UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
FACULDADE DE CIÊNCIAS ECONÔMICAS
PROGRAMA DE PÓS-GRADUAÇÃO EM ECONOMIA

Jeudi R. F. da Silva

STATISTICAL LEARNING AND CAUSAL EFFECTS: AN APPLICATION TO BOLSA FAMÍLIA PROGRAM

Porto Alegre
2023

Jeudi R. F. da Silva

STATISTICAL LEARNING AND CAUSAL EFFECTS: AN APPLICATION TO BOLSA FAMÍLIA PROGRAM

Dissertação submetida ao Programa de Pós-Graduação em Economia da Faculdade de Ciências Econômicas da UFRGS, como requisito para obtenção do título de Mestre em Economia, com ênfase em Economia Aplicada.

Orientador: Prof. Dr. Flávio Augusto Ziegelmann

Porto Alegre
2023

CIP - Catalogação na Publicação

Rufino Fernandes da Silva, Jeudi
Statistical learning and causal effects: an application to Bolsa Família Program / Jeudi Rufino Fernandes da Silva. -- 2023.
65 f.
Orientador: Flávio Augusto Ziegelmann.

Dissertação (Mestrado) -- Universidade Federal do Rio Grande do Sul, Faculdade de Ciências Econômicas, Programa de Pós-Graduação em Economia, Porto Alegre, BR-RS, 2023.

1. Statistical learning. 2. Machine learning. 3. Heterogeneous treatment effect. 4. Conditional Average Treatment effect. 5. Policy analysis. I. Augusto Ziegelmann, Flávio, orient. II. Título.

Elaborada pelo Sistema de Geração Automática de Ficha Catalográfica da UFRGS com os dados fornecidos pelo(a) autor(a).

JEUDI RUFINO FERNANDES DA SILVA

STATISTICAL LEARNING AND CAUSAL EFFECTS:
AN APPLICATION TO BOLSA FAMÍLIA PROGRAM

Dissertação submetida ao Programa de Pós-Graduação em Economia da Faculdade de Ciências Econômicas da UFRGS, como requisito parcial para obtenção do título de Mestre em Economia, área de concentração: Economia Aplicada.

Aprovada em: Porto Alegre, 12 de maio de 2023.

BANCA EXAMINADORA:

Prof. Dr. Flávio Augusto Ziegelmann - Orientador

PPGE/UFRGS

Prof. Dr. Marcelo de Carvalho Griebeler

PPGE/UFRGS

Prof. Dr. Ronald Otto Hillbrecht

PPGE/UFRGS

Prof. Dr. Guilherme Pumi

PPGEst/UFRGS

ABSTRACT

In this work we employ a study of statistical and machine learning methods for heterogenous treatment effect. Heterogenous treatment effect is also called Conditional Average Treatment Effect (CATE). We analyze general methods (off-the-shelf methods) and those tailored for causal inference: causal tree, causal forest and generalized causal forest. Beyond that, we explore one of those methods in the context of an empirical application. We explore causal forest methods to search for the heterogeneous treatment effect of a public policy. We evaluate the CATE of *Bolsa Família Program*.

Keywords: Statistical learning. Machine learning. Heterogeneous treatment effect. Conditional Average Treatment effect. Causal trees. Causal Forest. Generalized Random Forest. Causal inference. Off-the-shelf methods. Policy analysis. Bolsa família program.

RESUMO

Neste trabalho realizamos o estudo de métodos de aprendizagem estatística *(“statistical learning methods”*) e métodos de aprendizado de máquina *(“machine learning methods”*) para efeitos heterogêneos de tratamento. Tais tipos de efeitos são também chamados de efeito médio de tratamento condicional. Estudamos métodos gerais (*“off-the-shelf methods”*) e aqueles ajustados para inferência causal: arvore causal, floresta causal e floresta aleatória causal. Além disso, exploramos um desses métodos em um contexto de aplicação empírica. Exploramos os métodos de florestas causais para procurar efeitos heterogêneos de tratamento de uma política pública. Avaliamos os efeitos heterogeneous do programa Bolsa Famíia.

Palavras-chave: Aprendizado estatístico. Aprendizado de Máquina. Efeitos heterogêneos. Inferência Causal. Causal Forest. Generalized Random Forest. Causal inference. Programa Bolsa família.

CONTENTS

1 INTRODUCTION

The goal of this work is to study statistical and machine learning methods for heterogeneous treatment effect (HTE), also called Conditional Average Treatment Effect (CATE). More precisely, it will evaluate the heterogeneous effects of the Bolsa Família program (BFP) on the labor market[1]. Causal Forest methods (WAGER; ATHEY, 2018; ATHEY; TIBSHIRANI; WAGER, 2019) will be employed for such analysis.

Machine learning (ML) methods focus on algorithms designed to get the best predictions or to classify and cluster variables of a given data set (ATHEY, 2018). Essentially, they refer to many approaches for the estimation of some functional form of the underlying population. The majority of the methods could fit into one out of two possible labels: supervised and unsupervised. Supervised learning focuses on making the most accurate prediction given some set of covariates[2] and an observed response while unsupervised try to find patterns on data and try to group covariates into clusters or estimate their joint distribution. In an unsupervised context, we do not have a particular response and we are not necessarily interested in prediction. So, we do not have specific ways to evaluate the results. Supervised methods, in turn, try to verify its performance using some data set distinct from those used to construct the model. Therefore, on one hand we do not assess the unsupervised methods using data different from that one used to fit the model, on the other hand, we assess the accuracy of supervised methods checking its predictions: we could observe its predictions errors (JAMES et al., 2021). Furthermore, they have a data driven approach for model selection, searching for the best functional form. Besides, they handle a large set of covariates well.

ML has received attention in econometric literature. Varian (2014) comments the possibilities of using many tools and great amounts of information available in each of the most common huge data sets to analyze economic problems. The work comments on many traditional ML methods (eg.: regression trees, regression forests, LASSO) and briefly explores interactions between those methods and econometric approaches. Mullainathan and Spiess (2017) argued that many economic problems could be cast as prediction problems and some empirical studies already try to apply this approach. Imbens and Athey (2021) makes an analysis of the proposed two cultures (algorithmic models and data modeling) in economics/econometrics tradition.

A meaningful branch of causal inference is to estimate the effect of some treatment given a certain set of characteristics. The researcher wants to obtain insights about HTE, more than to estimate the average treatment effect (ATE). On many occasions, the policy maker wants to know how treatment effects differ among treated units. Some individuals could have a treatment effect better (or worse) than the average. The estimates of CATE/HTE could help to design more efficient policies (WAGER; ATHEY, 2018; ATHEY; TIBSHIRANI; WAGER, 2019).

Causal inference using machine learning is a challenge. Due to the "fundamental problem of causal inference" (HOLLAND, 1986) we do not observe all the potential outcomes. This

[1] The *PNAD contínua*, 2012, will be used

[2] In ML literature it is common to use "features" instead of covariates, predictors of regressors.

is a challenge that many traditional econometric methods (for causality) face. Note, however, that ML methods use observed responses to train/build their models and in the causal context there is no dataset to train, "to teach", the method being used. Furthermore, there is no test set to evaluate if the predictions made are correct and to proceed the model selection.[3] These two kinds of procedures are essential in supervised ML methods. Therefore, given the procedure of ML methods, not observing the (causal) responses make the use of the methods challenging. Nevertheless, there is a growing literature focused on causal effects using machine learning methods, specially on CATE. In the same vein of Varian (2014), the work of (ATHEY; IMBENS, G. W., 2019) reviews recent advances and the possibilities of the use of ML methods in economics. It explores not only traditional ML methods for prediction as the work of 2014, but also causal inference using these methods as well. Knaus, Lechner, and Strittmatter (2021) explores the performance of many ML algorithms and its estimators for causal inference. Using simulated data based on observational settings, they evaluate the performance using many data generation processes (DGP). They argue that their simulations try to approximate real observational settings. They also argue that above all there is no one best method, it will depend on the context. Chernozhukov, Demirer, et al. (2020), Semenova and Chernozhukov (2020) and Nie and Wager (2021) explore assymptotic properties for casual inference using estimators builded with methods. ML for causal effects could bring data-driven processes to estimate HTE due to their flexibility and their capability of handling high dimensional feature spaces. Additionally, they also could recover complex interactions between covariates and outcomes.

ML methods for HTE estimation could be classified into two labels: generic machine learning methods and modified machine learning methods. The first category is a set of traditional machine learning methods (or "off-the-shelf machine learning methods"[4]) used to estimate heterogeneous effects. There are no modifications in the used methods. The second category congregates ML methods with modifications to address HTE estimation. The first group uses machine learning for estimation of intermediate parameters ("nuisance parameters") in an initial stage. After that, these estimates are used in the estimation process of HTE. All those processes use a traditional ML method [5]. In the second group, there are causal trees and causal forests. These methods were built modifying traditional methods to estimate and infer causal effects. The tree splits are adjusted to capture heterogeneity; they favored bigger differences in the treatment effects.

Causal forest is a special case of generalized random forest (ATHEY; TIBSHIRANI; WAGER, 2019). This method relies on "honest estimation". First, the sample is divided into two. Using one subsample, we will grow tree stratifying covariate space, applying recursive partitioning until the tree has leaves with few treated and untreated units. Using the structure of the generated tree, we will compute CATE into the other subsample. This process will be

[3] In machine learning literature, supervised methods use a dataset to build the model and another one to evaluate the predictions of the built models. The first dataset is called "training set", the second one is called "test set".

[4] See Athey and Guido W. Imbens (2019) and Athey (2018).

[5] Each stage could use different methods.

repeated for some desired number of trees. Once finished, all trees will be ensembled and their predictions will be averaged. In short: for each tree of the forest, part of the sample is used for the estimation of the tree structure and the remainder is used to estimate the treatment effects in tree leafs. (WAGER; ATHEY, 2018; ATHEY; TIBSHIRANI; WAGER, 2019).

In order to explore the causal forest method, we will analyze BFP effects on labor supply. We will estimate propensity score (PS) employing regression forests and use these estimates to construct comparable groups. The use of PS is common for the analysis of BFP impacts (not only on labor supply). Ribeiro, Shikida, and Hillbrecht (2017) commented on some of the reasons for that. The BFP was not designed to be evaluated by the use of randomized control trials at any moment. Few sources that allow the use of Instrumental variables are found. There is evidence of manipulation to participate in the BFP, not allowing the use of discontinuity regression design: participants with an income that is close to the cutoff of being a beneficiary reduce their labor supply in order to continue participating. There is no longitudinal data of control and treatment group, not permitting to use differences in differences method. Consequently, many works about BFP impacts (in several dimensions) surveyed by the authors use propensity scores. We will use propensity scores in conjunction with the casual forest to have the CATE estimates of the BFP. We try to understand how It will be possible to evaluate, for a given set of characteristics, what the effect of BFP is on labor supply, via a data driven process. It is relevant to use a data driven approach because the search for heterogeneities do not stay restricted to only what researchers think is relevant to study. The ML method could bring light to aspects that would otherwise be ignored. Beyond that, the found heterogeneities can serve as a guidance to better treatment targeting, using in a more efficient way the BFP resources.

BFP is a kind of conditional cash transfer (CCT). This type of policy aims to bring social protection for the poor people. It provides income to beneficiaries given they accomplish some conditionalities. It is used for governments but also for humanitarian aid organizations. It could impact many outcomes: health, education and labor, for example. In 2018, cash transfer policies were used by 117 countries, CCT type was used in 52 of them (BAIRD; MCKENZIE; ÖZLER, 2018). In Latin America, many countries adopted this kind of transfer policy (Chile, Colombia, Equator, Mexico are some examples). The Bolsa Familia program was the main program of CCT in Brazil, lasting for almost 20 years. The program lasted until 2021, and it was replaced by *Auxílio Brasil* (Brazil Aid). In 2015, almost a quarter of the population was covered by BFP (RIBEIRO; SHIKIDA; HILLBRECHT, 2017).

CCT could impact economic outcomes. Economic theory predicts that agents who receive money will work less, given leisure being a normal good, but in lower/mid income contexts this is not what necessarily happens when we study the impact of this type of transfers. Baird, McKenzie, and Özler (2018) comment on some ways that CCT could generate a result different from that predicted by the standard model. When there are liquidity constraints and incomplete insurance, individuals are not capable of managing to invest and they do not take risks related to investments. The cash transfers could alleviate these constraints, generating changes in labor

supply. Also, CCT could change the relative prices because of its conditionalities. It could happen that, in order to receive the program transfer, the beneficiaries need to start a business (changing the relative prices of labor and leisure). Another condition could be some level of children caring, such as making more frequent visits to doctors or assuring that the kids go to the school (affecting the child and, potentially, adult labor). The main reason to worry about adverse effects on labor is the possibility that the household continues in the poverty cycle, staying in the so-called "poverty trap" (RIBEIRO; SHIKIDA; HILLBRECHT, 2017; BAIRD; MCKENZIE; ÖZLER, 2018). Despite the fact that there are many studies about cash transfers on health and education, evidence about effects on labor markets is limited (BAIRD; MCKENZIE; ÖZLER, 2018).

We contribute to the literature applying machine learning methods for heterogeneous effects in observational data: there are few applied studies in these settings (eg: KNAUS; LECHNER; STRITTMATTER, 2020; BAIARDI; NAGHI, 2021). As far as we know, there are no studies which apply statistical learning methods for causality to estimate effects of BFP. Our work finds effects on adult labor market outcomes due to the program. We provide empirical evidence to the debate about BFP effects on labor supply. We find evidence of HTE: we study quintiles of treatment effects finding distinct magnitudes of impacts given a set of characteristics. Our findings using causal forests show impacts mainly on female workers: in both occupation and formal sector participation there is a reduction in labor supply. These findings dialogue with the literature. On the other hand, men are those who are less likely to reduce their labor supply in the formal sector. Also, there is an increase in the occupation of male.

The sections of this work are divided in the following manner. First, a review of applied works for HTE using statistical learning models is made. After, we comment on BFP: we talk about some conceptual issues related to this intervention and its effect on the labor force, as well as summarize findings of recent works about BFP and its effects on adult labor. Afterwards, we build on the basic conceptual framework that allows us to talk about causality and comment on the first set of statistical learning models for causal inference. Then, we (briefly) explore modified statistical learning methods for causal inference: causal forest in its original context and in the context of generalized random forest. Finally, the applied analysis is conducted: we try to explore the method of causal forest studying the effect of BFP on the adult labor supply.

2 LITERATURE REVIEW

In this section we review some applied works using ML methods, specially those for CATE. Afterwards, we will employ a review of recent works on BFP.

2.1 APPLIED WORKS USING STATISTICAL LEARNING FOR CATE

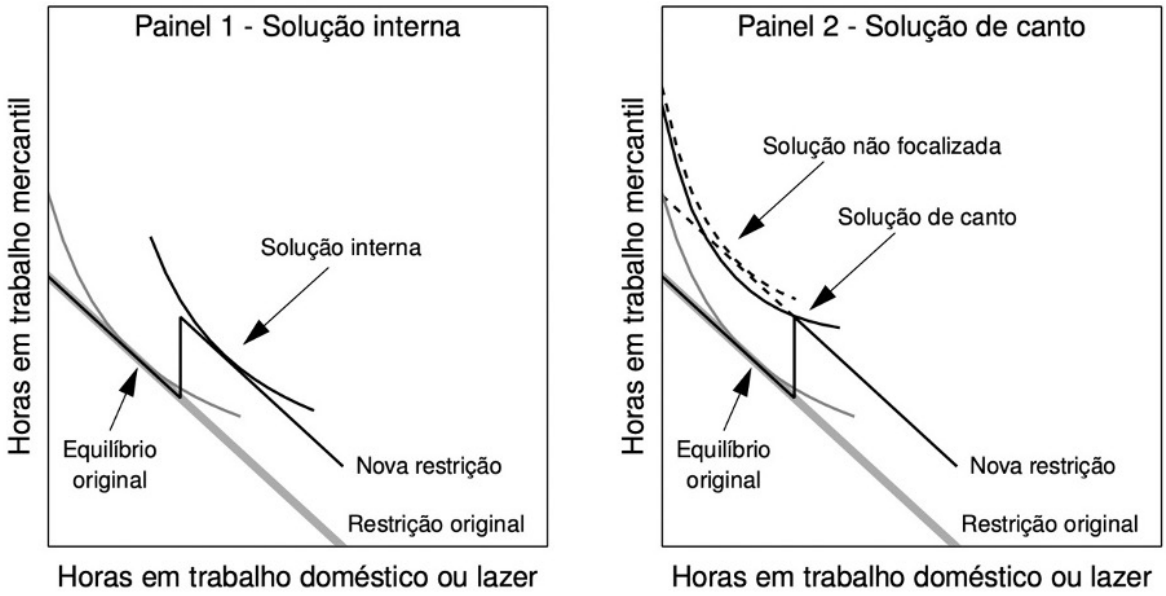
Andini et al. (2018) commented on some early applications of ML when economic problems are cast as prediction problems[1]: assistance for judges to decide if arrestees should be detained or released until adjudication; prediction of highest risk youth for anti-violence interventions; definition of restaurants for hygiene inspections; decision of hiring police officers with low chance to be violent and effectiveness of teachers (in terms of value added). Also, Andini et al. (2018) studied which units would be adequate to receive tax rebate in Italy in 2014 in a government program aiming the boosting of household consumption.

Recently, researchers explored causal effect estimation using statistical/machine learning methods, going beyond casting economic problems as prediction problems. Some studies have started to use these methodologies. Empirical studies for heterogeneous treatment effects with machine learning methods are still sparse, but growing.

Davis and Heller (2017) was one of the first applications of causal forest methods. They used causal forest to estimate effects of two randomized control trials (RTC) of the summer jobs program in Chicago. They showed evidence that the forests could identify effects that often traditional methods cannot capture (even with adjustment for multiple testing). Bertrand et al. (2017) studied public work programs from Ivory Coast. They analyzed the effects from an experiment. Using causal forest, they found evidence of a strong heterogeneity of program effects in both moments: contemporaneous and post-intervention effects. Also, they estimated the improvement of cost-effectiveness if the program had used predicted effects by machine learning methods for better program targeting. Ascarza (2018) used insights from Athey and Guido Imbens (2016) (but did not apply directly their method) to evaluate marketing intervention effects aiming to manage customer churn. Strittmatter (2019) analyzed the experimental data from Jobs First program in Connecticut with the double machine learning approach proposed by Chernozhukov, Chetverikov, et al. (2018). They showed evidence that supports labor supply predictions and contrasted their results' estimates with local constant model's estimates and quantile treatment effects. Knaus, Lechner, and Strittmatter (2020) focused on the heterogeneous treatment effects of job search programs in Switzerland. As the majority of the applied studies applying ML used RTC data, they proposed to apply it for causal investigation using observational data (TIAN et al., 2014; CHEN et al., 2017). In order to show gains with machine learning methods, Baiardi and Naghi (2021) revisited applied studies using machine learning and compared their results with the original ones. They analyzed both average treatment and heterogeneous treatment effects, causal forest was one of the methods employed.

[1] See their study for references.

Figure 1 – Labor supply and *bolsa família* program: comparative statics



Source: Oliveira and Soares (2012).

2.2 *BOLSA FAMÍLIA* PROGRAM

Theoretical background

In this section we follow Oliveira and Soares (2012) and Ribeiro, Shikida, and Hillbrecht (2017). The effects of the increase of worker income when he/she receives cash transfers without conditioning is the same as expected when there is any other increment of income: the worker could supply more hours or reduce it. All of this depends on the agent's preferences. It is not possible to assert if the policy of non-conditional cash transfers per se is an incentive or disincentive to work. When there is a cash transfer with conditionalities related to some threshold in income (as in BFP) an incentive is created: workers potentially could want to manipulate its income to be eligible to the program. There are two possible consequences of the conditional cash transfer: (1) the new equilibrium has an internal solution and we have the same as in non-conditional transfer, or (2) a corner solution happens. In (2) there is an incentive to reduce labor supply (in comparison with the non-conditional context). Among BFP beneficiaries there is room for both scenarios: for (1), we have those with so low income that there is no risk of not receiving the transfer. For (2), there are those whose income could potentially cross the threshold of the program.

There is a challenge for this static model: few workers have the power to freely choose how many hours they will offer, the majority of contracts have pre established hours of work. However, it is worth noting that poor workers could access some labor markets that have the possibility of managing how many hours to work (in this case, the workers work on their own). Beyond that, labor opportunities could be hard to find even for a willing worker. In addition, the search for work could bring some costs that could be high for low income workers. In this case,

transfers would help the search process and, eventually, help to increase the labor supply.

It is also necessary to consider the relationship between adult household members and the child labor force of the family. On one hand, it could be the case that labor supply of children and adults are substitutes, so the reduction of child labor could increase the adult labor supply. On the other hand, it is also possible that there is a reduction of adult labor supply (specially between women) because of the need to take care of children that, now, stay at home and not working anymore.

Therefore, only theoretical models are not enough. It is necessary empirical analysis to shed light on this problem. Until 2012, there was a certain consensus in empirical studies about effects of BPF on labor supply. More recently, new works appear to indicate that there are some different effects on labor supply (RIBEIRO; SHIKIDA; HILLBRECHT, 2017).

