Counterfactual Analysis Based on Grouped Data: Application to Poverty and Material Deprivation.

Minghai Mao[*]
Liaoning University

Antonio Raiola[1]
Universidad Carlos III de Madrid

June 7, 2024

Abstract

We propose inferences on the counterfactual mean: the mean of some population outcome, pretending that the related characteristics are distributed according to the distribution of another population. The purpose is to compare the mean of an outcome in two populations, holding constant other related factors that may distort the comparison. Once the data is partitioned into a number of classes, the counterfactual mean outcome is identified as the weighted sum of the outcome conditional expectation in one population, given that the related characteristics take values in each class, with weights the probability that the characteristics take values in each class in the other population. The relative counterfactual mean estimator follows by the analogy principle. The procedure can be applied using data of any kind, without specifying a model for the conditional expectation. The asymptotic properties of the proposed estimator are unaffected when the partitions are data-dependent with classes converging to a fixed limit. The main application of the procedure consists of decomposing the difference between sample means into a composition term, pretending that the conditional expectation functions are identical in each population, and a residual term, pretending that the distribution of the related characteristics in the two populations are identical. The proposed methodology is applied to decompose the effects of the great recession on Spanish poverty indices. The results suggest that differences in the distribution of characteristics over time, resulting from labor market disruptions and demographic changes, were responsible for the increase in poverty rates. While job losses pushed the rates higher, the out-migration of poor foreigners decreased the rates.

[*]Liaoning University, China. Email: maominghai@lnu.edu.cn

[1] Universidad Carlos III de Madrid, Calle Madrid 126, 28903 Getafe (Madrid), Spain. Email: araiola@eco.uc3m.es

1 Introduction

When comparing an outcome observed in two populations, it is necessary to hold related factors constant to avoid distorting the comparison. For example, when comparing poverty rates before and after the Great Recession, related characteristics such as race, household composition, or age should be kept constant to prevent misleading conclusions. The counterfactual distribution of the outcome in one population, pretending that the related characteristics are distributed according to the distribution of another population, is a key tool for making insightful comparisons.

The mean of the counterfactual distribution is typically identified by assuming a model for the regression function, such as a parametric model (Oaxaca 1973 and Blinder 1973, hereafter OB), a semiparametric model (Machado and Mata 2005, Chernozhukov, Fernández-Val, and Melly 2013), or a nonparametric model (DiNardo, Fortin, and Lemieux 1996 or Rothe 2010). Once a model for the regression, or the conditional distribution itself, is specified, the counterfactual mean is determined by the convolution between the regression in one population and the marginal distribution of related characteristics in the other population. Estimates of the counterfactual mean are the basis for decomposing the difference between sample means into a compositional component, explained by the difference between the marginal distributions of related characteristics in the two populations, and a residual component, explained by the difference between the regression functions.

When the data is grouped in a number of categories, as is the case when it is discrete, there is no need to specify a regression model, as it has been noticed by Neison (1844), Kitagawa (1955), or Ñopo (2008). The counterfactual mean, in this case, is the weighted sum of the mean outcome in each category in one population, with weights the marginal probability of each category in the other population.

Our proposal consists of partitioning the characteristics domain into a number of classes. Thus, the counterfactual mean is the weighted sum of the conditional expectation of the outcome in one population given that the characteristics belongs to each class, with weights the probability that the characteristics take values in each class in the other population. Consequently, the natural estimator of the counterfactual mean is the

weighted sum of the conditional mean of the observed outcome in each class in one population, weighted by the relative frequencies of each class in the other population. This approach can be implemented with any kind of data, without assuming a specification for the regression function. Furthermore, the asymptotic distribution of the counterfactual mean estimator accounts for data dependent partitions with the random cells converging to fixed cells.

Formalizing this technique, Kitagawa (1955) proposed decomposing the difference between crude rates into composition and residual effects. Each effect, given by the difference between the crude rate and the corresponding counterfactual rate, offers counterfactual interpretations. The composition effect measures the difference that would have been observed pretending that the populations solely differed in their characteristic distributions, while the residual effect represents the difference that would have been observed pretending that the characteristic distributions in the two populations had been identical.

In economics, OB popularized the decomposition by analyzing male-female and white-black wage gaps. In their model, the composition effect quantifies fair discrimination due to different characteristics (e.g., higher education levels), while the residual effect gauges the extent of discrimination in the market. Since then, the decomposition has found extensive application in analyzing wage disparities across various demographic groups, including gender (Oaxaca and Ransom 1999), race (Melly 2005), and comparisons between immigrants and residents (Chiquiar and Hanson 2005), to mention only a few. Other applications include decomposing gender differences in smoking behavior (Bauer, Göhlmann, and Sinning 2007), and poverty rates (Biewen and Jenkins 2005). Refer to Fortin, Lemieux, and Firpo (2011) for a comprehensive review of decomposition applications in economics.

The estimation of the counterfactual mean relies on estimating the regression function, typically achieved through a parametric specification of the true regression model. OB employ ordinary least squares (OLS) to estimate the conditional mean wage given observed characteristics, effectively assuming a linear regression model. Linearity allows for a detailed decomposition, revealing the contribution of each characteristic to the two effects. However, when the regression function is nonlinear, OB method is severely biased.

More flexible approaches rely on nonparametric estimates of the regression function,

such as kernel regressions (Stock 1989), or manipulate a standardized feature of the outcome distribution as a weighted average, paired with nonparametric estimates of the weighting factor - see Barsky, Bound, Charles, and Lupton (2002) for the mean and DiNardo, Fortin, and Lemieux (1996) for the entire distribution. Alternative methods estimates the counterfactual cumulative density function (CDF) through a reweighting of the conditional CDF estimates. Rothe (2010) use kernel estimates of the conditional CDF, while Machado and Mata (2005) and Chernozhukov, Fernández-Val, and Melly (2013) explore various semiparametric specifications of the conditional CDF, such as quantile or distributional regression models.

These decomposition, based on semiparametric and nonparametric specifications of the regression function, do not necessarily return zero effects when the two populations share the same empirical distribution. Detailed decompositions similar to OB are not available for these methods, except for Machado and Mata (2005), which offers a detailed decomposition of the residual component akin to OB. Furthermore, these methods are not suitable for settings with non-ordinal characteristics and sparse data, as is the case, for instance, in the literature on inequality of opportunity (see, e.g., Brunori, Peragine, and Serlenga 2019) or poverty analysis (e.g., Bourguignon, Ferreira, and Leite 2008).

In the final section, we apply our methods to study the impact of the Great Recession on poverty indices in Spain. Our findings highlight compositional changes as the main driver of shock effects, while residual effects are mostly negative, indicating potential reduction in post-recession poverty rates if no changes in composition had occurred. A detailed analysis of the groups reveals labor market disruptions exacerbating crisis consequences, while demographic changes and migratory movements alleviate recession impacts.

The rest of the article is organized as follows. In Section 2, we introduce the notation and in Section 3 we present the estimation method and discuss alternative grouping procedures. Section 4 provides the asymptotic distribution of the estimator under minimal regularity conditions and data dependent partitions. The application to the decomposition of poverty indices before and after the great recession can be found in Section 6.

2 Setting

We consider two populations, 1 and 0, and an outcome of interest distributed as the random variables $Y^{(1)}$ and $Y^{(0)}$, respectively, in the two populations. The related characteristics are distributed as the random vectors $X^{(1)}$ and $X^{(0)}$, with domains $\mathcal{X}^{(1)} \subset \mathbb{R}^p$ and $\mathcal{X}^{(0)} \subset \mathbb{R}^p$, for populations 1 and 0, respectively. The joint distribution of $(Y^{(j)}, X^{(j)\prime})'$ is denoted by $F^{(j)}_{Y,X}$, $j = 0, 1$, which can be expressed as

$$F^{(j)}_{Y,X}(y, x) = \int\limits_{\{\bar{x} \leq x\}} F^{(j)}_{Y|X}(y|\bar{x}) F^{(j)}_{X}(d\bar{x}), \tag{1}$$

where $F^{(j)}_{Y|X}$ is the conditional distribution of $Y^{(j)}$ given $X^{(j)}$ and $F^{(j)}_{X}$ is the marginal distribution of $X^{(j)}$. The mean of $Y^{(j)}$ is denoted as,

$$\mu^{(j)}_{Y} = \int_{\mathbb{R}^p} \left[\int_{\mathbb{R}} y F^{(j)}_{Y|X}(dy|x) \right] F^{(j)}_{X}(dx). \tag{2}$$

To simplify the discussion, we focus on experiments where the the two populations share similar characteristics.

Assumption 1 (Overlapping Support)

$$\mathcal{X}^{(0)} = \mathcal{X}^{(1)}$$

Assumption 1 is standard in the literature (e.g., Fortin, Lemieux, and Firpo 2011, Assumption 4). When the assumption fails, such as when $X^{(1)}$ is a linear transformation of $X^{(0)}$, but $\mathcal{X}^{(0)} \subset \mathcal{X}^{(1)}$, the counterfactual experiment is performed over the common support. In this case, the difference $\mu^{(1)} - \mu^{(0)}$ can be decomposed using a four-component approach, as in Ñopo (2008) (also see Black, Haviland, Sanders, and Taylor 2008), rather than the two-component decomposition discussed below.

