	0.09
Household Type: Couple w/o children	0.26	0.00	0.10	0.22	0.19	0.08	0.11	0.37	0.37	0.15
Heterogeneous Policy Effect	0.0022	-0.0638	0.0020	-0.0048	0.0110	0.0125	0.0240	0.0144	0.0283	0.0700
N	17,7440	8,200	16,830	10,090	12,030	7,365	16,460	11,660	47,140	10,360

Table A.5.6: Descriptive statistics of the heterogeneous groups identified for the Almere area (Top) and the Utrecht area (Bottom).

Group	1	2	3	4	5	6	7	8	9	10	11	12	13
Age	80.53	40.56	31.49	20.00	54.96	57.81	38.60	38.96	40.23	39.65	41.21	44.72	42.56
Sex	0.54	0.49	0.52	0.52	0.49	0.52	0.49	0.56	0.49	0.00	1.00	0.53	0.51
Income: % of soc. min.	258.49	539.07	280.31	275.19	328.72	218.66	-1.67	58.39	110.54	60.40	62.45	127.81	55.26
Migrant Background: Native Dutch	0.84	0.79	0.65	0.54	0.73	0.66	0.28	0.00	0.17	1.00	1.00	0.00	0.00
Migrant Background: Non-western	0.06	0.10	0.23	0.39	0.15	0.24	0.35	0.00	0.83	0.00	0.00	1.00	1.00
Migrant Background: Western	0.10	0.11	0.12	0.07	0.11	0.11	0.37	1.00	0.00	0.00	0.00	0.00	0.00
Household Type: Single parent	0.03	0.03	0.04	0.24	0.06	0.08	0.00	0.06	0.00	0.03	0.07	0.32	0.13
Household Type: Single person	0.39	0.08	0.37	0.03	0.22	0.29	0.00	0.69	0.00	0.74	0.75	0.35	0.63
Household Type: Unconventional	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
Household Type: Other multi-person	0.04	0.02	0.05	0.08	0.03	0.05	0.00	0.03	0.00	0.04	0.03	0.13	0.05
Household Type: Couple w children	0.05	0.54	0.28	0.60	0.40	0.25	0.00	0.07	1.00	0.00	0.00	0.00	0.00
Household Type: Couple w/o children	0.49	0.33	0.25	0.05	0.29	0.32	0.00	0.15	0.00	0.19	0.15	0.21	0.19
Heterogeneous Policy Effect	-0.0168	0.0003	0.0026	0.0098	0.0062	0.0082	-0.0386	0.0080	0.0136	0.0307	0.0237	0.0258	0.0331
N	16,650	159,765	111,910	13,175	45,055	44,345	20,010	18,335	18,005	34,520	47,180	11,780	20,705

	Consistent set		The Hague		Amsterdam		Rotterdam		Almere		Utrecht	
Variable	Estimate	S.E.	Estimate	S.E.	Estimate	S.E.	Estimate	S.E.	Estimate	S.E.	Estimate	S.E.
Intercept*	0.01376	0.00001	-0.02376	0.00009	-0.00292	0.00008	0.00380	0.00007	0.02841	0.00009	0.01402	0.00007
Migrant Background: Native Dutch	0.00705	0.00000	0.00246	0.00003	0.00313	0.00003	0.00043	0.00003	0.00678	0.00003	0.00531	0.00003
Migrant Background: Non-western	-0.00401	0.00001	0.00469	0.00003	0.01453	0.00003	0.00349	0.00003	-0.00132	0.00003	0.00572	0.00003
Household Type: Single parent	-0.00329	0.00001	-0.00097	0.00005	-0.00008	0.00004	0.00035	0.00003	-0.00006	0.00004	0.00202	0.00004
Household Type: Single person	0.00539	0.00000	0.00275	0.00003	-0.00178	0.00003	-0.00133	0.00003	0.00029	0.00003	0.00047	0.00003
Household Type: Unconventional	0.00262	0.00001	-0.03439	0.00007	-0.00527	0.00006	-0.06738	0.00006	-0.08252	0.00008	-0.06605	0.00006
Household Type: Other multi-person	0.00333	0.00001	-0.00029	0.00006	-0.00160	0.00005	-0.00113	0.00005	-0.00030	0.00005	0.00080	0.00005
Household Type: Couple w. children	0.00184	0.00000	-0.00114	0.00003	-0.00152	0.00003	-0.00112	0.00003	-0.00017	0.00002	-0.00224	0.00002
Sex	0.00063	0.00000	-0.00031	0.00002	-0.00056	0.00002	-0.00077	0.00002	0.00056	0.00002	-0.00096	0.00002
Age	-0.00005	0.00000	0.00165	0.00000	0.00065	0.00000	0.00091	0.00000	-0.00002	0.00000	0.00045	0.00000
Age squared	0.00000	0.00000	-0.00002	0.00000	-0.00001	0.00000	-0.00001	0.00000	0.00000	0.00000	0.00000	0.00000
Income: 5% - 15% of soc. min.	0.00261	0.00012	-0.00039	0.00060	0.03120	0.00051	0.00026	0.00058	0.00212	0.00067	-0.00301	0.00056
Income: 15% - 25% of soc. min.	0.00203	0.00006	-0.00079	0.00033	0.03091	0.00026	0.00078	0.00029	0.00229	0.00032	-0.00245	0.00032
Income: 25% - 35% of soc. min.	0.00718	0.00008	0.01555	0.00042	0.03416	0.00035	0.00092	0.00034	-0.00003	0.00047	0.00040	0.00038
Income: 35% - 45% of soc. min.	0.00884	0.00005	0.01244	0.00025	0.03282	0.00020	0.00065	0.00020	0.00121	0.00026	-0.00336	0.00023
Income: 45% - 55% of soc. min.	0.01131	0.00007	0.01126	0.00036	0.03349	0.00029	0.00065	0.00031	-0.00021	0.00037	-0.00359	0.00035
Income: 55% - 65% of soc. min.	0.01091	0.00003	0.01226	0.00017	0.03464	0.00014	0.00079	0.00014	0.00115	0.00019	-0.00198	0.00017
Income: 65% - 75% of soc. min.	0.01061	0.00005	0.01376	0.00024	0.03619	0.00020	0.00144	0.00020	0.00062	0.00027	-0.00241	0.00024
Income: 75% - 85% of soc. min.	0.00949	0.00002	0.01218	0.00012	0.03722	0.00010	0.03924	0.00009	0.00124	0.00014	-0.00087	0.00011
Income: 85% - 95% of soc. min.	0.00940	0.00003	0.01255	0.00015	0.03763	0.00013	0.03872	0.00012	0.00303	0.00017	-0.00120	0.00014
Income: 95% - 105% of soc. min.	0.01035	0.00001	0.01486	0.00006	0.03758	0.00005	0.03349	0.00005	0.00412	0.00007	-0.00132	0.00005
Income: 105% - 115% of soc. min.	0.01089	0.00002	0.00936	0.00011	0.03314	0.00009	0.00423	0.00008	-0.00677	0.00011	-0.00485	0.00009
Income: 115% - 125% of soc. min.	0.00949	0.00001	0.00502	0.00007	0.01940	0.00007	-0.00578	0.00006	-0.00700	0.00008	-0.00437	0.00006
Income: 125% - 135% of soc. min.	0.00918	0.00002	-0.00191	0.00011	0.01204	0.00010	-0.01090	0.00009	-0.00689	0.00011	-0.00470	0.00010
Income: 135% - 145% of soc. min.	0.00913	0.00001	-0.00169	0.00008	0.01197	0.00007	-0.01083	0.00006	-0.01583	0.00008	-0.00444	0.00006
Income: 145% - 155% of soc. min.	0.00943	0.00002	-0.00169	0.00011	0.01199	0.00010	-0.01071	0.00009	-0.01546	0.00011	-0.00438	0.00010
Income: 155% - 165% of soc. min.	0.00946	0.00001	-0.00158	0.00008	0.01177	0.00007	-0.01066	0.00006	-0.01559	0.00008	-0.01080	0.00006
Income: 165% - 175% of soc. min.	0.00952	0.00002	-0.00121	0.00011	0.01155	0.00010	-0.01061	0.00009	-0.01559	0.00011	-0.02401	0.00010
Income: 175% - 185% of soc. min.	0.00969	0.00001	-0.00110	0.00008	0.01148	0.00007	-0.01057	0.00006	-0.01549	0.00008	-0.02346	0.00006
Income: 185% - 195% of soc. min.	0.00990	0.00002	-0.00079	0.00012	0.01120	0.00011	-0.01044	0.00009	-0.01542	0.00010	-0.02286	0.00010
Income: 195% - 205% of soc. min.	0.00955	0.00001	-0.00070	0.00008	0.00650	0.00007	-0.01055	0.00006	-0.01542	0.00008	-0.02307	0.00006
Income: 205% - 215% of soc. min.	0.00861	0.00002	-0.00870	0.00011	-0.00233	0.00010	-0.01057	0.00009	-0.01523	0.00010	-0.02300	0.00009
Income: 215% - 225% of soc. min.	0.00075	0.00001	-0.00852	0.00008	-0.00232	0.00007	-0.01043	0.00006	-0.01519	0.00007	-0.02306	0.00006
Income: 225% - 235% of soc. min.	0.00079	0.00002	-0.00850	0.00011	-0.00250	0.00011	-0.01043	0.00009	-0.01523	0.00010	-0.02286	0.00009
Income: 235% - 245% of soc. min.	-0.00589	0.00001	-0.00852	0.00008	-0.00260	0.00007	-0.01048	0.00006	-0.01626	0.00007	-0.02271	0.00006
Income: 245% - 255% of soc. min.	-0.00923	0.00002	-0.00858	0.00011	-0.00256	0.00011	-0.01046	0.00009	-0.01875	0.00010	-0.02266	0.00009
Income: 255% - 265% of soc. min.	-0.00935	0.00001	-0.00860	0.00008	-0.00263	0.00007	-0.01589	0.00006	-0.01870	0.00007	-0.02249	0.00006
Income: 265% - 275% of soc. min.	-0.00946	0.00002	-0.00855	0.00012	-0.00273	0.00011	-0.01871	0.00009	-0.03048	0.00010	-0.02243	0.00009
Income: 275% - 285% of soc. min.	-0.00953	0.00001	-0.00853	0.00008	-0.00272	0.00007	-0.01882	0.00006	-0.03029	0.00007	-0.02476	0.00006
Income: 285% - 295% of soc. min.	-0.00963	0.00002	-0.00845	0.00012	-0.00266	0.00011	-0.01880	0.00010	-0.03038	0.00010	-0.02464	0.00009
Income: 295% - 305% of soc. min.	-0.01815	0.00001	-0.00842	0.00008	-0.00570	0.00007	-0.01895	0.00006	-0.03052	0.00007	-0.02449	0.00006
Income: 305% - 315% of soc. min.	-0.01810	0.00002	-0.00853	0.00012	-0.01182	0.00011	-0.01899	0.00010	-0.03034	0.00010	-0.02439	0.00010
Income: 315% - 325% of soc. min.	-0.01809	0.00001	-0.00838	0.00008	-0.01166	0.00007	-0.01903	0.00006	-0.03042	0.00007	-0.02433	0.00006
Income: 325% - 335% of soc. min.	-0.01808	0.00002	-0.00850	0.00013	-0.01161	0.00012	-0.01913	0.00010	-0.03040	0.00010	-0.02421	0.00010
Income: 335% - 345% of soc. min.	-0.01805	0.00001	-0.00840	0.00008	-0.01158	0.00008	-0.01911	0.00007	-0.03055	0.00007	-0.02421	0.00006
Income: 345% - 355% of soc. min.	-0.01806	0.00002	-0.00843	0.00013	-0.01160	0.00012	-0.01914	0.00010	-0.03067	0.00010	-0.02401	0.00010
Income: 355% - 365% of soc. min.	-0.01805	0.00001	-0.00827	0.00009	-0.01145	0.00008	-0.01913	0.00007	-0.03058	0.00007	-0.02398	0.00006
Income: 365% - 375% of soc. min.	-0.01804	0.00002	-0.00830	0.00014	-0.01152	0.00012	-0.01918	0.00011	-0.03074	0.00010	-0.02391	0.00010
Income: 375% - 385% of soc. min.	-0.01804	0.00001	-0.00824	0.00009	-0.01136	0.00008	-0.01914	0.00007	-0.03068	0.00008	-0.02389	0.00006
Income: 385% - 395% of soc. min.	-0.01804	0.00002	-0.00805	0.00014	-0.01125	0.00013	-0.01910	0.00011	-0.03068	0.00011	-0.02666	0.00010
Income: 395%+ of soc. min.	-0.01814	0.00001	-0.00789	0.00005	-0.01061	0.00005	-0.01889	0.00004	-0.03103	0.00006	-0.02635	0.00004
N	13,872,919	13,872,919	851,604	851,604	1,421,221	1,421,221	1,040,025	1,040,025	317,575	317,575	561,434	561,434
R^2	0.845621	0.845621	0.648269	0.648269	0.694146	0.694146	0.846871	0.846871	0.897064	0.897064	0.823862	0.823862

* Reference group are Western migrant couples without children with a household income < 5% of the social minimum or unknown.

Table A.5.7: Regression results decomposing policy effects by demographics for the consistent municipalities and each of the alternate municipalities.

Sample	N	Native Dutch	Western	Non-western
Netherlands	17,282,163	76.3%	10.3%	13.4%
Almere	207,904	57.1%	10.9%	32.0%
Amsterdam	862,965	45.5%	18.8%	35.7%
Rotterdam	644,618	48.4%	13.0%	38.6%
The Hague	537,833	45.4%	18.6%	36.0%
Utrecht	352,866	64.7%	11.9%	23.4%

Source: CBS Statline (2019)

Table A.5.8: Percentage of population by migrant background for The Netherlands and each of the alternate municipalities.

Variable	Estimate	S.E.
Intercept	0.1762	0.0006
Treatment (2016)	0.0004	0.0001
Age	-0.0115	0.0000
Age Squared	0.0001	0.0000
Migrant Background: Non-Western	-0.0152	0.0002
Migrant Background: Western	-0.0168	0.0002
Sex: Female	0.0197	0.0001
Household Type: Single person	0.0513	0.0003
Household Type: Unconventional	0.1925	0.0006
Household Type: Other multi-person	-0.0090	0.0005
Household Type: Couple with children	0.0035	0.0003
Household Type: Couple without children	-0.0376	0.0003
Income: 12.5 - 37.5 of soc. min.	0.0608	0.0019
Income: 37.5 - 62.5 of soc. min.	0.0411	0.0014
Income: 62.5 - 87.5 of soc. min.	0.0801	0.0009
Income: 87.5 - 112.5 of soc. min.	0.1869	0.0005
Income: 112.5 - 137.5 of soc. min.	0.1341	0.0005
Income: 137.5 - 162.5 of soc. min.	0.0807	0.0005
Income: 162.5 - 187.5 of soc. min.	0.0485	0.0005
Income: 187.5 - 212.5 of soc. min.	0.0304	0.0005
Income: 212.5 - 237.5 of soc. min.	0.0213	0.0005
Income: 237.5 - 262.5 of soc. min.	0.0173	0.0005
Income: 262.5 - 287.5 of soc. min.	0.0166	0.0005
Income: 287.5 - 312.5 of soc. min.	0.0188	0.0005
Income: 312.5 - 337.5 of soc. min.	0.0210	0.0005
Income: 337.5 - 362.5 of soc. min.	0.0215	0.0005
Income: 362.5 - 387.5 of soc. min.	0.0220	0.0005
Income: 387.5 - 412.5 of soc. min.	0.0210	0.0005
Income: 412.5 - 437.5 of soc. min.	0.0214	0.0005
Income: 437.5 - 462.5 of soc. min.	0.0202	0.0006
Income: 462.5 - 487.5 of soc. min.	0.0198	0.0006
Income: 487.5 - 512.5 of soc. min.	0.0189	0.0006
Income: 512.5 - 537.5 of soc. min.	0.0187	0.0007
Income: 537.5 - 562.5 of soc. min.	0.0178	0.0008
Income: 562.5 - 587.5 of soc. min.	0.0173	0.0008
Income: 587.5 - 612.5 of soc. min.	0.0175	0.0009
Income: 612.5 - 637.5 of soc. min.	0.0167	0.0010
Income: 637.5 - 662.5 of soc. min.	0.0168	0.0011
Income: 662.5 - 687.5 of soc. min.	0.0160	0.0012
Income: 687.5 - 712.5 of soc. min.	0.0161	0.0013
Income: 712.5 - 737.5 of soc. min.	0.0160	0.0014
Income: 737.5 - 762.5 of soc. min.	0.0161	0.0015
Income: 762.5 - 787.5 of soc. min.	0.0156	0.0016
Income: 787.5 - 812.5 of soc. min.	0.0148	0.0018
Income: 812.5 - 837.5 of soc. min.	0.0148	0.0019
Income: 837.5 - 862.5 of soc. min.	0.0164	0.0020
Income: 862.5 - 887.5 of soc. min.	0.0150	0.0022
Income: 887.5 - 912.5 of soc. min.	0.0143	0.0023
Income: 912.5 - 937.5 of soc. min.	0.0158	0.0024
Income: 937.5 - 962.5 of soc. min.	0.0113	0.0026
Income: 962.5 - 987.5 of soc. min.	0.0136	0.0028
Income: 987.5 - 1012.5 of soc. min.	0.0106	0.0009
N	10,850,260	
R^2	0.2090	

Table A.5.9: ATE setting 2016 as the treatment year and 2015 as the control year using covariate balancing.

Sample	N	Missing education	Proportion	Avg. age population	Avg. age of missing
Entire sample	22,454,609	7,229,938	32.20%	40.98	60.57
Adults (18+)	18,064778	7,158,182	39.623%	48.82	61.16

Table A.5.10: Missingness of the education variable for the entire sample as well as for adults (18+), including the average age of the entire sample versus those with missing education.

A.6 Chapter 6

[This section is deliberately left blank]

T Technical appendices

In this technical appendix, I provide background information and intuition into the methods used throughout this thesis. This part of the thesis is simply a reference manual for methods described elsewhere. I include the key references on which the exposition here is based as well as additional reading for each of the methods described here.