Recent Works

In the appendix, we provide a full review of recent (from 2010 onwards) works on the effects of BFP on the adult labor force. Here, we briefly highlight some salient aspects of the reviewed papers. Our goal here in this section is to bring to our work what we already learn from previous evaluations of the *Bolsa Fam*

More recent works show effects of BFP in adult women. de Brauw et al. (2015) shows that rural women (at the individual level) reduce their labor supply, while this does not happen with men. In addition, when we observe from a household level perspective, female members also reduce their labor supply in urban areas. Firpo et al. (2014) discoveries indicate that women, independent of their familiar structure, decrease labor force participation and weekly hours supplied to work. Tavares (2010) was the only recent work that goes in a different direction (she finds increment in the supply), despite recognizing that as the quantity of money transferred to mothers grows, their weekly hours and participation in the labor market reduce.

Another salient noticeable is the relationship of formal and informal sectors when examining the BFP in the adult labor supply. Evidence from recent works does not bring a consensus. On one hand, de Brauw et al. (2015) found the following effects: workers exchange 8 hours between sectors. On the other hand, de Holanda Barbosa and Corseuil (2013) does not find any effects. It is possible that the local nature of this last work prevents it from concluding something more general. de Brauw et al. (2015) study has a more general scope and this, perhaps, gives it more strength in this topic. This debate brings the challenge (given scant data) of adequately capturing the informal sector in Brazil.

Finally, there is a work suggesting that there are some effects of the BPF in reduction of labor supply, and this seems to occur because of program eligibility: Firpo et al. (2014) confirmed that adult working force manage their choice to continue getting the benefit (something suggested by de Holanda Barbosa and Corseuil (2013)).

3 METHODS

All models lay down their structure based on Potential Outcome Framework (SPLAWA-NEYMAN; DABROWSKA; SPEED, 1990; RUBIN, 1974). So, first of all, we start describing this approach to causality.

3.1 BASIC NOTATIONS AND DEFINITIONS

We start briefly describing the Potential Outcome framework (RUBIN, 1974). Let $i \in \{1,..,N\}$ be the index of the unit that could receive the treatment. We have Y_i as the observed outcome of unit i and X_i representing the covariates. The possible treatment is indicated by the binary $W_i \in \{0,1\}$. If $W_i = 1$, then the unit i received the treatment (sometimes called "intervention"), if $W_i = 0$, then the unit i did not receive it. Therefore, our observed data is a tripe (Y_i, W_i, X_i). The potential outcomes are two possible scenarios (states of the world): there is the outcome of the unit i when receiving the treatment and the outcome of the unit i when it does not receive. We only observe one outcome. The causal effect (or treatment effect) is defined by a comparison between these two scenarios. Formally, the causal effect of unit i is given by

$$\tau_i = Y_i^1 - Y_i^0 \quad ,$$

which is not observed. The observed outcome of a unit could be expressed by $Y_i = Y_i^1 W_i + Y_i^0(1 - W_i)$.

Although one does not observe τ_i, it is possible to estimate the average treatment effect (ATE) of an intervention. In order to do so, some assumptions are needed: (1) the treatment assignment and the potential outcomes must be independent, i.e., $(Y_i^0, Y_i^1) \perp W_i \ , \ \forall i$, (2) the intervention respects the stable unit treatment value assumption (SUTVA). The SUTVA means three things: (i) there is only one version of treatment, that is, the treatment is homogeneous across units, (ii) the treatment of a particular unit does not affect another unit potential outcome, and (iii) only W_i affects Y_i (not the other way around).

Under these assumptions, we have

$$ATE = \mathbb{E}(\tau_i) = \tau = \mathbb{E}(Y_i^1 - Y_i^0) \ .$$

The heterogeneous treatment effects given some covariates (or conditional ATE) is expressed in the following way:

$$HTE = CATE = \tau(x) = \mathbb{E}(Y_i^1 - Y_i^0 | X_i = x) \ .$$

Note that CATE is not the individual specific treatment, it is an average effect in a more targeted group characterized by covariates X_i. In order to allow identification of treatment effects when dealing with observational data, the machine learning literature usually assumes unconfoundedness:

$$W_i \perp (Y_i^1 - Y_i^0) | X_i \ .$$

We also assume unconfoundedness in this work.

We could summarize the probability of receiving the treatment in a scalar using relevant (observable) covariates and use it to make comparison between control and treated group. We define the propensity score as $e(X_i) := \mathbb{E}(W_i|X_i)$ as the propensity of some unit to receive the treatment given its observable covariates. In an experimental context, the $e(X_i)$ is known. In an observational setting we must estimate it. There are many ways of using this propensity (see Cunningham (2021) and Imbens and Rubin (2015), for example) to make comparisons between similar units. In the Methodology section we explain how we will use the propensity score. Together with unconfoundedness assumption, we also assume that propensity scores are not so much close to the extremes, i.e., close to 1 or 0. Formally: for some $\nu > 0$ it is true that $\nu < e(X_i) < (1-\nu)$. This assumption is called overlap. This means that whatever are the values of the features we observe, we could find a group of units in both control and treatment groups that have features with these values. This allows us to compare treatment and control groups for each value of the (observable) covariates.

Finally, we have the conditional mean of the outcome due to the covariates as $\mu(x) := \mathbb{E}(Y_i|X_i = x)$. Also, we have the conditional mean given a particular treatment and covariates, i.e., $\mu(w,x) := \mathbb{E}(Y_i|W_i = w, X_i = x)$.

3.2 GENERIC METHODS

This section will approach generic methods. Herein, we define as generic those methods that use traditional machine learning to estimate nuisance parameters. These methods do not make changes in off-of-shelf machine learning estimators (KNAUS; LECHNER; STRITTMATTER, 2021). They proceed to CATE estimation in two steps. In the first one, we estimate the nuisance parameters, those components that will be used as inputs to make estimates of CATE. These nuisance parameters typically are the propensity score, conditional mean given X_i, and the conditional mean given a particular treatment and covariates i.,e, $e(x)$, $\mu(x)$ and $\mu(w,x)$. In many practical contexts, we do not have access to these parameters, so it is necessary to make estimates of them in a first step. In all these methods, in observational studies, it is assumed that unconfoundedness assumption holds[1].

Transformed Outcome Methods

Here we describe estimators based on Inverse Probability Weighting (IPW) (e.g. HORVITZ; THOMPSON, 1952; HIRANO; IMBENS; RIDDER, 2003) and Double Robust (DR) (ROBINS; ROTNITZKY, 1995) estimators. Both use the propensity score. DR also uses $\mu(w,x)$. In this context we have:

[1] It is worth noting that Knaus, Lechner, and Strittmatter (2021) make some additional interesting comments about the generic models described below. We recommend reading their paper.

$$Y^*_{i,IPW} = Y_i \frac{W_i - e(X_i)}{e(X_i)(1-e(X_i))}$$

$$Y^*_{i,DR} = \mu(1,X_i) - \mu(0,X_i) + \frac{W_i(Y_i - \mu(1,X_i))}{e(X_i)} - \frac{(1-W_i)(Y_i - \mu(0,X_i))}{(1-e(X_i))}$$

and it is possible to show that for both $Y^*_{i,DR}$ and $Y^*_{i,IPW}$ it is true $E(Y^*_{i,t}|X=x) = \tau(x)$, where $t \in \{DR, IPW\}$ (KNAUS; LECHNER; STRITTMATTER, 2021). Then, one could estimate $e(X_i)$ and $\mu(w,x)$ with a convenient machine learning method and use the equations above. Semenova and Chernozhukov (2020) studied asymptotic properties of $\hat{\tau}(x)$ estimated with least squares.

R Learning

Using a decomposition proposed by Robinson (1988) and applying the process of estimating debiased machine learning estimators, Chernozhukov, Chetverikov, et al. (2018) and Nie and Wager (2021) [2] showed that the problem of to estimate CATE could be posed in the following manner:

$$\min_\tau \left\{ \frac{1}{N} \sum_{i=1}^{N} \left[(Y_i - \mu(X_i)) - (W_i - e(X_i))\tau(X_i) \right]^2 \right\}.$$

In this setup, the nuisance parameters are $\mu(X_i)$ and $e(X_i)$.

3.3 MODIFIED METHODS

Modified methods change the traditional methods aiming to estimate and infer causal effects. A fundamental concept in these methods is the honest estimation. "Honest" has a specific meaning[3], which is: part of the training sample is used to grow trees and another is used to estimate treatment effects within each leaf of those trees, that is, part of the training sample is used to place splits on trees focusing on treatment effects, while the other is used to make estimates. Given these two separate parts to each purpose, honest estimation eliminates bias but has a cost of reducing precision.

The empirical analysis employed in this work explores generalized random forest for heterogeneous treatment effect estimation using causal forest. It is a data-driven process to define heterogeneity, thus restricting researchers to focus on in some relevant dimensions indicated by data. We will give an overview about this method. After that, we describe the intersection between generalized random forest and causal forest.

[2] Knaus, Lechner, and Strittmatter (2021) informed that Chen et al. (2017) proposed the same, but with another name: A-learner.

[3] In reality there is another possibility of honest estimation in the context of forest. (WAGER; ATHEY, 2018) posed in the following manner: the observable outcomes of each unit of sample are not used to place splits into a tree, only the treatment variable are used.

Causal Forest

Based on a random forest method (BREIMAN, 2001), causal forest grows many deep trees instead of just one well pruned tree. Given that trees have high variance, it is better to have an averaged prediction of many random causal trees than trying to know and prune the most "adequate" (WAGER; ATHEY, 2018).

Wager and Athey (2018) proposes a modification in an established estimator (random forest) to estimate CATE. They proved consistency for causal forest, as well as it being asymptotically unbiased and Gaussian. They do not rely on estimated propensity scores to get their results. Intuitively, when building deep trees, leafs are defined by a narrow set of characteristics, so individuals that fall in the same leafs look alike in terms of observational characteristics. If unconfoundedness assumptions are assumed, then units are very close, except for the treatment status.

Causal forests use honest estimation as well. It splits the sample into two parts. Using one subsample, it grows a tree stratifying feature space X, applying recursive partitioning until the tree has leaves with few treated and untreated units. Then, it estimates treatment effects in the other subsample. This way, the estimation of CATE is given by

$$\hat{\tau}(x) = \frac{1}{(i : W_i = 1, X_i \in l(x))} \sum_{(i:W_i=1, X_i \in l(x))} Y_i \; - \; \frac{1}{(i : W_i = 0, X_i \in l(x))} \sum_{(i:W_i=0, X_i \in l(x))} Y_i$$

where $l(x)$ is a leaf containing x. Once finished, all B trees are ensembled and their predictions are averaged: $\hat{\tau}(x) = B^{-1} \sum_{b=1}^{B} \hat{\tau}(x)$, where $\hat{\tau}(x)$ is an estimate of one tree.

The authors noticed that although one could criticize honest estimation by arguing it is inefficient (because half of the sample is not used, in principle) causal forests method does not suffer from such inefficiency. The reason: forest subsampling mechanism. It is true that each training point shall only produce splits for calculating estimates in a given tree, but this is just for one tree. In every new tree of the forest, the data point shall participate in one option of subsamples. Each point will contribute for splitting into some trees, while for treatment estimation in others. Hence, all data will be used.

Generalized random forest (GRF)

Focusing on allowing forest to be applied in more general contexts than only those covered by averaging estimates of each member of ensemble, Athey, Tibshirani, and Wager (2019) cast forests as a type of adaptive locally weighted estimators:

> "The benefit of using forest-based weighting is that it is more effective in choosing which dimensions are important in determining closeness. Meanwhile, in methods like k-nearest neighbors every dimension is given equal importance. This becomes crucial for mitigating the curse of dimensionality in high-dimensional cases, since if

> much of the variation in treatment effects is found among only a few dimensions, k-nn will do a poor job at giving more weight to important observations."

This method includes problems that seek to estimate $\theta(x)$ defined by the local moment condition

$$\mathbb{E}[\Psi_{\theta(x),\nu(x)}(O_i)|X_i = x] = 0 \quad \text{for all} \quad x \in \mathbb{X}$$

given data $(X_i, O_i) \in \mathbb{X} \times \mathbb{O}$, function $\Psi(.)$ and the nuisance parameter $\nu(.)$. As an example of problems that could be encompassed by this method, Athey, Tibshirani, and Wager (2019) cite instrumental variable, quantile regression, local maximum likelihood and, of course, causal forests commented above. The classic random forest (BREIMAN, 2001) is a special case, as well as the causal forest described above. This method also has a more general splits criterion that seeks to find high heterogeneity for different $\theta(x)$ of interest. Beyond all that, generalized random forests have desirable asymptotic properties: their estimates are consistent and follow a normal distribution. Also, the paper shows a way to construct confidence intervals using variance builded using the bootstrap of little bags.

Intuitively, the method uses the many grown trees to find how close two points are. The relevance of each training observation to fit $\theta(.)$ given point x is defined using the frequency of appearance of this training point in the same leaf of x in each of the many trees of the forest. This frequency is a kind of measure of how close i-th training and x are. Thus, the more frequently a point falls in the same leaf of x, more weight is given to this training observation in the process to predict $\theta(x)$. Visually, it may be illustrated by Figure 2.

Figure 2 – GRF's processs

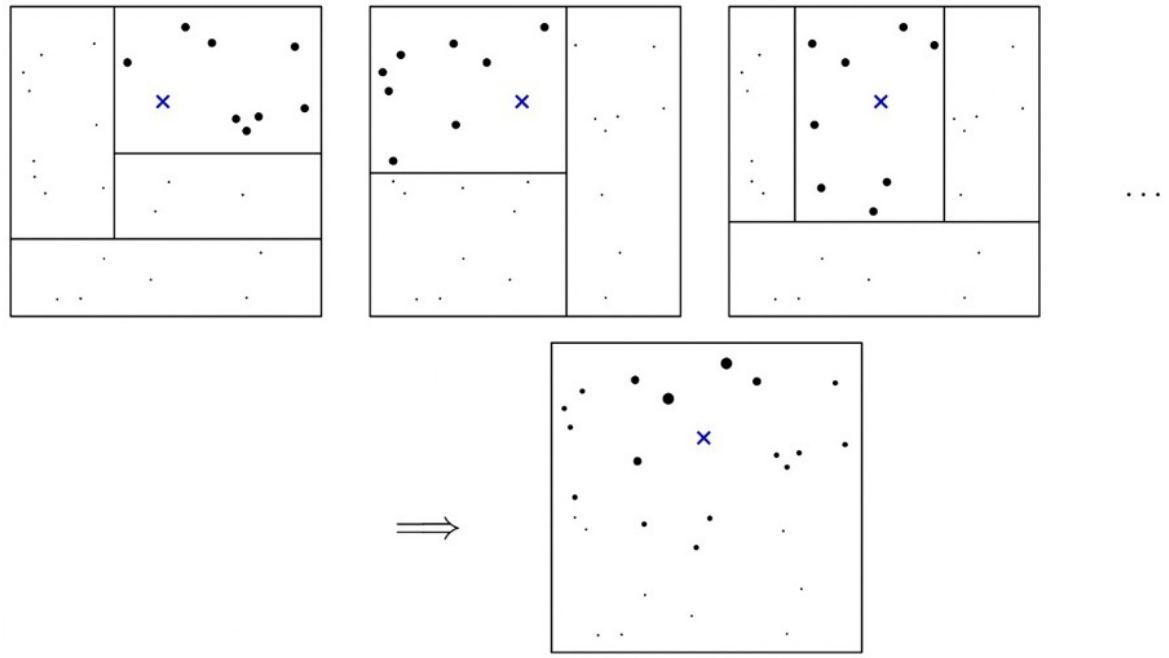


Illustration of the random forest weighting function. Each tree starts by giving equal (positive) weight to the training examples in the same leaf as our test point x of interest, and zero weight to all the other training examples. Then the forest averages all these tree-based weightings, and effectively measures how often each training example falls into the same leaf as x.

Source: Athey, Tibshirani, and Wager (2019).

Formally, the method tries to estimate the empirical version of the equation

$$(\hat{\theta}(x), \hat{\nu}(x)) \in \text{argmin}_{\theta,\nu} \left(|| \sum_{i=1}^{n} \alpha_i(x) \Psi_{\theta,\nu}(O_i) ||_2 \right)$$

and the weights $\alpha_i(x)$ are defined as

$$\alpha_i(x) = \frac{1}{\sum_{b=1}^{B} \alpha_{bi}(x)}, \text{ with } \alpha_{bi}(x) = \frac{\mathbf{1}(X_i \in l_b(x))}{|l_b(x)|} \ .$$

Remember: l_b is a leaf of a b tree, and $1, ..., B$ are the built trees. As expected, the weights sum up to 1.

The algorithm grows many trees, placing splits into trees using a criterion that maximizes as much heterogeneity as possible. This criterion is a gradient-based approximation of an ideal criterion that is expensive computationally. With all trees computed it is possible to produce estimates for observations of interest.

Causal forest and GRF

In order to make clearer the process of GRF algorithm in practice, herein we describe (briefly) the basic process of use of causal forest by using the generalized random forest. We

build these comments following Athey and Wager (2019)[4]. That work explores an application in an observational setting[5].

First, we get the weights from each observation using the trees built by the GRF process. Thereafter, using results from Robinson and the approach proposed by R Learner (ROBINSON, 1988), it is possible to estimate the treatment effect using the following:

$$\hat{\tau} = \frac{\sum_{i=1}^{n} \alpha_i(x) \left[Y_i - \hat{\mu}^{(-i)}(X_i) \right] \left[W_i - \hat{e}^{(-i)}(X_i) \right]}{\sum_{i=1}^{n} \alpha_i(x) [W_i - \hat{e}^{(-i)}(X_i)]^2} \ .$$

Remember that the propensity score and conditional mean of outcome given X_i are $e(x) := \mathbb{E}(W_i|X_i = x)$ and $\mu(x) := \mathbb{E}(Y_i|X_i = x)$ respectively. Besides, $(-i)$ indicates that the estimate for observation i results from "out-of-bag" predictions. That is: predictions for the observation i using trees that did not use that observation when they were built.

Procedurally, "the GRF implementation of causal forests starts by fitting two separate regression forests to estimate $\hat{\mu}(.)$ and $\hat{e}(.)$. It then makes out-of-bag predictions using these two first-stage forests, and uses them to grow a causal forest via [the last equation above]".

[4] Specifically pages 40-42.

[5] To be more precise, the researchers received the information that the dataset was from an experimental setting. But, when they analyzed the data more carefully, they concluded that it was more appropriate to treat the data from an observational setting then from a randomized one (see p.38).

4 EMPIRICAL APPLICATION

The exploration of causal forest in practice will be used to evaluate the heterogeneous effects of the Bolsa Família program on the labor market. Two aspects will be examined: first, we explore the impact on the status of occupation - do beneficiaries leave their occupations in the labor market given the CCT program? Second, we check if there are changes between the formal and informal sectors - do receivers of BFP work less in the formal sector? Before we start these analyses, we study the balance of the sample.

4.1 DATA AND SAMPLE DEFINITION

We conduct our analysis using data from Brazilian national household survey, the PNAD *contínua*. It substitutes a survey called only PNAD that was conducted on an annual basis. In October 2011 the PNAD *contínua*, which periodicity is on a quarterly basis, was experimentally implemented. From 2012 onwards, the survey was implanted in definitive character. About 211,000 households are surveyed. Each selected household is visited five times during five consecutive quarters. The survey aims to follow the evolution of the work force and to gather information on the socioeconomic development in Brazil. It was planned to produce information in a nationwide perspective, as well other levels of geographic dimension. The PNAD *contínua* is generated by the Brazilian Institute of Geography and Statistics (IBGE). Every quarter, in addition to the state of the workforce from households, the PNAD *contínua* also gathers information about basic characteristics from household residents: sex, age, color, race and education. We use PNAD *contínua* from 2012 in our work. Specifically, we use data from the first visit.

4.2 BALANCING AND PROPENSITY SCORE

In Table 1 we see some descriptive statistics from the control and treatment group surveyed in PNAD *contínua*. It is possible to see how there are much more BFP beneficiaries living in rural areas when compared with the other group: almost 80% of this last group leave in urban areas, while only just a little more than half of the BFP recipients leave in urban areas. Also, the control group has more residents than the treatment group, at least one more member. In terms of geographic distributions, almost 60% of beneficiaries live in the northeast of Brazil. North and southwest comes next with a much smaller volume of beneficiaries in each one. Both treatment and control group have similar proportions in terms of sex. In terms of race, we have much more non white people in the treatment group. Finally, note how different both groups are in terms of education: the treatment group is less educated.

Balance analysis

As we saw in the beginning of Section 4.2 covariate distributions from analyzed groups are very distinct, i.e., there is no covariate balance. The group of treated (those who participate in

Table 1 – Descriptive statistics

Variables	BF = 0	BF = 1
Bolsa Família benefit value (Avg., in R$)	0	134.44
Number of members (Avg.)	3.41	4.72
Age: between 25 and 65 (Avg.).	42.94	39.72
Years of schooling (Avg.)	9.20	5.79
Variables	**BF = 0 (%)**	**BF = 1 (%)**
Sex: Male	49.25	47.04
Race: White	49.24	24.47
Urban area	81.71	56.93
Metropolitan area	43.74	23.43
Region: North	10.98	16.25
Region: Northeast	24.16	58.52
Region: Southeast	32.30	13.16
Region: South	20.99	6.15
Region: Central West	11.56	5.93
Observations	169,038	46,388

Source: PNAD/IBGE (2012).
Note: in Brazil the questionnaire allows individuals to identify their race as white, black, yellow, indigenous and "parda".
Years of schooling: in Brazil, the high-school equivalent starts in 10th year. The college degree starts at 13th.