The counterfactual distribution of $Y^{(1)}$ pretending that $F^{(1)}_{X} = F^{(0)}_{X}$ is,

$$F^{(1,0)}_{Y}(y) = \int\limits_{\mathcal{X}^{(0)}} F^{(1)}_{Y|X}(y|\bar{x}) F^{(0)}_{X}(d\bar{x}). \tag{3}$$

Thus, the counterfactual mean of $Y^{(1)}$ pretending that $F_X^{(1)} = F_X^{(0)}$ is,

$$\mu_Y^{(1,0)} = \int_{\mathcal{X}^{(0)}} \left[\int_{\mathbb{R}} y F_{Y|X}^{(1)}(dy|x) \right] F_X^{(0)}(dx) = \int_{\mathcal{X}^{(0)}} h^{(1)}(x) F^{(0)}(dx), \tag{4}$$

where $h^{(j)}$ is the conditional mean (regression function) of $Y^{(j)}$ given $X^{(j)}$, $j = 0, 1$. This allows to decompose $\mu^{(1)} - \mu^{(0)}$ into two components, one reflecting differences in the marginal distributions of X, and the other differences in the conditional expectation (or distribution) of Y given X, within the two populations,

$$\mu_Y^{(1)} - \mu_Y^{(0)} = \underbrace{[\mu_Y^{(1)} - \mu_Y^{(1,0)}]}_{\textbf{Composition effect } \Delta^C} + \underbrace{[\mu_Y^{(1,0)} - \mu_Y^{(0)}]}_{\textbf{Residual effect } \Delta^R} \tag{5}$$

where,

$$\Delta^C = \int_{\mathcal{X}^{(0)}} h^{(1)}(x) \left[F^{(1)}(dx) - F^{(0)}(dx) \right] \qquad \Delta^R = \int_{\mathcal{X}^{(0)}} \left[h^{(1)}(x) - h^{(0)}(x) \right] F^{(0)}(dx) \tag{6}$$

In practice, it is often useful to focus on similar counterfactual experiments performed on groups in the data, rather than on (4) and (5). Notice that for any partition of $\mathbb{R}^p$, $\mathbb{C} = \{C_l\}_{l=1}^L$ say, and any $j = 0, 1$,

$$F_Y^{(j)}(y) = \int_{\mathcal{X}^{(j)}} F_{Y|X}^{(j)}(y|x) F_X^{(j)}(dx) = \sum_{l=1}^{L} F_{Y|X}^{(j)}(y|X \in C_l) F_X^{(j)}\{C_l\},$$

where

$$F_{Y|X}^{(j)}(y|X \in A) = \frac{1}{F_X^{(j)}\{A\}} \int_{\{x \in A\}} F_{Y|X}^{(j)}(y|dx) F_X^{(j)}(dx),$$

and

$$F_X^{(j)}\{A\} = \int_{\{x \in A\}} F_X^{(j)}(dx).$$

Accordingly,

$$\mu_Y^{(j)} = \int_{\mathcal{X}^{(j)}} \left[\int_{\mathbb{R}} y F_{Y|X}^{(j)}(dy|x) \right] F_X^{(j)}(dx) = \sum_{l=1}^{L} h^{(j)}\{C_l\} F_X^{(j)}\{C_l\},$$

where

$$h^{(j)}\{A\} = \frac{1}{F_X^{(j)}\{A\}} \int\limits_{\{x \in A\}} h^{(j)}(x) F_X^{(j)}(dx).$$

This forms the basis of the counterfactual decomposition introduced in the next section.

3 Decomposition with Aggregated Data

Based on Kitagawa (1955), rather than assuming a parametric model for $h^{(j)}$, we propose to perform the decomposition using the averaged outcome in each cell of some partition of the data. In particular, let $\mathbb{C} = \{C_l\}_{l=1}^L$ be a partition of $\mathbb{R}^p$ such that $\bigcup_{l=1}^L C_l = \mathbb{R}^p$, $C_l \bigcap C_j = \emptyset \ \forall \ l \neq j$, and $F_X^{(1)}\{C_l\} > a$ for all l and some $a > 0$. The counterfactual distribution of $Y^{(1)}$ using the averaged data in each cell is,

$$F_{Y,\mathbb{C}}^{(1,0)}(y) = \sum_{l=1}^{L} F_{Y|X}^{(1)}(y|X \in C_l) F_X^{(0)}\{C_l\}, \tag{7}$$

with the corresponding counterfactual mean using the integrated regression in each cell,

$$\mu_{Y,\mathbb{C}}^{(1,0)} = \sum_{l=1}^{L} h_l^{(1)}\{C_l\} F_X^{(0)}\{C_l\}. \tag{8}$$

Accordingly, the counterfactual decomposition is given by,

$$\mu_Y^{(1)} - \mu_Y^{(0)} = \underbrace{[\mu_Y^{(1)} - \mu_{Y,\mathbb{C}}^{(1,0)}]}_{\textbf{Composition effect } \Delta_{\mathbb{C}}^C} + \underbrace{[\mu_{Y,\mathbb{C}}^{(1,0)} - \mu_Y^{(0)}]}_{\textbf{Residual effect } \Delta_{\mathbb{C}}^R}, \tag{9}$$

where,

$$\Delta_{\mathbb{C}}^C = \sum_{l=1}^{L} \underbrace{h^{(1)}\{C_l\} \left(F_X^{(1)}\{C_l\} - F_X^{(0)}\{C_l\} \right)}_{\textbf{Composition contribution of the } l \textbf{ group}} \tag{10}$$

$$\Delta_{\mathbb{C}}^R, = \sum_{l=1}^{L} \underbrace{\left(h^{(1)}\{C_l\} - h^{(0)}\{C_l\} \right) F_X^{(0)}\{C_l\}}_{\textbf{Residual contribution of the } l \textbf{ group}}, \tag{11}$$

naturally define the contributions each group has on the composition and residual effects. Here, the composition and residual effects, $\Delta_{\mathbb{C}}^{R}$ and $\Delta_{\mathbb{C}}^{C}$, bear a similar interpretation to their ungrouped version in (6). $\Delta_{\mathbb{C}}^{C}$ indicates the difference observed pretending that the populations only differ in their group compositions, given by $(F_X^{(j)}\{C_1\}, ..., F_X^{(j)}\{C_L\})$ for $j = 0, 1$. Conversely, $\Delta_{\mathbb{C}}^{R}$ measures the difference observed pretending that the group compositions are identical.

Estimates of $\mu_{Y,\mathbb{C}}^{(1,0)}$ follow naturally by the analogy principle,

$$\hat{\mu}_{Y,\mathbb{C}}^{(1,0)} = \sum_{l=1}^{L} \hat{h}^{(1)}\{C_l\}\hat{p}^{(0)}\{C_l\} = \sum_{l=1}^{L} \hat{h}_l^{(1)}\hat{p}_l^{(0)} \tag{12}$$

where

$$\hat{p}_l^{(j)} = \frac{1}{n_j}\sum_{i=1}^{n_j} \mathbb{I}\{X_i^{(j)} \in C_l\}, \tag{13}$$

$$\hat{h}_l^{(j)} = \frac{1}{n_j \hat{p}_l^{(j)}}\sum_{i=1}^{n_j} Y_i^{(j)}\mathbb{I}\{X_i^{(j)} \in C_l\} \tag{14}$$

estimate $p_l^{(j)} = F_X^{(j)}\{C_l\}$ and $h_l^{(j)} = h^{(j)}\{C_l\}$, for $j = 0, 1$ and all l, respectively. Here n_j denotes the sample size of population j, while $\mathbb{I}\{A\}$ is the indicator function taking value 1 if condition A holds and 0 otherwise.

It is worth noticing that $\hat{\mu}_{Y,\mathbb{C}}^{(j,j)} = \hat{\mu}_Y^{(j)}$, where $\hat{\mu}_Y^{(j)} = n_j^{-1}\sum_{i=1}^{n_j} Y_i^{(j)}$, $j = 0, 1$, regardless of $\mathbb{C}$. Therefore, when the empirical distributions of $(Y^{(1)}, X^{(1)})$ and $(Y^{(0)}, X^{(0)})$ are identical, both composition and residual effects are zero. This feature is expected to hold for any proposal. However, it is not guaranteed in general that a parametric or semiparametric estimate of $\mu_Y^{(1,0)}$, $\hat{\mu}_Y^{(1,0)}$ say, would satisfy this property, particularly if the estimated regression residuals in each population, $Y_i^{(j)} - \hat{h}^{(j)}(X_i^{(j)})$, have sample mean different from zero.

This method complements the classic OB decomposition and can be applied regardless of the underlying regression model. Unlike methods based on parametric or semiparametric specifications of the regression function, which are sensitive to assumptions and smoothing, the decomposition with aggregated data depends only on the partition of the data. In general, $\mu_{Y,\mathbb{C}}^{(1,0)} \neq \mu_Y^{(1,0)}$, but the difference $\mu_{Y,\mathbb{C}}^{(1,0)} - \mu_Y^{(1,0)}$ approaches zero as the

size of each C_l shrinks. When the data is discrete, and $\mathbb{C}$ is the finest partition of the data, $\mu_{Y,\mathbb{C}}^{(1,0)} = \mu_Y^{(1,0)}$. In this case, the estimator $\hat{\mu}_{Y,\mathbb{C}}^{(1,0)}$ is identical to the counterfactual estimator in Ñopo (2008). If the data, instead, is already provided in contingency tables, the estimator $\hat{\mu}_{Y,\mathbb{C}}^{(1,0)}$ is identical to the one in Kitagawa (1955).

With continuous variables, one can split the sample into statistically equivalent blocks using quantiles of each element in $X^{(1)}$ (see Gessaman 1970, for an example). Alternatively, statistical learning theory offers a plethora of methods to discern meaningful groupings in the data (Hastie, Tibshirani, and Friedman 2009), such as k-means clustering (MacQueen et al. 1967). In more general settings, including non-ordinal characteristics, Classification and Regression Trees (CART) (Breiman, Friedman, Olshen, and Stone 1984) serve as effective grouping tools. The criterion for aggregating the data is based on how well the step function $\hat{h}^{(1)}(\cdot) = \sum_{l=1}^{L} \mathbb{I}\{\cdot \in C_l\}\hat{h}_l^{(1)}$, derived from the grouping, fits $h^{(1)}(\cdot)$. Trees offer the advantage of not relying on measures of distance in the characteristics' space and allow for grouping with non-ordered categorical variables. In all these cases, and many others, the partitioning choice is influenced by the data - i.e., the partition is data-dependent.

Notice that the counterfactual estimator $\hat{\mu}_{Y,\mathbb{C}}^{(1,0)}$ often coincides with a semiparametric estimator of $\mu_Y^{(1,0)}$. With equally-sized cells, for instance, $\hat{\mu}_{Y,\mathbb{C}}^{(1,0)}$ is a weighted average of the Cattaneo and Farrell (2013) partitioning estimator for the regression function, using a first-order polynomial basis approximation. Similarly, when the data is partitioned by CART, which is inherently a nonparametric estimation method.

In the next section, we provide fixed-L inference for the grouped estimator $\hat{\mu}_{Y,\mathbb{C}}^{(1,0)}$ and the relative decomposition components across various data-dependent partitioning schemes.