T.1 Misspecification bias

Functional form misspecification can be seen more generally as a form of omitted variable bias. This is most clear when some variable is associated with the outcome in a non-linear fashion, for instance a polynomial relationship, and the association is specified to be linear. In this case, the higher order terms are effectively 'missing'.

Omitted variable bias is the more general problem where some variable may be completely missing from the specification as well — rather than just its non-linear component, as was the case in the above example. Assume the true DGP is:

$$y = X\beta + z\alpha + \epsilon, \tag{T.1}$$

where regression of y on the matrix X and, for simplicity, vector of scalars z leads to consistent estimation of the parameters β and α.

If we do *not* include z into the model, it is effectively included in the error terms: $u = (z\alpha + \epsilon)$. From this, it becomes clear that the key assumption of exogeneity — that regressors are uncorrelated with the error term — will be violated whenever z is correlated with any of the covariates in X, as in that case it follows that $\mathrm{E}[u|X] \neq 0$. Consistency of the standard OLS estimator fundamentally relies on exogeneity as per definition $\beta_{\mathrm{OLS}} = \beta + (X'X)^{-1}X'u$ (Cameron and Trivedi, 2005, Chapter 4).

Clearly, if the true association between X and y is $g(\cdot)$, which is some non-linear function, then assuming linearity will lead to errors of the form $u = g(X) - f(X) + \epsilon$ where $f(\cdot)$ is the assumed linear specification. Generally $\mathrm{E}[u|X] \neq 0$ in this case. In other words, misspecification is a form of omitted variable bias that

guarantees exogeneity to be violated. Assuming linearity does reflect the best linear approximation to the possibly non-linear functional form. Unfortunately, this needn't be of any use for inference if the true relationship is highly non-linear.

Misspecification and omitted variable bias is treated in most standard textbooks. The exposition here is based on the discussion in Chapter 4 of Cameron and Trivedi (2005) and Long and Trivedi (1992), but similar discussion can be found in Wooldridge (2010) and Morgan and Winship (2015), amongst many others.

T.2 Model fit statistics

The goal of model fit statistics is to provide a metric of a model's ability to accurately fit the outcomes of interest y_i, based on information on covariates x_i. Typically, model fit metrics are used to compare different models with one another although the substantive fit of any one model can be of substantive interest as well (as discussed in Chapters 2 and 3). On the most general level, we are interested in the error, Err, of some model $\hat{f}(\cdot)$ fitted to a training dataset $\mathcal{D}^{\text{train}}$:

$$\text{Err}_{\mathcal{D}^{\text{train}}} = \text{E}_{(y',x')}[L(y', \hat{f}(x')|T)]. \tag{T.2}$$

This is the expected error according to some loss function $L(y, \hat{y})$ of a new pair of observations (y', x') drawn from the same distribution as $\mathcal{D}^{\text{train}}$. In what follows, I will distinguish between in-sample metrics to estimate (T.2), which usually include some level of correction for using the same data to estimate and evaluate a model, and out-of-sample statistics based on cross-validation. The same loss functions can typically be transferred from the in-sample to the out-of-sample context. This is particularly the case when the fit statistic is based on fitted values of the model $\hat{y}$. In some cases, fit statistics are based on other elements of the modeling process like the log-likelihood in the context of maximum likelihood estimation, in which case they may only apply to in-sample statistics or statistics within a certain class of functions.

T.2.1 In-sample statistics

The most simple and commonly reported metric of fit within the linear additive framework is R^2, the coefficient of multiple determination. Its popularity stems from its ability to be interpreted as the percentage of variance explained by the model and is defined as

$$R^2 = 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}}, \tag{T.3}$$

where SS_{res} stands for the sum of squared errors remaining after fitting some model $f(\cdot)$ to the data: $SS_{\text{res}} = \sum_{i=1}^{n}(y_i - f(\hat{x}_i))^2$. SS_{res} stands for the overall sum of squared errors of the outcome with respect to its mean: $SS_{\text{res}} = \sum_{i=1}^{N}(y_i - \bar{y})^2$.[3] In terms of (T.2) the R^2 metric scales with a loss function that is based on squared error loss. That is, the numerator in (T.3) follows (T.2) with $L = \sum_{i \in \mathcal{D}^{\text{train}}}(y_i - \hat{f}(x_i|\mathcal{D}^{\text{train}}))^2$ but using the same pairs of (y_i, x_i) to evaluate the loss function that were also used to estimate the model $\hat{f}(\cdot)$. Typically, using in-sample metrics will over-estimate (T.2) when in fact using new observations.

The difference in Err when using the same data for estimation and evaluation as compared to using unseen data is called the *optimism*. The optimism, ω, typically varies for different models $\hat{f}(\cdot)$ and in-sample fit metrics approximate the optimism of a model and add it to the in-sample $\bar{\text{err}}$ to obtain a metric that can be used to compare models in terms of (T.2). For most loss functions, the optimism relies on:

$$\omega \equiv \text{E}_{\mathbf{y}}(\text{optimism}) = \frac{2}{N}\sum_{i=1}^{N}\text{Cov}(\hat{y}_i, y_i). \tag{T.4}$$

In other words, the optimism scales directly with the correlation between the observed outcome of a single observation, y_i, and its prediction through the model $\hat{f}(\cdot) = \text{E}_{\mathcal{D}^{\text{train}}}(y|X = x_i)$. Thus, as the model is increasingly dependent on individual observations the optimism increases. This makes intuitive sense, as a model $\hat{f}(\cdot)$ that is highly dependent on the inclusion of a single observation will likely overfit

[3]Note that the R^2 cannot always be computed, for example in binomial or multinomial models. Various in-sample pseudo-R^2 metrics have been proposed as alternatives (Cameron and Trivedi, 2005, Chapter 8).

the model to the sample at hand. For linear models, (T.4) reduces to the number of covariates k times the remaining error in the model, leading to $\omega = 2\frac{k}{N}\sigma_\epsilon^2$. Plugging in an estimation for σ_ϵ^2 from the data leads to Mallow's C_p statistic that is frequently used to compare linear models estimated by OLS. The adjusted R^2 similarly scales the standard R^2 metric with the number of covariates in the model:

$$R^2_{\text{adj}} = 1 - (1 - R^2)\frac{n-1}{n-k}. \tag{T.5}$$

The C_p statistic for linear models can be generalized to any type of maximum likelihood based statistic through the Akaike information criterion (AIC):

$$\text{AIC} = \frac{2}{N}\text{E}[\text{loglik} + 2\frac{k}{N}], \tag{T.6}$$

where loglik reflects the log-likelihood of the model. In a general sense, assume α to be some tuning parameter or other model choice relevant to $\hat{f}_\alpha(\cdot)$. Then we can recalculate (T.6) and use it to select the best fitting model. Things become challenging when estimation of the effective number of parameters $k(\alpha)$ is no longer straightforward. This can be because a model is intrinsically difficult to define in terms of effective parameters, as is the case in tree-based models (Section T.5) or when the model $\hat{f}_\alpha(\cdot)$ is already determined adaptively based on fit.[4]

In cases where determining $k(\alpha)$ is not straightforward, we can try to approximate the number of effective parameters using more complex methods. An example is the Vapnik-Chervonenkis dimension which provides a general theory for setting upper bounds on the optimism of any non-linear model which is roughly based on determining the extent to which a function can completely divide observations which are assigned to either of two classes into pure sets.[5] Another, much more

[4]This is for example the case in step-wise models where the number of parameters k is selected using an algorithm. Pretending that the number of effective parameters is k effectively underestimates the true number of parameters used (Janson et al., 2015). In other words, the model $\hat{f}_\alpha(\cdot)$ might contain $k(\alpha)$ parameters but these were already selected out of a larger set of $\mathcal{K}$ possible parameters making the effective parameters larger than $k(\alpha)$.

[5]For example, any three observations in $\mathcal{R}^2$ assigned to one of two classes can be divided into two sets consisting of only observations of either class by a single straight line (as long as the observations aren't completely collinear). However, a straight line cannot always divide four observations into two such sets although a class of non-linear functions might. A function that can split four observations then has a higher VC dimension than one that cannot.

simple approach is to use cross-validation to directly assess model fit out-of-sample, which does not require the optimism to be estimated and added to the in-sample fit.

T.2.2 Out-of-sample statisics

Out-of-sample statistics of model fit have a long history in model selection (Efron, 1983; Kohavi et al., 1995). As described above, fit statistics require correcting for optimism when estimated in-sample. As models become more complex, optimism becomes increasingly difficult to estimate (Janson et al., 2015). Even within the context of in-sample statistics, they may not be valid under repeated estimation of different models (Gelman and Loken, 2013).[6] A standard idea is to assess a model's fit on data not used for estimation instead, as this provides an indication of model fit without the need to calculate the number of effective parameters in the model. The most commonly implemented approach is K-fold cross validation:

Algorithm T.1: K-fold cross validation

Split $\mathcal{D}$ into K equal-sized partitions $\mathcal{D}^{(k)}$ of size N/K;
for *i in 1:K* **do**
  Estimate model $\mathcal{M}^{(k)}$ excluding $\mathcal{D}^{(k)}$ from $\mathcal{D}$;
  Make predictions using data on $\mathcal{D}^{(k)}$ using the estimated model $\hat{\mathcal{M}}^{(k)}$;
  Store predictions $\hat{y}_{i \in \mathcal{D}}$ into vector $\hat{\mathbf{y}}$;
end
Evaluate $\text{Err}(\hat{\mathbf{y}}, y)$

Again, many different functions to assess Err can be used in the last step. In addition, there are many variants to Algorithm T.1 as discussed in more detail in Chapter 2. Most variations on the above have the same estimand (Yousef, 2020). Interestingly, it has been shown that an out-of-sample fit metric like Err_{cv} is a consistent estimator for the average error of the model $\mathcal{M}$ when fit to new datasets drawn from the same distribution as the original data: $\mathcal{D}^* \sim F$. This nuance also holds for in-sample statistics like Mallow's C_p statistic described earlier, and is

[6] Illustrative is the recommendation by Cameron and Trivedi in their 2005 textbook that it is good practice regardless to 'use only part of the sample ("training sample") for specification searches and model selection, and then report results using the preferred model estimated using a completely separate part of the sample ("estimation sample") (Cameron and Trivedi, 2005, p. 286).'

mostly relevant for applications where the onus is truly on accurately establishing the out-of-sample predictive performance of a model (Bates et al., 2021; Efron and Hastie, 2016). More importantly, out-of-sample statistics based on K-fold cross validation have been shown to under-estimate confidence intervals when directly evaluating the variability in Err_{cv} based on K samples. Similar in spirit to the optimism in (T.4), this problem inflates when the estimated model $\hat{\mathcal{M}}^{(k)}$ depends more strongly on what fold $\mathcal{D}^{(k)}$ is omitted from the model (Bengio and Grandvalet, 2004; Bates et al., 2021). To circumvent this problem, scaling parameters have been suggested but the most principled approach is to simply estimate Err on a set of data completely separated from the estimation routine. The downside to this approach is of course that data is lost for estimating $\mathcal{M}$ (Kohavi et al., 1995).

The above exposition is based mostly on Hastie et al. (2009), Chapter 7 and Efron and Hastie (2016), Chapter 12, as well as the early review paper on cross validation by Kohavi et al. (1995) and follow-up papers by Bengio and Grandvalet (2004) and Bates et al. (2021) who note some of the nuances to the cross-validation estimand and confidence intervals. Throughout this thesis, I implement the estimation of fit statistics myself but various packages in `R` and `Python` exist for this purpose.

T.3 The ordered probit model

When the dependent variable is categorical — i.e., $y_i = m$ with $m \in [1, \ldots, M]$ — but follows a natural ordering, an ordered response model like the ordered probit model can be used. Following Maddala (1986), the model can be estimated by assuming a latent variable y_i^* underlying the observed ordered outcomes:

$$y_i^* = X_i\beta + \epsilon_i, \tag{T.7}$$

where the domain of y^* can then be partitioned into M sections via $M - 1$ cut-off points μ_m according to:

$$\begin{aligned} y_i &= 1 \qquad \text{if} \qquad y_i^* \leq \mu_1, \\ y_i &= 2 \qquad \text{if} \qquad \mu_1 < y_i^* \leq \mu_2, \\ &\vdots \\ y_i &= M \qquad \text{if} \qquad \mu_{M-1} < y_i^*. \end{aligned} \tag{T.8}$$

By assuming ϵ_i to be standard-normally distributed and setting $\mu_1 = 0$, the above leads to the ordered probit model:

$$\begin{aligned} \Pr[y_i = 1] = &\quad \Phi(\mu_1 - X_i\beta) \\ \Pr[y_i = 2] = &\quad \Phi(\mu_2 - X_i\beta) - \Phi(\mu_1 - X_i\beta) \\ &\vdots \\ \Pr[y_i = M] = &\quad 1 - \Phi(\mu_{M-1} - X_i\beta). \end{aligned} \tag{T.9}$$

Estimation proceeds via maximum likelihood according to:

$$L(Y, \beta, \mu) = \prod_{i=1}^{N} f(y_i|X_i) = \prod_{i=1}^{N} \prod_{m=1}^{M} p_{im}^{\mathrm{d}_{im}}, \tag{T.10}$$

where d_{im} indicates whether we observed outcome $y_i = m$ for observation i. The above then yields the associated log-likelihood:

$$\mathcal{L}(Y, \omega, \mu) = \sum_{i=1}^{N} \sum_{j=1}^{M} \mathrm{d}_{im} \ln(p_{im}), \tag{T.11}$$

where optimization occurs over cutoff points μ and parameters β. Plugging in p_{im} as given in (T.9), the likelihood function becomes:

$$\mathcal{L}(y, \omega, \mu) = \sum_{i=1}^{N} \sum_{m=1}^{M} \mathrm{d}_{im} \ln(\Phi(X_i\beta - \mu_m) - \Phi(X_i\beta - \mu_{m-1})). \tag{T.12}$$

As has been established for general categorical outcome models, assumptions on the variance of the error term like that above lead to complications in interpreting and comparing coefficient estimates (Mood, 2010; Breen et al., 2018). The problem is that the error term in (T.7) is impossible to estimate as we only observe the binary outcome, but σ_ϵ will generally differ between models or the same model estimated to sub-samples. However, the true coefficient in the model scales with σ_ϵ thus making comparisons across models complicated as the scaling term will likely

differ. When wanting to make substantive comparisons between coefficients across models or estimated to separate samples, various solutions have been proposed in the literature (see Breen et al. (2018) for a review).

T.3.1 Multilevel setup

The latent variable model (T.7) can be extended to include random effects in the same way as a standard linear additive model with a continuous outcome:

$$y_i^* = \alpha^{[j]} + \beta_1^{[j]} x_{i1} + \ldots + \beta_k^{[j]} x_{ik} + \epsilon_i, \tag{T.13}$$

following the notation in Gelman and Hill (2006).

In terms of estimation, adding multilevel elements like in (T.13) combined with the non-linearity inherent in the ordered probit model requires random effects to be integrated out to obtain consistent estimates. In linear additive models this is not required. The log-likelihood becomes:

$$\mathcal{L}_j(Y_j, \omega, \mu) = \sum_{i=1}^{n_j} \sum_{m=1}^{M} \mathrm{d}_{im} \ln[(p_{im}|\mathbf{u}_j, \mu) p(\mathbf{u}_j|\omega)]. \tag{T.14}$$

Here, $p(\mathbf{u}_j|\omega)$ is the probability of observing a vector of random effects $\mathbf{u}_j$. Integrating out these random effects yields the following log-likelihood:

$$\mathcal{L}_j(Y_j, \omega, \mu) = \int \sum_{i=1}^{n_j} \sum_{m=1}^{M} \mathrm{d}_{im} \ln[(p_{im}|\mathbf{u}_j, \mu) p(\mathbf{u}_j|\omega)] d\mathbf{u}_j. \tag{T.15}$$

Estimation of the above will rely on numerical optimization (Skrondal and Rabe-Hesketh, 2004; Christensen, 2015). After estimation, modes of the distribution of the random effects can be extracted including confidence intervals. These modes can be used as point estimates for the random effects in (T.13).

The above exposition is based on the setup by Maddala (1986) as also described in Breen et al. (2018). I rely on the review by Breen et al. (2018) and article by Mood (2010) for the discussion on complexities in the interpretation of coefficients in non-linear probability models. Gelman and Hill (2006) provide an overview

of hierarchical modeling from a Bayesian perspective and Skrondal and Rabe-Hesketh (2004) and Christensen (2015) provide discussions of the specific case of the multilevel ordered probit model. Other textbooks include Agresti (2010) and Long and Freese (2006). I use the `R` package `MASS` (Ripley et al., 2013) to estimate the standard ordered probit model and `ordinal` (Christensen, 2015) to include random effects to the model.

T.4 The Jackknife and the Bootstrap

Some of the most quintessential computer age statistical devices are the Jackknife and Bootstrap. They allow for a computationally-driven way of estimating standard errors of a statistic $\hat{\tau} = T(\mathcal{D})$ with $\mathcal{D} \sim F$ with F a potentially unknown distribution. Formulations of standard errors typically only exist for certain estimates and assumptions on F, for example sample means or linear regression. Attempting to compute standard errors for more complicated statistics either requires laborious Taylor series approximations or is generally infeasible (Efron and Hastie, 2016, Chapter 10). Instead, the Jackknife relies on re-estimating the statistic $\hat{\tau}$ many times on slightly perturbed versions of $\mathcal{D}$. Specifically, the estimation is repeated N times whilst excluding observation i from the data: $\hat{\tau}_{(i)} = T(\mathcal{D}_{(i)})$. The Jackknife estimate of the standard error of $\hat{\tau}$ then becomes:

$$\hat{\text{se}}_{\text{jackknife}} = \frac{n-1}{n} \sum_{i}^{N} (\hat{\tau_{(i)}} - \sum_{i}^{N} \frac{\hat{\tau}_{(i)}}{N})^2. \tag{T.16}$$

This approach can be applied without knowledge of F, as long as the algorithm $T(\cdot)$ is available to (re-)estimate $\hat{\tau}(\mathcal{D})$. In effect, equation (T.16) relies on the variation in $\hat{\tau}$ with respect to individual observations i. The downside of the Jackknife is that it relies on such local derivatives and can behave erratically for non-smooth statistics like the median or step-wise relationships in data (Efron and Hastie, 2016, Chapter 10).