BFP) and untreated units are different on average in their observable characteristics. To examine the balancing in a little more depth between the two groups, we are going to proceed with an analysis of standardized mean differences and also show the graphs of each feature. First, we are going to verify the balancing of the two groups using the following statistic:

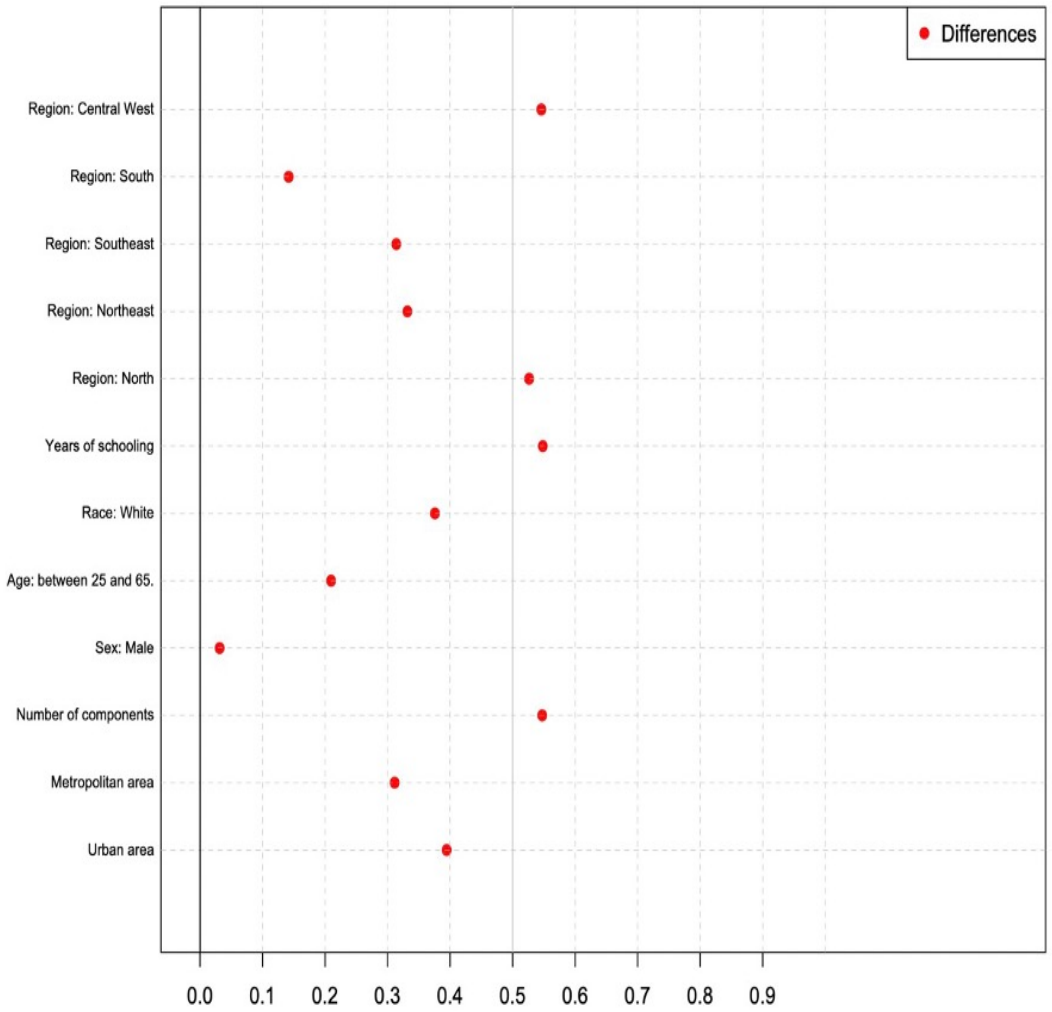
$$\frac{|\bar{Z}_1 - \bar{Z}_0|}{\sqrt{s_1^2 + s_0^2}} \ ,$$

where $i \in \{0,1\}$, Z_i is a function of feature X_i and $\bar{Z}_i$ is the mean of this function in the groups. s_i is the standard deviation of Z_i. We have $i = 1$ for the treated group and $i = 0$ for the control group. The above statistic informs us the distance of the means of studied groups. Lower values of the statistic indicate groups more balanced in the evaluated characteristic. The closer to zero, the more similar are the groups in relation to the observed characteristic. Figure 3 shows the standardized absolute mean differences. As expected, due to what we already saw in Table 1, the features of the groups are quite different. The only exception is sex. Figure 4 shows the densities of propensity scores of treated and control groups. As one could see, the two densities are (very) distinct. There is considerable volume of propensities close to zero in the control group, while

the propensities from the treatment group are well spread. This also suggests that these groups are, indeed, distinct. Recall that the propensity score is defined by the observable features of the units.

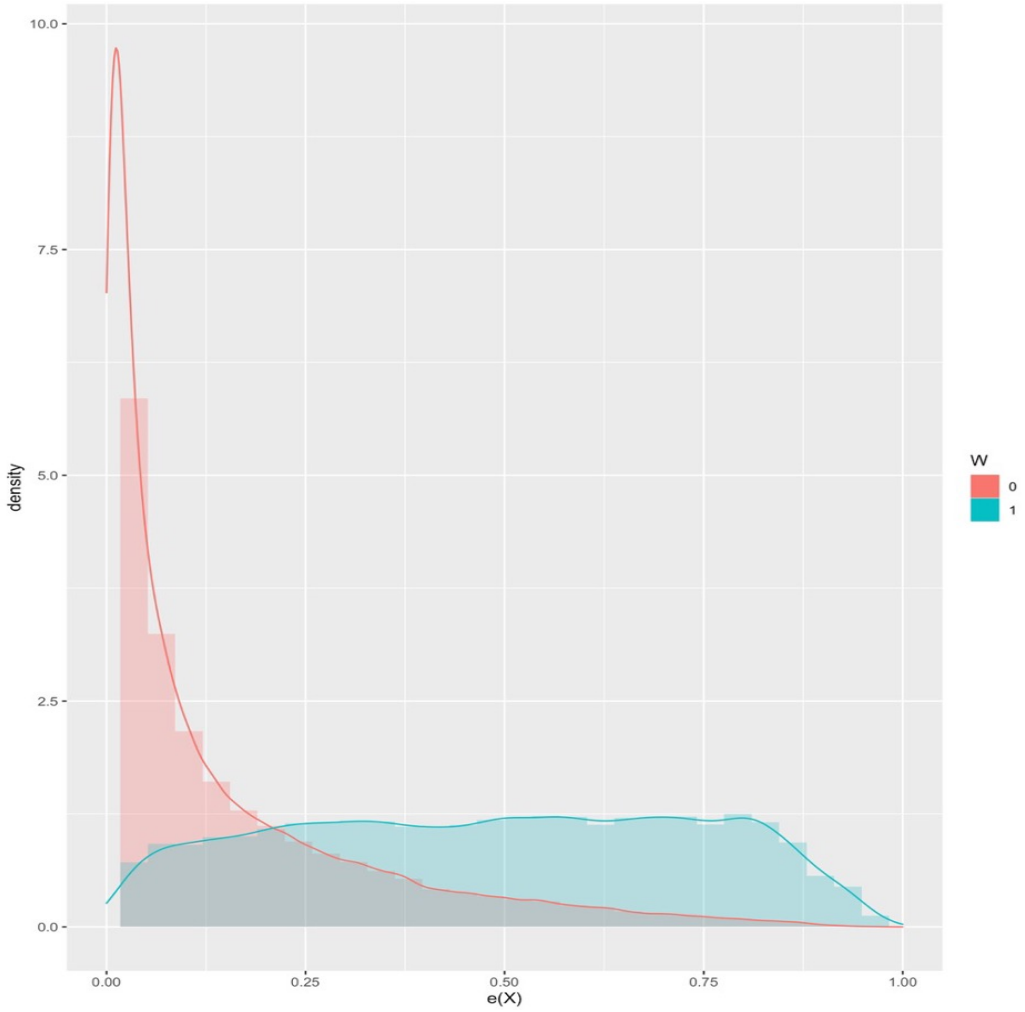
It is also possible to verify the unbalance between the groups verifying histograms. In appendix we have Figures 1, 2, 3, 4 showing the covariates histograms for the two groups. In Figure 1 we could see how the distributions of the number of components in the household are distinct: treatment groups have more people living in the same house. Also in Figure 1 we have where each group lives, if in rural or urban areas: what we saw in descriptive statistics also appears here. In Figure 2 we could observe the years of schooling. The control group has much more years of schooling than those who receive the benefit of BFP. There is a set of units in the control group that attend high school, and a set of those who attend college. In Figure 2 we also can note that between the beneficiaries of the program there is more non white than in the control group. Figure 3 and Figure 4 describe how control and treatment groups are distributed in the Brazilian regions. Beside that, in Figure 4 there is a histogram considering all regions altogether. Note: we have, for regions: 1 = south, 2 = southwest, 3 = midwest, 4 = northwest, 5 = north.

Figure 3 – Standardized absolute mean differences



Source: authors' elaboration.

Figure 4 – Propensity scores distribution



Source: authors' elaboration. Note: propensity scores via GRF's regression forest using the full sample.

Balancing the groups

As we are working in an observational context, we proceed using propensity score to get groups balanced. Beyond that, with the propensity score estimated it will be possible to evaluate if there is enough overlap. We use a regression forest to generate estimates from propensity scores.[1] We follow recommendations from Tibshirani et al. (2022) and Lab (2021) to generate propensities[2]. The process used the generalized random forest package ("GRF package"). With the propensity score estimated, we can evaluate if there is enough overlap, that is, if the estimates are bounded away from zero and one (formally: for some $\nu > 0$ it is true that $\nu < e(X_i) < 1 - \nu$), i.e, it is possible to find units in control and treatment groups for each level of features. This

[1] Estimates were generated with a forest with 2500 trees. Estimating propensity score used all computational resources available. As the sample size grows, the process of generating estimates using the GRF package becomes computationally demanding. From package documentation: "As a point of reference a causal forest with 1,000,000 observations and 30 continuous covariates takes around 1 hour to train (using default settings on a machine with 24 cores and 150 GB RAM), and the final forest takes up ca. 25 GB of memory (this is expected to scale linearly in the number of trees)".

[2] See also Imbens and Rubin (2015).

allows us to compare the group of beneficiaries and non-beneficiaries at each attribute level. Intuitively, it will be possible to find an amount of people in both control and treatment groups for all kinds of individuals in the analyzed population. It will be possible to compare between the analyzed groups. The overlap is also relevant because it allows us to adequately proceed the inverse propensity weighting (more on that below).

Together with overlap analysis, and aiming to choose an appropriate range of propensity scores distant from extremes, we verify how well calibrated, that is, if the model adequately captures the heterogeneity (in Section 4.3 section we get into details about this calibration). Also, we check if there is a balance between the two groups after weighting sample examples. We employ inverse propensity weighting for such tasks.

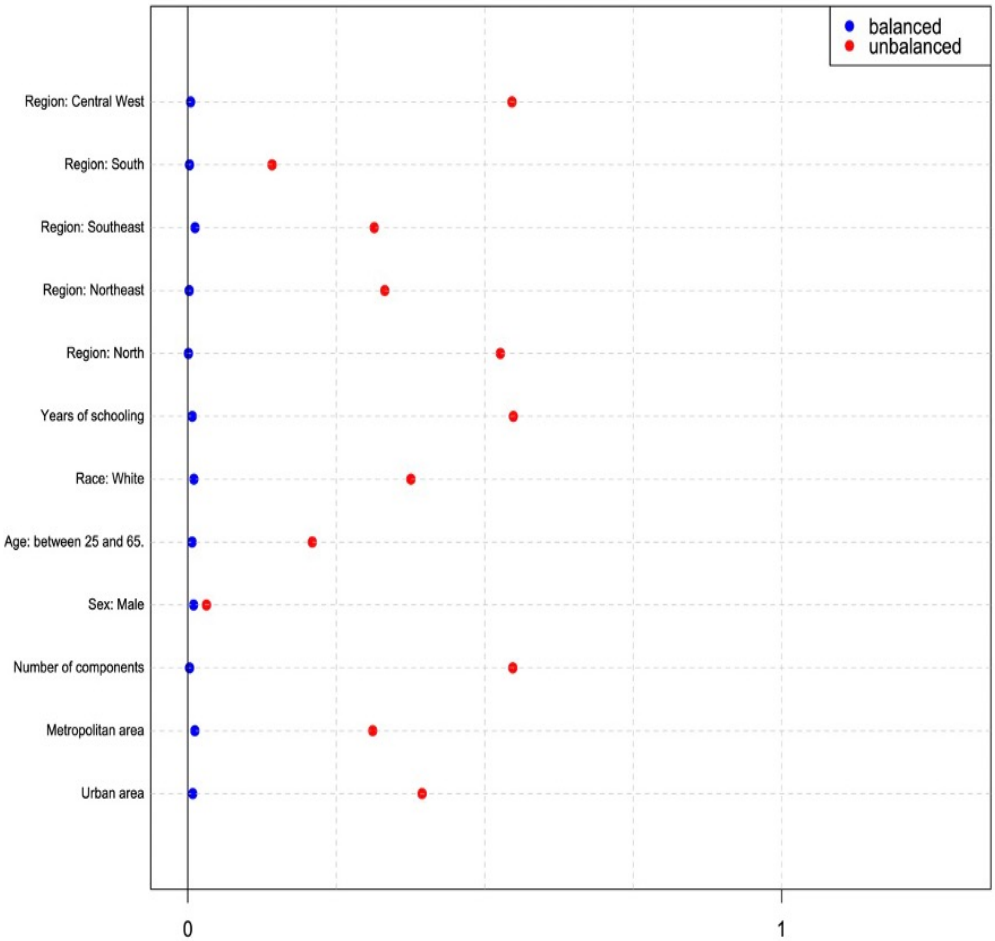
When we use the whole PNAD sample and its respective estimated propensity scores, we get many values close to zero and one. In addition to not respecting the overlap hypothesis, this brings some problems into the weighting process. Beyond that, when we use all this information to generate causal forest we get a model inadequately calibrated. So, in this context, we started our analysis (for impacts in occupations outcome) discretionarily choosing estimates that fall between 0.2 and 0.8. However, it seemed too arbitrary. Therefore, to choose a more data driven approach, we evaluated a series of ranges of propensity scores and verified how well calibrated the causal forest was generated in each case, as well how many observations were lost when we trimmed observations outside the range. Details of data driven processes are described in the Section 4.3. The final choice was to use the range of propensity scores between 0.2 and 0.67. With this range we get around 62 thousand observations. With this adequate range we have support for the overlap assumption. Besides, we get a well calibrated model. In Figure 6 we could see that support/overlap and estimates are from zero and one in the propensity score distribution. Our balance analysis and effects study are built using this selected range.

We are ready to see the balancing of the groups after weighting using propensity score. If the two groups get balanced after the process, we can proceed to the further steps of our analysis. Recall that we aim to achieve covariate balance, i.e., we aim to get similar covariate distributions after reweighting. We want both beneficiaries and non beneficiaries groups, after having their characteristics weighted using propensity score, to be similar when it comes to their observable characteristics. We proceed the balancing using inverse propensity weighting (IPW) in a similar manner that we commented in Section 3.2. We have $Y^*_{i,IPW} = Y_i \dfrac{W_i - e(X_i)}{e(X_i)(1-e(X_i))}$, where the $Y^*_{i,IPW}$ in this context refers to a (weighted) characteristic. After proceeding the weighting using estimates of $e(X_i)$ we employ a comparison between control and treatment groups. We get both groups with quite similar characteristics. The graph in Figure 5 indicates the balance using standardized mean differences. In addition to the differences of balanced characteristics, we plot the results before the weighting.

Now, we check the differences after weighing using the propensity score. As we said, if the groups are similar, then we see the differences close to zero. Therefore, we should get differences in this manner if the propensity scores are well calibrated. One can see that differences

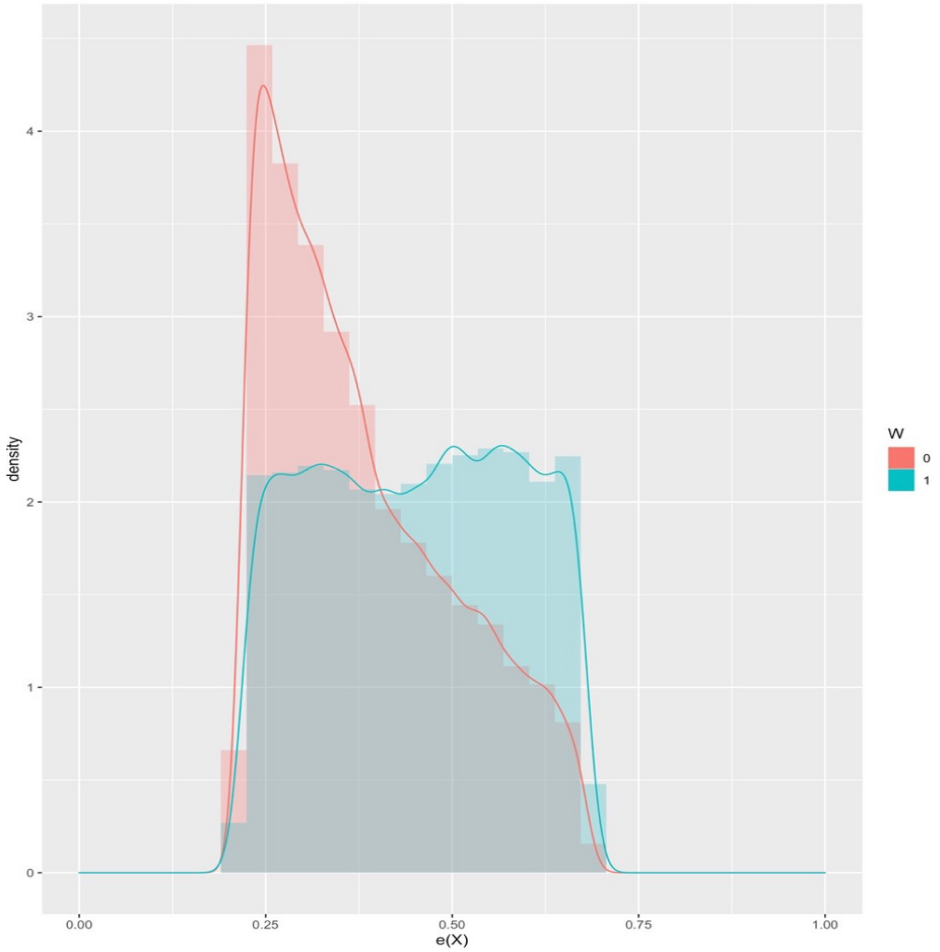
strongly reduce. Note how different the characteristics are before, and how close to zero are the absolute standardized mean differences after the weighing. Besides that, Figure 5 allows us to compare with pre-weighing characteristics. Note that the groups now have more balanced characteristics even in those features that before had great differences. In the appendix, the Figures 5, 6, 7 and 8 show the balance for each covariate after the balancing process. Now, one can see in Figure 5 that the distribution of the number of components in the household are similar in treated and untreated groups. In Figure 5 it is possible to verify that the proportions of those living in rural and urban areas are like. In Figure 2 years of schooling are, now, closer. Figures 7 and 8 show the distribution in each Brazilian region separately and when considering all regions together after the weighting. Remember that we have: 1 = south, 2 = southwest, 3 = midwest, 4 = northwest, 5 = north.

Figure 5 – Standardized absolute mean differences - Balanced and Unbalanced



Source: authors' elaboration.

Figure 6 – Propensity scores distribution adjusted



Source: authors' elaboration. Note: propensity scores between 0.22 and 0.68. Propensity scores are learned via GRF's regression forest.

4.3 RESULTS

Lab (2021) points out that a naive and inadequate form of evaluation of the results is to only observe the CATE histogram and/or evaluate the variable importance (measured by how often a variable is used to place split into trees of the forest). On one hand, the histogram could be concentrated in only one point, indicating that the model is underpowered (the effect would be more trackable and identified if there were more information and data). On the other hand, if the histogram was too much spread out, this could signal that it has been producing noisy estimates of the true CATE. In the case of variable importance, what could happen is that trees could split into a not so important variable in the true data generating process even though it is highly correlated with another one that is actually important to understand the heterogeneity. That is, we could ascribe some features as salient to produce heterogeneities when this is not adequate: it is not the case that the picked variable for the tree split is responsible for (at least directly) the observed CATE. When exploring the heterogeneity of CATE we will follow closely the approach indicated by Lab (2021) and Athey and Wager (2019). First, we employ an analysis about the impact of treatment on occupation. We verify if the *Bolsa Familia* benefit reduces/increases the

probability of getting occupied. After the first analysis, we study the effects of the treatment on a specific type of occupation: whether beneficiaries shift their workforce from the formal market.

Table 2 describes the possible types of occupation of the main job and Table 3 shows the average earning of each type of occupation. Those who do not receive the benefit are more likely placed in the formal sector than the program beneficiaries. The proportion of those in the control group who work in the formal private sector is at least twice as large as the proportion of the control group in the same sector. Besides, while in the private sector the untreated group works mainly in the formal jobs, the division between formal and informal private sector for treated group is almost 50%-50%. Observe how the two groups have a proportion of self employed workers close to 20%. Beyond that, there are many more employers in the control group (something we would expect). Table 3 informs us how the groups we are comparing have pronounced differences in terms of the earnings. Control group earns considerably more than the treatment group. The main earnings per hour from the first group is about R$ 9 (373,35/40,94) whereas the beneficiaries earn about R$ 4 (147,32/36,7). Observe that there is no huge difference in working hours per week, although control groups work more.