Then, in Section 5, we explore the semiparametric aspect of $\hat{\mu}_{Y,\mathbb{C}}^{(1,0)}$ with data-driven grouping and diverging L through Monte Carlo simulations. In particular, we compare the performance of $\hat{\mu}_{Y,\mathbb{C}}^{(1,0)}$ as a semiparametric estimator of $\mu_Y^{(1,0)}$ under data-driven groupings with other semiparametric estimation methods for $\mu_Y^{(1,0)}$.

4 Inference with Data-dependent Partitions

We discuss fixed-L inference for the counterfactual estimator $\hat{\mu}^{(1,0)}_{Y,\mathbb{C}}$. First, we provide the limit distribution of the statistics under a fixed partitioning scheme, and then extend the inference to data-dependent partitions, subject to certain restrictions on the partitioning algorithm. Similar results for the estimates of the decomposition effects, $\hat{\Delta}^{C}_{\mathbb{C}} = \hat{\mu}^{(1)}_{Y} - \hat{\mu}^{(1,0)}_{Y,\mathbb{C}}$ and $\hat{\Delta}^{R}_{\mathbb{C}} = \hat{\mu}^{(1,0)}_{Y,\mathbb{C}} - \hat{\mu}^{(0)}_{Y}$, are provided in the appendix.

Before doing so, is propaedeutical, following Rothe (2010), to distinguish between two different type of data generating processes defining the nature of the relationship between $(Y^{(1)}, X^{(1)})$ and $(Y^{(0)}, X^{(0)})$.

Assumption 2 (Data)

(a) The data $\{Y^{(1)}_i, X^{(1)}_i, Y^{(0)}_i, X^{(0)}_i\}_{i=1}^{n}$ *are jointly i.i.d.*

(b) The data $\{Y^{(1)}_i, X^{(1)}_i\}_{i=1}^{n_1}$ *and* $\{Y^{(0)}_i, X^{(0)}_i\}_{i=1}^{n_0}$*, with* $n_1/n_0 = \gamma + o(1)$ *for some* $\gamma > 0$*, are i.i.d. and mutually independent.*

Assumption 2 (a) models two-sample analysis, where the sample sizes (are assumed to) grow proportionally. While Assumption 2 (b) model dependence relationship between two sub-populations in panel data. Throughout, we adopt the notation relative to the former structure since it encompasses the latter as the special case $n_1 = n_0 = n$.

Under a fixed partition (i.e., $\mathbb{C}$ is non-random), the asymptotic distribution of $\hat{\mu}^{(1,0)}_{Y,\mathbb{C}}$ is as outlined in the following proposition.

Proposition 1 (Asymptotic Distribution with Fixed Partition)
Let $\sigma^{(j)}_l = \mathbb{V}(Y^{(j)}|X^{(j)} \in C_l)/p^{(j)}_l$*, under Assumptions 1, and 2 (a),* $\hat{\mu}^{(1,0)}_{Y,\mathbb{C}}$ *converges to a normal distribution:*

$$\sqrt{n}\left(\hat{\mu}^{(1,0)}_{Y,\mathbb{C}} - \mu^{(1,0)}_{Y,\mathbb{C}}\right) \xrightarrow{d} N\left(0, V_A\right),$$

$$V_A = \sum_{l=1}^{L}\left((p^{(0)}_l)^2\sigma^{(1)}_l + (h^{(1)}_l)^2 p^{(0)}_l(1-p^{(0)}_l)\right) - \sum_{l\neq f} h^{(1)}_l h^{(1)}_f p^{(0)}_l p^{(0)}_f + 2\sum_{l=1}^{L}\sum_{f=1}^{L} h^{(1)}_l c_{(l,f)} p^{(0)}_f,$$

where

$$c_{(l,f)} = \frac{q^{(1)}_{(l,f)} - h^{(1)}_f p_{(l,f)}}{p^{(1)}_f},$$

and,

$$q_{(l,f)}^{(j)} = \mathbb{E}\left[Y^{(j)}\mathbb{I}\{X^{(0)} \in C_l\}\mathbb{I}\{X^{(1)} \in C_f\}\right] \quad p_{(l,f)} = \mathbb{E}\left[\mathbb{I}\{X^{(0)} \in C_l\}\mathbb{I}\{X^{(1)} \in C_f\}\right].$$

Under Assumptions 1, and 2 (b)

$$\sqrt{n_1}\left(\hat{\mu}_{Y,\mathbb{C}}^{(1,0)} - \mu_{Y,\mathbb{C}}^{(1,0)}\right) \xrightarrow{d} N\left(0, V_A\right),$$

where:

$$V_A = \sum_{l=1}^{L}\left((p_l^{(0)})^2\sigma_l^{(1)} + \frac{1}{\gamma}(h_l^{(1)})^2 p_l^{(0)}(1-p_l^{(0)})\right) - \frac{1}{\gamma}\sum_{l\neq f} h_l^{(1)} h_f^{(1)} p_l^{(0)} p_f^{(0)}.$$

Proof. *Appendix.* ■

The result extends the matching estimator distribution of Ñopo (2008) to any (fixed) matching window and interdependent data.

In most cases, however, the grouping choice is influenced by the data, such as with equally-sized cells or with data-driven methods like k-means clustering. In this scenario, the asymptotic theory employed in Proposition 1 is no longer suitable. Instead, set-specific averages are modeled as empirical measures indexed by sets, and the asymptotic distribution of the counterfactual estimator follows from uniform convergence results (see, for example, Pollard 1990, Section 12).

The data-dependent partitions, based on Pollard (1979) and Andrews (1988), are modeled as random functions over a class of properly restricted measurable sets, $\mathcal{D}$ say, where $\mathcal{D}$ is a class of sets in $\mathbb{R}^p$ from which the cells of each partition are selected. Denote as $\mathcal{C}$ the class of partitions of $\mathbb{R}^p$ consisting of L sets from $\mathcal{D}$ (with L fixed for all n),

$$\mathcal{C} = \left\{\mathbb{C} = (C_1, ..., C_L) \in \mathcal{C}^L : \bigcup_{l=1}^{L} C_l = \mathcal{X},\ C_l \bigcap C_f = \emptyset,\ \forall l \neq f\right\}. \tag{15}$$

Equip $\mathcal{D}$ with the topology generated by the $L^2(F_X)$ semi-norm and give $\mathcal{C}$ the corresponding product topology, where $F_X = (\gamma/(1+\gamma))\, F_X^{(1)}/ + (1/(1+\gamma))\, F_X^{(0)}$ is the joint distribution of characteristics in the two populations and γ is as defined in Assumption

2. This means that two sets C_1 and C_2 in $\mathcal{X}$ are close if $F_X\{C_1 \Delta C_2\}$ is small, Δ being the symmetric difference operator, $C_1 \Delta C_2 = (C_1 \cup C_2) \backslash (C_1 \cap C_2)$.

For each sample size n (alternatively, n_1 or n_0), the data-dependent partition $\hat{\mathbb{C}} = (\hat{C}_1, ..., \hat{C}_L)$ is a measurable mapping from the underlying probability space to $\mathcal{C}$, such that $\hat{\mathbb{C}}$ converges in probability to some fixed partition in $\mathcal{C}$. Specifically, $\hat{\mathbb{C}} \xrightarrow{p} \mathbb{C}$ if and only if for all $\epsilon > 0$,

$$P(F_X\{\hat{C}_l \Delta C_l\} > \epsilon) \to 0, \quad \text{for all } l = 1, ..., L.$$

Assumption 3 $\hat{\mathbb{C}} \xrightarrow{p} \mathbb{C}$ *for some fixed set of cells* $\mathbb{C} \in \mathcal{C}$.

Assumption 3 is a standard requirement for the convergence of empirical processes indexed by sets. It is fulfilled by a wide range of partitioning algorithms, including k-means clustering, or by any partitioning algorithm where the splitting points depend continuously on estimated parameters with a constant probability limit (Andrews 1988).

Importantly, to obtain the limit distribution of $\hat{\mu}^{(1,0)}_{Y,\hat{\mathbb{C}}}$, it is crucial to limit the complexity of the partitioning algorithm. This limitation is established by assuming that the cells are selected from a Vapnik-Cervonenkis (VC) class.

Assumption 4 $\mathcal{C}$ *is a VC class of sets.*

This assumption is independent of the data distribution and sufficiently general for our purposes. For instance, algorithms generating cells with a finite number of straight edges and the class of hyper ellipsoids fall within the VC classes. The property is preserved under set operations such as unions, intersections, differences, and complements, as discussed by Andrews (1988) and Pollard (1984); further insights can be found in Section 2.6 of Van Der Vaart (1996).

All the partitioning methods discussed in the previous section have finite VC class, including CART when the number of cells grow at logarithmic rate (see, e.g., Athey and Wager 2021 pag. 143). Assumption 3, instead, is a high-level condition for regression tree methods.

When the above regularity conditions are fulfilled, the asymptotic distribution of the counterfactual estimator is defined by the limit partition.

Theorem 1 (Asymptotic Distribution with Random Partition)

Under Assumptions 1, 2, 3, and 4,

$$\sqrt{n}\left(\hat{\mu}^{(1,0)}_{Y,\hat{\mathbb{C}}} - \mu^{(1,0)}_{Y,\mathbb{C}}\right) \xrightarrow{d} N\left(0, V_A\right)$$

where V_A is defined as in Proposition 1.

Analogous results for the decomposition effects estimators are available in the appendix.

5 Monte Carlo Simulations

We conduct Monte Carlo simulations to investigate the effectiveness of the counterfactual estimator $\hat{\mu}^{(1,0)}_{Y,\hat{\mathbb{C}}}$ as a semiparametric estimator of $\mu^{(1,0)}$ in finite samples. To aggregate the data we generate partitions using CART and the k-means clustering algorithm. The performance assessment of the estimators is based on the mean squared error (MSE) and the mean absolute error (MAE) of the counterfactual mean estimates compared to the counterfactual mean $\mu^{(1,0)}$. The simulations compare the counterfactual estimator using aggregated data with both parametric and nonparametric estimators of the counterfactual mean. Parametric approaches considered include the OB decomposition (linear parametrization) and the method proposed by Machado and Mata (2005), using a linear specification of the conditional quantiles. Non-parametric estimators are based on kernel regression (Rothe 2010) and matching (Ñopo 2008).