The non-parametric bootstrap is an even more computationally burdensome but more flexible version of the Jackknife. The main idea behind a standard error is the uncertainty of $\hat{\theta}$ with respect to a different sample $\mathcal{D}$ drawn from the same

distribution, $\mathcal{D}^* \sim F$. Bootstrapped standard errors rely on 'Bootstrap samples' which are drawn by placing sampling weights equal to $\frac{1}{N}$ on every observation in $\mathcal{D}$ and sampling with replacement a new dataset $\mathcal{D}^*$. This new dataset $\mathcal{D}^*$ follows the empirical probability distribution of F as estimated through the observed $\mathcal{D}$. The Bootstrap estimate of the standard error of $\hat{\theta}$ then becomes:

$$\hat{\text{se}}_{\text{boot}} = [\sum_{j=1}^{B} \frac{(\hat{\theta}^{(j)} - \sum_1^B \frac{\theta^{(j)}}{B})^2}{B-1}]^{\frac{1}{2}}, \tag{T.17}$$

which is the empirical standard error when observing B estimates of $\hat{\tau}$ using B different datasets $\mathcal{D}^*$. The Bootstrap shares the same advantages as the Jackknife in that it is fully automatic and only requires the availability of some algorithm $T(\cdot)$ to generate estimates of $\hat{\tau}$ and no further assumptions on F. It improves on the Jackknife in that the deviations in different estimates $\hat{\tau}^{(i)}_{\text{boot}}$ are no longer dependent on the derivative with respect to a single observation i as was the case in $\hat{\tau}^{(i)}_{\text{jack}}$. Note that the weighting function to generate $\mathcal{D}^*$ which is implicit in the Bootstrap is a Uniform distribution where each observation i has the same resampling probability. Other sampling schemes are also possible, for example if one wants to retain a certain hierarchical structure within each $\mathcal{D}^*$.

The Bootstrap can also be used to estimate more challenging statistics than the standard error of $\hat{\tau}$. For example cases where we are interested in confidence intervals. There are many situations where the standard confidence intervals defined by $\hat{\tau} \pm 1.96\hat{\text{se}}_\tau$ do not hold, for example when asymptotic normality has not yet been reached or for distributions that are skewed by nature like a Poisson. In such cases, the Bootstrap can be used as well by simply ordering all B Bootstrap estimates $\hat{\tau}_b$ and selecting the estimates at the 2.5th and 97.5th percentile of this ordering when interested in a 95% interval. This is the empirical confidence interval rather than one implied via normality of the estimate. The key idea behind the percentile approach is that there always exists some transformation $m(\cdot)$ of the estimate $\hat{\tau}$ that normalizes it: $\hat{\phi} = m(\hat{\tau})$ with $\hat{\phi} \sim N(\phi, \sigma^2)$. Furthermore, this transformation is invariant to the ordering of the B Bootstrap estimates. Thus, the same estimate will lie at the 2.5% and 95.5% — or $\frac{\alpha}{2}$ for arbitrary confidence

level $(1-\alpha)$ — whether considering $\hat{\tau}$ or $\hat{\phi}$. Thus, we don't need to know $m(\cdot)$ to make use of the percentile method.

The exposition above relies mostly on the discussion in Efron and Hastie (2016), Chapters 10 and 11. The key references are (Efron, 1992, 1987) and other treatments can be found in e.g., Cameron and Trivedi (2005) Chapter 11 and (Efron and Tibshirani, 1993). Throughout this thesis, I implement the Bootstrap myself although the `R` package `bootstrap` (Efron and Tibshirani, 1993) allows for the implementation of a wide number of Jackknife and Bootstrap implementations.

T.5 Tree-based methods for prediction

Assume we have a set of data $\mathcal{D}$ consisting of N pairs (x_i, y_i) where y_i is the outcome of interest and x_i a multidimensional vector of explanatory variables or covariates. Whenever some model $f(\cdot)$ is estimated to $\mathcal{D}$, for example the ordered probit model above, we obtain parameter values θ and a general 'rule' $r_{\mathcal{D}}(x)$ relating covariates to the outcome. This rule has various possible uses. The typical approach in sociology is to use $r_{\mathcal{D}}$ for explanation. That is, to identify the relative contribution of individual elements in x on the outcome y, typically by interpreting the coefficients θ substantively. This explanation perspective typically requires $r_{\mathcal{D}}$ to be defined in such a way that θ is substantively meaningful. Alternatively, one can use $r_{\mathcal{D}}$ to make predictions of y without wanting to make inferences into θ. A predictive approach generally gives way to more flexible algorithms to model y as a function of x as an explanatory focus tends to put constraints on θ for interpretation.

The number of predictive algorithms available to generate $r_{\mathcal{D}}$ is large and has grown considerably with the exponential increase in the availability of computational resources. In addition, most of the standard statistical models typically used for explanatory work can equally be used to generate predictions, but tend to underperform those that exploit computational resources. In this thesis, tree-based methods are used as a general class of computationally intensive algorithms that model data with a predictive focus. The reasons for preferring tree-based methods

for sociological questions are discussed in Chapter 4. In what follows below, a more technical discussion of tree-based methods is provided.

T.5.1 Decision tree

The goal of a decision or regression tree is to divide $\mathcal{D}$ into m regions R_m in covariate space, which are mutually exclusive and add up to the total covariate space $\mathcal{R}^{\mathcal{K}}$. In other words, every observation i belongs to a single region R_m, and to that region only. Each region is then assigned its own prediction, c_m. The set of all regions R_m is called $\Pi = [R_1, \ldots, R_M]$, where Π reflects the tree structure of a decision tree. The prediction rule $r_{\mathcal{D}}$ is then defined as:

$$f(x_i) = \sum_{m=1}^{M} c_m \mathrm{I}(x_i \in R_m), \tag{T.18}$$

where $\mathrm{I}(x_i \in R_m)$ is 1 if $x_i \in R_m$ and 0 otherwise.

The optimal $\hat{c}_m$ conditional on Π depends on the loss function chosen. If we choose squared error loss, as is customary for regression, it follows that $\hat{c}_m = \frac{1}{N_{R_m}} \sum_{i \in R_m} y_i$, which is simply the mean of the outcomes observed in region R_m. In other words, the fitted value $\hat{y}_i$ associated with a vector of covariates x_i is the mean of the outcomes of observations that are close by with respect to Π. This makes trees akin to adaptive nearest neighbors methods, although instead of general distance metrics the regions R_m can be much wider in some dimensions and much more narrow in others. This allows regions R_m to be sensitive to dimensions where there is much variability in the outcome.

The complexity in estimating a tree lies in deciding how to determine the regions R_m. Decision trees generate these regions via recursive binary partitioning. This means that x_i belongs to region R_m based on a succession of binary partition rules of the sort $x_j > s$ where s is some number that can be used to split the covariate x_j into two partitions. Finding the optimal set of regions R_m is computationally infeasible as considering all partitions of $\mathcal{D}$ into R_m explodes computationally. Instead, some form of greedy algorithm is typically used to decide partitions. The most popular is the CART algorithm (Breiman et al., 2017).

Starting with the entire dataset $\mathcal{D}$, a split-point s along variable j defines two partitions $R_L(j,s) = \{X|x_j \leq s\}$ and $R_R(j,s) = \{X|x_j > s\}$. The search algorithm assesses the optimal j and s according to:

$$\min_{j,s}[\min_{c_L} \sum_{i \in R_L(j,s)} (y_i - c_L)^2 + \min_{c_R} \sum_{i \in R_R(j,s)} (y_i - c_R)^2], \tag{T.19}$$

which is that combination of variable and split-point that minimizes the loss function across both partitions after the split.

Repeatedly evaluating (T.19) on each subsequent partition leads to a fully grown tree. The tree could continue making partitions up until the point that a region R_m might even uniquely identify specific pairs of data. In that case, $\hat{c}_m$ under squared error loss would be exactly equal to the observed y_i in R_m and the loss would be 0.[7] Likely, letting the tree grow to this depth will lead to overfitting of the sample and poor generalization of the tree to new data.

To decide the depth of the tree, a combination of pruning and cross validation (Section T.2.2) is typically used. Pruning reflects the process of collapsing regions R_m from a fully grown tree together. Take $|R|$ to be the number of terminal nodes in the fully grown tree. Further define the mean squared error loss in node m as $Q_m(R) = \frac{1}{N_m} \sum_{i \in R_m} (y_i - \hat{c}_m)^2$. The standard approach is then to define a loss function that penalizes tree complexity through $|R|$ in the following way:

$$C_\alpha(R) = \sum_{m=1}^{|R|} N_m Q_m(R) + \alpha |R|. \tag{T.20}$$

This approach is similar in spirit to LASSO regression which penalizes the absolute size of coefficients in a linear regression context. Setting $\alpha = 0$ effectively means retaining the full tree, whereas increasing α leads to a sequence of smaller trees R_α. For two distinct pruned trees $R_{\alpha'}$ and $R_{\alpha'}$ where $\alpha' > \alpha$ note that $R_{\alpha'}$ is a sub tree of R_α (Hastie et al., 2009, Chapter 9). To decide the value of α, cross validation is usually applied to assess the out-of-sample performance of Π_α.

[7] Having a single observation in a partition of course requires that there are no other pairs $(y_{i'}, x_{i'})$ for which $x_i = x_{i'}$, or that (y_i, x_i) was not part of some partition $R_{m'}$ which was already 'pure' in the sense that all $y_i, i \in R_{m'}$ were equal.

The type of decision tree discussed above is a so-called regression tree where $y \in \mathcal{R}$. Moving to discrete outcomes like a binary outcome $y \in [0, 1]$ requires adapting squared error loss with a suitable alternative. Assume there are k discrete outcomes, define

$$\hat{\mathrm{p}}_{m,k} = \frac{1}{N_m} \sum_{i \in R_m} \mathrm{I}(y_i = k), \tag{T.21}$$

which is the proportion of observations in R_m belonging to category k. Classification occurs based on $k(m) = \mathrm{argmax}_k(\hat{\mathrm{p}}_{m,k})$, that is the class with the highest probability. Instead of squared error loss the misclassification rate is typically used to grow the tree, which is defined as

$$\frac{1}{N_m} \sum) i \in R_m \mathrm{I}(y_i \neq k(m)) = 1 - \hat{\mathrm{p}}_{m,k}. \tag{T.22}$$

Various alternative types of loss also exist, like the Gini index:

$$\sum_{k \neq k'} \hat{\mathrm{p}}_{m,k} \hat{\mathrm{p}}_{m,k'} = \sum_{k=1}^{K} \hat{\mathrm{p}}_{m,k}(1 - \hat{\mathrm{p}}_{m,k}), \tag{T.23}$$

or cross-entropy:

$$-\sum_{k=1}^{K} \hat{\mathrm{p}}_{m,k} \mathrm{log} \hat{\mathrm{p}}_{m,k}. \tag{T.24}$$

The latter two loss functions prioritize purity of leaves in addition to correct classification. For example, assume a dataset with 400 observations of two classes each: $(400, 400)$. Assume the data can be split in two ways: the first generates $R_L = (300, 100)$ and $R_R = (100, 300)$, whereas the second generates $R_L = (400, 200)$ and $R_R = (0, 200)$. Both examples lead to the same misclassification rate of 0.25 but both the Gini index and cross-entropy prefer the latter partition as it weighs the additional purity in R_R greater than the loss of purity in R_L thus preferring the second partition over the first.[8]

The major downside of regression trees is that they can be sensitive to changes in the data. This is primarily due to the hierarchical nature of the algorithm. For example, in case there are two candidate pairs of variable and split value (j, s) that

[8]This example is borrowed from Hastie et al. (2009), Section 9.2.

have almost the same improvement in loss at the first partition, a small change in the data can lead to a different initial split which might have consequences for the rest of the tree. In other words, there is considerable variance in tree estimates. Similarly, a tree consists of 'hard' boundaries due to the nature of binary partitioning. When compared to for instance a linear additive functional form the regression surface will not be smooth. Both problems are addressed by expanding the number of decision trees that are estimated. These developments are discussed next.

Note that as suggested by Athey and Imbens (2016) and proven by Wager and Athey (2018) (Appendix B), extreme values of the outcome ('outliers') can lead to inconsistencies in estimated leaf effects. Specifically, in case of considerable variability, regression trees start to fit to outliers and in doing so misrepresent the underlying structure of the data (Wager, 2014). This might not matter from a predictive perspective, but it does when substantively interested in the point estimates of the conditional mean function of a decision tree. In that case it is advised to use a separate sample for determining the split pairs (j, s) that define the regions R_m from the sample used to estimate $\hat{c}_m$ in the regression tree context. This approach is referred to as 'honest' estimation. This is discussed further in Section T.8. The issue with honest estimation is that less data is used in generating partitions, which can have implications for the performance of the tree as a whole (Denil et al., 2014).

T.5.2 Random forest

The simplest solution to the relatively high variance of trees is estimating many i.i.d. trees into a forest and averaging predictions amongst them. This ensures the forest has the same bias properties as each individual tree, but variance might reduce. Averaging B trees reduces variance proportionally up to the correlation amongst the individual trees.[9] Therefore, de-correlating the trees is a crucial element in the variance-reduction properties of forests. This de-correlation is done by bootstrap

[9]Taking the mean of B i.i.d. random variables with variance σ^2 leads to variance $\frac{1}{B}\sigma^2$. If random variables are correlated according to some parameter ρ then variance decreases as $\sigma^2(\rho + \frac{1-\rho}{B})$.

resampling the original dataset $\mathcal{D}$ as well as selecting only a subset of the total number of variables $m \leq p$ of the p variables in the model.

The approach of generating de-correlated models and averaging over each of them works well for non-linear methods as the common correlation amongst models, ρ, quickly decreases. For linear models, this process of 'bagging' (bootstrap aggregating) tends to perform less well, as linear models typically retain stronger correlations across bootstrapped resamples of the data. Random forests also effectively smooth the regression surface through averaging across a large number of trees and typically improve strongly in terms of out-of-sample performance as compared to single decision trees.

T.5.3 Gradient boosting

Boosting differs from bagging approaches like random forests by iteratively combining a large amount of 'weak learners' rather than averaging a large amount of 'full' models. The definition of a weak learner is a model which only marginally improves on random guessing the outcome. Averaging over a large amount of such weak learners will lead to similarly weak performance, as every individual model has high bias. The point of boosting is instead to repeatedly apply some weak learner $G_m(x)$ M times to modified versions of the data. Each weak learner makes a set of prediction which are combined through a weighted average to generate the full model's predictions according to a B-dimensional weighting vector α_B. Progress is made by modification of the datasets to which the weak learner is repeatedly estimated. Specifically, the data is inversely weighted based on the residual of the weak learner. Whereas each observation is weighted equally during the first application of $G_1(x)$ — i.e., $w_i = \frac{1}{N}$ — new weights are assigned for each subsequent estimation routine.

Dependent on the weak learner used, different weighting functions for the dataset are calculated. One of the most popular approaches to boosting is so-called gradient boosting where the gradient of the loss function is explicitly estimated with respect to each individual prediction made by the model $G_{m-1}(x_i)$. In other words, those

observations for which the loss function would improve the most if they could be changed are weighted the strongest:

Algorithm T.2: Gradient Boosting

$G_0(x) = h(y;\theta)$;
for *i in 1:M* **do**
  Calculate gradients $r_{im} = -\frac{\delta L(y_i, G_{m-1}(x))}{\delta G_{m-1}(x)}, \quad i = 1, ..., N$;
  Estimate tree $h_m(r_{im};\theta) = \text{argmin}_\theta \text{L}(r_{im}, \theta)$;
  Update $G_m(x) = G_{m-1} + \rho h_m(r_{im};\theta)$
end
Output $\hat{G}(x) = G_M(x)$

In the first step $G_0(x_i)$, a tree $h(\cdot)$ is fitted to the data. As the tree $h(\cdot)$ is a weak learner, fit will be relatively poor. For every individual prediction of the initial weak learner the gradient with respect to the loss function is then estimated. If the loss function is defined as $L(y; F(x)) = \sum_i \frac{1}{2}(y_i - F(x_i))^2$ then this gradient is equal to the residual of the prediction. In the next step, a similar weak learner is estimated to these residuals. In the final two steps, this new learner is added to the original learner using weighting factor ρ_m which is typically a tuning parameter. In essence, gradient boosting is thus a weighted additive function where each tree 'specializes' on different parts of the dataset. The idea of repeatedly estimating weak learners to those observations that are fitted poorly in earlier runs of the data has led to considerable improvements in model fit, and gradient boosting is a popular approach in many predictive tasks (Chen et al., 2015).

The iterative nature of boosting type approaches — as compared to bagging approaches that estimate independent models to the data and averages them — means that they are difficult to determine statistical properties for. Whereas breakthroughs in statistical theory have been developed for random forests, these remain largely elusive to this date for boosting methods (see Section T.8). An extremely popular approach to gradient boosting is the `XGBoost` algorithm (`https://xgboost.readthedocs.io/en/stable/`) which follows the same principles of the gradient boosting approach outlined above but incorporates additional

regularization of the model and incorporates a number of additional tricks to speed up computation (for details see Chen et al. (2015)).

The exposition above is based mainly on Hastie et al. (2009), Chapters 9 and 10, as well as Efron and Hastie (2016), Chapter 17. The text by Friedman (2001) gives an overview of the boosting approach and the text by Breiman et al. (2017) introduces tree-based models more generally. The CART algorithm for standard decision and regression trees is available in `R` through the `caret` package (Kuhn et al., 2020) which includes a number of tools for tree-based modeling. The `ranger` (Wright and Ziegler, 2015) and `randomForest` (Liaw et al., 2002) packages provide implementations of the random forest algorithm. Both standard gradient boosting as well as XGBoost are implemented in the `R` packages `gbm` (Greenwell et al., 2019) and `xgboost` (Chen et al., 2015), respectively. The various tree-based methods listed above are all part of the `sklearn` library in `Python` (Pedregosa et al., 2011a). I apply standard tree-based methods for prediction like those discussed here in Chapter 4.

T.6 Super Learner

The basic idea underlying a Super Learner is that different modeling approaches might capture different structure in the data better. By combining different models a better overall model can be obtained. In ML, such an approach is typically called an 'ensemble method.' The Super Learner effectively incorporates a large number of possible algorithms $\mathcal{K}$ and uses cross-validation to select the best algorithm to fit the data. A key result for a cross-validation based Super Learner is that it produces fitted values that perform at least as good as the fitted values from the best algorithm that is part of the ensemble (Van der Laan et al., 2007).