Table 2 – Types of occupation

Type	BF = 0	BF = 1	BF = 0 (%)	BF = 1 (%)
Private Sector				
Formal	43,841	5,805	28.19	12.79
Informal	10,105	4,797	6.5	10.57
Public sector				
Formal	2,229	214	1.43	0.47
Informal	3,029	976	1.95	2.15
Domestic				
Formal	2,828	535	1.82	1.18
Informal	4,651	2,507	2.99	5.52
Others				
Employer	5,942	321	3.82	0.71
Self-employed	29,108	10,433	18.72	22.98
Non informed	53,778	19,806	34.58	43.63

Source: PNAD/IBGE (2012).
Note: occupation refers to the kind to the main job.

It is useful to analyze how the heterogeneity of treatment is distributed in the analyzed subpopulation we are dealing with. As we said before, sometimes the average effect of treatment could not show how, in different subgroups, the treatment effect occurs. Therefore, we will find groups and evaluate the treatment effect in stratas of the analyzed sample. We will order groups according to CATE estimates. After constructing quintiles, we will pay attention to the average

Table 3 – Earnings

Type	BF = 0	BF = 1
Main		
Month (Avg.)	R$ 1,493.46	R$ 589.27
Week (Avg.)	R$ 373.36	R$ 147.32
Worked hours (Avg., weekly)	40.94	36.7
Total		
Month (Avg.)	R$ 1,562.84	R$ 529.94
Week (Avg.)	R$ 390.71	R$ 132.49
Worked hours (Avg., weekly)	41.54	37.24

Source: PNAD/IBGE (2012).
Note: Main refers to earnings (in money or products) from main work from people with age greater than 14 in the household.
Total refers to earnings of all possible sources: money or products as payment from work or other sources. E.g: pensions for retirement, unemployment insurance, donations.

effect in each group. Furthermore, we study the joint distribution of characteristics of each group, i.e., we will also study the features of each group.

When building the quintiles, to avoid a biased ranking between estimated effects, we split the sample in a way that it allows us to rank sample observations adequately. We must guarantee that when ranking the estimates from given two observations, the estimates are produced with a model constructed without using any of the two observations. We split the sample in K folds and fit the model for CATE estimates in $K-1$ folds, then we use the model to produce estimates for the held-out fold. After, in each fold, estimates are ranked for the least to the greatest found effect. Finally, we join the rankings and precede our analysis. To make this process more concrete, suppose there are $\{1,...,K\}$ folds. Fix a fold k. We build a causal forest where part of the trees of the forest do not use data from fold k. With these trees we get predictions for observations in k. Then, observations of fold k are ranked in quintiles according to CATE estimates. This process is repeated for the remaining $K-1$ folds. After that, the rankings are concatenated.

Occupation

Causal forest and calibration of the model

Before analyzing the results we pay attention to the model calibration, that is, we assess if it adequately captures the heterogeneity. Lab (2021) suggested a way to try to evaluate this, it recommends using a linear approximation and seeing if, with this simple setting, the model is capable of detecting heterogeneity. It is possible to evaluate if the model is well calibrated to capture heterogeneity (if there is some) by using the held out data and the following linear

approximation:

$$\hat{\tau}_i^{ho}(X_i) = \alpha\bar{\tau} + \beta\left(\hat{\tau}^{-i}(X_i) - \bar{\tau}\right) + \varepsilon \ , \tag{4.1}$$

where $-i$ indicates estimates for observation $-i$ results from "out-of-bag" predictions. Recall that this kind of prediction refers to those for observation i using trees from causal forest that did not use this observation when they were constructed. In the 4.1 the quantity $\hat{\tau}_i^{ho}(X_i)$ is the CATE prediction in held out data, $\hat{\tau}^{-i}(X_i)$ is the predicted treatment effect and $\bar{\tau} := n^{-1}\sum_{i=1}^{n}\hat{\tau}^{-i}(X_i)$ is the average of treatment effect predictions. In this approximation, we decompose $\hat{\tau}_i^{ho}(X_i)$ between the average of (conditional) treatment effect predictions and the deviations of $\hat{\tau}^{-i}(X_i)$ from this average. The term $(\hat{\tau}^{-i}(X_i) - \bar{\tau})$ is a kind of measure of how different is the treatment effect of unit i from the other observations, it captures how diverse is treatment effect of unit i. The α coefficient indicates how well our model recovers the average effect. The β suggests how well the model captures the heterogeneity. When both coefficients are significant and close to one, this suggests that our model is well calibrated: the results from causal forest adequately capture the heterogeneity of the treatment effect. Lab (2021) explains the rationale for evaluation of the coefficients:

> "The coefficients α and β allow us to evaluate the performance of our estimates. If $\alpha = 1$, then the average prediction produced by the forest is correct. Meanwhile, if $\beta = 1$, then the forest predictions adequately capture the underlying heterogeneity."

We proceed with an analysis of the model calibration for both aspects we are studying: for impacts on occupation and formal labor sector participation. When we evaluate the model calibration for the outcome occupation considering the full PNAD sample we get a model that is not well calibrated.To deal with this, we build a data driven analysis to choose a more adequate model. We proceeded searching for distinct intervals of propensity scores. We carry out the following process: we run a causal forest for some intervals of the propensity score, get their results and save for posterior analysis of the model calibration. After the process is finished, we pay attention to how well calibrated the model is using saved outputs. The main goal in the process was to find an adequate setting to employ our analysis. Adequate here means a setting that allows for proper coefficients (i.e. close to one and statistically different from zero) in the equation 4.1, producing a well calibrated model. Furthermore, the setting would be considered adequate if it allowed maintenance for a reasonable part of the sample after trimmed observations, given the chosen interval. In Table 5 we have the analyzed coefficients. In this work, we choose to use the propensity scores between 0.2 and 0.67. This allows us to work with approximately 62 thousand observations. The coefficients of this interval for equation 4.1 are shown in Table 4.

$\hat{\tau}_i^{ho}(X_i)$ is the CATE prediction in held out data. * p<0.1; ** p<0.05; *** p<0.01.
Robust standard errors in parenthesis.

Table 4 – Test calibration - Occupation

	$\hat{\tau}_i^{ho}(X_i)$
α	1.002^{*}
	(0.094)
β	0.706^{***}
	(0.050)

Source: authors' elaboration.
Note: $\hat{\tau}_i^{ho}(X_i)$ is the CATE prediction in held out data.
Robust standard errors in parenthesis.
* p < 0.1; ** p<0.05; *** p<0.01.

Table 5 – Calibration tests - Intervals and coefficients

Intervals	Num. Obs.	α	α s.d.	α p-value	β	β s.d.	β p-value
0.21-0.67	60,166	1.020	0.680	0.070	0.690	0.050	0
0.22-0.67	58,032	0.990	0.740	0.090	0.700	0.050	0
0.23-0.67	55,693	1.040	0.730	0.080	0.690	0.050	0
0.24-0.67	53,626	0.990	0.740	0.090	0.680	0.050	0
0.25-0.67	51,635	1.010	0.780	0.100	0.670	0.050	0
0.26-0.67	49,721	1.010	0.740	0.090	0.660	0.060	0
0.27-0.67	47,749	1.050	0.770	0.090	0.690	0.060	0
0.28-0.67	45,943	1.020	0.920	0.140	0.690	0.060	0
0.29-0.67	44,173	1.490	2.830	0.300	0.700	0.060	0
0.3-0.67	42,456	1.330	3.240	0.340	0.690	0.060	0
0.2-0.66	61,604	1.070	0.620	0.040	0.700	0.050	0
0.21-0.66	59,350	1.070	0.670	0.060	0.690	0.050	0
0.22-0.66	57,216	1.050	0.760	0.080	0.690	0.050	0
0.23-0.66	54,877	0.990	0.700	0.080	0.690	0.050	0
0.24-0.66	52,810	1.060	0.770	0.080	0.680	0.050	0
0.25-0.66	50,819	0.990	0.790	0.100	0.670	0.050	0
0.26-0.66	48,905	1.050	0.700	0.070	0.680	0.060	0
0.27-0.66	46,933	1.060	0.780	0.090	0.700	0.060	0
0.28-0.66	45,127	1.140	1.060	0.140	0.700	0.060	0
0.29-0.66	43,357	0.830	1.810	0.320	0.680	0.060	0
0.3-0.66	41,640	0.830	2.140	0.350	0.680	0.060	0
0.2-0.65	60,785	1.070	0.570	0.030	0.690	0.050	0
0.21-0.65	58,531	0.970	0.570	0.050	0.690	0.050	0
0.22-0.65	56,397	1.100	0.730	0.070	0.690	0.050	0
0.23-0.65	54,058	1.110	0.730	0.070	0.690	0.050	0
0.24-0.65	51,991	0.950	0.630	0.070	0.680	0.050	0
0.25-0.65	50,000	1.100	0.740	0.070	0.680	0.060	0
0.26-0.65	48,086	1	0.640	0.060	0.670	0.060	0
0.27-0.65	46,114	1.060	0.710	0.070	0.670	0.060	0
0.28-0.65	44,308	0.970	0.810	0.120	0.690	0.060	0
0.29-0.65	42,538	0.980	1.660	0.280	0.680	0.060	0
0.3-0.65	40,821	1.280	2.410	0.300	0.670	0.060	0

Source: authors' elaboration. Note: for each interval we run a causal forest and note the calibration test result. We also pay attention to the number of observations. Causal forests had been built with 1000 trees due to computational costs. Heteroskedasticity-robust (HC3) standard errors.

Treatment effects

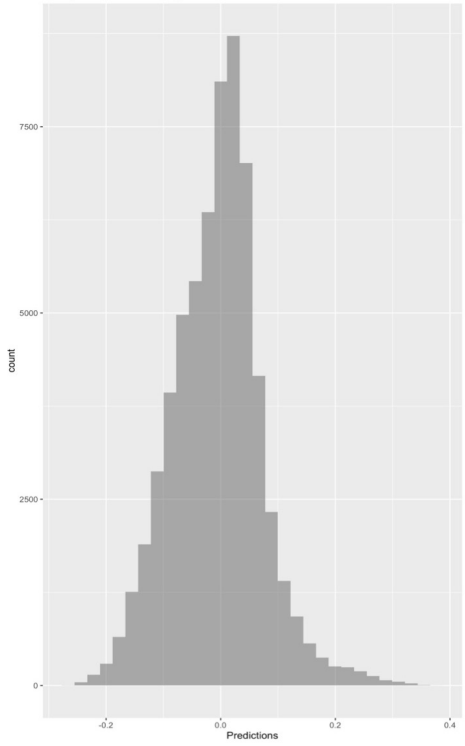
Before we proceed with the heterogeneity analysis, it is possible to see what the estimate of ATE is in occupation. Table 6 reports the estimate. One could see that there is little effect, on average, on occupation. Although it is not correct to make an analysis of heterogeneity using only histograms, we could use it to get an idea of how the estimates are spread out (despite the caveats we already made). Figure 7 displays the histogram of CATE estimates. It suggests that although on average the effects are practically null, there are distinct effects on those who are treated. In this case, a deeper heterogeneity analysis is helpful. Our next step is to conduct such analysis.

Table 6 – Occupation - ATE

estimate	std.err
−0.007	0.004

Source: authors' elaboration.
Note: estimated average treatment effect for the scores between 0.2 and 0.67.

Figure 7 – Occupation: CATE histogram

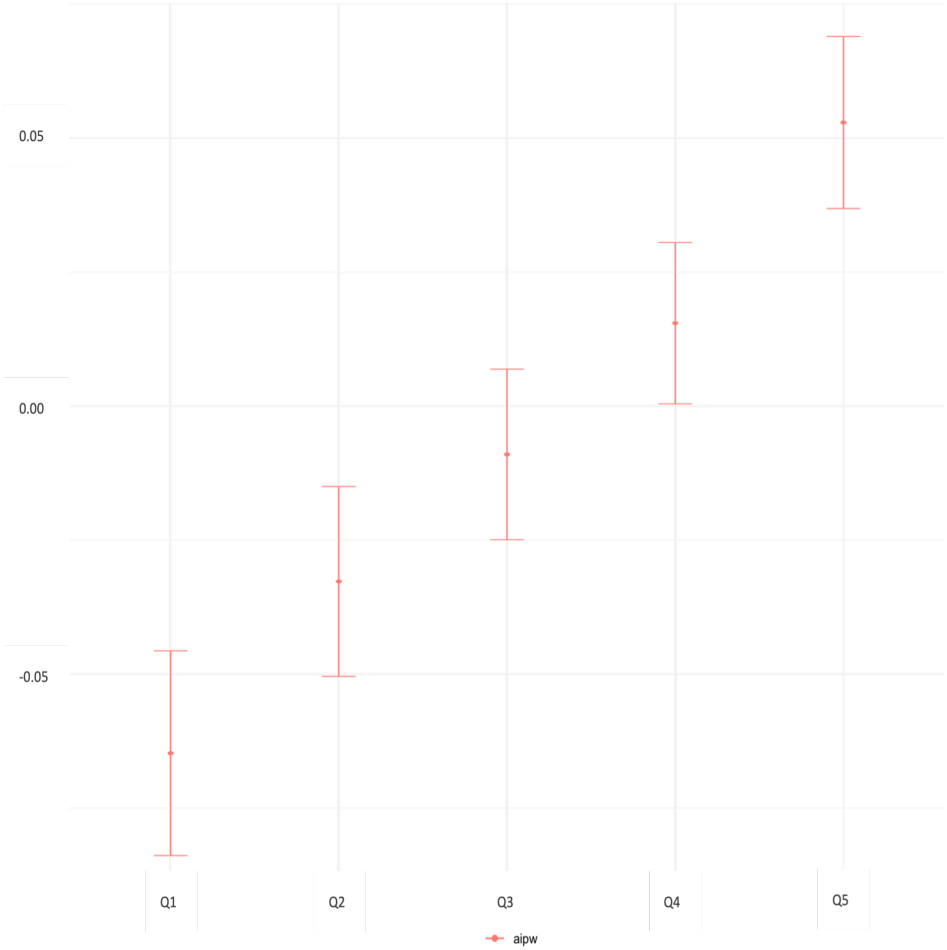


Source: authors' elaboration.

As commented in the first part of Section 4.3, we cluster treatment effects, organizing them in quintiles. Then, we calculate an average of the effect in each quintile. The Table 7, transcribed below, contains these estimates and it is possible to see that the graph in Figure 8 shows a monotonic pattern. This, as argued by Lab (2021), indicates that the model adequately

captures the heterogeneity. The first quintile has negative effects, and the second, positive effects. The differences between them surpass 10%. One could see in quintiles 2 and 4 something similar to those extreme quintiles, albeit with less magnitude. In the majority of quintiles, the impact on occupation is negative, although one could argue that the effect on third quintiles are marginally negligible. A valid question is if these effects, described in the table and illustrated in the graph, are really distinct. We conducted tests checking if there were differences of the averages in each quintile and we found that, in fact, there are distinctions between quintiles (with exception of first and second quintiles). Tables 1, 2, 3, 4, 5 in appendix show the tests results. All of this suggests the existence of distinct effects in those groups and the existence of heterogeneity. There is a group that reduces their occupation in face of treatment. There are those recipients who do not alter their behavior when receiving the benefit. And, finally, there is a part of the sample who gets more occupied.

Figure 8 – Occupation - Average CATE within each ranking



Source: authors' elaboration. Note: Ranking defined by predicted CATE. 95% confidence intervals.

Covariate Analysis

We conduct an analysis of how the groups (each quintile) are distinct in respect to each of the covariates. In previous sections, we verify how the treatment effect on occupation varies

Table 7 – Occupation - ATE within each quintile

ranking	estimate	std.err
Q1	−0.065	0.010
Q2	−0.033	0.009
Q3	−0.009	0.008
Q4	0.015	0.008
Q5	0.053	0.008

Source: authors' elaboration. Note: average of CATE in each quintile. Quintiles defined according to CATE estimates. Informed robust standard errors.

in the five analyzed groups. Now, we want to pay attention to the observable characteristics of these groups. Do the groups look similar in terms of the covariates? If the groups are different, the difference is due to which covariables? This could be helpful to policy makers, for example, when using the estimates of treatment effect as a tool to improve their treatment assignment. We build the heatmap indicated in Figure 9. It helps to evaluate how the joint distribution of features changes in each group.

First, it is worth noting things that are similar in all groups. The majority of individuals in all these groups identified themselves as non white (in Brazil, the PNAD questionnaire allows individuals to identify themselves as white, black, yellow, Indigenous and "parda"). Between 20% and 30%of this population reports as white. The major part of those groups live in urban areas: in the first quintile approximately 78% lives in this type of area and from the second quintile until the fifth one, the proportion goes from around 70% to 60%. All groups are little educated: all of them have fewer years of study. The years of schooling range from approximately 8 years to almost 4 years. Great part of the analyzed sample has only 6 years of schooling. In Brazil, the high-school equivalent starts in 10th year, so all groups (on average) do not access this level of education.

Let's focus on the differences. Although a major part of all groups live in urban areas, the difference between first quintile and the fifth quintile in the proportion of those living in rural areas is considerable, the first group has 20% living in this condition, the second has around 40%. Besides, those in the first quintile have more people per household than those in the fifth quintile. In reality, one can see that in almost all groups the number of components of households is (on average) at least four people. This changes in the fifth quintile. It is the only group that can have less than four people per household. If one rounds these numbers up, we can say that all groups have five components per household, except the fifth quintile, which has four people per family. Furthermore, there is a feature of the first quintile well different from any other group: it is younger than other quintiles. They have, on average, at least 35 years, while the age in all other groups revolves around 40-41 years.

Note, also, that there is a considerable difference between men and women in the groups. When one compares the first quintile with the fifth quintile, one will see a striking difference: a significant part of those in the first group are women, and the last one have around 80% of males.

As we move from the first quintile to the fifth quintile, passing by those between the extremes, we see an increasing movement: the more positive the effect, greater is the number of men in the group. This movement goes in agreement with the findings in literature: often the group who stops working when the household receives the benefit are women.

Therefore, we could identify the following pattern: those who left some occupation live mostly in urban and metropolitan areas. Their households have more people living in the same residence. Although they have more years of education when compared with the others (those with increase in occupation), they tend not to attend high school. The groups with an increase in occupation have these features following a different pattern. Both groups are poorly educated. Finally, both groups are mostly non white.

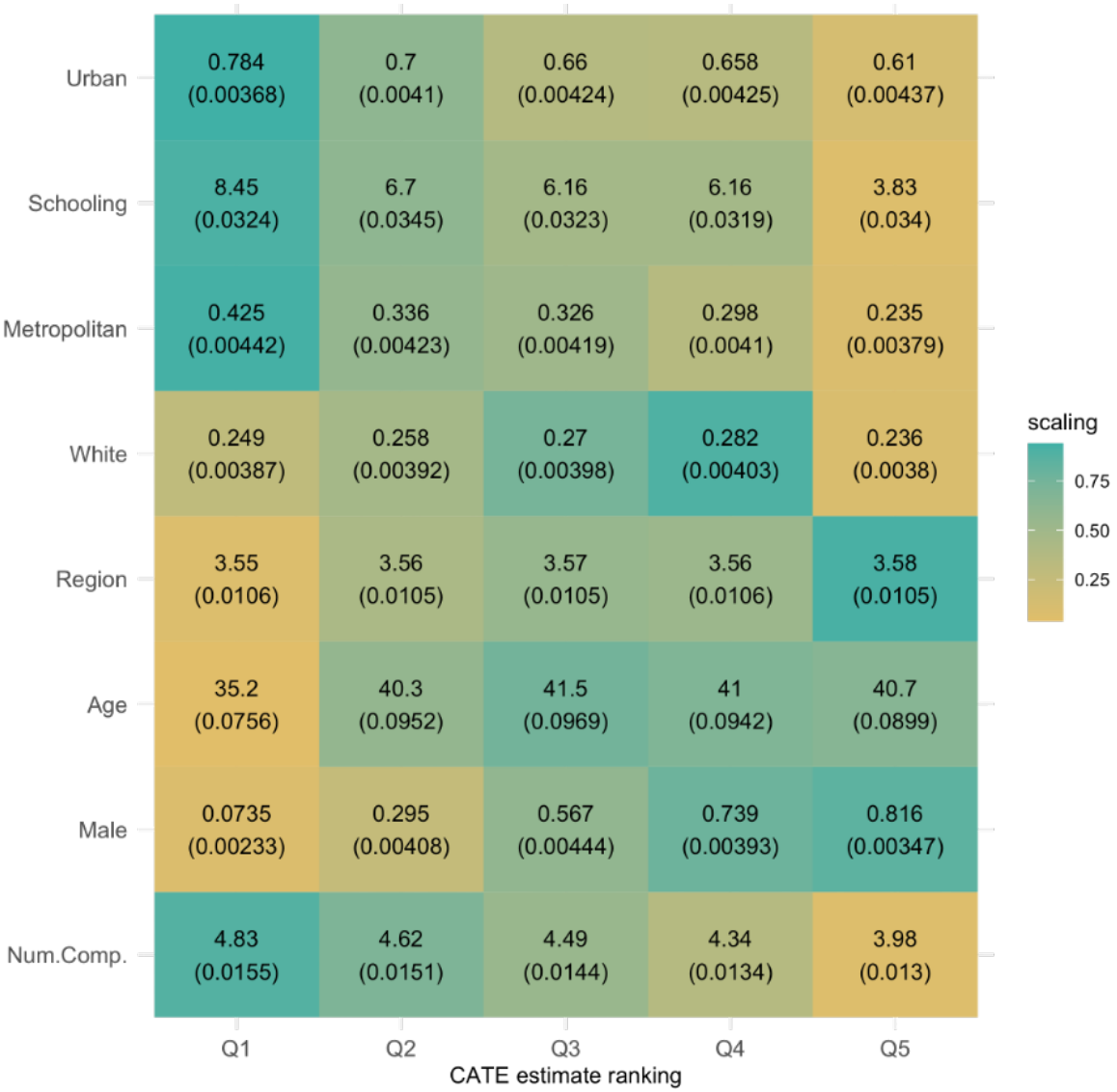
One hypothesis of why those in the first and second quintile left some occupation could be formulated. It could be the case that living in more urban (and metropolitan) areas could bring more social security due to the access to public goods (allied with the number of people). So, when there is a relief in the budget constraint, people could stop working because they have more income and could use the facilities given the urban regions. On the other hand, those who live in rural areas have a more precarious context (due to the lower public goods supply).