We divide the simulation study into two parts, first discussing the case when the covariate vector $X^{(j)}$ consists of continuous variables, and then when $X^{(j)}$ consists of both discrete and continuous variables. In the first case, we consider three different specifications for the regression function $h^{(1)}(\cdot)$, and in the second, four. The simulated samples have sizes $n_1 = n_0 = n$ with $n \in \{1000, 5000\}$. For each model and estimator we perform 1000 Monte Carlo replications. Throughout the simulations, outcomes, $Y^{(j)}$, are generated according to

$$Y^{(j)} = 5 + h^{(j)}(X^{(j)}) + V^{(j)}$$

for $j \in \{0, 1\}$, where $V^{(g)}$ follows a truncated normal distribution $N(0, (1+h^{(j)}(X^{(j)}))^2, -3, 3)$.

In the first simulation design, $X^{(j)}$ consists of K continuous variables, with $K \in \{2, 5, 10\}$, each independently distributed as an exponential r.v. $f(x, \lambda_j) = \lambda_j e^{-\lambda_j x}$ truncated at 1 with the decay parameter $\lambda_0 = 3$ and $\lambda_1 = 4.5$. We consider three sparse specifications of $h^{(1)}(\cdot)$ depending only on the first two covariates as following:

$$\begin{aligned} \text{Linear}: \quad & h^{(1)}(X^{(1)}) = X_1^{(1)} + X_2^{(1)} \\ \text{Nonlinear}: \quad & h^{(1)}(X^{(1)}) = \sin(\pi X_1^{(1)}) + \sin(\pi X_2^{(1)}) \\ \text{Discontinuous}: \quad & h^{(1)}(X^{(1)}) = \mathbf{1}(X_1^{(1)} \geq 0.5) + \mathbf{1}(X_2^{(1)} \geq 0.5) \end{aligned} \tag{16}$$

Whereas $h^{(0)}(X^{(1)}) = X_1^{(0)} + X_2^{(0)}$ for all DGP settings.

In the second part of the Monte Carlo study, we consider characteristics $(X_1^{(j)}, X_2^{(j)})$ where $X_1^{(j)}$ is a continuos r.v. distributed as one of the $X^{(j)}$ in the first simulation design, while $X_2^{(j)}$ takes values in an unordered set $\{a_1, a_2, ..., a_P\}$, with $P \in \{5, 10\}$. We equivalently write $X_2^{(j)}$ as P dummy variables $\{Z_t^{(j)} = \mathbb{I}(X_2^{(j)} = t)\}_{t=1}^P$. The probabilities $\mathbb{P}(Z_t^{(0)} = 1) = 1/P$ and $\mathbb{P}(Z_t^{(1)} = 1) = F(F^{-1}(t/P, \lambda_0), \lambda_1) - F(F^{-1}((t-1)/P, \lambda_0), \lambda_1)$, where $F(x, \lambda_j)$ is the CDF of $f(x, \lambda_j)$. We consider four specifications of $h^{(1)}(\cdot)$:

$$\begin{aligned} \text{Linear}: \quad & h^{(1)}(X^{(1)}) = X_1^{(1)} + \sum_{t=1}^{L} \beta_t Z_t^{(1)} \\ \text{Nonlinear}: \quad & h^{(1)}(X^{(1)}) = \sin(2\pi X_1^{(1)}) + \sum_{t=1}^{L} \beta_t Z_t^{(1)} \\ \text{Interaction}: \quad & h^{(1)}(X^{(1)}) = \sin(2\pi X_1^{(1)}) + \sum_{t=1}^{L} \beta_t Z_t^{(1)} + \sum_{t=1}^{L} \beta_t Z_t^{(1)} \sin(2\pi X_1^{(1)}) \\ \text{Discontinuous}: \quad & h^{(1)}(X^{(1)}) = \mathbf{1}(X_1^{(1)} \leq 0.5) + \sum_{t=1}^{L/2} Z_t^{(1)}. \end{aligned} \tag{17}$$

Whereas the model for population 0 is $h^{(0)}(X^{(0)}) = X_1^{(0)} + \sum_{t=1}^{L-1} \beta_t Z_t^{(0)}$.

In the first part, the bandwidth for the Kernel estimator is $h_b = 1.5\sigma_{x_k} n^{-(2K+1)}$ (Rothe, 2010), where σ_{x_k} is the standard deviation of the respective covariates. While in the second part, the kernel estimator using the unfeasible optimal bandwidth of Racine and Li (2004). Ñopo (2008) matching estimator only uses the first nearest neighbor. The minimum size of each partition, for both CART and k-means, is $h_c = 2n_1^{\frac{1}{3}}$. in the CART algorithm, we

set the initial penalty term to 0 cp = 0, while in the k-means clustering algorithm we set the number of partitions as $n/2h_c$ and the maximum number of iteration to 100.

Tables 1 and 2 present the results of the first and second sets of simulations, respectively. As expected, in both set of simulations, the parametric estimators exhibit the lowest MAE and RMSE in the linear models. However, even with continuos covariates, nonparametric approaches demonstrate significant advantages over parametric estimators in the nonlinear models, particularly when $K = 2$. In sparse models ($K = 5$ and $K = 10$), only the CART algorithm maintains strong performance, while other nonparametric methods suffer from the curse of dimensionality. In the mixed covariates case, the matching estimator keeps good performance even when $P = 10$, while CART performs better than matching only if the DGP involves the interaction term between continuous and discrete variables. On the other hand, in the discontinuous models, the CART algorithm outperforms other nonparametric methods for all K in both continuos and mixed covariates settings.

6 Analysis of AROPE Index in Spain

We investigate the impact of the Great Recession on poverty rates in Spain, analyzing how changes in the distribution of poverty risk factors, induced by the economic shock, influenced the surge in poverty rates. This analysis contributes to a large literature employing OB decomposition to decompose poverty rates across different groups, such as regions (Ayala, Jurado, and Pérez-Mayo 2011), ethnic groups (Gradín 2012), and rural and urban areas (Ayala, Jurado, and Pérez-Mayo 2021). See Biewen and Jenkins (2005), Bourguignon and Ferreira (2005), and Bourguignon, Ferreira, and Leite (2008) for an application to the household income distribution.

Taking 2008 as the population of reference (standard), corresponding to the recession's onset, we decompose the difference between this baseline and the rates in subsequent years $t \in \{2009, ..., 2014\}$ into composition and residual effects.

The data is grouped using the CART algorithm with the *Rpart* package in R. The number of groups is determined by minimizing the cross-validated mean squared error, as detailed in part D of the appendix. The grouped decomposition offers crucial insights

Table 1: **Monte Carlo Simulations: Continous Covariates**

Model	K	n	OB	Quantile	Kernel	Matching	KM	CART
				MAE				
Linear	2	1000	0.0360	0.0426	0.0398	0.0494	0.0412	0.0581
	2	5000	0.0158	0.0331	0.0176	0.0229	0.0176	0.0224
	5	1000	0.0389	0.0457	0.0723	0.0538	0.0623	0.0605
	5	5000	0.0183	0.0307	0.0578	0.0291	0.0436	0.0283
	10	1000	0.0452	0.0508	0.1008	0.0746	0.0972	0.0677
	10	5000	0.0204	0.0319	0.0924	0.0535	0.0670	0.0286
Nonlinear	2	1000	0.0501	0.0805	0.0412	0.0556	0.0416	0.0416
	2	5000	0.0333	0.0756	0.0192	0.0249	0.0188	0.0191
	5	1000	0.0503	0.0764	0.0563	0.0543	0.0697	0.0415
	5	5000	0.0333	0.0738	0.0419	0.0256	0.0455	0.0193
	10	1000	0.0589	0.0811	0.0714	0.0615	0.1041	0.0431
	10	5000	0.0354	0.0736	0.0618	0.0341	0.0696	0.0188
Discontinuous	2	1000	0.0379	0.0372	0.0396	0.0505	0.0400	0.0368
	2	5000	0.0209	0.0174	0.0193	0.0221	0.0191	0.0168
	5	1000	0.0416	0.0395	0.0956	0.0590	0.0654	0.0368
	5	5000	0.0212	0.0190	0.0727	0.0288	0.0374	0.0163
	10	1000	0.0474	0.0457	0.1466	0.0971	0.1117	0.0374
	10	5000	0.0238	0.0213	0.1297	0.0688	0.0697	0.0175
				RMSE				
Linear	2	1000	0.0447	0.0526	0.0502	0.0620	0.0510	0.0738
	2	5000	0.0199	0.0376	0.0220	0.0286	0.0222	0.0281
	5	1000	0.0493	0.0568	0.0822	0.0664	0.0738	0.0743
	5	5000	0.0230	0.0363	0.0613	0.0359	0.0484	0.0341
	10	1000	0.0558	0.0627	0.1086	0.0889	0.1057	0.0826
	10	5000	0.0257	0.0381	0.0943	0.0593	0.0707	0.0348
Nonlinear	2	1000	0.0625	0.0934	0.0515	0.0692	0.0521	0.0523
	2	5000	0.0390	0.0790	0.0240	0.0312	0.0234	0.0239
	5	1000	0.0628	0.0904	0.0679	0.0680	0.0823	0.0512
	5	5000	0.0399	0.0783	0.0468	0.0317	0.0508	0.0242
	10	1000	0.0736	0.0976	0.0836	0.0762	0.1143	0.0542
	10	5000	0.0425	0.0792	0.0654	0.0413	0.0741	0.0238
Discontinuous	2	1000	0.0479	0.0465	0.0503	0.0632	0.0505	0.0463
	2	5000	0.0260	0.0217	0.0243	0.0278	0.0241	0.0211
	5	1000	0.0520	0.0495	0.1042	0.0727	0.0769	0.0456
	5	5000	0.0262	0.0234	0.0751	0.0355	0.0426	0.0202
	10	1000	0.0590	0.0573	0.1519	0.1107	0.1193	0.0473
	10	5000	0.0297	0.0272	0.1310	0.0737	0.0735	0.0219

MAE and RMSE of the counterfactual mean for different models and estimators. OB stands for the Oaxaca-Blinder decomposition, Quantile for the method proposed by Machado and Mata (2005), Kernel for the Rothe (2010) estimator, Matching for the Ñopo (2008) matching estimator, KM and CART stand for the partitioning estimator $\hat{\mu}^{(1,0)}_{\hat{\mathbb{C}}}$ using k-means and CART, respectively.