Assume $|\mathcal{K}|$ methods are chosen to model some dataset $\mathcal{D}$. Models are understood here in the broadest sense and could include linear additive models, random forests, or neural networks and each with different parameter values — i.e., a random forest estimated using 500 trees or the same random forest estimated using 2000 trees would constitute two different models. Next, the standard approach to cross validation is implemented (Section T.2.2) to assess each models' predictive

performance. The discrete Super Learner simply uses the best performing model. The complete Super Learner approach includes a second modeling step where the outcome is modeled as a weighted average of the K models estimated, for instance using a standard linear regression to obtain weight w_k for each model k. For further reading and a full discussion including simulation results see Van Der Laan and Dudoit (2003) or Van der Laan et al. (2007).

T.6.1 Oracle result for cross-validated model selection

The attractiveness of the Super Learner stems from a number of theoretical guarantees, most notably the so-called 'oracle result.' This result states that using a Super Learner to select amongst various candidate algorithms $\mathcal{K}$ will work asymptotically as well as knowing the best (selection of) algorithm(s) in $\mathcal{K}$ a-priori in case the true parametric model is *not* included in $\mathcal{K}$. In case the true parametric model *is* included the Super Learner model converges at a $\frac{\log n}{n}$ rate, which is considered an efficient price for the flexibility the Super Learner provides. However, we typically reside in the first of the two cases.

Assume $\hat{k}$ to be the model resulting from the Super Learner routine — where a weighted model based on the complete Super Learner is treated as a model as well. Take the a-priori best model k^* in $\mathcal{K}$, which leads to $m_{k*}(x) = \mathrm{E}(Y|X = x)$. We typically don't know $m_{k^*}(x)$ because it requires knowledge of the true underlying DGP. The model $m_{k^*}(x)$ is therefore considered the 'oracle' result. Instead, we want to use the Super Learner to approximate $m_{k^*}(x)$. Specifically, we are interested in how well the selected model $m_{\hat{k}}(X)$ compares to $m_{k^*}(x)$ and the price we have to pay for including models in $\mathcal{K}$.

If we implement cross validation dividing the dataset into B blocks to select $m_{\hat{k}}(x)$, we obtain the following estimate for the squared difference between $m(x)$ and $\hat{m}_{k^*}(x)$:

$$R_n^*(k) = \frac{1}{B}\sum_{b=1}^{B} \mathrm{E}_X[(m(X) - \hat{m}_{k^*}(X|x \notin B))^2]. \tag{T.25}$$

The oracle result due to Van Der Laan and Dudoit (2003) shows that if the outcome Y and estimator $\hat{m}_{k^*}$ are both bounded, then:

$$\mathrm{E}[R_n^*(\hat{k})] \leq (1+2\epsilon)\mathrm{E}[R_n^*(k)] + 2C_\epsilon(\frac{1+\log|K|}{n/B}), \tag{T.26}$$

for arbitrary $\epsilon > 0$, C_ϵ a constant depending on ϵ and $|K|$ the number of models evaluated in the Super Learner. This inequality bounds the extent to which the selected model $\hat{k}$ deviates from the optimal model k^*. This deviance increases in the number of models considered through $|K|$ in the denominator and decreases with n. Asymptotically, if the true DGP is not included in $\mathcal{K}$, which is the most realistic case, then the ratio between the oracle and the Super Learner selected model converges at the same rate. If not, the Super Learner converges at the $\frac{\log n}{n}$ rate which is considered as almost the parametric rate of convergence. This means that the price to pay for using a Super Learner is low. For proofs and simulation results, the reader is referred to Van Der Laan and Dudoit (2003); Van der Vaart et al. (2006).

Although a Super Learner is very simple to set up, there are software packages available like `SuperLearner` (Polley et al., 2021) in `R`. Ensemble methods that rely on the same cross validation approach are part of the `sklearn` library in `Python` (Pedregosa et al., 2011a).

T.7 Shapley values

In game theory, Shapley values are defined as follows. Consider a game $\mathcal{G}$ consisting of $\mathcal{M}$ possible players. Further assume that the game has some outcome $v(\mathcal{S})$ where $\mathcal{S} \subseteq \mathcal{M}$. Thus, there might be a subset of the total set of players who play the game. Shapley values have been developed as a way to allocate each of the players in $\mathcal{S}$ a part of the game's outcome $v(\mathcal{S})$. Clearly, there are many ways to divide $v(\mathcal{S})$ amongst $\mathcal{S}$ but Shapley values are the only division that satisfy a number of properties discussed below. Shapley values are estimated as follows:

$$\phi_j(v(\mathcal{S})) = \phi_j = \sum_{\mathcal{S}\subseteq\mathcal{M}/j} \frac{|\mathcal{S}|!(M-|\mathcal{S}|-1)!}{M!}(v(\mathcal{S}\cup j) - v(\mathcal{S})), \quad j = 1,\ldots,M. \tag{T.27}$$

Where each player j gets their part of the payoff, equal to ϕ_j.

The Shapley value ϕ_j for player j should be seen as a weighted sum of the differences in the outcome of the game $v(\cdot)$ given some set of players S when including player j to the set of players, versus the payoff without including player j.

Take as example $\mathcal{M} = [1,2,3,4]$, then the exact Shapley values are:

$$\begin{aligned}
\phi_1 =& \frac{1}{4}(v([1,2,3,4]) - v([2,3,4])) + \frac{1}{12}[(v([1,3,4]) - v([3,4]))+ \\
& (v([1,2,4]) - v([2,4])) + (v([1,2,3]) - v([2,3])) + (v([1,4]) - v([4]))+ \\
& (v([1,2]) - v([2])) + (v([1,3]) - v([3]))] + \frac{1}{4}(v([1]) - v([\oslash])) \\
\phi_2 =& \frac{1}{4}(v([1,2,3,4]) - v([1,3,4])) + \frac{1}{12}[(v([2,3,4]) - v([3,4]))+ \\
& (v([1,2,4]) - v([1,4])) + (v([1,2,3]) - v([1,3])) + (v([1,2]) - v([1]))+ \\
& (v([2,4]) - v([4])) + (v([2,3]) - v([3]))] + \frac{1}{4}(v([2]) - v([\oslash])) \\
\phi_3 =& \frac{1}{4}(v([1,2,3,4]) - v([1,2,4])) + \frac{1}{12}[(v([2,3,4]) - v([2,4]))+ \\
& (v([1,3,4]) - v([1,4])) + (v([1,2,3]) - v([1,2])) + (v([1,3]) - v([1]))+ \\
& (v([2,3]) - v([2]))] + \frac{1}{4}(v([3]) - v([\oslash])) \\
\phi_4 =& \frac{1}{4}(v([1,2,3,4]) - v([1,2,3])) + \frac{1}{12}[(v([2,3,4]) - v([2,3]))+ \\
& (v([1,3]) - v([1,3,4])) + (v([1,2,4]) - v([1,2])) + (v([1,4]) - v([1]))+ \\
& (v([2,4]) - v([2]))] + \frac{1}{4}(v([4]) - v([\oslash])).
\end{aligned} \tag{T.28}$$

Note that by adding all four Shapley values together and defining $\phi_0 = v(\oslash)$ we see that $\phi_0 + \phi_1 + \phi_2 + \phi_3 + \phi_4 = v([1,2,3,4])$. In other words, the Shapley values sum up to the outcome of the game. This is the so-called completeness property. Three other properties of Shapley values are:

1. If $v(\mathcal{S}\cup i) = v(\mathcal{S}\cup j)$ for all coalitions $\mathcal{S}$ then $\phi_i = \phi_j$.

2. If including player j to the coalition $\mathcal{S}$ never changes $v(\mathcal{S})$ then $\phi_j = 0$.

3. If two games with outcomes $v(\cdot)$ and $w(\cdot)$ are combined then $\phi_j(v + w) = \phi_j(v) + \phi_j(w)$ and for any real number a, $\phi_j(av) = a\phi_j$.

As shown by Shapley (1953) and Young (1985) the Shapley values as defined in (T.27) are the only distribution function for which these properties hold.

To adapt Shapley values as a method to explain a prediction $\hat{y}_i$ made by some model $\hat{f}(\cdot)$, we exchange $v(\mathcal{S})$ with $\hat{f}(x_i)$ and define:

$$\hat{f}(x_i) = \phi_0 + \sum_{k=1}^{K} \phi_k, \tag{T.29}$$

where $\phi_0 = \mathrm{E}(\hat{f}(x))$. In other words, the Shapley values ϕ_k capture the difference between $\hat{f}(x_i)$ and the overall mean prediction ϕ_0. As was the case for the game theoretic setup, Shapley values for predictions are the only additive method that satisfy the above mentioned properties (Lundberg and Lee, 2017).

To calculate Shapley values for the prediction case, we substitute $\hat{f}(\cdot)$ for $v(\cdot)$ where we consider sub-sets $\mathcal{S} \subseteq \mathcal{K}$ of covariates instead of players. In other words, each player's contribution to the game's outcome in the game theoretical setup now represents each covariate's contribution to the prediction of a model. This requires evaluating the function $\hat{f}(\cdot)$ whilst excluding some covariates from the model. That is, replacing $v(\mathcal{S})$ in (T.27) with $\mathrm{E}[\hat{f}(x|x = x_{\mathcal{S}})]$. Generally speaking, calculating exact Shapley values is computationally challenging as the number of sets $\mathcal{S}$ is $2^{|\mathcal{K}|}$.

T.7.1 Kernel SHAP

A computationally efficient way of estimating Shapley values has been developed by Lundberg and Lee (2017). Their approach, Kernel SHAP, relies on reformulating the problem of estimating ϕ_j for $j = 0, \dots K$ as the following weighted least squares (WLS) problem:

$$\sum_{\mathcal{S} \subseteq \mathcal{K}} (\mathrm{E}[\hat{f}(x|x = x_{\mathcal{S}})] - (\phi_0 + \sum_{k \in \mathcal{S}} \phi_k))^2 \mathrm{k}(\mathcal{K}, \mathcal{S}). \tag{T.30}$$

In this setup, the function $\mathrm{k}(\cdot)$ reflects the so-called Shapley kernel weights: $\mathrm{k}(\mathcal{K}, \mathcal{S}) = \frac{|\mathcal{K}|-1}{\binom{|\mathcal{K}|}{|\mathcal{S}|}|\mathcal{S}|(\mathcal{K}-|\mathcal{S}|)}$.[10] If we define $\mathbf{Z}$ to be a $2^{\mathcal{K}} \times (\mathcal{K}+1)$ matrix consisting of zeroes and ones indicating the inclusion of covariate k into the set $\mathcal{K}$ with a leading column of ones for ϕ_0 and if we then define $\mathbf{W}$ to be a $2^{\mathcal{K}} \times 2^{\mathcal{K}}$ matrix with the Shapley kernel weights defined above, and finally $\mathbf{f}$ to be a vector containing $\hat{f}(\mathcal{K}) = \mathrm{E}[\hat{f}(x|x = x_{\mathcal{S}})]$ for all $\mathcal{S}$ and ordered in line with $\mathbf{Z}$, the setup in (T.30) can be rewritten as:

$$(\mathbf{f} - \mathbf{Z}\phi)^{\mathrm{T}}\mathbf{W}(\mathbf{f} - \mathbf{Z}\phi), \tag{T.31}$$

which leads to the standard WLS solution $\phi = (\mathbf{Z}^{\mathrm{T}}\mathbf{W}\mathbf{Z})^{-1}\mathbf{Z}^{\mathrm{T}}\mathbf{W}\mathbf{f}$. The benefit of this formulation is that the matrix multiplication $\mathbf{R} = (\mathbf{Z}^{\mathrm{T}}\mathbf{W}\mathbf{Z})^{-1}\mathbf{Z}^{\mathrm{T}}\mathbf{W}$ only has to be calculated once as it does not depend on $\mathcal{K}$. Calculating $\mathbf{R}$ still requires a considerable matrix inversion, but by sampling a subset $\mathcal{K}'$ of $\mathcal{K}$ according to the Shapley kernel weights and including only the associated elements in $\mathbf{f}'$ and $\mathbf{Z}'$ allows for a considerably more computationally tractable approximation to (T.30). This is the first key innovation by Lundberg and Lee (2017) and makes Kernel SHAP a computationally feasible approximation to the exact Shapley values in (T.27).

The second component required to calculate Shapley values is the vector $\mathbf{f}$ which consists of the function evaluation under all $2^{\mathcal{K}}$ possible combinations of covariates — or $2^{\mathcal{K}'}$ in case of the sampling scheme mentioned above. To assess $\hat{f}(\mathcal{S}) = \mathrm{E}[\hat{f}(x|x = x_{\mathcal{S}})]$, one approach is to integrate out all covariates that are not included in $\mathcal{S}$ from the expectation:

$$\mathrm{E}[\hat{f}(x)|x = x_{\mathcal{S}}] = \int \hat{f}(x_{\bar{\mathcal{S}}}, x_{\mathcal{S}})\mathrm{p}(x_{\bar{\mathcal{S}}}|x = x_{\mathcal{S}})d\mathbf{x}_{\bar{\mathcal{S}}}, \tag{T.32}$$

where $\mathrm{p}(x_{\bar{\mathcal{S}}}|x = x_{\mathcal{S}})$ is the conditional distribution of the covariates not included in $\mathcal{S}$, indicated with $\bar{\mathcal{S}}$. In principle, this density is conditional on the observed values of those covariates which *are* included in $\mathcal{S}$: $x_{\mathcal{S}}$. The standard Kernel

[10] Note that the Shapley kernel weights are not defined for $\mathrm{k}(\mathcal{K}, \mathcal{K})$ as the denominator is zero. This can be addressed by setting $k(\mathcal{K}, \mathcal{K})$ to some high constant or setting $\phi_0 = \hat{f}(\oslash)$ and $\sum_{k=0}^{\mathcal{K}} = \hat{f}(\mathcal{K})$ as additional constraints instead (Lundberg and Lee, 2017, p. 6).

SHAP by Lundberg and Lee (2017) assumes independence amongst features and thus replaces the conditionality in the distribution and simply implements $p(x_{\bar{\mathcal{S}}})$, the empirical density of the variables. In case such independence assumptions are unlikely, various alternative solutions have been developed to incorporate the conditionality between $x_{\mathcal{S}}$ and $x_{\bar{\mathcal{S}}}$ (Aas et al., 2021).

T.7.2 Tree SHAP

An alternative estimation approach to $\hat{f}(\mathcal{S}) = \mathrm{E}[\hat{f}(x|x = x_{\mathcal{S}})]$ is available for tree-based methods, Tree SHAP, due to Lundberg et al. (2020). By exploiting tree structures, exact Shapley values can be calculated instead of relying on the sampling approach applied in Kernel SHAP. Whereas exact Shapley value calculation would require a summation over all possible sets $\mathcal{S}$, Tree SHAP only evaluates the decision paths for a set $x_{\mathcal{S}}$ throughout the tree and is thus effectively capped through the number of partitions in the tree. The basic approach to estimate $\hat{f}(\mathcal{S}) = \mathrm{E}[f(x|x = x_{\mathcal{S}})]$ using a tree is the following recursive algorithm:

Algorithm T.3: Tree SHAP algorithm

```
G(j) :=;
Receive node j;
if node j is a leaf then
 | return fitted value of leaf
else
 | if node's split variable x_j ∈ x_S then
 | | if x_j <= s_j then
 | | | return G(j_L)
 | | else
 | | | return G(j_R)
 | | end
 | else
 | | return (G(j_L)·r_L + G(j_R)·r_R) / (r_L + r_R)
 | end
end
```

This algorithm $G(\cdot)$ is initialized at the root node of the tree and first evaluates whether the node is a leaf — i.e., a terminal node. If so, the leaf estimate is simply returned as $\hat{f}(\mathcal{S})$. If the node is not a leaf, the algorithm assesses whether the

split variable x_j used to split the node further into $[j_L, j_R]$ is in the set $\mathcal{S}$. If so, it is evaluated whether the observed value of x_j for the observation for which a prediction is being explained lies in j_L or j_R and the algorithm is repeated on the child node within which the observation falls. If $x_j \notin \mathcal{S}$ the algorithm is applied to both nodes, and a weighted average of both is used based on r_L or r_R, which are the proportions of data points falling in either of the two nodes. In other words, recursively applying $G(\cdot)$ the algorithm follows the effective tree structure until it reaches a node j for which the split variable is not part of $\mathcal{S}$ and from there onwards uses a weighted average of $G(\cdot)$ applied to both nodes.

A key development in Lundberg et al. (2020) is a much more computationally efficient version of Algorithm T.3 that elegantly stores information of paths that have already been followed within previous passes of the algorithm. The reader is referred to the discussion in Lundberg et al. (2020) (p. 65) and the source code documentation, located at `https://github.com/slundberg/shap`. The advantage of the Tree SHAP approach is that the assumption of covariate independence is relaxed in the process of determining $\hat{f}(\mathcal{S}) = \mathrm{E}[\hat{f}(x|x = x_{\mathcal{S}})]$, as the approach is conditional on the tree Π to generate estimates of $\hat{f}(\mathcal{S})$ whilst using the information available in $\mathcal{S}$ (cf. Aas et al. (2021)).