There is another hypothesis that could justify the decrease in occupation in the first and second quintiles[3] while there is an increase in the fourth/fifth quintile.

Groups that reduce their occupation tend to have more people at home, especially the first quintile. This fact, together with the finding that there are more women in these groups, may suggest that they stop working to be able to stay at home and take care of their children. If, beyond that, most of the people in the family are in working age (or close to this age), this can contribute to the woman's decision to stay at home, as there are people who can continue to work. In both the scenarios the benefit provides relief in family budget constraint allowing the woman to make the decision to stay at home.

[3] In the third quintile the effect was almost zero, so we do not pay attention to them in this hypothesis

Figure 9 – Occupation - Average covariate value within groups



Source: authors' elaboration.
Note: this graph shows the average of each covariate. The color indicates the normalized distance between the value from the covariate in each group and the mean of this covariate in all analyzed sample, i.e.: $\Phi^{-1}\left((\widehat{\mathrm{E}}[X_i|Q_i]-\widehat{\mathrm{E}}[X_i])/\widehat{\mathrm{Var}}(\widehat{\mathrm{E}}[X_i|Q_i])\right)$. Rows are ordered by variation: from highest (first row, up to down) to lowest (last row, up to down).The measure of variation is given by: $\widehat{\mathrm{Var}}(\widehat{\mathrm{E}}[X_i|Q_i])/\widehat{\mathrm{Var}}(X_i)$. Standard errors in parenthesis. Urban: 1 if people live in urban area, 0 otherwise. Schooling: 1 to 16. Metropolitan: 1 if people live in a metropolitan area, 0 otherwise. White: 1 if race = white, 0 otherwise. Region: 1 = south, 2 = southwest, 3 = midwest, 4 = northwest, 5 = north. Age: 25 up to 65. Male: 1 if sex = male, 0 otherwise.
Rankings defined by predicted CATE. Robust s.e. in parenthesis.

Formal sector

Treatment effects

In this section we complement our previous analysis turning our attention to the impact of the BFP in the participation of beneficiaries in the formal sector. As before, we will first evaluate the best interval of propensity score to build our analysis (see Section 4.3 for details). Given that the same setting used to evaluate the effect in occupation also produces a well calibrated model when the output variable is formal sector participation, then it will be also used here to employ analysis. We have the found coefficients in Table 8.

Table 8 – Test calibration - Formal Sector

	$\hat{\tau}_i^{ho}(X_i)$
α	0.998^{***}
	(0.052)
β	0.738^{***}
	(0.058)

Source: authors' elaboration.
Note: $\hat{\tau}_i^{ho}(X_i)$ is the CATE prediction in held out data.
Robust standard errors in parenthesis.
* p $<$ 0.1; ** p$<$0.05; *** p$<$0.01.

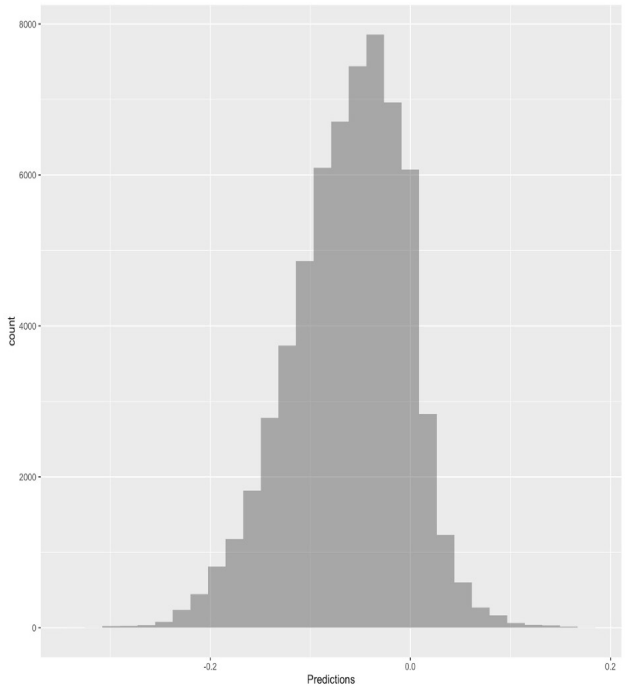
When it comes to ATE estimate in the context of the formal sector, one could see that there is, on average, a reduction in formal labor market participation. Estimates are in Table 9 and the histogram of conditional treatment effects is on the Figure 10 (recall the caveats already commented on this kind of graph). As made before, we looked for differences of treatment effect (on average) between quintiles. We find differences between them. There is a negative effect occurring in all quintiles (see Table 10). Graphically we may see it in Figure 11. All groups reduce their labor supply in the formal sector. In all groups, we see that participation in formal jobs reduces, although in the fifth and fourth quintile there is a weaker movement. Almost all differences between the stratas are statistically significant, except when we compare the effects between the third and fourth quintiles. Therefore, there is a detectable heterogeneity in the effects. Tables 6, 7, 8, 9 and 10 in appendix summarize the test of differences between quintiles.

Table 9 – Formal sector - ATE

estimate	std.err
-0.061	0.003

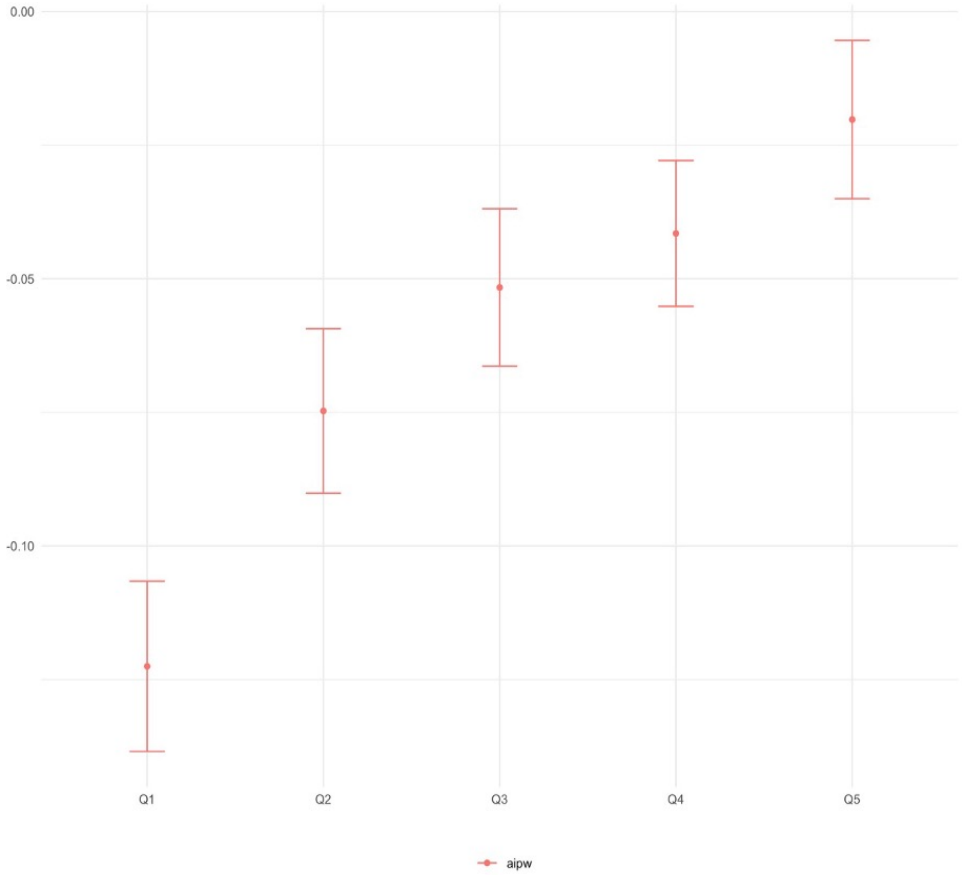
Source: authors' elaboration.
Note: estimated average treatment effect for the scores between 0.2 and 0.67.

Figure 10 – Formal sector participation: CATE histogram



Source: authors' elaboration.

Figure 11 – Formal sector - average CATE within each ranking



Source: authors' elaboration. Note: Ranking defined by predicted CATE. 95% confidence intervals.

Table 10 – Formal Sector - Average of treatment within each quintile

ranking	estimate	std.err
Q1	−0.123	0.008
Q2	−0.075	0.008
Q3	−0.052	0.007
Q4	−0.042	0.007
Q5	−0.020	0.007

Source: authors' elaboration.
Note: Average of CATE in each quintile. Quintiles defined according to CATE estimates. Informed standard errors are robust.

Covariate Analysis

We conduct an analysis of how each quintile is distinct from the other in respect to each of the covariates. Previous section suggested that there is reduction in participation in the formal sector. Now, we want to study the observable characteristics of these groups. Do the groups look different in terms of these covariates? How similar or distinct are they due to their covariables? This could be helpful to policy makers, for example, when using the estimates of treatment effect as a tool to improve their treatment assignment. The heatmap indicated in Figure 12 is designed to help to evaluate the joint distribution of the features of those groups. As before (in the occupation case), we could pay attention to similar characteristics in all strata. Again, a major part of the analyzed group live in urban areas and close in region, but do not live in metropolitan areas. Around 76% of those in the first quintile live in urban areas, while in the other groups this number revolves around 65%. Consequently, the number of people living in rural areas ranges from 25%-35%. Also, a great part of the population in all quintiles are predominantly non white. The quintiles with more white people have only 29% and 27% (fourth and third quintile, respectively) of this segment.

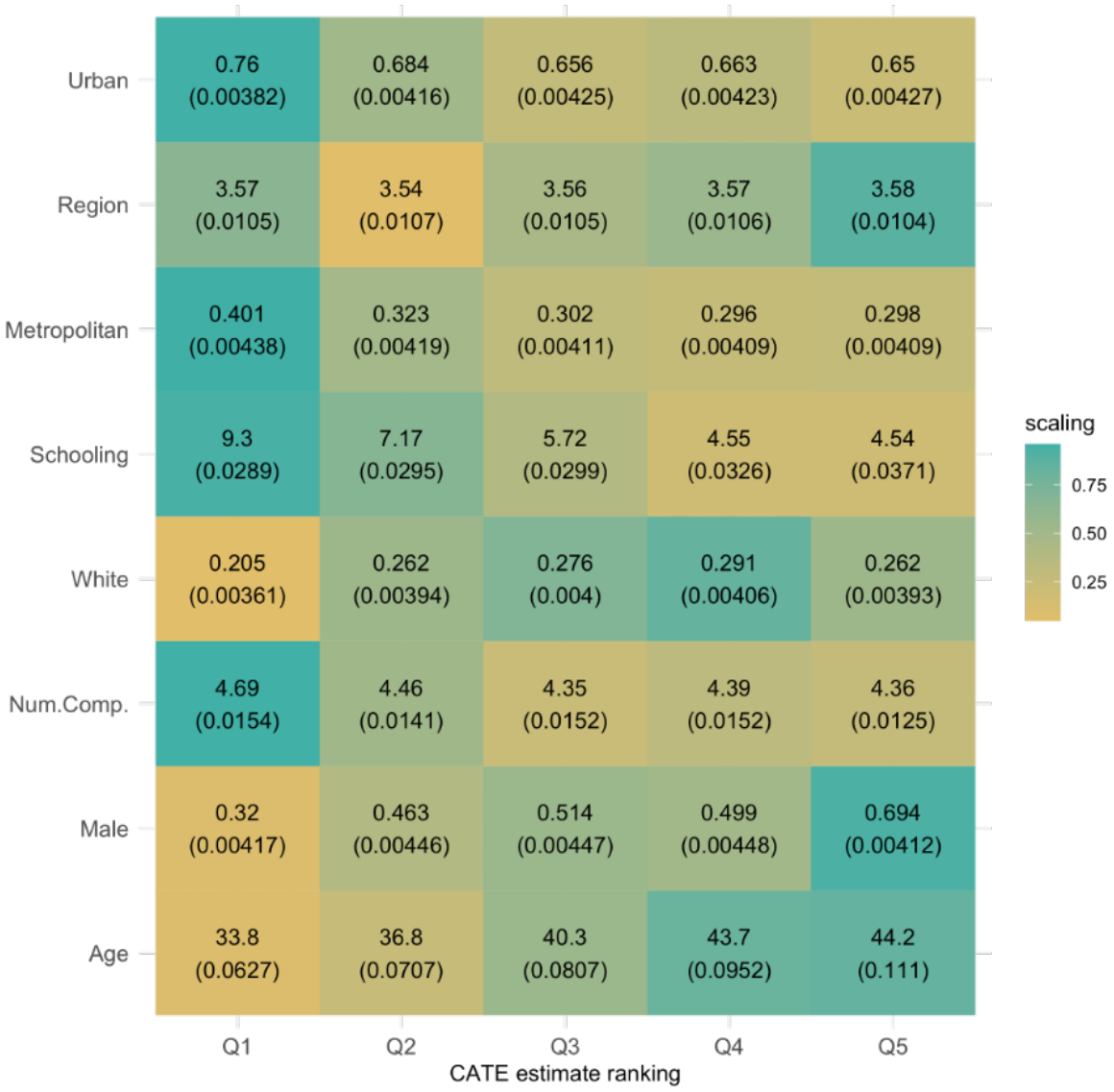
Women also are the major part of the group with more beneficiaries leaving the formal sector. The proportion of women in the first quintile is almost 70

It is puzzling to try to make some hypothesis for the differences in the magnitude of observed reduction in quintiles using what we know about covariates in each quintile. We could try to provide some hypothesis for the distinction between the striking difference in effects between the first quintiles and the last ones in the labor supply for the formal labor market: it could be happening because there are greater opportunities, in more urban areas, for informal jobs (supposing that those who reduce the their participation in formal sector could start to work in the informal sector).

A variation/extension of the previous argument: it is possible that, due to the fact that the fifth quintile has more people living in rural areas, this quintile has individuals that work in areas where the government control on the formal and informal sectors is easier: there are fewer firms to inspect. However, in urban areas (where the majority of people of the first quintile live) there

are much more firms and much more possible labor functions, bringing some difficulties for law officers to do their job, thus making government supervision more fragile. If we assume that part of the reduction of the formal sector participation is due to the change of worker to the informal sector, then it is possible that when receiving the benefit, those in the first quintiles choose to reduce their participation in the formal market to ensure the maintenance of benefit. Meanwhile, they choose to engage in an informal market for complementing their earnings.

Figure 12 – Formal Sector - Average covariate value within groups



Source: authors' elaboration.
Note: this graph shows the average of each covariate. The color indicates the normalized distance between the value from the covariate in each group and the mean of this covariate in all analyzed sample . I.e.: $\Phi^{-1}\left((\widehat{\mathrm{E}}[X_i|Q_i] - \widehat{\mathrm{E}}[X_i])/\widehat{\mathrm{Var}}(\widehat{\mathrm{E}}[X_i|Q_i])\right)$. Rows are ordered by variation: from highest (first row, up to down) to lowest (last row, up to down).The measure of variation is given by: $\widehat{\mathrm{Var}}(\widehat{\mathrm{E}}[X_i|Q_i])/\widehat{\mathrm{Var}}(X_i)$. Standard errors in parenthesis. Urban: 1 if people live in urban area, 0 otherwise. Schooling: 1 to 16. Metropolitan: 1 if people live in a metropolitan area, 0 otherwise. White: 1 if race = white, 0 otherwise. Region: 1 = south, 2 = southwest, 3 = midwest, 4 = northwest, 5 = north. Age: 25 up to 65. Male: 1 if sex = male, 0 otherwise.
Rankings defined by predicted CATE. Robust s.e. in parenthesis.

5 CONCLUSION

In this work, we review some models built on machine learning and statistical learning techniques for the analysis of heterogeneous treatment effects. We briefly describe models that use off-the-shelf machine learning methods and, after that, models tailored for direct estimation of conditional average treatment effects. We use a forest based method to an applied analysis: causal forest. We describe this method and its intersection with generalized random forest[1]. Also, we commented on how honest estimation is used in this method, as well. The causal forest model allowed us to discover the average effect of the treatment and, mainly, the heterogeneity in the treatment effect for the covariates of the analyzed sample. Afterwards, we explore this method using an empirical application: we explore the impact of a conditional cash transfer in Brazil, the BFP.

It is important to comment on the existence of two important challenges when using causal forest: how to verify whether the detected heterogeneity actually exists and how to understand the detected heterogeneity. In the process of checking if the model correctly detected heterogeneity, the variable's importance and CATE of histograms are not enough and could be misleading. Some processes are suggested to perform in the literature and in this work we proceed through a linear approximation. The process of studying estimates using such approximation is called calibration. As for the study of the detected heterogeneity, the suggested procedure is to gather the observations into groups according to the ranking of the CATE estimates. In this way we have mechanisms to evaluate the different groups (in terms of treatment effects). This grouping allows us to study the joint distribution of the features of the groups. This evaluation of the characteristics of the different groups is relevant for the policy maker to decide to whom he should, in the next interventions, assign the treatment. Therefore, the ability to correctly detect heterogeneity and subsequently produce adequate strata of the analyzed population is of great relevance. A whole body of literature have been developing on how to utilize the heterogeneity of causal effects in the optimal treatment assignment process (see Manski (2000, 2004) for classic works, and Kitagawa and Tetenov (2018) and Kasy and Sautmann (2021) for a more recent material).

It is also worth mentioning that, as in other works that use propensity score, it is also necessary to pay attention to the overlap hypothesis. The use of the causal forest does not exempt the researcher from this. Therefore, the limitations that the use of the propensity score brings to empirical works also occur here. In our work, we had to use sample smaller than the original one in order to obtain the overlap.

In exploring the causal forest in an applied context we study the *Bolsa Família* program literature on the effects of that policy in labor supply. We studied the intervention effects in both occupation and formal labor market participation. We commented on the theoretical aspects of this intervention, briefly described the condition of cash transfer programs and explained the

[1] We saw that this method is demanding power computation, but has desirable asymptotic properties.

historical background of the program as well as its institutional aspects and settings. Afterwards, we commented on how previous works studied the effects of the program and its main impacts. Then, we approach the *Bolsa Família* policy using the causal forests. We study how the method explores heterogeneities in the effects of the policy.

We found distinct levels of effects of the BFP. In general (on average), the effect on occupation is almost nonexistent although there is a considerable reduction in formal labor market participation.

When one considers the features of the five groups we split our analyzed sample, there is a nuanced picture. The non-white population are the majority in all groups. The majority of the groups live outside metropolitan areas. On one hand, we see a strong impact on female workers. That effect is pronounced in both scenarios: be it in the occupation analysis or in the analysis of formal sector participation. These findings dialogue with literature evidence. On the other hand, when one analyzes occupation, there is an increase in men's participation. Men are, also, an expressive part of the group with least reduction in formal sector participation. An important point to pay attention to and worth of future analysis is to better understand how and why these heterogeneities happen. Also, we highlight that these heterogeneities would not be found if we only used the ATE estimates. Therefore, a causal forest shows to be a valuable tool to study the treatment effects with more details. Some questions for future research on BFP can be proposed. It would be interesting to analyze why there is so much difference in the years of study between quintiles (especially the extremes). Beyond that, the reason why there are so many younger adults reducing their participation in the formal sector when receiving the program benefit could be valuable. Understanding how the quantity of people per household influences the behavior of families when they receive the values of the *Bolsa Família* also seems to be relevant for policy decisions. All these aspects can be related with the behavior of agents. Also, these aspects could be components of what drives their decisions in relation to occupation and the formal sector.

Finally we note that we were not able to evaluate the impacts of the BFP over time. We proceed an analysis using only one point in time. Therefore, we are not capable of assessing the evolution of effects of the program over time.

REFERENCES

ANDINI, Monica et al. Targeting with Machine Learning: An Application to a Tax Rebate Program in Italy. **Journal of Economic Behavior & Organization**, Netherlands, v. 156, p. 86–102, Dec. 2018. Available from: https://linkinghub.elsevier.com/retrieve/pii/S016726811830252X. Visited on: 27 Aug. 2021.