Table 2: **Monte Carlo Simulations: Mixed Covariates**

Model	P	n	OB	Quantile	Kernel	Match	KM	CART
				MAE				
Linear	5	1000	0.0378	0.0440	0.0654	0.0552	0.0473	0.0632
		5000	0.0173	0.0338	0.0249	0.0232	0.0178	0.0241
	10	1000	0.0372	0.0419	0.0895	0.0553	0.0733	0.0657
		5000	0.0173	0.0328	0.0357	0.0247	0.0329	0.0268
Nonlinear	5	1000	0.0447	0.0644	0.0572	0.0549	0.0517	0.0546
		5000	0.0257	0.0575	0.0221	0.0218	0.0187	0.0290
	10	1000	0.0456	0.0624	0.0736	0.0567	0.0716	0.0562
		5000	0.0260	0.0568	0.0316	0.0231	0.0351	0.0290
Interaction	5	1000	0.0479	0.0690	0.0443	0.0538	0.0552	0.0440
		5000	0.0316	0.0655	0.0176	0.0212	0.0176	0.0197
	10	1000	0.0492	0.0681	0.0581	0.0540	0.0771	0.0430
		5000	0.0315	0.0633	0.0210	0.0223	0.0394	0.0197
Discontinuous	5	1000	0.0400	0.0494	0.0527	0.0567	0.0540	0.0392
		5000	0.0187	0.0384	0.0205	0.0230	0.0194	0.0169
	10	1000	0.0411	0.0501	0.0743	0.0575	0.0868	0.0408
		5000	0.0196	0.0375	0.0268	0.0253	0.0406	0.0177
				RMSE				
Linear	5	1000	0.0476	0.0548	0.0758	0.0697	0.0582	0.0766
		5000	0.0216	0.0388	0.0303	0.0288	0.0226	0.0303
	10	1000	0.0468	0.0531	0.0978	0.0705	0.0847	0.0797
		5000	0.0217	0.0382	0.0405	0.0306	0.0387	0.0329
Nonlinear	5	1000	0.0557	0.0771	0.0689	0.0695	0.0641	0.0672
		5000	0.0313	0.0619	0.0270	0.0275	0.0231	0.0348
	10	1000	0.0562	0.0751	0.0842	0.0713	0.0839	0.0697
		5000	0.0317	0.0613	0.0365	0.0292	0.0413	0.0351
Interaction	5	1000	0.0601	0.0824	0.0547	0.0677	0.0683	0.0543
		5000	0.0369	0.0691	0.0222	0.0268	0.0222	0.0246
	10	1000	0.0621	0.0824	0.0684	0.0683	0.0911	0.0532
		5000	0.0368	0.0671	0.0260	0.0280	0.0454	0.0247
Discontinuous	5	1000	0.0499	0.0604	0.0640	0.0712	0.0663	0.0489
		5000	0.0236	0.0436	0.0257	0.0286	0.0244	0.0212
	10	1000	0.0514	0.0614	0.0860	0.0717	0.1011	0.0515
		5000	0.0246	0.0434	0.0325	0.0319	0.0470	0.0221

MAE and RMSE of the counterfactual mean for different models and estimators. OB stands for the Oaxaca-Blinder decomposition, Quantile for the method proposed by Machado and Mata (2005), Kernel for the Rothe (2010) estimator, Matching for the Ñopo (2008) matching estimator, KM and CART stand for the partitioning estimator $\hat{\mu}^{(1,0)}_{\hat{\mathbb{C}}}$ using k-means and CART, respectively.

for targeting pro-poorness policies, such as directing resources to address unemployment within specific age ranges.

Individuals in poverty conditions are identified by the AROPE (At-risk-of-poverty-or-social-exclusion) measure, a multi-dimensional index adopted by the European Commission to assess poverty. The AROPE classifies an individual as at risk of poverty or social exclusion if they face at least one of the following situations:

- At risk of Poverty (AROP60): The individual lives in a household with an equivalized disposable income[1] below 60% of the national median equivalized disposable income.
- Severe material deprivation (MD): The individual cannot afford at least 4 out of 9 predefined material items considered by most people to be desirable or even necessary to live an adequate life.
- Low work intensity (LJ): The individual lives in a household where the adults worked a working time equal to or less than 20% of their total combined work-time potential during the previous year.

The proportion of individuals falling into the AROPE classification determines the AROPE rate.

The data used is from the Survey on Living Condition (*Encuesta de condiciones de vida* or ECV) elaborated by the Spanish Statistical National Institute (INE). It is a yearly survey collecting harmonized data on income, poverty, social exclusion, and living conditions. We consider only individuals older than 16 years old and exclude the autonomous cities of Ceuta and Melilla. Furthermore, we drop from the sample observations with missing values. The drop affects a percentage of individuals smaller than 3% of the whole sample. The survey furnishes population weight for the adult population (16+) that we employ in the estimation of the statistics and relative standardization.

As potential drivers of poverty, we choose a set of characteristics considered relevant factors of risk for being poor or materially deprived. These include information both at the household and individual levels. At the household level, we report the household's type (single person, couple with or without children, single-parent households, etc.) and the

[1]The equivalized disposable income is the main welfare measure adopted by Eurostat. This is equal to the total household income divided by the OECD scale of family size.

Table 3: **Descriptive Statistics**

	Sex	Age	Nat	Cbirth	Educ	WorkS	H-type	HH-sex	HH-age	H-syze
2008										
mean	1.524	61.747	1.090	1.124	2.465	3.514	9.764	1.358	67.900	3.192
median	2	61	1	1	2	3	9	1	67	3
q-25	1	47	1	1	1	1	8	1	57	2
q-75	2	76	1	1	3	5	12	2	79	4
std.dev	0.499	18.715	0.396	0.463	1.584	2.756	2.787	0.479	14.693	1.312
2014										
Mean	1.522	57.359	1.081	1.139	2.590	3.988	9.558	1.363	63.661	3.067
Median	2	57	1	1	2	5	9	1	63	3
p-25	1	43	1	1	1	1	8	1	53	2
p-75	2	71	1	1	5	6	12	2	74	4
St.dev	0.499	18.643	0.374	0.487	1.662	2.577	2.858	0.480	14.776	1.299

Descriptive statistics of the characteristics in 2008 and 2014. The table reports mean, median, 25th, 75th percentiles, and standard deviation. The characteristics considered are: sex (sex), age (age), nationality (nat), country of birth (cbi), education (educ), work status (workS), household type (H-type), and size (H-syze), household head sex (HH-sex), and age (HH-age). Detailed description of the variables is in the appendix.

household head's gender and age. At the individual level, we choose a set of variables that characterizes age, education, employment, sex, the country of birth, and the nationality of each individual in the sample. In the appendix, we report a detailed breakdown of the variables used in this analysis.

The proportion of individuals in AROPE experienced a dramatic increase during the recession, rising from 22.6% of the total (adult) population in 2008 to a peak of 28% in 2014. Even after 2014, despite the recovery of the Spanish economy, AROPE rates remained around 26% of the total population. The individual AROPE components followed similar trends, although, in proportion, the effect on the MD and LJ indices was much higher than on the AROPE and the income-based measure. During the 2008-2014 period, the proportion of individuals in low work intensity and material deprivation conditions more than doubled.

To understand the dynamic behind this rapid surge in poverty rates, we decompose the AROPE rate and its three components along the 2008-2014 period, using the 2008 population as a reference. Figure 1 illustrates the decomposition of the four rates, the indices time series are depicted in red, while the counterfactual rates are in blue. Each graph reports a dashed black line corresponding to the indices level in 2008, the differences between the red and blue line and the blue and black line determine the compositional

Table 4: **National Poverty Rates**

Year	AROPE	AROP60	MD	LJ
2008	22.6	18.5	3.1	5.3
2009	23.6	19.0	4.0	5.9
2010	25.1	19.4	4.4	8.4
2011	25.8	19.4	4.3	10.2
2012	26.4	19.7	5.5	10.9
2013	26.5	19.2	5.6	12.0
2014	28.0	20.7	6.5	13.0
2015	27.7	20.9	5.8	11.8
2016	27.1	21.1	5.5	11.3
2017	25.7	20.3	4.8	9.6
2018	25.5	20.6	5.1	8.1
2019	24.3	19.3	4.4	8.0
2020	25.4	19.6	6.4	7.4

Evolution of AROPE rate and its three components in Spain from 2008 to 2020. The values indicate percentage of the population.

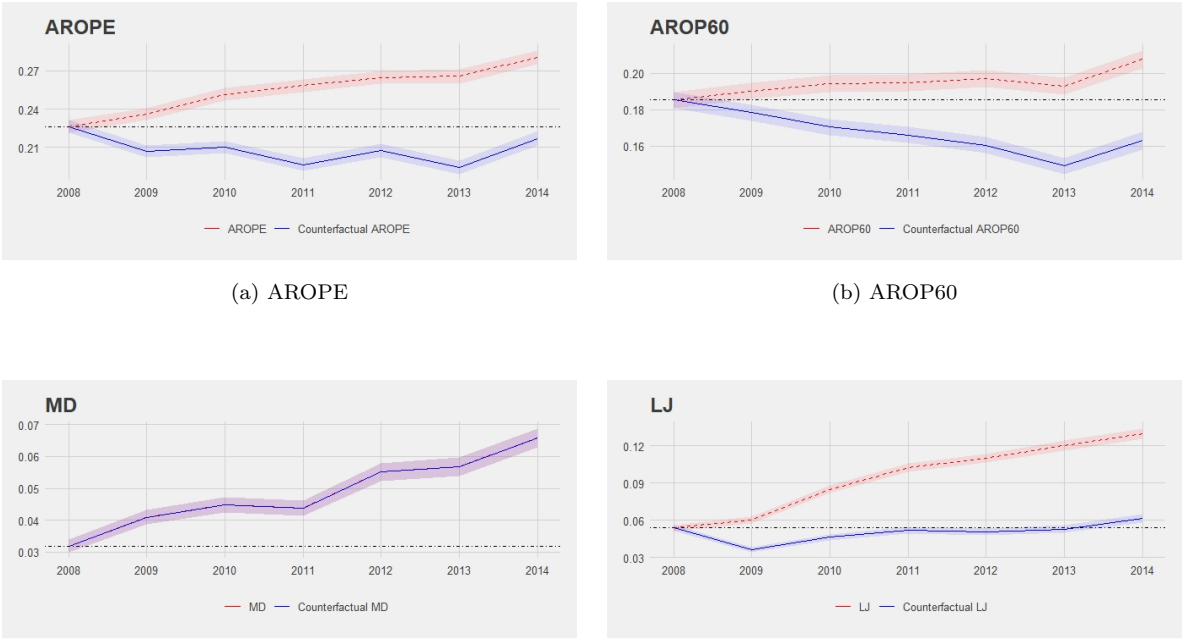
and residual components, respectively. Table 4, instead, reports the decompositions of the four indices 2008-2014 differentials, together with the number of groups determined by the CART algorithm.