T.7.3 Pairwise Shapley values

The dramatic decrease in computation time required to estimate Tree SHAP as compared to Kernel SHAP allows for the calculation of exact Shapley values, but also for calculation of local interaction effects through pairwise Shapley values. These pairwise Shapley values go beyond the more simple linear decomposition of a prediction along the K covariates as reflected in (T.29). SHAP interaction values calculate the difference between a Shapley value ϕ_k for some observation i when covariate j is present versus when it is absent. Specifically, it is estimated through:

$$\Phi_{j,k}(f, x) = \sum_{\mathcal{S} \subseteq \mathcal{K}\ k,j} = \frac{|\mathcal{S}|!(|\mathcal{K}| - |\mathcal{S}| - 2)!)}{2(|\mathcal{K}| - 1)!} \nabla_{j,k}(f, x, \mathcal{S}), \tag{T.33}$$

for $j \neq k$ where

$$\nabla_{j,k}(f, x, \mathcal{S}) = f_x(\mathcal{S} \cup j, k) - f_x(\mathcal{S} \cup j) - f_x(\mathcal{S} \cup k) + f_x(\mathcal{S}). \tag{T.34}$$

Here, $f_x(\cdot)$ is the standard formula to estimate a Shapley value as defined in (T.27). Effectively, the direct Shapley value for variable j is $\Phi_{j,j}(f, x)$ and is defined as the standard Shapley value $\phi_j(f, x)$ minus the interaction values $\sum_{j \neq k} \Phi_{j,k}(f, x)$. The intuition behind (T.34) is that the pairwise Shapley value between variable k and variable j is a weighted average of all sets of covariates excluding k and j: $\mathcal{S}/\{j, k\}$. For each set the Shapley value is first estimated when including *both* k and j — the first term in (T.34). From this, the Shapley values when *either* including k or j are subtracted — the second and third terms in (T.34). Finally, the Shapley value from including *neither* k and j is added — the last term in (T.34). Again, using the same efficient implementation as discussed above allows for computationally tractable estimation of these pairwise values (Lundberg et al., 2020, p. 66).

Shapley value estimation using the Kernel SHAP approach as well as the Tree SHAP approach have been implemented in `Python` through the `Shap` package (Lundberg and Lee, 2017) and have been incorporated in various packages using popular tree-based methods (`https://github.com/slundberg/shap`). The package `shapr` is available for `R` and includes various methods to deal with feature dependence in the Kernel SHAP setup (Sellereite and Jullum, 2020). The general introduction of Shapley values and the Kernel SHAP approach is based on the discussions in Aas et al. (2021) and Lundberg and Lee (2017). The discussion of the Tree SHAP approach and pairwise Shapley values is based on Lundberg et al. (2020). The proofs for the properties of Shapley values within the original game theoretical approach can be found in Shapley (1953) and Young (1985). The analogue proofs for the prediction case can be found in Lundberg and Lee (2017). I implement the standard Tree SHAP throughout Chapter 4, and estimate pairwise Shapley values in the GSS example therein.

T.8 Tree-based methods for effect heterogeneity

Adapting tree-based methods discussed in Section T.5 to estimate other estimands than the outcome y provides a considerable challenge to the standard estimation routine for tree-based methods. Specifically, in standard tree-based methods cross-validated model fit is used to prevent overfitting (Section T.2.2 and T.5.1). Using out-of-sample model fit metrics is typically not possible for estimands that are not explicitly observed — and if they were, they could simply be used as the outcome of interest within a standard decision tree framework. The first challenge is thus to ensure that the tree does not overfit to the data without using OOS fit metrics. A second challenge is that when using trees to estimate specific estimands, it typically requires statistical properties like consistency as well as uncertainty estimates like confidence intervals for meaningful inference. These have been known to be difficult to estimate for tree-based methods, with most standard decision trees not even pointwise consistent (Wager and Athey, 2018; Wasserman and Roeder, 2009).

T.8.1 Honest estimation

In standard trees, both the need to adjust for overfitting and bias in leaf estimands stem from the same general issue. This issue is that the same sample is repeatedly used throughout the tree-building process. Assume we evaluate the fit of a partition through a mean squared error (MSE) statistic, then the MSE will strictly improve as the tree becomes deeper and deeper. However, the expected MSE (EMSE) on a new sample will likely increase after some level of depth. This is because both the partition as well as the estimand is becoming overfit to the sample at hand. The tree will start grouping together extreme values of the outcome leading to the possibility for bias in the leaf estimates when compared to applying the same decision rule $x_i \in R_j$ to a new sample.

A solution to this problem is honest estimation (Athey and Imbens, 2016). Define Π to be a tree consisting of J regions R_j. The entire sample is $\mathcal{S}^{\text{tr}}$ and this sample can be divided into $\mathcal{S}^{\text{tr}} = [\mathcal{S}^{\text{est}}, \mathcal{S}^{\text{te}}]$. The estimand function is $\mu(x; \Pi)$ which in the standard case is simply the function assigning the average outcome

in region R_j as the predicted outcome for all $x \in R_j$. In order to address the problem of using the same sample to estimate both the tree Π and $\mu(x;\Pi)$ we adapt the standard MSE, which can be written as:

$$-\operatorname{MSE}_\mu(\mathcal{S}^{\text{tr}},\Pi) = \frac{1}{N^{\text{tr}}}\sum_{i\in\mathcal{S}^{\text{tr}}}(y_i - \hat{\mu}^2(X_i;\mathcal{S}^{\text{tr}},\Pi))^2. \tag{T.35}$$

This metric can be adapted into the EMSE which instead estimates the MSE for a dataset that is drawn from the same distribution as $\mathcal{S}^{\text{tr}}$, rather than the MSE of the very same sample. This is done by using $\mathcal{S}^{\text{est}}$ and $\mathcal{S}^{\text{te}}$:

$$-\operatorname{EMSE}_\mu(\Pi) \equiv \operatorname{E}_{\mathcal{S}^{\text{te}},\mathcal{S}^{\text{est}}}[\operatorname{MSE}_\mu(\mathcal{S}^{\text{te}},\mathcal{S}^{\text{est}},\Pi)]. \tag{T.36}$$

Following Athey and Imbens (2016) the MSE metric in (T.35) can be adjusted by subtracting Y_i^2 from the standard MSE, as it does not rely on an estimator and thus does not affect the ordering of estimators in terms of loss:

$$-\operatorname{MSE}_\mu(\mathcal{S}^{\text{est}},\mathcal{S}^{\text{te}},\Pi) = \frac{1}{N^{\text{te}}}\sum_{i\in\mathcal{S}^{\text{te}}}[Y_i - \hat{\mu}(X_i;\mathcal{S}^{\text{est}},\Pi)^2 - Y_i^2], \tag{T.37}$$

where $\mathcal{S}^{\text{est}}$ and $\mathcal{S}^{\text{te}}$ are independent from one another. This then allows for the following decomposition:

$$\begin{aligned}
-\operatorname{EMSE}_\mu(\Pi) =& -\operatorname{E}_{\mathcal{S}^{\text{est}}}[(Y_i - \mu(X_i;\Pi))^2 - Y_i^2] \\
&- \operatorname{E}[\hat{\mu}(X_i;\mathcal{S}^{\text{est}},\Pi) - \mu(X_i;\Pi)^2] \\
=& E_{X_i}[\mu^2(X_i;\Pi)] - \operatorname{E}_{\mathcal{S}^{\text{est}},X_i}[\operatorname{V}(\hat{\mu}^2(X_i;\mathcal{S}^{\text{est}},\Pi))].
\end{aligned} \tag{T.38}$$

The key is that $\operatorname{E}_{\mathcal{S}}(\hat{\mu}(x;\mathcal{S},\Pi)$ is consistent for $\mu(x;\Pi)$ as long as the sample used to estimate the estimand is independent from the sample used to build the tree Π. The intuition behind (T.38) is that the EMSE goes down as the variability of $\mu(x;\Pi)$ increases (first term in the final line), which means that there is a lot of variability between regions R_j. The metric decreases, however, if the variance within each leaf goes up (the second term).

Using the EMSE instead of the MSE means it is no longer needed to address optimism in standard tree-building processes via tuning parameters controlling the complexity of the tree as the loss criterion is no longer biased. The price paid is that a smaller sample is available for tree-building which can lead to more variance.

Note, however, that (T.38) is only consistent for fixed Π. When repeatedly using the same sample to estimate partition rules in Π, this again can lead to optimism in the goodness-of-fit. Here, cross-validation can come in handy by using a cross-validation sample $\mathcal{S}^{\text{tr, cv}}$ which has not been used in making splits to evaluate candidate splits when building the tree (Athey and Imbens, 2016, p. 7356).

T.8.2 Causal trees

The key problem with the honest estimation approach outlined above is that when we are interested in an estimand like the treatment effect $\tau_i = Y_i^{(1)} - Y_i^{(0)}$ we do not observe this outcome in practice. Therefore, we cannot plug values for Y_i into (T.38). However, as Athey and Imbens (2016) show, it is possible to estimate it nonetheless. Assuming we are interested in the ATE conditional on observables $X_i = x$ we have:

$$\tau(x) = \mathrm{E}[Y_i^{(1)} - Y_i^{(0)} | X_i = x], \tag{T.39}$$

which is the general framework for estimating treatment effects assuming unconfoundedness conditional on observables. This allows us to simply compare the outcomes of those that were treated conditional on $X_i = x$ with those untreated. In the tree context, we can thus use:

$$\tau(x) = \mathrm{E}[Y_i^{(1)} - Y_i^{(0)} | X_i \in R_j]. \tag{T.40}$$

That is, the difference between the treated and untreated in region R_j. We can simply plug $\hat{\tau}(x)$ into (T.38) instead of $\hat{\mu}(x)$ which leads to:

$$-\mathrm{EMSE}_\tau(\Pi) = E_{X_i}[\tau^2(X_i; \Pi)] - \mathrm{E}_{\mathcal{S}^{\text{est}}, X_i}[\mathrm{V}(\hat{\tau}^2(X_i; \mathcal{S}^{\text{est}}, \Pi))]. \tag{T.41}$$

Again, the criterion rewards heterogeneity in $\hat{\tau}$, which means it rewards leaves that differ from one another. At the same time, the criterion rewards leaves that estimate the treatment effect more precisely — i.e., with lower variance. Note that in the standard framework in (T.38), both terms typically lead to splits that reduce the MSE. However, in (T.41) a split might be placed that does not lead to better approximation of the heterogeneity, but instead in more precisely estimated treatment effects. This is somewhat akin to the loss functions discussed for binary outcomes that reward leaf purity (Section T.5.1).

T.8.3 Causal forests

The same rationale motivating the use of random forests with respect to decision trees also hold for causal trees. Whereas multiple causal trees might have equal performance, making it difficult to decide which one is 'best', a causal forest estimates B different causal trees to samples of the data and averages them to obtain $\hat{\tau}(x) = \frac{1}{B}\sum_{b=1}^{B}\hat{\tau}_b(x)$. The causal forest still leads to an estimate of the individual-level treatment effect $\hat{\tau}(x_i)$ but exploiting many de-correlated trees to improve precision (Section T.5.2). A key development in the literature on causal forests which sets them apart from other ML techniques that attempt to identify treatment heterogeneity is that under relatively mild assumptions causal forest predictions of treatment effects follow an asymptotic normal distribution (Wager and Athey, 2018). Normality relies on frequent resampling of the dataset for different trees B according to a sub-sample of size $s \ll n$, rather than also sampling from the feature space as is customary in random forests. The general idea is to rely on variation in trees B to provide bootstrap-type confidence intervals by setting the number of trees B sufficiently large that variation in the estimates is solely due to sampling variation. Since the statistical properties of forest-based predictions are not applied in this thesis, I refer the reader to the paper by Wager and Athey (2018) for further reading on this topic. However, the consistency of causal trees and forests under honest estimation and ability to define uncertainty bounds for the latter are a key benefit of tree-based effect heterogeneity over alternative ML methods with a similar goal.

Various versions of the causal tree algorithm have been implemented in the `causalTree` (Athey and Imbens, 2016) package in `R`. Causal forests have been implemented in the `grf` (Tibshirani et al., 2022) package. The above exposition relies on the paper by Athey and Imbens (2016) and Wager (2014) for the discussion on honest estimation and the causal tree algorithm. The discussion on causal forests relies on the paper by Wager and Athey (2018). I implement both causal trees and causal forests in Chapter 5 using the `causalTree` and `grf` packages.

T.9 Code and packages

All code and instructions to download data underlying the analyses in this thesis can be found at `https://github.com/MarkDVerhagen/DPhil_Thesis_Computer_Age_Sociology`. Not all data can be publicly shared, or requires permission by the data provider in which case this is noted in the documentation in each chapter's repository. The analyses throughout this thesis rely on the open source programming languages `R` (4.2.1) (R Core Team, 2022) and `Python` (3.8.10) (Van Rossum and Drake Jr, 1995). Table T.1 provides a list of all the packages used throughout this thesis, including a brief description and documentation.

Package name	Version	Description	Documentation	Reference
R packages				
tidyverse	1.3.1	Convenience for data wrangling.	https://www.tidyverse.org	(Wickham, 2017a)
data.table	1.14.2	Faster data wrangling.	https://cran.r-project.org/web/packages/data.table/index.html	(Dowle et al., 2019)
assertthat	0.2.1	General unit testing.	https://cran.r-project.org/web/packages/assertthat/index.html	-
haven	2.5.0	Convenient loading of .dta files.	https://cran.r-project.org/web/packages/haven/index.html	-
xgboost	1.6.0.1	Implementation of extreme gradient boosting.	https://cran.r-project.org/web/packages/xgboost/xgboost.pdf	(Chen and Guestrin, 2016)
caret	6.0.93	CART regression and decision tree estimation.	https://cran.r-project.org/web/packages/caret/index.html	(Kuhn et al., 2020)
randomForest	4.7.1.1	Implementation of random forest.	https://cran.r-project.org/web/packages/randomForest/index.html	(Liaw et al., 2002)
ggsci	2.9	Colour schemes for plotting.	https://cran.r-project.org/web/packages/ggsci/index.html	(Xiao et al., 2018)
reshape2	1.4.4	Reshaping data from wide to long.	https://cran.r-project.org/web/packages/reshape2/index.html	(Wickham, 2017b)
patchwork	1.1.1	Arranging ggplot objects.	https://cran.r-project.org/web/packages/patchwork/index.html	(Pedersen, 2019)
cowplot	1.1.1	Formatting ggplot objects.	https://cran.r-project.org/web/packages/cowplot/index.html	(Wilke et al., 2019)
scales	1.2.0	Formatting ggplot axes.	https://cran.r-project.org/web/packages/scales/index.html	(Wickham et al., 2016)
grid	4.2.1	Formatting ggplot objects.	https://cran.r-project.org/web/packages/grid/index.html	(Murrell, 2002)
gridExtra	2.3	Formatting ggplot objects.	https://cran.r-project.org/web/packages/gridExtra/index.html	(Auguie et al., 2017)
MetBrewer	0.2.0	Colour schemes for plotting.	https://cran.r-project.org/web/packages/MetBrewer/index.html	-
lubridate	1.8.0	Handling of date objects.	https://cran.r-project.org/web/packages/lubridate/index.html	(Spinu, 2016)
MASS	7.3.58	General statistical convenience functions.	https://cran.r-project.org/web/packages/MASS/index.html	(Ripley et al., 2013)
ordinal	2019.12.10	Ordered models with unobserved heterogeneity.	https://cran.r-project.org/web/packages/ordinal/ordinal.pdf	(Christensen, 2015)
lme4	1.1.30	Estimating hierarchical models.	https://cran.r-project.org/web/packages/lme4/index.html	(Bates et al., 2014)
texreg	1.38.6	Output tables to LaTeXformat.	https://cran.r-project.org/web/packages/texreg/index.html	(Leifeld, 2013)
fastDummies	1.6.3	Generate dummies of categorical variables.	https://cran.r-project.org/web/packages/fastDummies/index.html	(Kaplan, 2020)
SuperLearner	2.0.28	Implementation of the Super Learner.	https://cran.r-project.org/web/packages/SuperLearner/index.html	(Van der Laan et al., 2007)
doParallel	1.0.17	Package to enable parallel computing.	https://cran.r-project.org/web/packages/doParallel/index.html	(Calaway et al., 2015)
parallel	4.2.1	Package to enable parallel computing.	https://cran.r-project.org/web/views/HighPerformanceComputing.html	-
lmtest	0.9.40	Package to perform tests on linear models.	https://cran.r-project.org/web/packages/lmtest/index.html	(Hothorn et al., 2015)
ggpubr	0.4.0	Formatting ggplot objects.	https://cran.r-project.org/web/packages/ggpubr/ggpubr.pdf	(Kassambara, 2020)
vtable	1.3.4	Generating tables efficiently.	https://cran.r-project.org/web/packages/vtable/index.html	(Huntington-Klein, 2020)
WeightIt	0.13.1	Convenience functions for balancing.	https://cran.r-project.org/web/packages/WeightIt/index.html	(Greifer, 2019)
ebal	0.1.8	Convenience package for entropy balancing.	https://cran.r-project.org/web/packages/ebal/index.html	(Hainmueller, 2022)
cobalt	4.4.0	Visualizing covariate balance	https://cran.r-project.org/web/packages/cobalt/index.html	(Greifer, 2020)
grf	2.1.0	Estimation of causal forests.	https://cran.r-project.org/web/packages/grf/index.html	(Wager and Athey, 2018)
causalTree	dev	Estimation of causal trees.	https://github.com/susanathey/causalTree	(Athey and Imbens, 2016)
rpart.plot	3.1.1	Visualization of trees.	https://cran.r-project.org/web/packages/rpart.plot/index.html	(Milborrow, 2019)
rpart	4.1.16	Estimating trees.	https://cran.r-project.org/web/packages/rpart/index.html	(Therneau et al., 2015)
Python packages				
pandas	1.4.4	Convenience for data wrangling	https://pandas.pydata.org/	(pandas development team, 2020)
numpy	1.23.3	Handling of arrays.	https://numpy.org/	(Harris et al., 2020)
shap	0.41.0	Estimate Shapley values.	https://github.com/slundberg/shap	(Lundberg and Lee, 2017)
sklearn	1.1.2	Library for machine learning.	https://scikit-learn.org/stable/	(Pedregosa et al., 2011b)
statsmodels	0.13.2	Convenience for classical model inference.	https://www.statsmodels.org/stable/index.html	(Seabold and Perktold, 2010)
joblib	1.2.0	Functionality for parallel computing.	https://joblib.readthedocs.io/en/latest/	(Joblib Development Team, 2020)
xgboost	1.6.2	Implementation of extreme gradient boosting.	https://github.com/dmlc/xgboost	(Chen and Guestrin, 2016)

Table T.1: R (4.2.1) and Python (3.8.10) packages used throughout this thesis.

Bibliography

Kjersti Aas, Martin Jullum, and Anders Løland. Explaining individual predictions when features are dependent: More accurate approximations to Shapley values. *Artificial Intelligence*, 298, 2021.

Andrew Abbott. Transcending general linear reality. *Sociological Theory*, pages 169–186, 1988.