ASCARZA, Eva. Retention Futility: Targeting High-Risk Customers Might Be Ineffective. **Journal of Marketing Research**, United States, v. 55, n. 1, p. 80–98, Feb. 2018. Available from: http://journals.sagepub.com/doi/10.1509/jmr.16.0163. Visited on: 27 Aug. 2021.

ATHEY, Susan. The Impact of Machine Learning on Economics. *In*: AGRAWAL, Ajay; GANS, Joshua; GOLDFARB, Avi (ed.). **The Economics of Artificial Intelligence**: An Agenda. Chicago: University of Chicago Press, 2018. p. 507–547. Available from: http://www.nber.org/chapters/c14009. Visited on: 27 Aug. 2021.

ATHEY, Susan; IMBENS, Guido. Recursive Partitioning for Heterogeneous Causal Effects. **Proceedings of the National Academy of Sciences**, United States v. 113, n. 27, p. 7353–7360, July 2016. Available from: http://www.pnas.org/lookup/doi/10.1073/pnas.1510489113. Visited on: 18 Aug. 2021.

ATHEY, Susan; IMBENS, Guido W. Machine Learning Methods That Economists Should Know About. **Annual Review of Economics**, United States, v. 11, n. 1, p. 685–725, 2 Aug. 2019. Available from: https://www.annualreviews.org/doi/10.1146/annurev-economics-080217-053433. Visited on: 30 Sept. 2022.

ATHEY, Susan; TIBSHIRANI, Julie; WAGER, Stefan. Generalized Random Forests. **The Annals of Statistics**, United States v. 47, n. 2, Apr. 2019.Available from: https://projecteuclid.org/journals/annals-of-statistics/volume-47/issue 2/Generalized-random-forests/10.1214/18-AOS1709.full. Visited on: 20 Sept. 2021.

ATHEY, Susan; WAGER, Stefan. Estimating Treatment Effects with Causal Forests: An Application. **Observational Studies**, Pennsylvania, v. 5, n. 2, p. 37–51, 2019. ISSN 2767-3324. DOI: 10.1353/obs.2019.0001. Available from: https://muse.jhu.edu/article/793356. Visited on: 26 Aug. 2022.

BAIARDI, Anna; NAGHI, Andrea A. **The Value Added of Machine Learning to Causal Inference**: Evidence from Revisited Studies. 2021. Available from: http://arxiv.org/abs/2101.00878. Visited on: 20 Sept. 2021.

BAIRD, Sarah; MCKENZIE, David; ÖZLER, Berk. The Effects of Cash Transfers on Adult Labor Market Outcomes. **IZA Journal of Development and Migration**, Germany, v. 8, n. 1, p. 22, Dec. 2018. ISSN 2520-1786. DOI: 10.1186/s40176-018-0131-9. Available from: https://link.springer.com/10.1186/s40176-018-0131-9. Visited on: 20 Sept. 2021.

BERTRAND, Marianne *et al*. **Contemporaneous and Post-Program Impacts of a Public Works Program**: Evidence from Côte d'Ivoire. 2017. Available from: https://documents1.worldbank.org/curated/en/361281506439891614/pdf/120065-

WP-PUBLIC-141p-Public-Works-CIV-Bertrand-Crepon-Marguerie-Premand-Draft-May17.pdf. Visited on: 17 Sept. 2021.

BREIMAN, Leo. Random Forests. **Machine Learning**, United States, v. 45, n. 1, p. 5–32, 2001. DOI: 10.1023/A:1010933404324. Available from: http://link.springer.com/10.1023/A:1010933404324. Visited on: 17 Sept. 2021.

CAVALCANTI, Daniella Medeiros *et al.* Impacts of Bolsa Família Programme on Income and Working Offer of the Poor Famílies: An Approach Using the Treatment of Quantile Effect. **Economia Aplicada**, São Paulo, v. 20, n. 2, p. 173, 30 June 2016. Available from: https://doi.org/10.11606/1413-8050/ea130092. Visited on: 3 Oct. 2022.

CECHIN, Luis Antonio. O impacto das regras do Programa Bolsa Família sobre a fecundidade das beneficiárias. **Revista Brasileira de Economia**, Rio de Janeiro, v. 69, n. 3, 2015. Available from: https://doi.org/10.5935/0034-7140.20150014. Visited on: 3 Oct. 2022.

CHEN, Shuai *et al*. A General Statistical Framework for Subgroup Identification and Comparative Treatment Scoring. **Biometrics**, United States, v. 73, n. 4, p. 1199–1209, Dec. 2017. Available from: https://onlinelibrary.wiley.com/doi/10.1111/biom.12676. Visited on: 27 Aug. 2021.

CHERNOZHUKOV, Victor; CHETVERIKOV, Denis *et al*. Double/Debiased Machine Learning for Treatment and Structural Parameters. **Econometrics Journal**, United Kingdom, v. 21, n. 1, p. C1–C68, Feb. 2018. Available from: https://doi.org/10.1111/ectj.12097. Visited on: 16 Sept. 2021.

CHERNOZHUKOV, Victor; DEMIRER, Mert *et al*. **Generic Machine Learning Inference on Heterogenous Treatment Effects in Randomized Experiments**. 2020. Available from: http://arxiv.org/abs/1712.04802. Visited on: 15 Sept. 2021.

CHITOLINA, Lia; FOGUEL, Miguel Nathan; MENEZES-FILHO, Naercio Aquino. The Impact of the Expansion of the Bolsa Família Program on the Time Allocation of Youths and Their Parents. **Revista Brasileira de Economia**, Rio de Janeiro, v. 70, n. 2, 2016. Available from: https://doi.org/10.5935/0034-7140.20160009. Visited on: 2 Oct. 2022.

CUNNINGHAM, Scott. **Causal Inference**: The Mixtape. New Haven: Yale University Press, 2021.

DAVIS, Jonathan M.V.; HELLER, Sara B. Using Causal Forests to Predict Treatment Heterogeneity: An Application to Summer Jobs. **American Economic Review**, United States, v. 107, n. 5, p. 546–550, May 2017. Available from: https://pubs.aeaweb.org/doi/10.1257/aer.p20171000. Visited on: 24 Aug. 2021.

BRAUW, Alan de *et al*. Bolsa Família and Household Labor Supply. **Economic Development and Cultural Change**, Chicago, v. 63, n. 3, p. 423–457, Apr. 2015. Available from: https://www.journals.uchicago.edu/doi/10.1086/680092. Visited on: 3 Oct. 2022.

BARBOSA, Ana Luiza Neves de Holanda; CORSEUIL, Carlos Henrique L. **Bolsa Família, escolha ocupacional e informalidade no Brasil**. 2013. Available from: https://ideas.repec.org/p/ipc/opport/226.html.

FIRPO, Sergio et al. Evidence of Eligibility Manipulation for Conditional Cash Transfer Programs. **EconomiA**, Brasília, v. 15, n. 3, p. 243–260, Sept. 2014. Available from: https://doi.org/10.1016/j.econ.2014.09.001. Visited on: 3 July 2022.

GERARD, François; NARITOMI, Joana; SILVA, Joana. **Cash Transfers and Formal Labor Markets**: Evidence from Brazil. 2021. Available from: https://ideas.repec.org/p/cpr/ceprdp/16286.html. Visited on: 3 July 2022.

HIRANO, Keisuke; IMBENS, Guido W.; RIDDER, Geert. Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score. **Econometrica**, United States, v. 71, n. 4, p. 1161–1189, July 2003. Available from: https://doi.org/10.1111/1468-0262.00442. Visited on: 20 Sept. 2021.

HOLLAND, Paul W. Statistics and Causal Inference. **Journal of the American Statistical Association**, United States, v. 81, n. 396, p. 945–960, Dec. 1986. DOI: 10.1080/01621459.1986.10478354. Available from: http://www.tandfonline.com/doi/abs/10.1080/01621459.1986.10478354. Visited on: 20 Sept. 2021.

HORVITZ, D. G.; THOMPSON, D. J. A Generalization of Sampling Without Replacement from a Finite Universe. **Journal of the American Statistical Association**, United States, v. 47, n. 260, p. 663–685, Dec. 1952. DOI: 10.1080/01621459.1952.10483446. Available from: http://www.tandfonline.com/doi/abs/10.1080/01621459.1952.10483446. Visited on: 20 Sept. 2021.

IMBENS, Guido; ATHEY, Susan. Breiman's Two Cultures: A Perspective from Econometrics. **Observational Studies**, Pennsylvania, v. 7, n. 1, p. 127–133, 2021. DOI: 10.1353/obs.2021.0028. Available from: https://muse.jhu.edu/article/799753. Visited on: 5 July 2022.

IMBENS, Guido W.; RUBIN, Donald B. **Causal Inference for Statistics, Social, and Biomedical Sciences**: An Introduction. Cambridge: Cambridge University Press, 2015. Available from: https://doi.org/10.1017/CBO9781139025751. Visited on: 20 Sept. 2021.

JAMES, Gareth *et al*. An Introduction to Statistical Learning: With Applications in R. 2nd ed. New York: Springer, 2021. (Springer Texts in Statistics).

KASY, Maximilian; SAUTMANN, Anja. Adaptive Treatment Assignment in Experiments for Policy Choice. **Econometrica**, United States, v. 89, n. 1, p. 113–132, 2021. Available from: https://doi.org/10.3982/ECTA17527. Visited on: 24 Apr. 2023.

KITAGAWA, Toru; TETENOV, Aleksey. Who Should Be Treated? Empirical Welfare Maximization Methods for Treatment Choice. **Econometrica**, United States, v. 86, n. 2, p. 591–616, 2018. Available from: https://doi.org/10.3982/ECTA13288. Visited on: 24 Apr. 2023.

KNAUS, Michael C.; LECHNER, Michael; STRITTMATTER, Anthony. Heterogeneous Employment Effects of Job Search Programmes: A Machine Learning Approach. **Journal of Human Resources**, United States, p. 0718–9615R1, Mar. 2020. Available from: https://doi.org/10.3368/jhr.57.2.0718-9615R1. Visited on: 27 Aug. 2021.

KNAUS, Michael C.; LECHNER, Michael; STRITTMATTER, Anthony. Machine Learning Estimation of Heterogeneous Causal Effects: Empirical Monte Carlo Evidence. **The Econometrics Journal**, United Kingdom, v. 24, n. 1, p. 134–161, Mar. 2021. Available from: https://doi.org/10.1093/ectj/utaa014. Visited on: 27 Aug. 2021.

LAB, Golub Capital Social Impact. ML-based Causal Inference Tutorial. 2021. Available from: https://bookdown.org/content/3e3ee3cb-b53e-4956-b8d3-a3243e663162/vdTZdBd52/#.

MANSKI, Charles F. Identification Problems and Decisions under Ambiguity: Empirical Analysis of Treatment Response and Normative Analysis of Treatment Choice. **Journal of Econometrics**, Netherlands, v. 95, n. 2, p. 415–442, Apr. 2000. Available from: https://doi.org/10.1016/S0304-4076(99)00045-7. Visited on: 24 Apr. 2023.

MANSKI, Charles F. Statistical Treatment Rules for Heterogeneous Populations. **Econometrica**, United States, v. 72, n. 4, p. 1221–1246, July 2004. Available from: https://doi.org/10.1111/j.1468-0262.2004.00530.x. Visited on: 24 Apr. 2023.

MCCRARY, Justin. Manipulation of the running variable in the regression discontinuity design: a density test. **Journal of Econometrics**, Netherlands, v. 142, n. 2, p. 698–714, Feb. 2008. Available from: https://doi.org/10.1016/j.jeconom.2007.05.005. Visited on: 2 Oct. 2022.

MULLAINATHAN, Sendhil; SPIESS, Jann. Machine Learning: An Applied Econometric Approach. **Journal of Economic Perspectives**, United States, v. 31, n. 2, p. 87–106, May 2017. Available from: https://pubs.aeaweb.org/doi/10.1257/jep.31.2.87. Visited on: 19 Sept. 2021.

NASCIMENTO, Adriana Rosa do; KASSOUF, Ana Lúcia. Impacto do Programa Bolsa Família sobre as decisões de trabalho das crianças: uma análise utilizando os microdados da PNAD. **Análise Econômica**, Porto Alegre, v. 34, n. 66, Sept. 2016. Available from: https://doi.org/10.22456/2176-5456.54855. Visited on: 3 Oct. 2022.

NIE, X; WAGER, S. Quasi-Oracle Estimation of Heterogeneous Treatment Effects. **Biometrika**, United Kingdom, v. 108, n. 2, p. 299–319, May 2021. Available from: https://doi.org/10.1093/biomet/asaa076. Visited on: 16 Sept. 2021.

OLIVEIRA, Luis Felipe Batista de; SOARES, Sergei S. D. **O que se sabe sobre os efeitos das transferências de renda sobre a oferta de trabalho**. Brasília: IPEA, 2012. Available from: http://repositorio.ipea.gov.br/bitstream/11058/1161/1/TD_1738.pdf.

PONCZEK, Vladimir Pinheiro; MATTOS, Enlinson. **O efeito do estigma sobre os beneficiários de Programas de Transferência No Brasil**. 2010. Available from: https://ideas.repec.org/p/fgv/eesptd/226.html.

RIBEIRO, Felipe Garcia; SHIKIDA, Claudio; HILLBRECHT, Ronald Otto. Bolsa Família: um survey sobre os efeitos do programa de transferência de renda condicionada do Brasil. **Estudos Econômicos**, São Paulo, v. 47, n. 4, p. 805–862, Dec. 2017. Available from: https://doi.org/10.1590/0101-416147468fcr. Visited on: 27 Aug. 2021.

ROBINS, James M.; ROTNITZKY, Andrea. Semiparametric Efficiency in Multivariate Regression Models with Missing Data. **Journal of the American Statistical Association**, United States, v. 90, n. 429, p. 122–129, Mar. 1995. Available from: http://www.tandfonline.com/doi/abs/10.1080/01621459.1995.10476494. Visited on: 15 Sept. 2021.

ROBINSON, P. M. Root-N-Consistent Semiparametric Regression. **Econometrica**, United States, v. 56, n. 4, p. 931, July 1988. Available from: https://www.jstor.org/stable/1912697. Visited on: 16 Sept. 2021.

RUBIN, Donald B. Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies. **Journal of Educational Psychology**, United States, v. 66, n. 5, p. 688–701, 1974. Available from: https://doi.org/10.1037/h0037350 https://doi.org/10.1037/h0037350. Visited on: 20 Sept. 2021.

SEMENOVA, Vira; CHERNOZHUKOV, Victor. **Debiased Machine Learning of Conditional Average Treatment Effects and Other Causal Functions**. 2020. Available from: http://arxiv.org/abs/1702.06240. Visited on: 15 Sept. 2021.

SIMÕES, Patrícia; SOARES, Ricardo Brito. Efeitos do Programa Bolsa Família Na Fecundidade Das Beneficiárias. **Revista Brasileira de Economia**, Rio de Janeiro, v. 66, n. 4, p. 445–468, Dec. 2012. Available from: https://doi.org/10.1590/S0034-71402012000400004. Visited on: 3 Oct. 2022.

SOUZA, Wallace Patrick Santos de Farias *et al.* Trabalho Infantil e Programasd e Transferência de Renda: Uma Análise Do Impacto Do Programa Bolsa Família Nas Zonas Urbana e Rural Do Brasil. **Pesquisa e Planejamento Econômico**, Rio de Janeiro, v. 49, p. 131–164, n. 2 Aug. 2019. Available from: https://ppe.ipea.gov.br/index.php/ppe/article/view/2031.

SPLAWA-NEYMAN, Jerzy; DABROWSKA, D. M.; SPEED, T. P. On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9. **Statistical Science**, v. 5, n. 4, Nov. 1990. Available from: https://doi.org/10.1214/ss/1177012031. Visited on: 8 Sept. 2021.

STRITTMATTER, Anthony. What Is the Value Added by Using Causal Machine Learning Methods in a Welfare Experiment Evaluation? 2019. Available from: http://arxiv.org/abs/1812.06533. Visited on: 20 Sept. 2021.

TAVARES, Priscilla Albuquerque. Efeito Do Programa Bolsa Família Sobre a Oferta de Trabalho Das Mães. **Economia e Sociedade**, v. 19, n. 3, p. 613–635, Dec. 2010. Available from: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0104-

06182010000300008&lng=pt&tlng=pt. Visited on: 1 Oct. 2022.

TIAN, Lu *et al.* A Simple Method for Estimating Interactions Between a Treatment and a Large Number of Covariates. **Journal of the American Statistical Association**, United States, v. 109, n. 508, p. 1517–1532, Oct. 2014. Available from: http://www.tandfonline.com/doi/abs/10.1080/01621459.2014.951443. Visited on: 27 Aug. 2021.

TIBSHIRANI, Julie *et al*. Grf: **Generalized Random Forests**. manual. 2022. Available from: https://CRAN.R-project.org/package=grf.

VARIAN, Hal R. Big Data: New Tricks for Econometrics. **Journal of Economic Perspectives**, United States, v. 28, n. 2, p. 3–28, May 2014. Available from: https://pubs.aeaweb.org/doi/10.1257/jep.28.2.3. Visited on: 19 Sept. 2021.

WAGER, Stefan; ATHEY, Susan. Estimation and Inference of Heterogeneous Treatment Effects Using Random Forests. **Journal of the American Statistical Association, United States**, United States, v. 113, n. 523, p. 1228–1242, July 2018. Available from: https://www.tandfonline.com/doi/full/10.1080/01621459.2017.1319839. Visited on: 17 Sept. 2021.

APPENDIX A - *BOLSA FAMÍLIA* PROGRAM

1. Institutional background

In what follows, we review institutional aspects of Programa Bolsa Família building on Simões and Soares (2012); Firpo et al. (2014); Cechin (2015); Nascimento and Kassouf (2016); Cavalcanti et al. (2016); Chitolina, Foguel, and Menezes-Filho (2016); Ribeiro, Shikida, and Hillbrecht (2017) works, besides Brazilian legislation.

The BFP was a conditional cash transfer policy. It was created through a provisory policy in 2003 and turned into law in 2004. It integrated the Plano Brasil sem Miséria ("Brazil without misery plan") and replaced some assistencial programs that existed at that time: Food Card (Bolsa Alimentação, Programa Nacional de Acesso à Alimentação – PNAA - "Cartão Alimentação"), Food Allowance (Programa Nacional de Renda Mínima vinculado à saúde – "Bolsa Alimentação") and Gas Aid[2] ("Vale-Gás"). The main database used to control beneficiaries is the Cadastro Único para Programas Sociais (CADÚNICO) (Unified Registry for Social Programs). Monitoring compliance of conditionalities was a task delegated to Ministério do Desenvolvimento Social (MDS), Ministério da Educação (MEC) and Ministério da Saúde (MS). Also, local governments were responsible for the monitorization the program.

The program had some goals:

1. To promote access to a network of public services, specially health, education and social assistance;
2. To fight hunger and to promote nutritional safety;
3. To stimulate sustained emancipation of families that live in poverty or extreme poverty;
4. To fight poverty;
5. To promote synergy of social practices of the public sector.

There are two kinds of targeting strategies to CCT. One is called "means tested" and the other is "proxy means tested". The eligibility according to proxy means it is characterized by the household score computed given household characteristics[3]. In the case of "means tested" we have eligibility defined according to some threshold of income. BFP adopts the mean proxy targeting approach.

There were two targets for the policy: households in poverty or households in extreme poverty. The first group was eligible if its composition had nursing mothers, pregnant women or adolescents up to 17 years old. To extreme poverty, there was no household composition criteria. Besides basic benefit, families could receive an additional (variable) benefit if: (1) it had children with age equal or less than 15 years old or with pregnant or nursing mothers (at maximum of 3 additional benefits), (2) young people between 16 and 17 years old (up to 3 additional benefits).

[2] Some time after it was revoked.

[3] "Progresa" program in Mexico.

The benefits are called, respectively: Beneficio Variável (Variable Benefit) and Bolsa Variável Jovem: BVJ (Youth Variable Benefit).

The benefit was granted if families observed some conditionalities related to health and education. Families should meet some children and women (between 14 and 44 years old) vaccination requirements. Pregnant and nursing mothers should be accompanied by prenatal doctors. Also, children younger than 7 years old should have their growth monitored. Related to education, children in ages between 6 and 17 should be enrolled in school with a frequency of at least 85% (on monthly basis). If their age was between 16 and 17 years, then the frequency should be 75%. And, lastly, children should participate in social programs focused on the reduction of child labor.

The reason for these conditionalities was to enable some changes that allow the break of the poverty trap. They try to bring incentives for parents to reduce the number of children and invest in the existing ones (CECHIN, 2015). Investment in education helps children to have the best future opportunities (NASCIMENTO; KASSOUF, 2016). The BFP's conditionalities enable it to accumulate human capital (SOUZA et al., August-2019).

The program was replaced by Auxílio Brasil (Brazil Aid) in 2021. This last one has close conditionalities with BFP. The benefit was given to approximately 20 million households nationwide[4].

[4] In August of 2022.

2. Recent Works

In this section we review works about the BFP and its impact on the labor force. The main focus will be in recent research, i.e., from 2010 onwards.