Compositional variations appear as the main driver of the shock effects on the aggregate index. Similar patterns show in the decomposition of the LJ and AROP60 rates, which perturbations seem to drive the raise in AROPE rates. As unemployment is a primary risk factor for poverty, it is not surprising that variations in characteristics' composition had a massive impact on the increase of poverty rates. The shock drastically reduced the demand for labor, particularly in sectors like construction, thus, moving a big portion of the population from low-risk factors (full-time employment) to high-risk factors (part-time employment or unemployment).

The interpretation of residual components is more difficult than that of composition components, as they depend on unobserved factors. Loosely speaking, positive (negative) residual components implies that the poverty conditions of groups of individuals worsened (improved) (keeping constant the population composition). In Figure 1 (a-b), we observe negative residual components across most of the recession period, which suggests that if the population kept the same characteristics composition of 2008, the poverty ratio would have actually decreased.

Remarkably, the CART algorithm does not find useful splits for the MD index, as

Figure 1: Poverty Rates Decomposition



(a) AROPE

(b) AROP60

(c) MD

(d) LJ

The graphs illustrate the time series of the four rates, AROPE, AROP60, MD, LJ, in red together with the relative counterfactual rate, $\hat{\mu}_{Y,\mathbb{C}}^{(t,2008)}$, in blue for $t \in \{2008, ..., 2014\}$.

the predictive power of the characteristics is insufficient for this index. As a result the composition component is zero. This evidence, suggesting strong difference in risk factors of material deprivation and income-based measure, is consistent with previous studies showing that only a small percentage of individuals exiting income-based poverty also leave conditions of material deprivation (see, e.g., Ayala, Jurado, and Pérez-Mayo 2011).

The grouping with CART also allows to extrapolate information about groups of individuals and how they contributed to the rapid rise of poverty rates. For the sake

Table 5: **Decomposition of National Poverty Rates**

Index	Δ	**Comp. eff.**	**Res. eff.**	**Nr. Groups**
AROPE	5.3	6.1 (27.19)	-0.7 (-2.05)	24
AROP60	2.2	4.4 (27.8)	-2.2 (-6.47)	18
MD	3.3	0 (·)	3.3 (31.01)	1
LJ	7.6	6.8 (34.45)	0.7 (3.76)	21

Decomposition of 2008-2014 national poverty rates differentials. Reported statistics are multiplied by 100. Δ is the rate differential in the two period. The third and fourth columns report the composition and residual effects, respectively. **Nr. Groups** are the number of groups generated by the CART algorithm. T-ratio with 0 null hypothesis in parenthesis.

Table 6: **Groups Details**

Group	WorkS	Age	Educ	Cbirth	Nat	H-type	HH-age	H-size
1	1,5	≥ 67.5	0,1,2	1	-	-	-	-
2	6,9	< 67.5	3,4,5	1	-	8,9,10,11,12,13,14	≥ 69.5	≥ 3.5
3	1,5	-	3,4,5	1	-	-	-	-
4	2,4,8	≥ 67.5	-	-	-	-	-	-
5	6,9	< 67.5	$0, 1, 2$	-	-	-	-	-
6	6,9	< 67.5	3,4,5	1	-	8,9,10,11,12,13,14	< 69.5	-
7	1	< 67.5	0,1,2	1	-	-	-	-
8	6,9	< 67.5	3,4,5	1	-	8,9,10,11,12,13,14	≥ 69.5	< 3.5
9	2,4,8	< 67.5	2,3,4,5	1,2	-	-	-	-
10	1,5	-	5	2,3	-	10,11,12,13,14	-	-
11	6,9	≥ 67.5	-	-	-	6,7,8,9,10,11,12,13,14	-	-
12	1,5	-	-	2,3	-	1,2,3,4,5,6,7,8,9	-	-
13	6,9	< 67.5	3,4,5	2,3	-	8,9,10,11,12,13,14	-	-
14	4,8	< 67.5	0,1	-	-	-	-	-
15	5	< 67.5	0,1,2	1	-	-	-	-
16	1,5	-	0,1,2,3,4	2,3	3	10,11,12,13,14	-	-
17	2	< 67.5	2,3,4,5	3	-	-	-	-
18	6,9	< 67.5	3,4,5		-	1,2,3,4,5,6,7	-	-
19	2	< 67.5	0,1	-	-	-	≥ 63.5	-
20	2	< 67.5	0,1	-	-	-	< 63.5	-
21	1,5	-	0,1,2,3,4	2,3	1,2	10,11,12,13,14	-	< 4.5
22	6,9	≥ 67.5	-	-	-	1,2,3,4,5	-	-
23	4,8	< 67.5	2,3,4,5	3	-	-	-	-
24	1,5	-	0,1,2,3,4	2,3	1,2	10,11,12,13,14	-	≥ 4.5

Split rule for each group determined by the CART algorithm in the 2014-2008 differential decomposition of the AROPE rate. Variables for which there were no useful splits have been excluded.

of brevity, we report and comment only on the detailed decomposition of the 2008-2014 AROPE differential. The tree generated by the CART algorithm, returns an immediate overview of the generated groups and their contribution to the AROPE rise, as reported in Table 6 and 7. The former table reports the splitting rule determining each group, the latter reports the relative size of the groups and the within-group AROPE rate in the 2008 and 2014 populations. Increases (decreases) in the probability of being in these groups, $p_l^{(14)} - p_l^{(08)}$, produce positive (negative) composition contributions; increases (decreases) of the within-group AROPE rate, $h_l^{(14)} - h_l^{(08)}$, determine positive (negative) residual contribution. The magnitude of the contribution on the two term is determined by the interaction between these variations and the probability of being in AROPE in 2014 (composition) or in the group in 2008 (residual).

The group analysis unveils nuanced patterns in the data that would be overlooked when considering only aggregate statistics. Group 1, for instance, comprises elderly Spanish individuals, either employed full-time or retired, with lower educational levels. Conversely,

Figure 2: AROPE Tree: The terminal leafs report the predicted AROPE of each class in the 2014 population; green colors denote high within-group AROPE rate, while blue denotes the opposite.