Seye Abimbola, Leonard Baatiema, and Maryam Bigdeli. The impacts of decentralization on health system equity, efficiency and resilience: A realist synthesis of the evidence. *Health Policy and Planning*, 34(8):605–617, 2019.

Mayank Agrawal, Joshua C Peterson, and Thomas L Griffiths. Scaling up psychology via scientific regret minimization. *Proceedings of the National Academy of Sciences*, 117(16):8825–8835, 2020.

Alan Agresti. *Analysis of ordinal categorical data*. John Wiley & Sons, 2010.

Karl L Alexander, Doris R Entwisle, and Maxine S Thompson. School performance, status relations, and the structure of sentiment: Bringing the teacher back in. *American Sociological Review*, pages 665–682, 1987.

MA Allers. Decentralisatie en gemeentelijke opschaling. *Liberaal Reveil*, 54(2): 119–124, 2013.

Chris Anderson. The end of theory: The data deluge makes the scientific method obsolete. *Wired magazine*, 16(7):16–07, 2008.

Alberta Andreotti, Enzo Mingione, and Emanuele Polizzi. Local welfare systems: a challenge for social cohesion. *Urban Studies*, 49(9):1925–1940, 2012.

George Ansalone. Tracking: Educational differentiation or defective strategy. *Educational Research Quarterly*, 34(2):3, 2010.

Susan Athey. Beyond prediction: Using big data for policy problems. *Science*, 355 (6324):483–485, 2017.

Susan Athey and Guido Imbens. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360, 2016.

Susan Athey and Guido W Imbens. Machine learning methods that economists should know about. *Annual Review of Economics*, 11:685–725, 2019.

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

Baptiste Auguie, Anton Antonov, and Maintainer Baptiste Auguie. Package 'gridextra'. *Miscellaneous Functions for "Grid" Graphics*, 2017.

Valerio Baćak and Edward H Kennedy. Principled machine learning using the Super Learner: An application to predicting prison violence. *Sociological Methods & Research*, 48(3):698–721, 2019.

Thomas Bahle, Michaela Pfeifer, and Claus Wendt. Social assistance. In *The Oxford handbook of the welfare state*. 2010.

Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. Fitting linear mixed-effects models using lme4. *arXiv preprint arXiv:1406.5823*, 2014.

Stephen Bates, Trevor Hastie, and Robert Tibshirani. Cross-validation: what does it estimate and how well does it do it? *arXiv*, 2021. doi: 2104.00673.

Dominik Becker. The impact of teachers' expectations on students' educational opportunities in the life course: An empirical test of a subjective expected utility explanation. *Rationality and society*, 25(4):422–469, 2013.

Janina Beiser-McGrath and Liam F Beiser-McGrath. Problems with products? Control strategies for models with interaction and quadratic effects. *Political Science Research and Methods*, 8(4):707–730, 2020.

Yoshua Bengio and Yves Grandvalet. No unbiased estimator of the variance of k-fold cross-validation. *Journal of Machine Learning Research*, 5(Sep):1089–1105, 2004.

Richard A Berk and Justin Bleich. Statistical procedures for forecasting criminal behavior: A comparative assessment. *Criminology & Pub. Pol'y*, 12(3):513–544, 2013.

Julian R Betts. The economics of tracking in education. In *Handbook of the Economics of Education*, volume 3, pages 341–381. Elsevier, 2011.

Christopher M Bishop. *Pattern recognition and machine learning*. Springer, 2006.

Matthew Blackwell and Michael P Olson. Reducing model misspecification and bias in the estimation of interactions. *Political Analysis*, pages 1–20, 2021.

Harry Blair. Participation and accountability at the periphery: Democratic local governance in six countries. *World Development*, 28(1):21–39, 2000.

Ines Böhmer, Cornelia Gräsel, Sabine Krolak-Schwerdt, Thomas Hörstermann, and Sabine Glock. Teachers' school tracking decisions. In *Competence Assessment in Education*, pages 131–147. Springer, 2017.

Thijs Bol, Jacqueline Witschge, Herman G. van de Werfhorst, and Jaap Dronkers. Curricular tracking and central examinations: Counterbalancing the impact of social background on student achievement in 36 countries. *Social Forces*, 92(4): 1545–1572, 2014.

Simon Boone, Sarah Thys, Piet Van Avermaet, and Mieke Van Houtte. Class composition as a frame of reference for teachers? The influence of class context on teacher recommendations. *British Educational Research Journal*, 44(2):274–293, 2018.

Heather Booth. Demographic forecasting: 1980 to 2005 in review. *International Journal of Forecasting*, 22(3):547–581, 2006.

Pierre Bourdieu and Jean Claude Passeron. *The inheritors: French students and their relation to culture*. Univ of Chicago Pr, 1979.

Jennie E Brand, Jiahui Xu, Bernard Koch, and Pablo Geraldo. Uncovering sociological effect heterogeneity using tree-based machine learning. *Sociological Methodology*, 52(2):189–223, 2021.

Hennie Brandsma and Simone Doolaard. Differences in effectiveness between primary schools and their impact on secondary school recommendations. *School Effectiveness and School Improvement*, 10(4):430–450, 1999.

Richard Breen and John Ermisch. Sibling models, categorical outcomes, and the intra-class correlation. *European Sociological Review*, 37(3):497–504, 2021.

Richard Breen and Jan O Jonsson. Inequality of opportunity in comparative perspective: Recent research on educational attainment and social mobility. *Annual review of sociology*, pages 223–243, 2005.

Richard Breen, Kristian Bernt Karlson, and Anders Holm. Interpreting and understanding logits, probits, and other nonlinear probability models. *Annual Review of Sociology*, 44:39–54, 2018.

Leo Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996.

Leo Breiman. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science*, 16(3):199–231, 2001.

Leo Breiman, Jerome H Friedman, Richard A Olshen, and Charles J Stone. *Classification and regression trees*. Routledge, 2017.

Jere E Brophy and Thomas L Good. Teachers' communication of differential expectations for children's classroom performance: Some behavioral data. *Journal of Educational Psychology*, 61(5):365, 1970.

Mauricio Bucca and Daniela R Urbina. Lasso regularization for selection of log-linear models: An application to educational assortative mating. *Sociological Methods & Research*, 50(4):1763–1800, 2019.

Andreas Buja, Lawrence Brown, Richard Berk, Edward George, Emil Pitkin, Mikhail Traskin, Kai Zhang, and Linda Zhao. Models as approximations i: Consequences illustrated with linear regression. *Statistical Science*, 34(4):523–544, 2019a.

Andreas Buja, Lawrence Brown, Arun Kumar Kuchibhotla, Richard Berk, Edward George, and Linda Zhao. Models as approximations ii: A model-free theory of parametric regression. *Statistical Science*, 34(4):545–565, 2019b.

Andreas Buja, Arun Kumar Kuchibhotla, Richard Berk, Edward George, Eric Tchetgen Tchetgen, and Linda Zhao. Models as approximations—rejoinder. *Statistical Science*, 34(4):606–620, 2019c.

David T Burkam, Laura LoGerfo, Doug Ready, and Valerie E Lee. The differential effects of repeating kindergarten. *Journal of Education for Students Placed at Risk*, 12(2):103–136, 2007.

Rich Calaway, Revolution Analytics, Steve Weston, Dan Tenenbaum, and Maintainer Rich Calaway. Package 'doparallel'. 2015.

A Colin Cameron and Pravin K Trivedi. *Microeconometrics: Methods and applications*. Cambridge university press, 2005.

Ayse Can. Specification and estimation of hedonic housing price models. *Regional Science and Urban Economics*, 22(3):453–474, 1992.

CBS (accessed 7th December 2020). Statline Data. https://www.cbs.nl/nl-nl/dossier/nederland-regionaal/wijk-en-buurtstatistieken/kerncijfers-wijken-en-buurten-2004-2019. URL `https://www.cbs.nl/nl-nl/dossier/nederland-regionaal/wijk-en-buurtstatistieken/kerncijfers-wijken-en-buurten-2004-2019`.

Lars-Erik Cederman and Nils B Weidmann. Predicting armed conflict: Time to adjust our expectations? *Science*, 355(6324):474–476, 2017.

Chapter 2. *Computer Age Sociology.*

Chapter 3. *Computer Age Sociology.*

Chapter 4. *Computer Age Sociology.*

Daniele Checchi and Luca Flabbi. Intergenerational mobility and schooling decisions in Italy and Germany. In *IZA Discussion Paper 2876, Bonn, Germany*. 2006.

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

Tianqi Chen, Tong He, Michael Benesty, Vadim Khotilovich, Yuan Tang, Hyunsu Cho, et al. Xgboost: extreme gradient boosting. *R package version 0.4-2*, 1(4): 1–4, 2015.

Justin Cheng, Lada Adamic, P Alex Dow, Jon Michael Kleinberg, and Jure Leskovec. Can cascades be predicted? In *Proceedings of the 23rd international conference on World wide web*, pages 925–936. 2014.

Bin Chi, Adam Dennett, Thomas Oléron-Evans, and Robin Morphet. A new attribute-linked residential property price dataset for england and wales, 2011 to 2019. *UCL Open: Environment Preprint*, 2021.

Anna K. Chmielewski. An international comparison of achievement inequality in within- and between-school tracking systems. *American Journal of Education*, 120(3):293–324, 2014.

Björn Christensen and Sören Christensen. Are female hurricanes really deadlier than male hurricanes? *Proceedings of the National Academy of Sciences*, 111(34): E3497–E3498, 2014.

Rune Haubo Bojesen Christensen. Ordinal — regression models for ordinal data. *R package version*, 28:2015, 2015.

Dorien Claessen, Quirine Eijkman, and Majda Lamkaddem. Social support and access to justice at the kitchen table? An assessment of the legal capabilities of community social care professionals in The Netherlands. *European Journal of Social Work*, pages 1–14, 2019.

Johan Coenen and Chris Van Klaveren. Better test scores with a same-gender teacher? *European Sociological Review*, 32(3):452–464, 2016.

Joan Costa-Font and Joan Gil. Exploring the pathways of inequality in health, health care access and financing in decentralized Spain. *Journal of European Social Policy*, 19(5):446–458, 2009.

Skyler J Cranmer and Bruce A Desmarais. What can we learn from predictive modeling? *Political Analysis*, 25(2):145–166, 2017.

R Crook and Alan S Sverrisson. Does decentralization contribute to poverty reduction. *Changing paths: International development and the new politics of inclusion*, 2003.

JA Davis and TW Smith. *The NORC General Social Survey: A User Guide*. SAGE, 1991.

Jonathan MV Davis and Sara B Heller. Rethinking the benefits of youth employment programs: The heterogeneous effects of summer jobs. *Review of Economics and Statistics*, 102(4):664–677, 2020.

Hester de Boer, Roel J Bosker, and Margaretha P. C van Der Werf. Sustainability of teacher expectation bias effects on long-term student performance. *Journal of Educational Psychology*, 102(1):168, 2010.

Paul M De Graaf and Harry BG Ganzeboom. Family background and educational attainment in the Netherlands for the 1891-1960 birth cohorts. *Persistent inequality: Changing educational attainment in thirteen countries*, pages 75–99, 1993.

A de Wijs, F Kamphuis, F Kleintjes, and M Tomesen. Wetenschappelijke verantwoording: Spelling voor groep 3 tot en met 8. Cito, 2010.

Misha Denil, David Matheson, and Nando De Freitas. Narrowing the gap: Random forests in theory and in practice. In *International conference on machine learning*, pages 665–673. PMLR, 2014.

Dienst Uitvoering Onderwijs (accessed 10th December 2020). Open onderwijsdata. URL `https://duo.nl/open_onderwijsdata/databestanden/po/leerlingen-po/po-totaal/bo-gewicht-leeftijd.jsp`.

Ben Domingue. Evaluating the equal-interval hypothesis with test score scales. *Psychometrika*, 79(1):1–19, 2014.

Ben Domingue, Klint Kanopka, Sam Trejo, and Jeremy Freese. An ordinal model for analysis of years of education. *SocArXiv*, 2021.

David Donoho. 50 years of data science. *Journal of Computational and Graphical Statistics*, 26(4):745–766, 2017.

Jurgen A Doornik and David F Hendry. Statistical model selection with "Big Data". *Cogent Economics & Finance*, 3(1), 2015.

Reinhold August Dorwart. *The Prussian welfare state before 1740*. Harvard University Press, 2013.

Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv*, 2017. doi: 1702.08608.

Michael R Dougherty, Rick P Thomas, Ryan P Brown, Jeffrey S Chrabaszcz, and Joe W Tidwell. An introduction to the general monotone model with application to two problematic data sets. *Sociological Methodology*, 45(1):223–271, 2015.

Matt Dowle, Arun Srinivasan, Jan Gorecki, Michael Chirico, Pasha Stetsenko, Tom Short, Steve Lianoglou, Eduard Antonyan, Markus Bonsch, Hugh Parsonage, et al. Package 'data. table'. *Extension of 'data. frame*, 2019.

G Driessen, E Smeets, L Mulder, and H Vierke. De relatie tussen prestaties en advies: Onder- of overadvisering bij de overgang van basis-naar voortgezet onderwijs. *Inspectie van het Onderwijs, Onderadvisering in beeld*, pages 59–81, 2007.

Geert Driessen, Peter Sleegers, and Frederik Smit. The transition from primary to secondary education: Meritocracy and ethnicity. *European sociological review*, 24 (4):527–542, 2008.

J Dronkers, M Van Erp, M Robijn, and J Roeleveld. Krijgen leerlingen in de grote steden en met name in Amsterdam te hoge adviezen? *Tijdschrift voor Onderwijsresearch*, 23:17–30, 1998.

JW Duyvendak and E Tonkens. Decentralisaties lijden onder een democratisch tekort. *Sociale Vraagstukken*, 2017.

Bradley Efron. Estimating the error rate of a prediction rule: improvement on cross-validation. *Journal of the American statistical association*, 78(382):316–331, 1983.

Bradley Efron. Better bootstrap confidence intervals. *Journal of the American statistical Association*, 82(397):171–185, 1987.

Bradley Efron. Bootstrap methods: another look at the jackknife. In *Breakthroughs in statistics*, pages 569–593. Springer, 1992.

Bradley Efron and Trevor Hastie. *Computer age statistical inference*. Cambridge University Press, 2016.

Bradley Efron and Robert J Tibshirani. *An introduction to the bootstrap*. CRC press, 1993.

Benjamin Elbers. A method for studying differences in segregation across time and space. *Sociological Methods & Research*, page 0049124121986204, 2021.

J Paul Elhorst. Applied spatial econometrics: Raising the bar. *Spatial Economic Analysis*, 5(1):9–28, 2010.

Per Engzell, Arun Frey, and Mark D Verhagen. Learning loss due to school closures during the COVID-19 pandemic. *Proceedings of the National Academy of Sciences*, 118(17):e2022376118, 2021.

Nicholas Epley and Thomas Gilovich. The mechanics of motivated reasoning. *Journal of Economic Perspectives*, 30(3):133–40, 2016.

Gang-Zhi Fan, Seow Eng Ong, and Hian Chye Koh. Determinants of house price: A decision tree approach. *Urban Studies*, 43(12):2301–2315, 2006.

H Feenstra, F Kamphuis, F Kleintjes, and R Krom. Wetenschappelijke verantwoording: Begrijpend lezen voor groep 3 tot en met 6. Cito, 2010.

Simon Field, Malgorzata Kuczera, and Beatriz Pont. No more failures: Ten steps to equity in eduation. Technical report, OECD, 2007.

Peter Flora and Arnold Joseph Heidenheimer. *The development of welfare states in Europe and America*. Transaction Publishers, 1981.

Ginevra Floridi, Ludovico Carrino, and Karen Glaser. Socioeconomic inequalities in home-care use across regional long-term care systems in Europe. *The Journals of Gerontology: Series B*, 76(1):121–132, 2021.

Malcolm R Forster. Predictive accuracy as an achievable goal of science. *Philosophy of Science*, 69(S3):S124–S134, 2002.

Nicole Fortin, Thomas Lemieux, and Sergio Firpo. Decomposition methods in economics. In *Handbook of Labor Economics*, volume 4, pages 1–102. Elsevier, 2011.

Michael Freeden, Lyman Tower Sargent, and Marc Stears. *The Oxford handbook of political ideologies*. Oxford University Press, 2013.

David A Freedman. Statistical models and shoe leather. *Sociological Methodology*, pages 291–313, 1991.

David A Freedman. *Statistical models: theory and practice*. Cambridge university press, 2009.

Yoav Freund and Robert E Schapire. Experiments with a new boosting algorithm. In *icml*, volume 96, pages 148–156, 1996.

Jerome H Friedman. Multivariate adaptive regression splines. *The Annals of Statistics*, 4(3):1–67, 1991.

Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

Andrew Gelman and Jennifer Hill. *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press, 2006.

Andrew Gelman and Eric Loken. The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time. *Department of Statistics, Columbia University*, 348:1–17, 2013.

Alan S Gerber and Neil Malhotra. Publication bias in empirical sociological research: Do arbitrary significance levels distort published results? *Sociological Methods & Research*, 37(1):3–30, 2008.

Sara Geven, Anatolia Batruch, Herman van de Werfhorst, et al. Inequality in teacher judgements, expectations and track recommendations: A review study, 2018.

Sabine Glock and Sabine Krolak-Schwerdt. Does nationality matter? The impact of stereotypical expectations on student teachers' judgments. *Social Psychology of Education*, 16(1):111–127, 2013.

Sabine Glock, Sabine Krolak-Schwerdt, Florian Klapproth, and Matthias Böhmer. Beyond judgment bias: How students' ethnicity and academic profile consistency influence teachers' tracking judgments. *Social Psychology of Education*, 16(4): 555–573, 2013.

Sabine Glock, Sabine Krolak-Schwerdt, and Ineke M Pit-ten Cate. Are school placement recommendations accurate? The effect of students' ethnicity on teachers' judgments and recognition memory. *European Journal of Psychology of Education*, 30(2):169–188, 2015.

Richard M Golden, Steven S Henley, Halbert White, and T Michael Kashner. Generalized information matrix tests for detecting model misspecification. *Econometrics*, 4(4):46, 2016.

Bryce Goodman and Seth Flaxman. European union regulations on algorithmic decision-making and a "right to explanation". *AI magazine*, 38(3):50–57, 2017.