For the review of works before the periods investigated here see Oliveira and Soares (2012). That work concludes that, except for some demographics, overall, there is no “laziness effect” (income effect bigger than substitution effect). Those workers in the informal sector are more vulnerable and more sensitive to the transfers: it brings some protection for volatile times. Permeating many studies are the differences in responses between women and men. Women are more elastic to the benefit, reducing their labor supply given the transfer (possibly because of the need to take care of their children). Also, poorer women reduce their participation in the labor market given their BFP involvement. Likewise, women have a slightly smaller labor supply in total hours worked. Women, more intensively than men, change their work outside home for those related to domestic activities.

Tavares (2010) uses Heckman correction and propensity score matching to examine how mothers have their labor supply affected by BFP. The supply refers to hours offered and participation. Three types of mothers were used as control group: (i) mothers that were included in the program, although they did not receive the cash transfers, (ii) non-beneficiary mothers, but eligible to the program (from poor or extremely-poor families) and (iii) mothers who have low incomes close to the eligibility criteria (household per capita below the cutoff), but do not receive the benefit given they do not fully meet the criteria. It finds that participation in the program increases the mother’s labor supply. But, it also finds that as the transfers increase, a reduction in supply is observed. The results are found in all three control groups. Author’s suggested hypothesis of why those effects could happen together are: substitution effects, more available time for work to the mothers and "stigma effect", given the participation in the program. Women considered were householders or married mothers and the age gap between them and their children range from 12 and 50. This work used PNAD 2004.

She suggests that there is no "laziness effect" (p. 269): the income effect is dominated by the substitution effect. Besides, the possible explanation for increase in labor supply resides in the fact that children, given program conditionalities, stay more time in school and this alleviates the time constraint of mothers, enabling them to have more time to work. For its turn, the "stigma effect” refers to the discrimination suffered by women because they were the beneficiary of BFP. Tavares (2010) argues that this explanation dialogues with findings in Ponczek and Mattos (2010): those who received income delivered by the program felt embarrassment and this has a positive impact on the search for jobs, as well as it contributes for a reduction of unemployment. Note that the embarrassment could be, also, self perception.

de Holanda Barbosa and Corseuil (2013) analyzes the composition of labor sectors (formal and informal) and its relationship with BFP. The effect in hours worked in the informal sector also was studied. Using RDD (fuzzy), they found that the policy does not affect the occupational choice of beneficiaries. They studied the effect on the probability of the family head

occupation to be in the informal sector, as well as the secondary occupation of the family head or other main occupation of other members. There was no statistically significant effect. When analyzing the intensive margin, to see if there was some effect in the proportion of informal sector hours worked in relation with the total amount worked in families, no significant statistic different from zero are found. They used a discontinuity in eligibility rules concerning children's ages. The database used was PNAD of 2006.

The work asserts that the use of eligibility based on household income per capita is not adequate to study the effects of BFP on labor sector choice of households. Indeed, the reason is not appropriate because the adult working force of the household could manage their choice to continue participating in the program. So, a better variable to pay attention to is the age of the youngest son. This is not a feature that families are not able to manipulate. de Holanda Barbosa and Corseuil (2013) explores a discontinuity in eligibility rule of BFP. Families where the younger son completes 16 years old are no longer eligible to benefit, but the exclusion does not happen immediately, only during the next year. Taking advantage of this, the work compares households with the youngest child about to turn 16 years old on December 31 (2005) with households whose son already turned 16 until this same date. Twelve scenarios were analyzed: for each of the 3 samples, four bandwidths were studied. The inexistence of effects was common to all of them.

In Firpo et al. (2014) there is a study about the possibility of manipulation of eligibility criteria and about the BFP on labor supply. When conducting such an investigation they used the McCrary test (MCCRARY, 2008) for the first question, and used a fuzzy regression discontinuity design to study the last one. They found evidence of manipulation. There is discontinuity around the cutoff (income 120 reais), and this result is robust in many options of cutoff. This evidence is stronger among women. In relation to the second question, they focused on three dimensions: (i) if a person participated in the labor force, (ii) if the person was employed when the survey took place and (iii) the quantity of weekly hours worked. They found "suggestive evidence" (p.251, 256) that people who receive the benefit right below the cutoff tend to have lower labor force participation, have a major probability to be unemployed and to work less hours per week. Also, those associated with manipulation are the less participative ones in the labor force. Authors considered PNAD 2006.

The work investigates discontinuities in other possible values of cutoff: R$ 130, R$ 140 and R$ 150. Only in the last one there is no evidence of discontinuity in each studied group. The evidence was found in both household per capita income net and its logarithm. Also for robustness, they look for the same income variable in 1998 and 1999: the density is smooth without any jump around the cutoff.

Firpo et al. (2014) show that households below the cutoff are associated with a bigger probability of participating in BFP: around 11% for families formed by couples (family group 1). For families where mothers are single or divorced (family group 2), the probability found was around 10. All those families have at least one child younger than fifteen years old.

The researchers work with three stratas when exploring the "suggestive evidence" of the impact of BFP on labor supply decisions: married men (of group 1), married mothers (of group 1) and single or divorced mothers (from group 2). In all stratas there is a negative effect of BFP: in labor force participation, weekly hours worked and if the person was employed when the survey took place. Women (married or not) always have major impacts. The effect in this group is always bigger than those in other groups. To give an example: married men reduce their weekly hours by approximately 2.5, married women in 4.6 and single or divorced mothers by around 5 hours. It is worth to note that, given the verification of discontinuity, the work acknowledges that their estimation is not a classic Wald estimator (p. 251). Concluding the paper, another McCrary is made and finds some evidence that individuals less linked to the labor market also could manage to be eligible to BFP.

At the end, Firpo et al. (2014) makes two warnings. First, decrease in labor supply may occur by the income effect *per se* and not individuals manipulating their income. The authors suggest that both motives could coexist being hard to clearly distinguish their influence in labor supply. Second, it could be the case that people misreport their labor information (consciously, to be eligible to BFP), so there are no *real* effects in labor supply at all.

de Brauw et al. (2015) analyze some questions about the BFP effects exploring Avaliação de Impacto do Programa Bolsa Família (AIBF-1). They investigate how the BFP affected labor supply and allocation of labor time. They study these issues in formal and informal sectors and between locations (urban and rural areas). They use propensity score and (longitudinal) panel data. The period covered by the research was 2005 and 2009.

The analysis considering rural and urban areas, each with its specificities, is important because each environment is different, impacting how household labor supply responses. Besides that, differences between wages in formal and informal sectors are distinct in each area.

In the overall (individual[5]) sample of people in rural areas the probability of working reduces in 10.5%. Rural women reduce in 13.1%, no effect is found in rural men. In urban areas, the effect for the overall sample is not found and the same happens with urban women, although an increase of almost 9% is found for urban men.

When analyzing the overall data at household level, it is not found any changes in aggregate labor supply, but there is change in its composition. There are less weekly hours offered in the formal sector while there is an increase in the informal one (approximately 8 hours on both). These movements occur for both genders (they share approximately half of total effect). The exchange of formal sector for informal sector is leaded captained by urban areas.

Note that in both rural (individually) and urban (at household level) areas, women labor reduces. But only in rural areas men offset this decrease: they start working (a little) more. de Brauw et al. (2015) affirm that a possible reason for the effects in the formal/informal sector is the process to monitor who receives the benefits. The official available data primarily comprises the formal labor sector, thus creating incentives for reallocation of labor supply in the informal

[5] The work focuses on those who are between 18-55. They argue that it is representative of working class age.

sector.

Finally, the de Brauw et al. (2015) comments on about the implications in welfare of its findings. First of all, there are lower wages. Given the exchange, urban workers have worse wages than before. In addition, when the worker stays too long in the informal sector, their chance to come back to the formal sector reduces considerably. When it comes to the impact on women labor supply in rural areas, this could bring to her more vulnerability and loss of her autonomy. Beyond that, there is a fiscal cost: informal workers do not contribute to the social security system.

When studying intention to treat (ITT) given the expansion of the policy, Chitolina, Foguel, and Menezes-Filho (2016) used differences in differences estimator. The investigation was about the influence on time allocation of poor young adults (between 16 and 17 years old) and their parents' labor supply when these households started to be covered by the BFP (creation of Benefício Variável Jovem in 2007). Although impacts in education outcomes were studied, here we only focus on labor outcomes: from youths and other members of the family. The results do not show influence in labor supply of parents on the overall sample. Only mothers of rural areas had a different result: among them there was an increase of supply. Possible reason: more free time or effort to compensate for the reduction of work. This work studied both participation in the labor market and the number of hours worked. The work makes use of PNAD, years of 2006 and 2009.

Chitolina, Foguel, and Menezes-Filho (2016) studied only ITT because it is not possible to check if all the potential beneficiaries effectively get their transfers (after the expansion). The requirement was to follow some school attendance rules for youths with age between 16 and 17.

They present the effect in labor outcomes for (target) youth of the expansion of the program: there is an increase of approximately 4.4% in the probability of work while attending school. There is no increase in probability of only studying and not working at the same time. When looking at results by location, urban areas do not show any effect, but in rural areas an increment of 9.7% appears[6]. Beyond that, when analyzing effects on parents' labor supply, it is not found evidence that the program drives some "laziness effect" on members of the family.

Gerard, Naritomi, and Silva (2021) examines the effect of BPF in local labor markets. They consider a change in the methodology of federal administration on how to distribute funding of the program for municipalities. The reform of interest took place in 2009. Previously, the federal government, responsible for the BFP budget, distributed funds (quotas of national slots) for municipalities according to poverty estimates by Instituto Brasileiro de Geografia e Estatística (IBGE). The institute calculated the quantity of poor families based on a poverty threshold: half of the minimum wage per capita. After, using data from available Brazilian census IBGE calculates the total of households below the threshold. The funds are allocated following this poverty localization and distribution across municipalities. After the reform, IBGE changed the way they calculate the total of poor families: they started to use the poverty map

[6] They use a multinomial logit model when looking at these aspects.

methodology from the World Bank. After the reform, there was an increase of 17% in the number of households covered by the program.

A positive effect was found in local labor markets. The identification strategy of Gerard, Naritomi, and Silva (2021) uses the extent of increment of funds for BFP given more poor households identified in the municipalities according to 2009 methodology. Their findings show a positive increase in the formal private sector: employment (2%) and payroll (about 1.7%). When comparing those municipalities that had more increments with those that had less, the effects are more pronounced: the bigger the increase is in allocated funding (given more poor families in each municipality), a stronger effect is found. There is no evidence of effect in the public sector. They use longitudinal administrative data[7] and employ a difference-in-difference design. The period used in the estimation ranges from the beginning of 2007 until 2012.

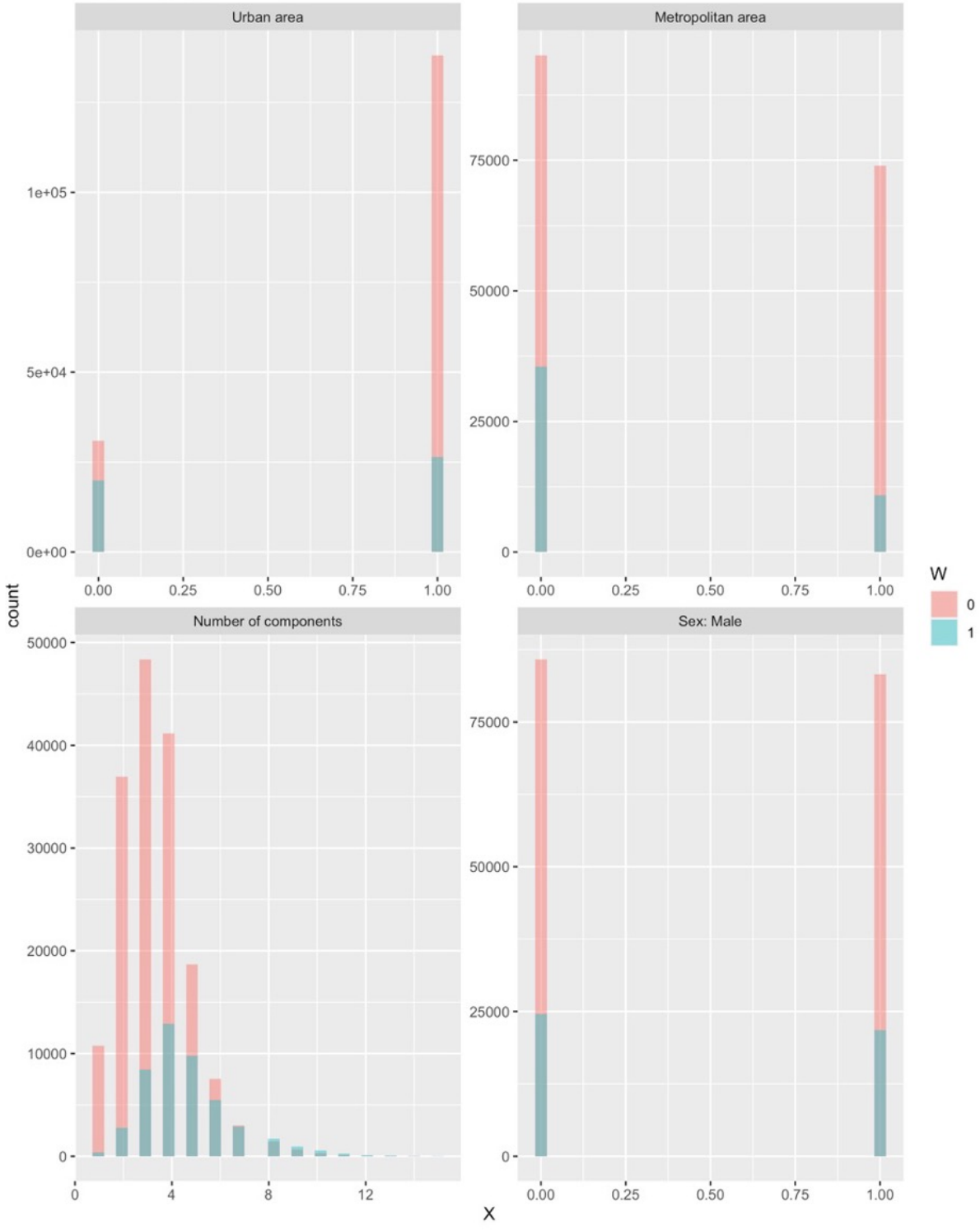
To complement what we describe here, we recommend readers to read Ribeiro, Shikida, and Hillbrecht (2017) for a survey about BFP impact in many more dimensions, either in those that the program has a goal to direct effect, or those indirectly affected.

[7] Relação Anual de Informações Sociais (RAIS), Bolsa Família Payment sheet and Cadastro Único (CadÚnico).

APPENDIX B - FIGURES

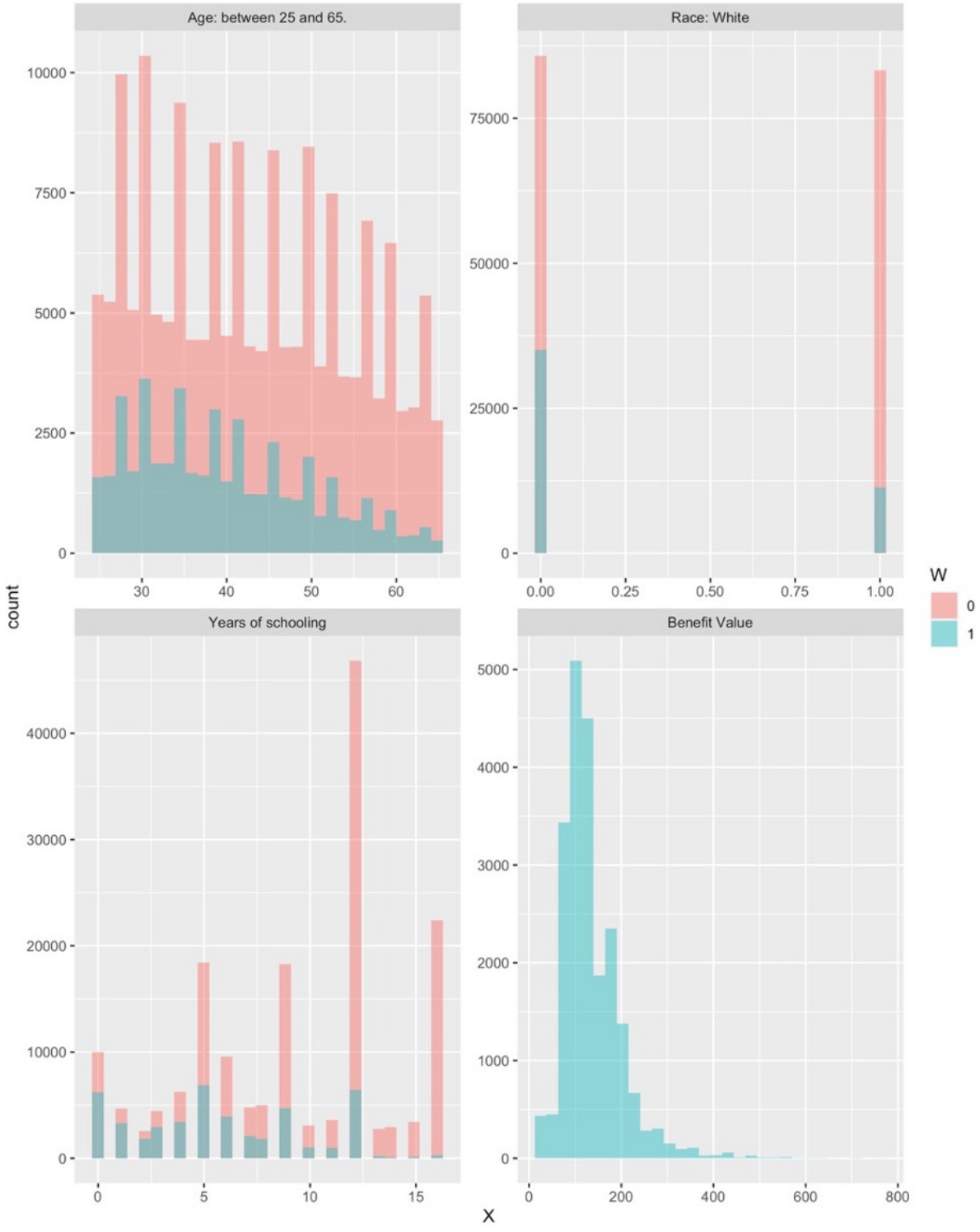
1. Unbalanced groups: Histograms

Figure 1 – Covariate histograms: urban and metropolitan area, number of components and sex



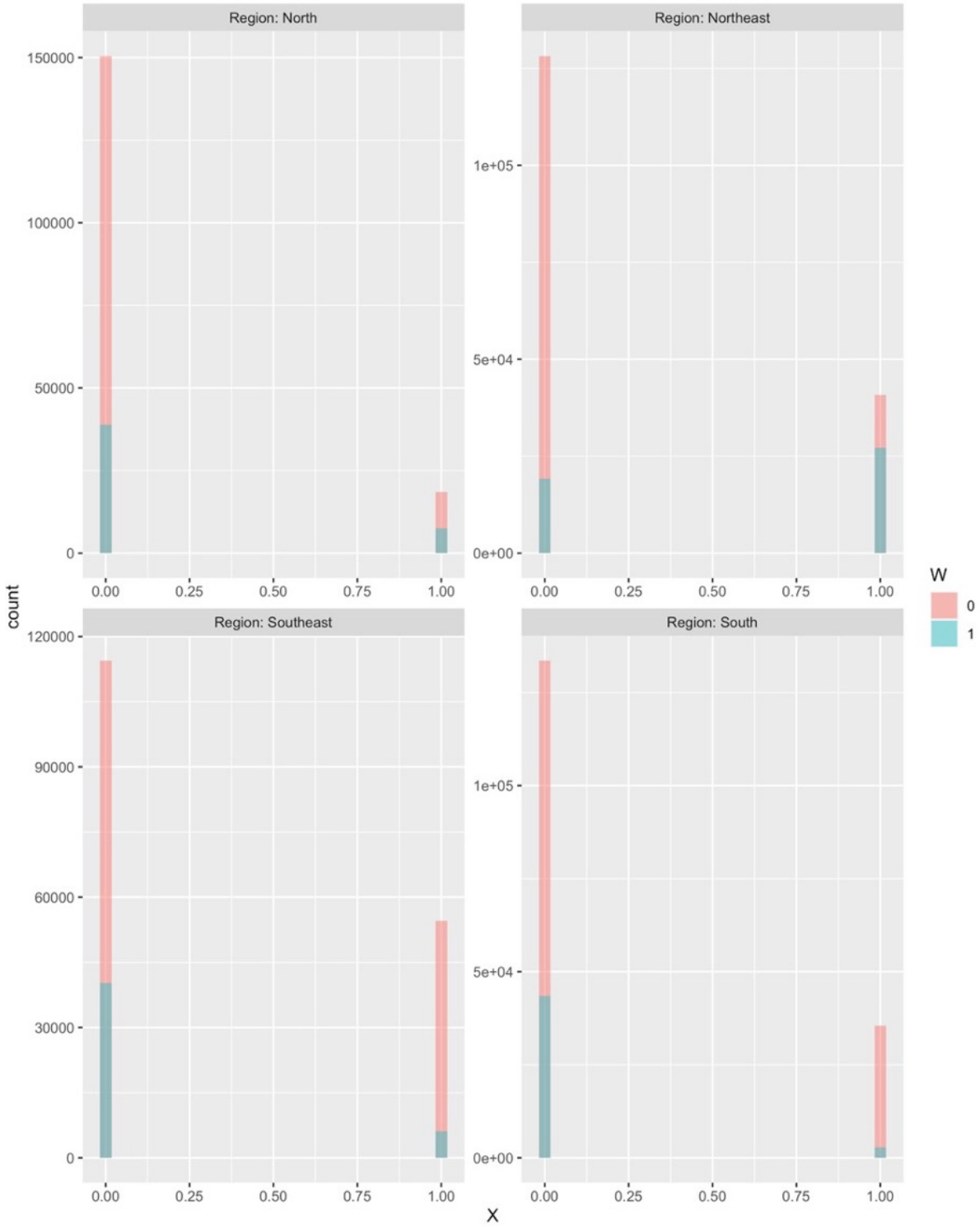
Source: authors' elaboration.