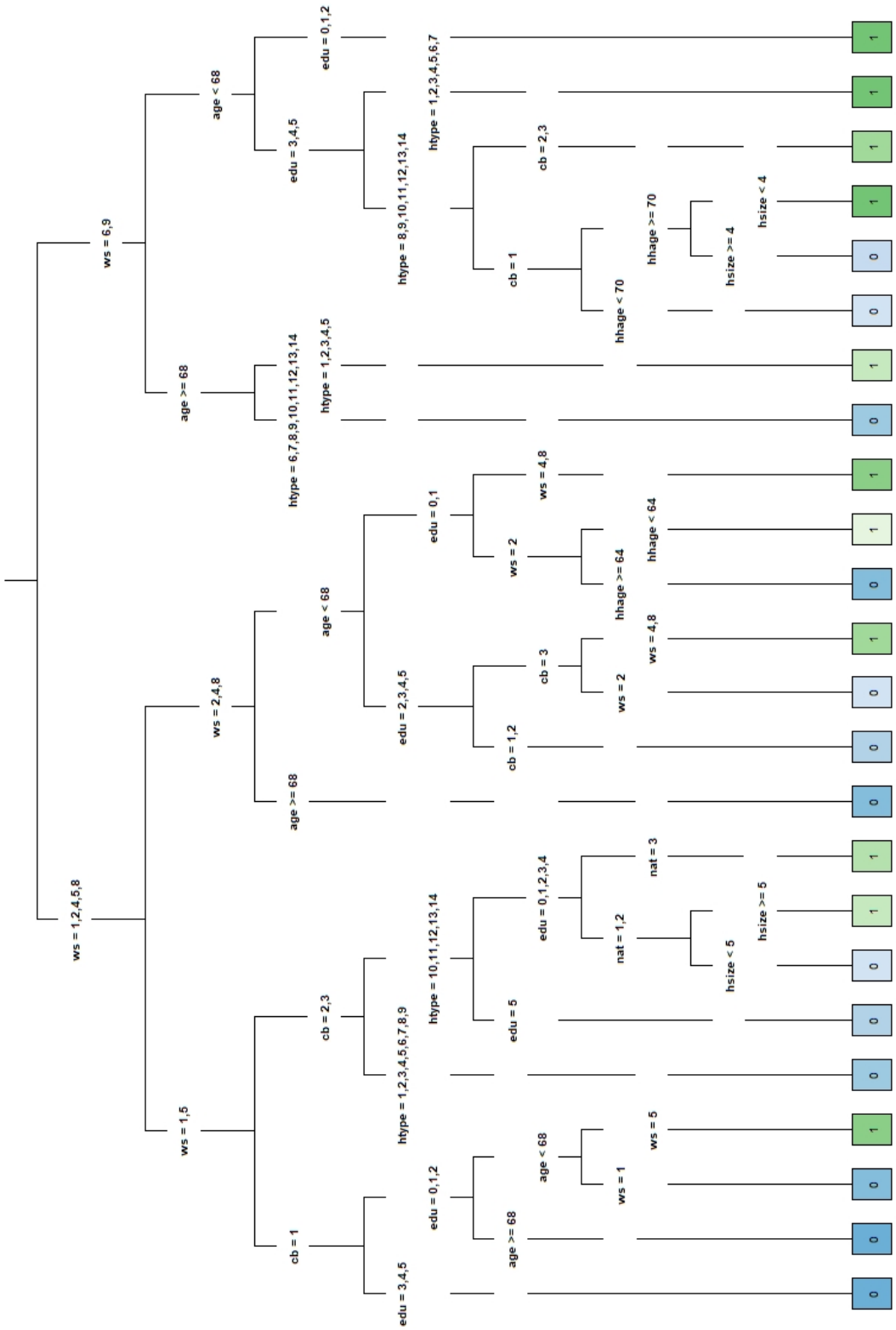


Table 7: **Groups Contribution**

Group	$h_l^{(14)}$	$p_l^{(14)}$	$h_l^{(08)}$	$p_l^{(08)}$	**ContrC**	**ContrR**
1	12.19	12.27	22.70	15.38	-37.91	-161.64
2	38.32	0.34	39.80	0.10	9.19	-0.14
3	6.85	25.36	7.03	29.36	-27.4	-5.28
4	18.01	7.90	32.05	7.94	-0.72	-111.47
5	71.61	10.74	55.86	2.28	605.82	35.91
6	37.13	4.73	21.00	0.62	152.60	10.00
7	22.79	9.11	20.33	15.49	-145.40	38.10
8	71.54	0.35	83.85	0.09	18.60	-1.10
9	29.63	14.47	23.23	11.50	88.00	73.60
10	27.55	0.71	30.32	1.04	-9.09	-2.88
11	26.15	2.30	35.22	3.54	-32.42	-32.10
12	28.63	2.68	30.16	3.98	-37.21	-6.08
13	64.18	1.43	56.66	0.18	80.22	1.35
14	69.63	1.64	50.79	2.08	-30.63	39.18
15	77.59	0.26	55.47	0.14	9.31	3.09
16	67.06	0.77	50.81	2.22	-97.23	36.07
17	48.89	0.48	49.98	0.66	-8.80	-0.71
18	79.77	0.95	83.06	0.07	70.19	-0.23
19	22.70	0.24	29.63	0.24	0.00	-1.66
20	51.67	0.43	36.97	0.48	-2.58	7.05
21	46.56	0.98	45.90	0.74	11.17	0.48
22	63.46	0.28	64.55	0.23	3.17	-0.25
23	61.15	1.19	65.32	0.99	12.23	-4.12
24	57.48	0.30	47.30	0.52	-12.64	5.29

Estimated frequencies and AROPE rates of each group, multiplied by 100. The last two columns shows the contribution of each group to composition (ContrC) and residual (ContrR) components.

group 16 consists of foreign-born individuals, also employed full-time or retired, with low education, and residing in households with dependent children. Interestingly, both groups exhibited a decrease in their incidence in 2014, contributing negatively to the composition component. Moreover, group 1 demonstrated a lower percentage of individuals in poverty in 2014 compared to 2008, indicating socio-economic dynamics that affected the groups differently across the two periods.

The reduction in the weight of group 16 provides evidence of the well-known migratory movements of foreign-born individuals to their country of origin during recession periods. However, the mechanisms behind group 1 are less clear. Potential explanations include generational effects, whereby the elderly in 2014 are more educated than those in 2008, and demographic changes, such as a smaller proportion of elderly individuals in 2014 compared to 2008.

On the other hand, group 5, comprising working-age individuals out of the labor market with low levels of education, experienced a substantial increase both in the weight

of the group across the entire population and in the percentage of individuals within this group facing AROPE conditions. Following the recession, the incidence of this group in the entire population surged from 2.28% in 2008 to 10.74% in 2014, with a staggering 71.61% experiencing AROPE conditions within the group in 2014. The intersection of these two factors resulted in the highest contribution to the composition component among all groups. In this case, it is evident that labor market perturbations consequent to the shock drove the variations in group frequency

7 Conclusion

We propose alternative counterfactual decomposition based on grouped data. This methodology does not require assuming a parametric model for the regression function and naturally provides a detailed decomposition of each group's contribution. Importantly, the decomposition ensures zero composition and residual effects, when the empirical distribution of the two populations is the same. When the data lacks natural clusters, a data-driven approach such as regression trees offers a valid alternative for grouping the data. We provide fixed-cell inference valid for a wide range of data-driven partitions. Monte Carlo evidence investigate the semiparametric nature of the estimator, showing good finite sample perfomance compared with other methods. In the final section, we apply the methodology to decompose the rise in poverty rates in Spain following the Great Recession. Our results indicate that variations in the composition of poverty risk factors, such as unemployment, explain most of the increase in AROPE rates, while the rise in material deprivation rates remains unexplained by variations in the composition of income-based poverty measures. The detailed decomposition reveals that perturbations in the distribution of the labor force within groups exacerbated poverty, while demographic changes worked in the opposite direction.

A Appendix

The appendix is organized as follows. In Section A, we provide the asymptotic distribution of the decomposition components $\hat{\Delta}^C_{\mathbb{C}}$ and $\hat{\Delta}^R_{\mathbb{C}}$, along with the relative extension to data-dependent partitions. Section B contains the proofs of all propositions and theorems in the paper. Section C presents the codification of the poverty-risk factors used in the empirical application. Finally, in Section D, we provide details about the cross-validation procedure used by CART.

Proposition 2 ***(Asymptotic Distribution*** $\hat{\Delta}^C_{\mathbb{C}}$***)***

Let $\hat{\Delta}^C_{\mathbb{C}} = \hat{A}^{(1)} - \hat{A}^{(1,0)}_{\mathbb{C}}$ *be the estimator of the composition component, under assumption 1, 2 (a):*

$$\sqrt{n}\left(\hat{\Delta}^C_{\mathbb{C}} - \Delta^C_{\mathbb{C}}\right) \xrightarrow{d} N(0, V_{\Delta_C}),$$

$$V_{\Delta_C} = \sum_{l=1}^{L}(p_l^{(1)} - p_l^{(0)})^2\sigma_l^{(1)} + \sum_{l=1}^{L}\sum_{f=1}^{L} h_l^{(1)} v_{(l,f)} h_f^{(1)} - 2\sum_{l=1}^{L}\sum_{f=1}^{L} h_l^{(1)} c_{(l,f)}(p_f^{(1)} - p_f^{(0)}) + \\ + 2\sum_{l=1}^{L}\sum_{f=1}^{L} h_l^{(1)}\left(p_{(l,f)} - p_l^{(1)}p_f^{(0)}\right) h_f^{(1)}$$

where,

$$v_{(l,f)} = \begin{cases} p_l^{(1)}(1-p_l^{(1)}) + p_l^{(0)}(1-p_l^{(0)}) & \text{if } l = f \\ -(p_l^{(1)}p_f^{(1)} + p_l^{(0)}p_f^{(0)}) & \text{if } l \neq f \end{cases},$$

and $p^{(1,0)}_{(l,f)}$ *and* $c_{(l,f)}$ *are defined as in proposition 1. Under Assumptions 1, and 2 (b), instead,*

$$\sqrt{n_1}\left(\hat{\Delta}^C_{\mathbb{C}} - \Delta^C_{\mathbb{C}}\right) \xrightarrow{d} N(0, V_{\Delta_C}),$$

$$V_{\Delta_C} = \sum_{l=1}^{L}(p_l^{(1)} - p_l^{(0)})^2\sigma_l^{(1)} + \frac{1}{\gamma}\sum_{l=1}^{L}\sum_{f=1}^{L} h_l^{(1)} v_{(l,f)} h_f^{(1)}$$

Proposition 3 ***(Asymptotic Distribution*** $\hat{\Delta}^R_{\mathbb{C}}$***)***

Let $\hat{\Delta}^R_{\mathbb{C}} = \hat{A}^{(1,0)} - \hat{A}^{(0)}_{\mathbb{C}}$ *be the estimator of the composition component, under assumption*

1, 2 (a):

$$\sqrt{n}\left(\hat{\Delta}_{\mathbb{C}}^{R}-\Delta_{\mathbb{C}}^{R}\right)\xrightarrow{d}N(0,V_{\Delta_R})$$

$$\begin{aligned}V_{\Delta_R}=&\sum_{l=1}\left(p_l^{(0)}\right)^2\left(\sigma_l^{(1)}+\sigma_l^{(0)}\right)+\sum_{l=1}^{L}\left(h_l^{(1)}-h_l^{(0)}\right)p_l^{(0)}(1-p_l^{(0)})\left(h_f^{(1)}-h_f^{(0)}\right)+\\&-\sum_{l\neq f}\left(h_l^{(1)}-h_l^{(0)}\right)(p_l^{(0)}p_f^{(0)})\left(h_f^{(1)}-h_f^{(0)}\right)+2\sum_{l=1}^{L}\sum_{f=1}^{L}\left(h_l^{(1)}-h_l^{(0)}\right)c_{(l,f)}p_f^{(0)}-\\&-2\sum_{l=1}^{L}\sum_{f=1}^{L}\frac{p_f^{(0)}}{p_f^{(1)}}\left(q_{(l,f)}^{(1,0)}-h_l^{(0)}q_{(f,l)}^{(1)}-h_f^{(1)}q_{(l,f)}^{(0)}+h_l^{(0)}h_f^{(1)}p_{(l,f)}\right)\end{aligned}$$

where $q_{(l,f)}^{(1,0)}=\mathbb{E}\left[Y^{(1)}Y^{(0)}\mathbb{I}\{X^{(1)}\in C_l\}\mathbb{I}\{X^{(0)}\in C_f\}\right]$. *Under Assumptions 1, and 2 (b),*

$$\sqrt{n_1}\left(\hat{\Delta}_{\mathbb{C}}^{R}-\Delta_{\mathbb{C}}^{R}\right)\xrightarrow{d}N(0,V_{\Delta_R}),$$

with

$$\begin{aligned}V_{\Delta_R}=&\sum_{l=1}\left(p_l^{(0)}\right)^2\left(\sigma_l^{(1)}+\frac{\sigma_l^{(0)}}{\gamma}\right)+\sum_{l=1}^{L}\left(h_l^{(1)}-h_l^{(0)}\right)\frac{p_l^{(0)}(1-p_l^{(0)})}{\gamma}\left(h_f^{(1)}-h_f^{(0)}\right)+\\&-\sum_{l\neq f}\left(h_l^{(1)}-h_l^{(0)}\right)(p_l^{(0)}p_f^{(0)})\left(h_f^{(1)}-h_f^{(0)}\right).\end{aligned}$$