Brian Greenhill, Michael D Ward, and Audrey Sacks. The separation plot: A new visual method for evaluating the fit of binary models. *American Journal of Political Science*, 55(4):991–1002, 2011.

Brandon Greenwell, Bradley Boehmke, Jay Cunningham, GBM Developers, and Maintainer Brandon Greenwell. Package 'gbm'. *R package version*, 2(5), 2019.

Noah Greifer. Package 'weightit', 2019.

Noah Greifer. Package 'cobalt'. *R Package Version, 4.0*, 2020.

Justin Grimmer, Margaret E Roberts, and Brandon M Stewart. Machine learning for social science: An agnostic approach. *Annual Review of Political Science*, 24: 395–419, 2021.

Jens Hainmueller. Package 'ebal'. 2022.

Michael J Hanmer and Kerem Ozan Kalkan. Behind the curve: Clarifying the best approach to calculating predicted probabilities and marginal effects from limited dependent variable models. *American Journal of Political Science*, 57(1):263–277, 2013.

Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, September 2020. doi: 10.1038/s41586-020-2649-2. URL `https://doi.org/10.1038/s41586-020-2649-2`.

Trevor Hastie and Robert Tibshirani. Generalized additive models: some applications. *Journal of the American Statistical Association*, 82(398):371–386, 1987.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, 2009.

Friedrich August Hayek. The use of knowledge in society. *The American economic review*, 35(4):519–530, 1945.

James J Heckman. Sample selection as a specification error. *Econometrica*, 47: 153–161, 1979.

James J Heckman. Econometric causality. *International statistical review*, 76(1): 1–27, 2008.

James J Heckman, John Eric Humphries, and Gregory Veramendi. Returns to education: The causal effects of education on earnings, health, and smoking. *Journal of Political Economy*, 126(S1):S197–S246, 2018.

Peter Hedström and Petri Ylikoski. Causal mechanisms in the social sciences. *Annual Review of Sociology*, 36:49–67, 2010.

Carl G Hempel and Paul Oppenheim. Studies in the logic of explanation. *Philosophy of Science*, 15(2):135–175, 1948.

Tom Heskes, Evi Sijben, Ioan Gabriel Bucur, and Tom Claassen. Causal shapley values: Exploiting causal knowledge to explain individual predictions of complex models. *Advances in Neural Information Processing Systems*, 33:4778–4789, 2020.

Steffen Hillmert and Marita Jacob. Selections and social selectivity on the academic track: A life-course analysis of educational attainment in germany. *Research in Social Stratification and Mobility*, 28(1):59–76, 2010.

Matthew Hindman. Building better models: Prediction, replication, and machine learning in the social sciences. *The Annals of the American Academy of Political and Social Science*, 659(1):48–62, 2015.

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

Jake M Hofman, Amit Sharma, and Duncan J Watts. Prediction and explanation in social systems. *Science*, 355(6324):486–488, 2017.

Jake M Hofman, Duncan J Watts, Susan Athey, Filiz Garip, Thomas L Griffiths, Jon Kleinberg, Helen Margetts, Sendhil Mullainathan, Matthew J Salganik, Simine Vazire, et al. Integrating explanation and prediction in computational social science. *Nature*, 595(7866):181–188, 2021.

Katja Hölsch and Margit Kraus. Poverty alleviation and the degree of centralization in European schemes of social assistance. *Journal of European Social Policy*, 14 (2):143–164, 2004.

Marc Hooghe and Jennifer Oser. Partisan strength, political trust and generalized trust in the united states: An analysis of the general social survey, 1972–2014. *Social Science Research*, 68:132–146, 2017.

Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366, 1989.

William C Horrace and Ronald L Oaxaca. Results on the bias and inconsistency of ordinary least squares for the linear probability model. *Economics Letters*, 90(3): 321–327, 2006.

Torsten Hothorn, Achim Zeileis, Richard W Farebrother, Clint Cummins, Giovanni Millo, and David Mitchell. Package 'lmtest'. *Testing linear regression models. https://cran. r-project. org/web/packages/lmtest/lmtest. pdf. Accessed*, 6, 2015.

Mieke Van Houtte and Peter A. J. Stevens. Study involvement of academic and vocational students: Does between-school tracking sharpen the difference? *American Educational Research Journal*, 46(4):943–973, 2009.

Nick Huntington-Klein. vtable: Variable table for variable documentation. r package version 1.2. 1, 2020.

Kosuke Imai and Marc Ratkovic. Estimating treatment effect heterogeneity in randomized program evaluation. *The Annals of Applied Statistics*, 7(1):443–470, 2013.

Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

Guido W Imbens and Jeffrey M Wooldridge. Recent developments in the econometrics of program evaluation. *Journal of economic literature*, 47(1):5–86, 2009.

Inspectie van het Onderwijs (accessed 7th December 2020). Onderwijsdata. https://www.onderwijsinspectie.nl/trends-en-ontwikkelingen/onderwijsdata. URL `https://www.onderwijsinspectie.nl/trends-en-ontwikkelingen/onderwijsdata`.

John PA Ioannidis. Why most published research findings are false. *PLoS medicine*, 2(8):e124, 2005.

Erik Jansen, Jana Javornik, Annica Brummel, and Mara A Yerkes. Central-local tensions in the decentralization of social policies: Street-level bureaucrats and social practices in the Netherlands. *Social Policy & Administration*, 55(7):1262–1275, 2021.

Lucas Janson, William Fithian, and Trevor J Hastie. Effective degrees of freedom: A flawed metaphor. *Biometrika*, 102(2):479–485, 2015.

Judith Janssen, Norman Verhelst, Ronald Engelen, and Floor Scheltens. Wetenschappelijke verantwoording: Rekenen-wiskunde voor groep 3 tot en met 8. Cito, 2010.

Dolores Jiménez-Rubio, Peter C Smith, and Eddy Van Doorslaer. Equity in health and health care in a decentralised context: evidence from Canada. *Health economics*, 17(3):377–392, 2008.

Joblib Development Team. Joblib: running python functions as pipeline jobs. 2020. URL `https://joblib.readthedocs.io/`.

Mark R Joslyn, Donald P Haider-Markel, Michael Baggs, and Andrew Bilbo. Emerging political identities? Gun ownership and voting in presidential elections. *Social Science Quarterly*, 98(2):382–396, 2017.

Paul Jungbluth. *De ongelijke basisschool. Etniciteit, sociaal mileu, sekse, verborgen differentiatie, segregatie, onderwijskansen en schooleffectiviteit.* [Sl]: Nijmegen: ITS, 2003.

Lee Jussim and Kent D Harber. Teacher expectations and self-fulfilling prophecies: Knowns and unknowns, resolved and unresolved controversies. *Personality and social psychology review*, 9(2):131–155, 2005.

J Kaplan. fastdummies: Fast creation of dummy (binary). columns and rows from categorical variables. r package version 1.6. 2, 2020.

Alboukadel Kassambara. Package 'ggpubr'. *R package version 0.1*, 6, 2020.

Maximilian Kasy. Of forking paths and tied hands: Selective publication of findings, and what economists should do about it. *Journal of Economic Perspectives*, 35 (3):175–92, 2021.

Timo M Kauppinen. Schools as mediators of neighbourhood effects on choice between vocational and academic tracks of secondary education in Helsinki. *European Sociological Review*, 24(3):379–391, 2008.

Yuri Kazepov. *Rescaling social policies: towards multilevel governance in Europe*, volume 38. Ashgate Publishing, Ltd., 2010.

Florian Klapproth. Biased predictions of students' future achievement: An experimental study on pre-service teachers' interpretation of curriculum-based measurement graphs. *Studies in Educational Evaluation*, 59:67–75, 2018.

Jon Kleinberg, Annie Liang, and Sendhil Mullainathan. The theory is predictive, but is it complete? An application to human perception of randomness. In *Proceedings of the 2017 ACM Conference on Economics and Computation*, pages 125–126, 2017.

Ron Kohavi et al. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *International Joint Conference on Artificial Intelligence*, volume 14, pages 1137–1145. 1995.

Satyapriya Krishna, Tessa Han, Alex Gu, Javin Pombra, Shahin Jabbari, Steven Wu, and Himabindu Lakkaraju. The disagreement problem in explainable machine learning: A practitioner's perspective. *arXiv*, 2022. doi: 2202.01602.

Max Kuhn, Jed Wing, Steve Weston, Andre Williams, Chris Keefer, Allan Engelhardt, Tony Cooper, Zachary Mayer, Brenton Kenkel, R Core Team, et al. Package 'caret'. *The R Journal*, 223:7, 2020.

Stein Kuhnle and Anne Sander. The emergence of the Western welfare state. *The Oxford handbook of the welfare state*, 1:63–82, 2010.

Joeke Kuyvenhoven and Willem R Boterman. Neighbourhood and school effects on educational inequalities in the transition from primary to secondary education in Amsterdam. *Urban Studies*, 2020.

Helen F Ladd and Edward B Fiske. Weighted student funding in the Netherlands: A model for the US? *Journal of Policy Analysis and Management*, 30(3):470–498, 2011.

Sebastian Lapuschkin, Stephan Wäldchen, Alexander Binder, Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Unmasking Clever Hans predictors and assessing what machines really learn. *Nature Communications*, 10(1):1–8, 2019.

David Lazer, Alex Pentland, Lada Adamic, Sinan Aral, Albert-Laszlo Barabasi, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, et al. Social science. computational social science. *Science (New York, NY)*, 323(5915):721–723, 2009.

David MJ Lazer, Alex Pentland, Duncan J Watts, Sinan Aral, Susan Athey, Noshir Contractor, Deen Freelon, Sandra Gonzalez-Bailon, Gary King, Helen Margetts, et al. Computational social science: Obstacles and opportunities. *Science*, 369 (6507):1060–1062, 2020.

Edward E Leamer. Let's take the con out of econometrics. *The American Economic Review*, 73(1):31–43, 1983.

Philip Leifeld. texreg: Conversion of statistical model output in r to latex and html tables. *Journal of Statistical Software*, 55(8):1–24, 2013.

Sigrid Leitner. Varieties of familialism: The caring function of the family in comparative perspective. *European societies*, 5(4):353–375, 2003.

Kimberley Martine Lek. *Teacher knows best? On the (dis)advantages of teacher judgments and test results, and how to optimally combine them.* PhD thesis, Utrecht University, 2020.

Thomas Lemieux. The "mincer equation" thirty years after schooling, experience, and earnings. In *Jacob Mincer a pioneer of modern labor economics*, pages 127–145. Springer, 2006a.

Thomas Lemieux. Increasing residual wage inequality: Composition effects, noisy data, or rising demand for skill? *American Economic Review*, 96(3):461–498, 2006b.

Yun Liao et al. Machine learning in macro-economic series forecasting. *International Journal of Economics and Finance*, 9(12):71–76, 2017.

Andy Liaw, Matthew Wiener, et al. Classification and regression by randomforest. *R news*, 2(3):18–22, 2002.

Torrin M Liddell and John K Kruschke. Analyzing ordinal data with metric models: What could possibly go wrong? *Journal of Experimental Social Psychology*, 79: 328–348, 2018.

Zachary C Lipton. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31–57, 2018.

Linda Lobao, Ron Martin, and Andrés Rodríguez-Pose. Rescaling the state: new modes of institutional–territorial organization. *Cambridge journal of regions, economy and society*, 2(1):3–12, 2009.

J Scott Long and Jeremy Freese. *Regression models for categorical dependent variables using Stata*, volume 7. Stata press, 2006.

J Scott Long and Pravin K Trivedi. Some specification tests for the linear regression model. *Sociological Methods & Research*, 21(2):161–204, 1992.

Guillem Lopez-Casasnovas, Joan Costa-Font, and Ivan Planas. Diversity and regional inequalities in the spanish 'system of health care services'. *Health economics*, 14 (S1):S221–S235, 2005.

Samuel R. Lucas. Effectively maintained inequality: Education transitions, track mobility, and social background effects. *American Journal of Sociology*, 106(6): 1642–1690, 2001.

Elke Lüdemann and Guido Schwerdt. Migration background and educational tracking. *Journal of Population Economics*, 26(2):455–481, 2013.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Proceedings of the 31st international conference on neural information processing systems*, pages 4768–4777. 2017.

Scott M Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. From local explanations to global understanding with explainable ai for trees. *Nature Machine Intelligence*, 2(1):56–67, 2020.

H Luyten and RJ Bosker. Hoe meritocratisch zijn schooladviezen? *Pedagogische Studiën*, 2004.

JAM Hans Maarse and PP Patrick Jeurissen. The policy and politics of the 2015 long-term care reform in the Netherlands. *Health Policy*, 120(3):241–245, 2016.

Gangadharrao S Maddala. *Limited-dependent and qualitative variables in econometrics*. Number 3. Cambridge university press, 1986.

Stephen Malpezzi. Hedonic pricing models: a selective and applied review. *Housing Economics and Public Policy*, 1:67–89, 2003.

Oscar Marcenaro-Gutierrez and Anna Vignoles. A comparison of teacher and test-based assessment for Spanish primary and secondary students. *Educational Research*, 57(1):1–21, 2015.

Sönke Hendrik Matthewes. Better together? Heterogeneous effects of tracking on student achievement. *The Economic Journal*, 131(635):1269–1307, 2021.

Elizabeth Aura McClintock. Occupational sex composition and gendered housework performance: Compensation or conventionality? *Journal of Marriage and Family*, 79(2):475–510, 2017.

Clark McKown and Rhona S Weinstein. Teacher expectations, classroom context, and the achievement gap. *Journal of School Psychology*, 46(3):235–261, 2008.

Gerhard Meisenberg. Verbal ability as a predictor of political preferences in the United States, 1974–2012. *Intelligence*, 50:135–143, 2015.

Richard Miech, Fred Pampel, Jinyoung Kim, and Richard G Rogers. The enduring association between education and mortality: the role of widening and narrowing disparities. *American Sociological Review*, 76(6):913–934, 2011.

Stephen Milborrow. Package 'rpart. plot'. *Plot'rpart'Models: An Enhanced Version of'plot. rpart*, 2019.

Mario Molina and Filiz Garip. Machine learning for sociology. *Annual Review of Sociology*, 2019.

Carina Mood. Logistic regression: Why we cannot do what we think we can do, and what we can do about it. *European Sociological Review*, 26(1):67–82, 2010.

Martine de Mooij, Dion Dieleman, and Sabrina de Regt. Jaarrapport integratie 2020. *CBS Jaarrapportages*, 2020.

Stephen L Morgan and Jiwon Lee. Social class and party identification during the clinton, bush, and obama presidencies. *Sociological Science*, 4:394–423, 2017.

Stephen L Morgan and Christopher Winship. *Counterfactuals and causal inference*. Cambridge University Press, 2015.

Ilaria Mosca. Is decentralisation the real solution?: A three country study. *Health policy*, 77(1):113–120, 2006.

Esther Mot, Ali Aouragh, Marieke de Groot, and Hein Mannaerts. *The Dutch system of long-term care*. Citeseer, 2010.

Sendhil Mullainathan and Jann Spiess. Machine learning: an applied econometric approach. *Journal of Economic Perspectives*, 31(2):87–106, 2017.

Alicia H Munnell, Geoffrey MB Tootell, Lynn E Browne, and James McEneaney. Mortgage lending in boston: Interpreting hmda data. *The American Economic Review*, pages 25–53, 1996.

John Muñoz and Cristobal Young. We ran 9 billion regressions: Eliminating false positives through computational model robustness. *Sociological Methodology*, 48 (1):1–33, 2018.

Paul Murrell. The grid graphics package. *R News*, 2(2):14–19, 2002.

Richard A Musgrave. *The theory of public finance; a study in public economy*. Kogakusha Co., 1959.

Martin Neugebauer, Marcel Helbig, and Andreas Landmann. Unmasking the myth of the same-sex teacher advantage. *European sociological review*, 27(5):669–689, 2011.

Wallace E Oates et al. Fiscal federalism. *Books*, 1972.

Onderwijsinspectie. *Onderwijsverslag 'De Staat van het Onderwijs'*. Ministerie van Onderwijs, Cultuur en Wetenschap, 2016. Available at `https://www.onderwijsinspectie.nl/documenten/publicaties/2016/04/13/staat-van-het-onderwijs-2014-2015`.

Onderwijsinspectie. *Onderwijsverslag 'De staat van het Onderwijs'*. Ministerie van Onderwijs, Cultuur en Wetenschap, 2019. Available at `https://www.onderwijsinspectie.nl/documenten/rapporten/2019/04/10/rapport-de-staat-van-het-onderwijs-2019`.

Open Science Collaboration. Estimating the reproducibility of psychological science. *Science*, 349(6251):aac4716, 2015.

Elinor Ostrom, Roy Gardner, James Walker, James M Walker, and Jimmy Walker. *Rules, games, and common-pool resources*. University of Michigan Press, 1994.

The pandas development team. pandas-dev/pandas: Pandas, February 2020. URL `https://doi.org/10.5281/zenodo.3509134`.

Panteia and E'til. Personeelsmonitor gemeenten 2020. 2020.

Harry Anthony Patrinos. School choice in the Netherlands. CESifo DICE Report 2/2011, IFO Institute for Economic Research, Munich, 2011.

Judea Pearl. *Causality*. Cambridge University Press, 2009.

Thomas Lin Pedersen. Package 'patchwork'. *R package http://CRAN. R-project. org/package= patchwork. Cran*, 2019.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011a.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011b.

Joshua C Peterson, David D Bourgin, Mayank Agrawal, Daniel Reichman, and Thomas L Griffiths. Using large-scale experiments and machine learning to discover theories of human decision-making. *Science*, 372(6547):1209–1214, 2021.

Alexander Peysakhovich and Jeffrey Naecker. Using methods from machine learning to evaluate behavioral models of choice under risk and ambiguity. *Journal of Economic Behavior & Organization*, 133:373–384, 2017.

Massimo Pigliucci. The end of theory in science? *EMBO reports*, 10(6):534–534, 2009.

Andy Pike, Andrés Rodríguez-Pose, John Tomaney, Gianpiero Torrisi, and Vassilis Tselios. In search of the 'economic dividend'of devolution: Spatial disparities, spatial economic policy, and decentralisation in the uk. *Environment and Planning C: Government and Policy*, 30(1):10–28, 2012.

Eric Polley, Erin LeDell, Chris Kennedy, Sam Lendle, and Mark van der Laan. Package 'superlearner', 2021.