Figure 2 – Covariate histograms: age, race, years of schooling and benefit value



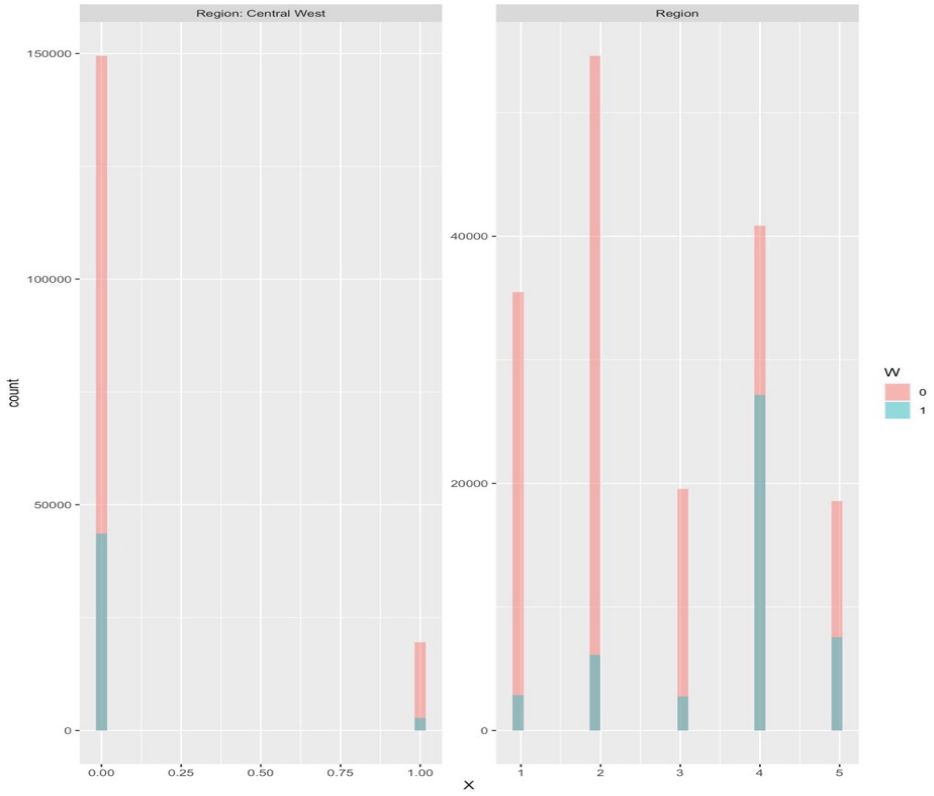
Source: authors' elaboration.

Figure 3 – Covariate histograms of regions



Source: authors' elaboration.

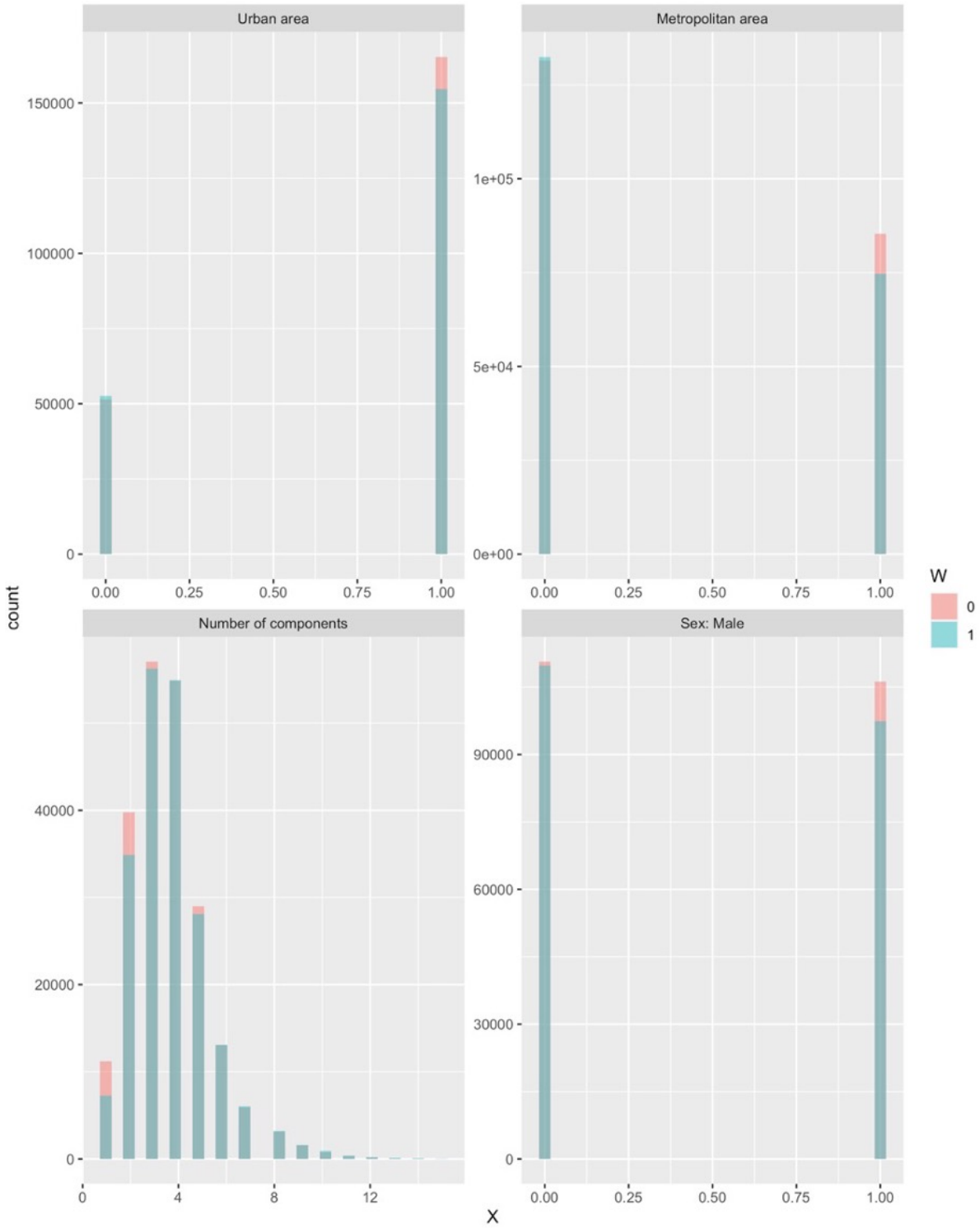
Figure 4 – Covariate histograms: region north and all regions together



Source: authors' elaboration.

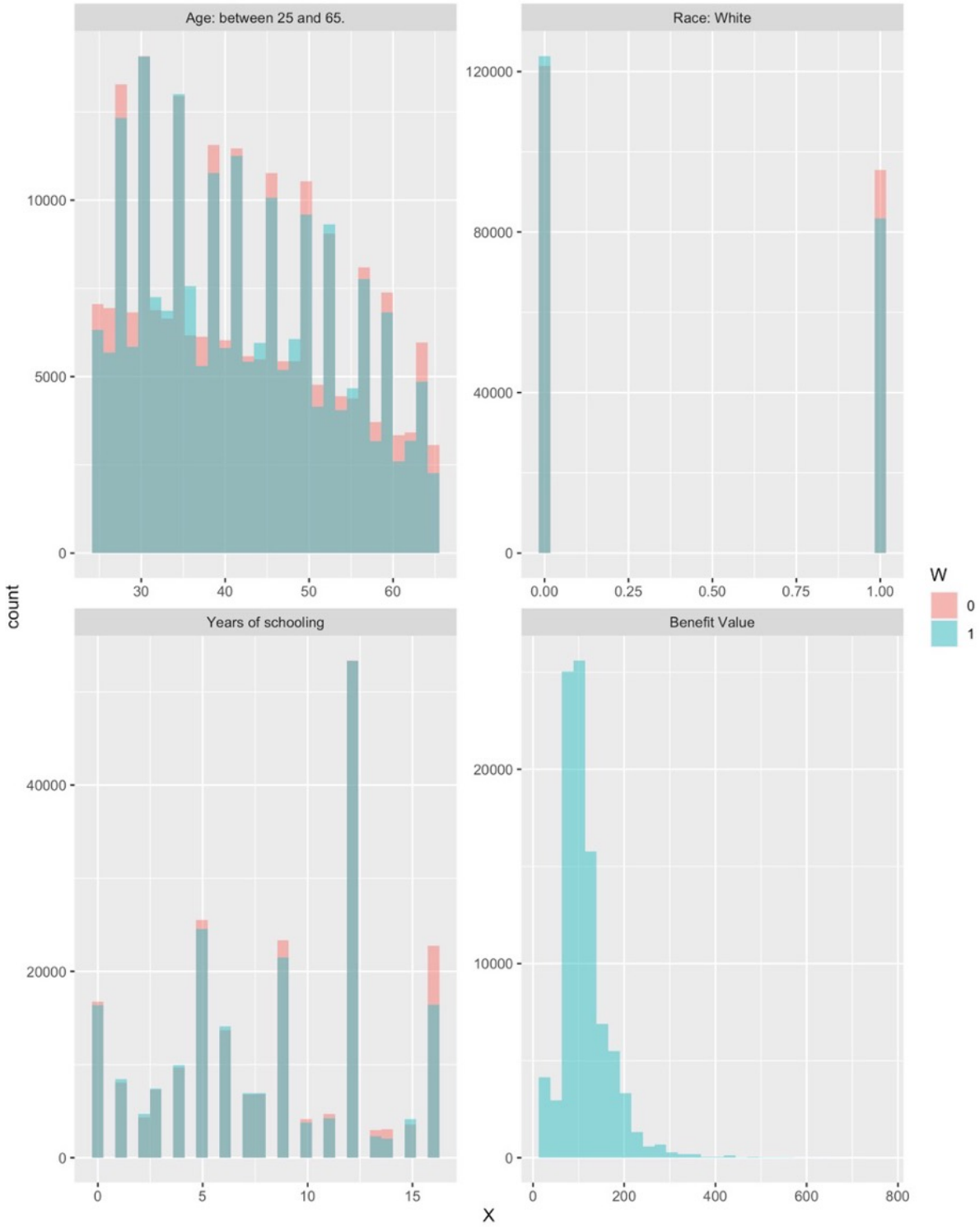
2. Balanced groups: Histograms

Figure 5 – Covariate histograms (adjusted): urban and metropolitan area, number of components and sex



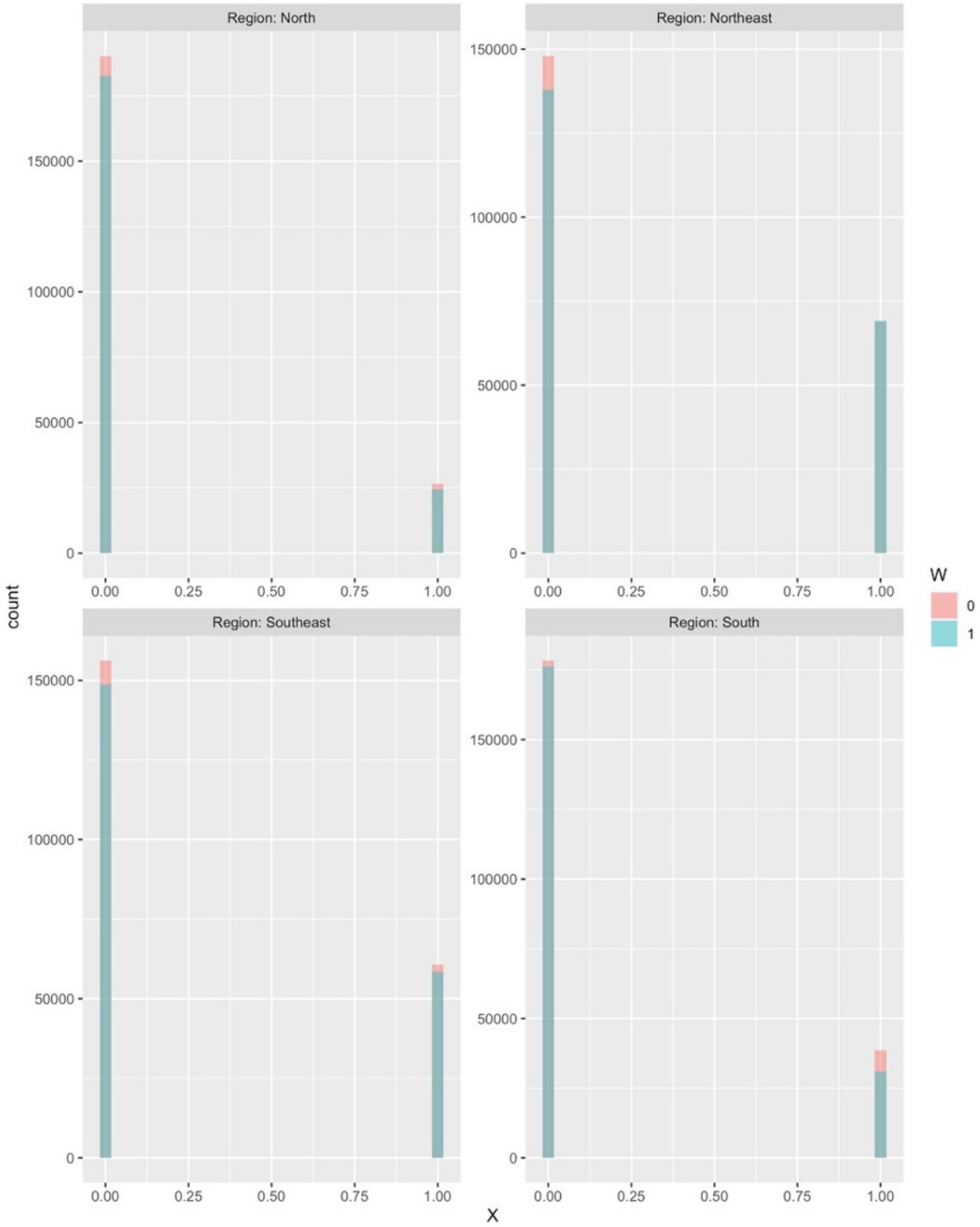
Source: authors' elaboration.

Figure 6 – Covariate histograms (adjusted): age, race, years of schooling and benefit value



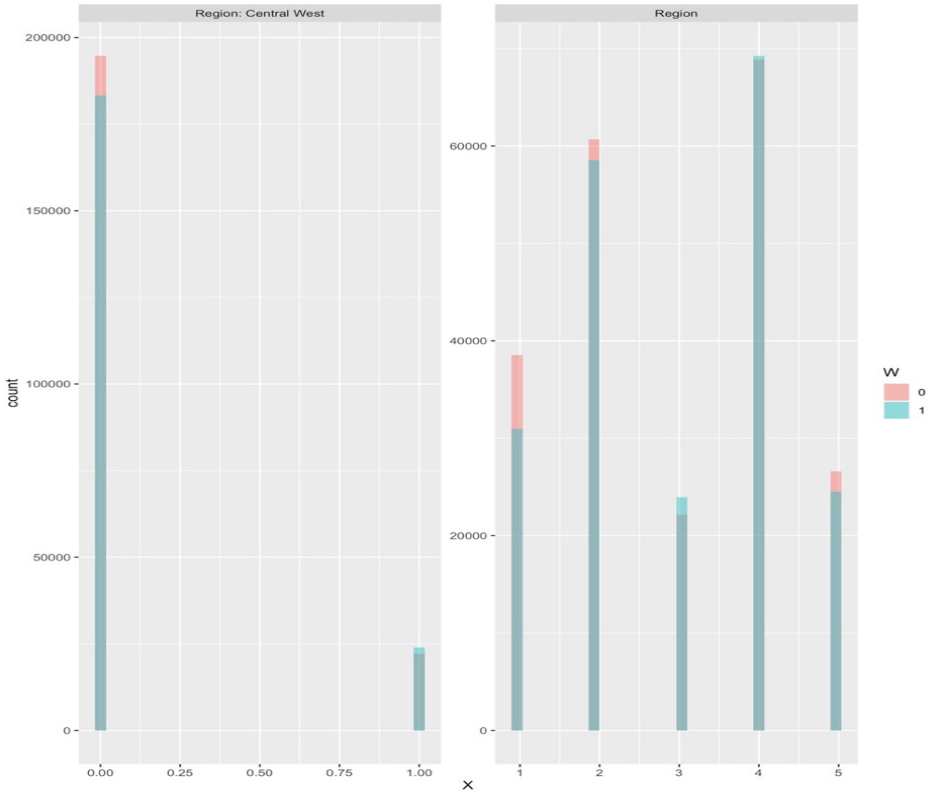
Source: authors' elaboration.

Figure 7 – Covariate histograms (adjusted) of regions



Source: authors' elaboration.

Figure 8 – Covariate histograms (adjusted): region north and all regions together



Source: authors' elaboration.

APPENDIX C - TABLES

1. Occupation: differences between CATE quintiles

Table 1 – Occupation - Differences with Q1

	Estimate	Std. Error	Orig. p-value	Adj. p-value
Q2 - Q1	0.032	0.012	0.007	0.007
Q3 - Q1	0.056	0.012	0.00000	0
Q4 - Q1	0.080	0.012	0	0
Q5 - Q1	0.118	0.012	0	0

Source: Authors' elaboration.

Table 2 – Occupation - Differences with Q2

	Estimate	Std. Error	Orig. p-value	Adj. p-value
Q3 - Q2	0.024	0.012	0.046	0.045
Q4 - Q2	0.048	0.012	0.0001	0.0002
Q5 - Q2	0.086	0.012	0	0
Q1 - Q2	−0.032	0.012	0.007	0.014

Source: Authors' elaboration.

Table 3 – Occupation - Differences with Q3

	Estimate	Std. Error	Orig. p-value	Adj. p-value
Q4 - Q3	0.024	0.012	0.040	0.073
Q5 - Q3	0.062	0.012	0.00000	0
Q1 - Q3	−0.056	0.012	0.00000	0
Q2 - Q3	−0.024	0.012	0.046	0.073

Source: Authors' elaboration.

Table 4 – Occupation - Differences with Q4

	Estimate	Std. Error	Orig. p-value	Adj. p-value
Q5 - Q4	0.037	0.012	0.002	0.003
Q1 - Q4	−0.080	0.012	0	0
Q2 - Q4	−0.048	0.012	0.0001	0.0005
Q3 - Q4	−0.024	0.012	0.040	0.036

Source: Authors' elaboration.

Table 5 – Occupation - Differences with Q5

	Estimate	Std. Error	Orig. p-value	Adj. p-value
Q1 - Q5	−0.118	0.012	0	0
Q2 - Q5	−0.086	0.012	0	0
Q3 - Q5	−0.062	0.012	0	0
Q4 - Q5	−0.037	0.012	0.002	0.002

Source: Authors' elaboration.

2. Formal Sector: differences between CATE quintiles

Table 6 – Formal Sector - Differences with Q1

	Estimate	Std. Error	Orig. p-value	Adj. p-value
Q2 - Q1	0.048	0.011	0.00001	0
Q3 - Q1	0.071	0.011	0	0
Q4 - Q1	0.081	0.011	0	0
Q5 - Q1	0.102	0.011	0	0

Source: Authors' elaboration.

Table 7 – Formal Sector - Differences with Q2

	Estimate	Std. Error	Orig. p-value	Adj. p-value
Q3 - Q2	0.023	0.011	0.029	0.030
Q4 - Q2	0.033	0.011	0.002	0.003
Q5 - Q2	0.055	0.011	0.00000	0
Q1 - Q2	−0.048	0.011	0.00001	0.0001

Source: Authors' elaboration.

Table 8 – Formal Sector - Differences with Q3

	Estimate	Std. Error	Orig. p-value	Adj. p-value
Q4 - Q3	0.010	0.011	0.338	0.338
Q5 - Q3	0.031	0.011	0.003	0.009
Q1 - Q3	−0.071	0.011	0	0
Q2 - Q3	−0.023	0.011	0.029	0.054

Source: Authors' elaboration.

Table 9 – Formal Sector - Differences with Q4

	Estimate	Std. Error	Orig. p-value	Adj. p-value
Q5 - Q4	0.021	0.011	0.043	0.082
Q1 - Q4	−0.081	0.011	0	0
Q2 - Q4	−0.033	0.011	0.002	0.004
Q3 - Q4	−0.010	0.011	0.338	0.341

Source: Authors' elaboration.

Table 10 – Formal Sector - Differences with Q5

	Estimate	Std. Error	Orig. p-value	Adj. p-value
Q1 - Q5	−0.102	0.011	0	0
Q2 - Q5	−0.055	0.011	0.00000	0
Q3 - Q5	−0.031	0.011	0.003	0.006
Q4 - Q5	−0.021	0.011	0.043	0.045

Source: Authors' elaboration.