Eric C Polley, Sherri Rose, and Mark J Van der Laan. Super learning. In *Targeted learning*, pages 43–66. Springer, 2011.

Hanneke Posthumus, Bart Bakker, Jan van der Laan, Martine de Mooij, Sander Scholtus, Mersiha Tepic, Jeroen van den Tillaart, and Nander de Vette. Herziening gewichtenregeling primair onderwijs: Fase I. Statistics Netherlands, 2016.

Hanneke Posthumus, Sander Scholtus, and Jaap Walhout. De nieuwe onderwijs achterstandenindicator primair onderwijs: Samenvattend rapport. Statistics Netherlands, 2019.

Tijana Prokic-Breuer and Jaap Dronkers. The high performance of Dutch and Flemish 15-year-old native pupils: Explaining country differences in math scores between highly stratified educational systems. *Educational Research and Evaluation*, 18(8):749–777, 2012.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2022. URL `https://www.R-project.org/`.

Charles Rahal, Mark D Verhagen, and David S Kirk. The rise of machine learning in the academic social sciences. *AI and Society*, 2022. doi: 10.1007/s00146-022-01540-w.

James Bernard Ramsey. Tests for specification errors in classical linear least-squares regression analysis. *Journal of the Royal Statistical Society: Series B (Methodological)*, 31(2):350–371, 1969.

Douglas D Ready and David L Wright. Accuracy and inaccuracy in teachers' perceptions of young children's cognitive abilities: The role of child background and classroom context. *American Educational Research Journal*, 48(2):335–360, 2011.

Bernd Rechel, Emily Grundy, Jean-Marie Robine, Jonathan Cylus, Johan P Mackenbach, Cecile Knai, and Martin McKee. Ageing in the European union. *The Lancet*, 381(9874):1312–1322, 2013.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144. 2016.

Daniel E Rigobon, Eaman Jahani, Yoshihiko Suhara, Khaled AlGhoneim, Abdulaziz Alghunaim, Alex "Sandy" Pentland, and Abdullah Almaatouq. Winning models for grade point average, grit, and layoff in the fragile families challenge. *Socius*, 5, 2019.

Gaston V Rimlinger. *Welfare policy and industrialization in Europe, America, and Russia*. John Wiley & Sons, 1971.

Brian Ripley, Bill Venables, Douglas M Bates, Kurt Hornik, Albrecht Gebhardt, David Firth, and Maintainer Brian Ripley. Package 'mass'. *CRAN R*, 538:113–120, 2013.

Jozef MM Ritzen, Jan Van Dommelen, and Frans J De Vijlder. School finance and school choice in the netherlands. *Economics of Education Review*, 16(3):329–335, 1997.

Peter M Robinson. Root-n-consistent semiparametric regression. *Econometrica: Journal of the Econometric Society*, 56(4):931–954, 1988a.

Peter M Robinson. Semiparametric econometrics: A survey. *Journal of Applied Econometrics*, 3(1):35–51, 1988b.

Jaap Roeleveld, Geert Driessen, Guuske Ledoux, Joris Cuppen, and Joost Meijer. Doelgroepleerlingen in het basisonderwijs. *Historische ontwikkeling en actuele situatie. Amsterdam/Nijmegen: Kohnstamm Instituut/ITS*, 2011.

Sherri Rose. Mortality risk score prediction in an elderly population using machine learning. *American Journal of Epidemiology*, 177(5):443–452, 2013.

Christine Rubie-Davies, John Hattie, and Richard Hamilton. Expecting the best for students: Teacher expectations and academic outcomes. *British Journal of Educational Psychology*, 76(3):429–444, 2006.

Christine M Rubie-Davies. *Teacher expectations in education*. Routledge, 2017.

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

Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1 (5):206–215, 2019.

Cynthia Rudin, Rebecca J Passonneau, Axinia Radeva, Haimonti Dutta, Steve Ierome, and Delfina Isaac. A process for predicting manhole events in manhattan. *Machine Learning*, 80(1):1–31, 2010.

Frida Rudolphi and Robrt Erikson. Social selection in formal and informal tracking in sweden. In *Models of Secondary Education and Social Inequality*. Edward Elgar Publishing, 2016.

Tobias Rüttenauer and Volker Ludwig. Fixed effects individual slopes: Accounting and testing for heterogeneous effects in panel data or other multilevel models. *Sociological Methods & Research*, 2020.

Xavier X Sala-i Martin. I just ran four million regressions. Technical report, National Bureau of Economic Research, 1997.

Matthew Salganik. *Bit by bit: Social research in the digital age*. Princeton University Press, 2019.

Matthew J Salganik, Ian Lundberg, Alexander T Kindel, and Sara McLanahan. Introduction to the special collection on the fragile families challenge. *Socius*, 5: 2378023119871580, 2019b.

Matthew J Salganik, Ian Lundberg, Alexander T Kindel, Caitlin E Ahearn, Khaled Al-Ghoneim, Abdullah Almaatouq, Drew M Altschul, Jennie E Brand, Nicole Bohme Carnegie, Ryan James Compton, et al. Measuring the predictability of life outcomes with a scientific mass collaboration. *Proceedings of the National Academy of Sciences*, 117(15):8398–8403, 2020.

Richard B Saltman. Decentralization, re-centralization and future european health policy. *European Journal of Public Health*, 18(2):104–106, 2008.

Wojciech Samek, Grégoire Montavon, Andrea Vedaldi, Lars Kai Hansen, and Klaus-Robert Müller. *Explainable AI: interpreting, explaining and visualizing deep learning*. Springer Nature, 2019.

Robert J Sampson, Christopher Winship, and Carly Knight. Translating causal claims: Principles and strategies for policy-relevant criminology. *Criminology & Pub. Pol'y*, 12:587, 2013.

Jaap Scheerens, Melanie Ehren, Peter Sleegers, and Renske de Leeuw. Country background report for the Netherlands. OECD Review on evaluation and assessment frameworks for improving school outcomes, 2012.

Philip A Schrodt. Seven deadly sins of contemporary quantitative political analysis. *Journal of peace research*, 51(2):287–300, 2014.

Frederik T Schut and Bernard van den Berg. Sustainability of long-term care financing in The Netherlands. In *International Conference on the Policies and Regulations governing the Costs of Health Care and Long-Term Care of the Elderly*, 2009.

Skipper Seabold and Josef Perktold. statsmodels: Econometric and statistical modeling with python. In *9th Python in Science Conference*, 2010.

Nikolai Sellereite and Martin Jullum. shapr: An R-package for explaining machine learning models with dependence-aware shapley values. 2020.

Lloyd S Shapley. Stochastic games. *Proceedings of the National Academy of Sciences*, 39(10):1095–1100, 1953.

Galit Shmueli. To explain or to predict? *Statistical Science*, 25(3):289–310, 2010.

Uri Simonsohn, Joseph P Simmons, and Leif D Nelson. Specification curve analysis. *Nature Human Behaviour*, 4(11):1208–1214, 2020.

Anders Skrondal and Sophia Rabe-Hesketh. *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models*. Crc Press, 2004.

Adam Slez. The difference between instability and uncertainty: Comment on Young and Holsteen (2017). *Sociological Methods & Research*, 48(2):400–430, 2019.

Steven Sloman. *Causal models: How people think about the world and its alternatives*. Oxford University Press, 2005.

Ed Smeets, Jos Kuijk, and Geert Driessen. *Handreiking bij het opstellen van het basisschooladvies*. 12 2014.

Nicole S Sorhagen. Early teacher expectations disproportionately affect poor children's high school performance. *Journal of Educational Psychology*, 105(2): 465, 2013.

Maintainer Vitalie Spinu. Package 'lubridate'. *Recuperado el*, 2016.

Mervyn Stone. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(2):111–133, 1974.

Mervyn Stone. An asymptotic equivalence of choice of model by cross-validation and akaike's criterion. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):44–47, 1977.

VS Subrahmanian and Srijan Kumar. Predicting human behavior: The next frontiers. *Science*, 355(6324):489–489, 2017.

Anna Südkamp, Johanna Kaiser, and Jens Möller. Accuracy of teachers' judgments of students' academic achievement: A meta-analysis. *Journal of Educational Psychology*, 104(3):743, 2012.

Anthony Mwinkaara Sumah, Leonard Baatiema, and Seye Abimbola. The impacts of decentralisation on health-related equity: A systematic review of the evidence. *Health Policy*, 120(10):1183–1192, 2016.

Laura Marie Tach and George Farkas. Learning-related behaviors, cognitive skills, and ability grouping when schooling begins. *Social Science Research*, 35(4): 1048–1079, 2006.

Siyi Tang, Amirata Ghorbani, Rikiya Yamashita, Sameer Rehman, Jared A Dunnmon, James Zou, and Daniel L Rubin. Data valuation for medical imaging using shapley value and application to a large-scale chest x-ray dataset. *Scientific reports*, 11(1):1–9, 2021.

Terry Therneau, Beth Atkinson, and Brian Ripley. Package 'rpart'. *Available online: cran. ma. ic. ac. uk/web/packages/rpart/rpart. pdf (accessed on 20 April 2016)*, 2015.

Lu Tian, Ash A Alizadeh, Andrew J Gentles, and Robert Tibshirani. A simple method for estimating interactions between a treatment and a large number of covariates. *Journal of the American Statistical Association*, 109(508):1517–1532, 2014.

Julie Tibshirani, Susan Athey, Rina Friedberg, Vitor Hadad, David Hirshberg, Luke Miner, Erik Sverdrup, Stefan Wager, Marvin Wright, and Maintainer Julie Tibshirani. Package 'grf'. 2022.

Nicole Tieben and Maarten HJ Wolbers. Transitions to post-secondary and tertiary education in the netherlands: a trend analysis of unconditional and conditional socio-economic background effects. *Higher education*, 60(1):85–100, 2010.

Charles M Tiebout. A pure theory of local expenditures. *Journal of political economy*, 64(5):416–424, 1956.

Anneke Timmermans, Hans Kuyper, and Greetje van der Werf. Schooladviezen en onderwijsloopbanen. *Voorkomen, risicofactoren en gevolgen van onder-en overadvisering. Groningen: GION*, 2013.

Anneke C Timmermans, Hans Kuyper, and Greetje van der Werf. Accurate, inaccurate, or biased teacher expectations: Do Dutch teachers differ in their expectations at the end of primary education? *British Journal of Educational Psychology*, 85(4):459–478, 2015.

Anneke C Timmermans, H De Boer, HTA Amsing, and MPC Van Der Werf. Track recommendation bias: Gender, migration background and ses bias over a 20-year period in the dutch context. *British Educational Research Journal*, 44(5):847–874, 2018.

Jochem Tolsma and Maarten HJ Wolbers. Onderwijs als nieuwe sociale scheidslijn? De gevolgen van onderwijsexpansie voor sociale mobiliteit, de waarde van diploma's en het relatieve belang van opleiding in nederland. 2010.

Herman G van de Werfhorst. Inequality in learning is a major concern after school closures. *Proceedings of the National Academy of Sciences*, 118(20), 2021.

Herman G Van de Werfhorst and Jonathan JB Mijs. Achievement inequality and the institutional structure of educational systems: A comparative perspective. *Annual Review of Sociology*, 36:407–428, 2010.

Linda van den Bergh, Eddie Denessen, Lisette Hornstra, Marinus Voeten, and Rob W. Holland. The implicit prejudiced attitudes of teachers: Relations to teacher expectations and the ethnic achievement gap. *American Educational Research Journal*, 47(2):497–527, 2010.

Lia van der Ham, Maaike den Draak, Wouter Mensink, Peggy Schyns, and Esther van de Berg. De Wmo 2015 in praktijk. 2018.

Jan van der Laan, Edwin de Jonge, Marjolijn Das, Saskia Te Riele, and Tom Emery. A whole population network and its application for the social sciences. *European Sociological Review*, 2022.

Mark J Van Der Laan and Sandrine Dudoit. Unified cross-validation methodology for selection among estimators and a general cross-validated adaptive epsilon-net estimator: Finite sample oracle inequalities and examples. *U.C. Berkeley Division of Biostatistics Working Paper Series*, 2003.

Mark J Van der Laan, Eric C Polley, and Alan E Hubbard. Super learner. *Statistical Applications in Genetics and Molecular Biology*, 6(1), 2007.

Michelle van der Tier, Koen Hermans, and Marianne Potting. Linking macro-level mechanisms to street-level accountability practices. a cross-national case study of street-level accountability of social workers in government funded homeless shelters. *Social Policy & Administration*, 55(1):191–205, 2021.

Aad W Van der Vaart, Sandrine Dudoit, and Mark J van der Laan. Oracle inequalities for multi-fold cross validation. *Statistics & Decisions*, 24(3):351–371, 2006.

Herman G. van der Werfhorst, Louise Elffers, Sjoerd Karsten, et al. *Onderwijsstelsels vergeleken: leren, werken en burgerschap*. Didactief onderzoek, 2015.

Eddy Van Doorslaer, Cristina Masseria, Xander Koolman, et al. Inequalities in access to medical care by income in developed countries. *Cmaj*, 174(2):177–183, 2006.

Marjolein Broese Van Groenou, Karen Glaser, Cecilia Tomassini, and Thérèse Jacobs. Socio-economic status differences in older people's use of informal and formal help: a comparison of four European countries. *Ageing & Society*, 26(5): 745–766, 2006.

Anne van Leest, Lisette Hornstra, Jan van Tartwijk, and Janneke van de Pol. Test-or judgement-based school track recommendations: Equal opportunities for students with different socio-economic backgrounds? *British Journal of Educational Psychology*, 91(1):193–216, 2021.

Guido Van Rossum and Fred L Drake Jr. *Python reference manual*. Centrum voor Wiskunde en Informatica Amsterdam, 1995.

Tyler J VanderWeele. Mediation and mechanism. *European journal of epidemiology*, 24(5):217–224, 2009.

Mark D Verhagen. A pragmatist's guide to using prediction in the social sciences. *Socius*, 8:23780231221081702, 2022.

Wouter Vermeulen. Decentralisation of social policy in The Netherlands. *Decentralisation of education, health and social protection: Issues and challenges*, page 127, 2015.

Gijsbert Vonk and Paul Schoukens. *Devolution and Decentralisation in Social Security: A comparative European Perspective*. Eleven International Publishing, 2019.

GJ Vonk and A Tollenaar. *Lokale verzorgingsstaat: nieuwe uitdagingen voor de sociale rechtsstaat*. Rijksuniversiteit Groningen, Vakgroep Bestuursrecht & Bestuurskunde, 2012.

Stefan Wager. Asymptotic theory for random forests. *arXiv*, 2014. doi: 1405.0352.

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

Adam Wagstaff, Pierella Paci, and Eddy Van Doorslaer. On the measurement of inequalities in health. *Social Science & Medicine*, 33(5):545–557, 1991.

M Mitchell Waldrop. The chips are down for moore's law. *Nature News*, 530(7589): 144, 2016.

Larry Wasserman and Kathryn Roeder. High dimensional variable selection. *Annals of statistics*, 37(5A):2178, 2009.

Penelope W. St. J. Watson, Christine M. Rubie-Davies, Kane Meissel, Elizabeth R. Peterson, Annaline Flint, Lynda Garrett, and Lyn McDonald. Teacher gender, and expectation of reading achievement in new zealand elementary school students: essentially a barrier? *Gender and Education*, 31(8):1000–1019, 2019.

Duncan J Watts. Common sense and sociological explanations. *American Journal of Sociology*, 120(2):313–351, 2014.

Duncan J Watts. Should social science be more solution-oriented? *Nature Human Behaviour*, 1(1):1–5, 2017.

Halbert White. Using least squares to approximate unknown regression functions. *International Economic Review*, 21(1):149–170, 1980.

Halbert White. Consequences and detection of misspecified nonlinear regression models. *Journal of the American Statistical Association*, 76(374):419–433, 1981.

Hadley Wickham. The tidyverse. *R package ver*, 1(1):1, 2017a.

Hadley Wickham. reshape2: flexibly reshape data: a reboot of the reshape package. *R package version*, 1(3), 2017b.

Hadley Wickham, Maintainer Hadley Wickham, and Imports RColorBrewer. Package 'scales', 2016.

Claus O Wilke, Hadley Wickham, and Maintainer Claus O Wilke. Package 'cowplot'. *Streamlined Plot Theme and Plot Annotations for 'ggplot2*, 2019.

Jeffrey M Wooldridge. *Econometric analysis of cross section and panel data*. MIT press, 2010.

Marvin N Wright and Andreas Ziegler. ranger: A fast implementation of random forests for high dimensional data in c++ and r. *arXiv preprint arXiv:1508.04409*, 2015.

Nan Xiao, Miaozhu Li, Maintainer Nan Xiao, and TRUE LazyData. Package 'ggsci'. 2018.

Adonis Yatchew. An elementary estimator of the partial linear model. *Economics Letters*, 57(2):135–143, 1997.

Adonis Yatchew. *Semiparametric regression for the applied econometrician*. Cambridge University Press, 2003.

Weiwen Yin, Weidong Huo, and Danyang Lin. The effects of state coercion on voting outcome in protest movements: a causal forest approach. *Political Science Research and Methods*, pages 1–9, 2021.

Cristobal Young. Model uncertainty and the crisis in science. *Socius*, 4, 2018.

Cristobal Young. The difference between causal analysis and predictive models: Response to "comment on Young and Holsteen (2017)". *Sociological Methods & Research*, 48(2):431–447, 2019. doi: 10.1177/0049124118782542. URL `https://doi.org/10.1177/0049124118782542`.

Cristobal Young and Katherine Holsteen. Model uncertainty and robustness: A computational framework for multimodel analysis. *Sociological Methods & Research*, 46(1):3–40, 2017.

Cristobal Young and Sheridan A Stewart. Functional form robustness: Advancements in multiverse analysis. *Unpublished*, 2021.

H Peyton Young. Monotonic solutions of cooperative games. *International Journal of Game Theory*, 14(2):65–72, 1985.

Waleed A. Yousef. A leisurely look at versions and variants of the cross validation estimator. *arXiv*, 2020. doi: 1907.13413v3.

Jianlong Zhou, Amir H Gandomi, Fang Chen, and Andreas Holzinger. Evaluating the quality of machine learning explanations: A survey on methods and metrics. *Electronics*, 10(5):593, 2021.