Corollary 1 (Asymptotic Distribution $\hat{\Delta}_{\hat{\mathbb{C}}}^{C}$ and $\hat{\Delta}_{\hat{\mathbb{C}}}^{R}$ with Random Partition) *Under Assumptions 1, 2, 3, and 4,*

$$\sqrt{n}\left(\hat{\Delta}_{\hat{\mathbb{C}}}^{C}-\Delta_{\mathbb{C}}^{C}\right)\xrightarrow{d}N\left(0,V_{\Delta_C}\right)$$

$$\sqrt{n}\left(\hat{\Delta}_{\hat{\mathbb{C}}}^{R}-\Delta_{\mathbb{C}}^{R}\right)\xrightarrow{d}N\left(0,V_{\Delta_R}\right)$$

where V_{Δ_C} *and* V_{Δ_R} *are defined as in Propositions 2 and 3, respectively.*

B Proofs

Proof of Proposition 1. Under Assumption 2 (a), $n_1=n_0=n$. Let $\hat{q}_l^{(1)}=\frac{1}{n}\sum_{i=1}^{n}g(Y_i^{(1)})\mathbb{I}\{X_i^{(1)}\in C_l\}$ such that $\hat{h}_l^{(1)}=\hat{q}_l^{(1)}/\hat{p}_l^{(1)}$ for all l, and let $\hat{q}^{(j)}=(\hat{q}_1^{(j)},...,\hat{q}_L^{(j)})$,

$\hat{h}^{(j)} = (\hat{h}_1^{(j)}, ..., \hat{h}_L^{(j)})$, and $\hat{p}^{(j)} = (\hat{p}_1^{(j)}, ..., \hat{p}_L^{(j)})$ for $j = 0, 1$. Furthermore, denote as $q^{(j)}$, $h^{(j)}$, and $p^{(j)}$ as the respective probability limit vectors, and define as $d(x)$ the diagonal matrix with main diagonal elements given by x, for any $\mathbb{R}^L$-valued vector x. Then, the vector:

$$\sqrt{n}\begin{bmatrix}\hat{h}^{(1)} - h^{(1)} \\ \hat{p}^{(0)} - p^{(0)}\end{bmatrix} = \sqrt{n}\begin{bmatrix} d(\hat{p}^{(1)}) & -d(\hat{p}^{(1)})^{-1}d(h^{(1)}) & 0 \\ 0 & 0 & I_L \end{bmatrix}\begin{bmatrix}\hat{q}^{(1)} - q^{(1)} \\ \hat{p}^{(1)} - p^{(1)} \\ \hat{p}^{(0)} - p^{(0)}\end{bmatrix}$$

has the following limit distribution:

$$\sqrt{n}\begin{bmatrix}\hat{h}^{(1)} - h^{(1)} \\ \hat{p}^{(0)} - p^{(0)}\end{bmatrix} \xrightarrow{d} N(0, V) \text{ with } V = \begin{bmatrix} V_h^{(1)} & V_c \\ V_c' & V_p^{(0)}\end{bmatrix},$$

$$V_h^{(1)} = \begin{bmatrix} \sigma_1^{(1)} & 0 & \cdots & 0 \\ 0 & \sigma_2^{(1)} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_L^{(1)}\end{bmatrix} V_p^{(0)} = \begin{bmatrix} p_1^{(0)}(1-p_1^{(0)}) & -p_1^{(0)}p_2^{(0)} & \cdots & -p_1^{(0)}p_L^{(0)} \\ -p_2^{(0)}p_1^{(0)} & p_2^{(0)}(1-p_2^{(0)}) & \cdots & -p_2^{(0)}p_L^{(0)} \\ \vdots & \vdots & \ddots & \vdots \\ -p_L^{(0)}p_1^{(0)} & -p_2^{(0)}p_L^{(0)} & \cdots & p_L^{(0)}(1-p_L^{(0)}),\end{bmatrix}$$

and

$$V_c = Cov\left(\sqrt{n}\hat{p}^{(0)}, \sqrt{n}\hat{q}^{(1)}\right) d(p^{(1)})^{-1} - Cov\left(\sqrt{n}\hat{p}^{(0)}, \sqrt{n}\hat{p}^{(1)}\right) d(p^{(1)})^{-1}d(h^{(1)}),$$

is an $L \times L$ matrix with typical elements

$$\left\{c_{(l,f)}\right\}_{l,f=1}^{L} = \left\{\frac{q_{(l,f)}^{(1)} - h_f^{(1)}p_{(l,f)}^{(0,1)}}{p_f^{(1)}}\right\}_{l,f=1}^{L}.$$

An application of the delta method to $\mu_{Y,\mathbb{C}}^{(1,0)} = h^{(1)'}p^{(0)}$ gives the result,

$$\sqrt{n_1}\left(\hat{\mu}_{Y,\mathbb{C}}^{(1,0)} - A_{\mathbb{C}}^{(1,0)}\right) \xrightarrow{d} N\left(0, [p^{(0)'}, h^{(1)'}]V[p^{(0)'}, h^{(1)'}]'\right) = N\left(0, V_A\right).$$

Under Assumption 2 (b), instead,

$$\sqrt{n_1}\begin{bmatrix}\hat{h}^{(1)}-h^{(1)}\\ \hat{p}^{(0)}-p^{(0)}\end{bmatrix}\xrightarrow{d}N(0,V)\ \text{ with }\ V=\begin{bmatrix}V_h^{(1)} & 0\\ 0 & \frac{V_p^{(0)}}{\gamma}\end{bmatrix},$$

and, thus, $V_A=\sum_{l=1}^{L}\left((p_l^{(0)})\sigma_l^{(1)}+\frac{1}{\gamma}(h_l^{(1)})^2p_l^{(0)}(1-p_l^{(0)})\right)-\frac{1}{\gamma}\sum_{l\neq f}h_l^{(1)}h_f^{(1)}p_l^{(0)}p_f^{(0)}$. ■

Proof of Proposition 2. Under Assumption 2 (a),

$$\sqrt{n}\begin{bmatrix}\hat{h}^{(1)}-h^{(1)}\\ \hat{p}^{(0)}-p^{(0)}\\ \hat{p}^{(1)}-p^{(1)}\end{bmatrix}=\sqrt{n}\begin{bmatrix}|\hat{p}^{(1)}| & -|\hat{p}^{(1)}|^{-1}|h^{(1)}| & 0\\ 0 & 0 & I_L\\ 0 & I_L & 0\end{bmatrix}\begin{bmatrix}\hat{q}^{(1)}-q^{(1)}\\ \hat{p}^{(1)}-p^{(1)}\\ \hat{p}^{(0)}-p^{(0)}\end{bmatrix}$$

converges to a normal distribution with mean 0 and variance V, $N(0,V)$, where

$$V=\begin{bmatrix}V_h^{(1)} & V_c & 0\\ V_c' & V_p^{(0)} & V_p^{(1,0)}\\ 0 & V_p^{(1,0)'} & V_p^{(1)}\end{bmatrix}.$$

Here, V_c is as in the proof of Proposition 1, while $V_p^{(1,0)}$ is an $L\times L$ matrix with typical elements given by $\{p_{(l,f)}-p_l^{(1)}p_f^{(0)}\}_{l,f=1}^{L}$. After noticing that $\Delta_{\mathbb{C}}^{C}=h^{(1)'}\left(p^{(1)}-p^{(0)}\right)$, the result follows from the delta method,

$$\sqrt{n}\left(\hat{\Delta}_{\mathbb{C}}^{C}-\Delta_{\mathbb{C}}^{C}\right)\xrightarrow{d}N(0,V_{\Delta_C}),$$

where

$$V_{\Delta_C}=(p^{(1)}-p^{(0)})'V_h^{(1)}+h^{(1)'}\left(V_p^{(1)}+V_p^{(0)}\right)+h^{(1)'}\left(V_p^{(1,0)}+V_p^{(1,0)'}\right)h^{(1)}-2h^{(1)'}V_c'(p^{(1)}-p^{(0)}).$$

Under Assumption 2 (b), instead,

$$\sqrt{n_1}\begin{bmatrix}\hat{h}^{(1)}-h^{(1)}\\ \hat{p}^{(0)}-p^{(0)}\\ \hat{p}^{(1)}-p^{(1)}\end{bmatrix}\xrightarrow{d}N(0,V)$$

with

$$V=\begin{bmatrix}V_h^{(1)} & 0 & 0\\ 0 & \frac{V_p^{(0)}}{\gamma} & 0\\ 0 & 0 & V_p^{(1)}\end{bmatrix}.$$

And it follows that,

$$V_{\Delta_C}=\Delta f' V \Delta f=\left(p^{(1)}-p^{(0)}\right)' V_h^{(1)}\left(p^{(1)}-p^{(0)}\right)+\frac{1}{\gamma}h^{(1)'}\left(\gamma V_p^{(1)}+V_p^{(0)}\right)h^{(1)}.$$

■

Proof of Proposition 3. Under Assumption 2 (a),

$$\sqrt{n}\begin{bmatrix}\hat{h}^{(1)}-h^{(1)}\\ \hat{h}^{(0)}-h^{(0)}\\ \hat{p}^{(0)}-p^{(0)}\end{bmatrix}=\sqrt{n}\begin{bmatrix}|\hat{p}^{(1)}| & -|\hat{p}^{(1)}|^{-1}|h^{(1)}| & 0 & 0\\ 0 & 0 & |\hat{p}^{(0)}| & -|\hat{p}^{(0)}|^{-1}|h^{(0)}|\\ 0 & 0 & 0 & I_L\end{bmatrix}\begin{bmatrix}\hat{q}^{(1)}-q^{(1)}\\ \hat{p}^{(1)}-p^{(1)}\\ \hat{q}^{(0)}-q^{(0)}\\ \hat{p}^{(0)}-p^{(0)}\end{bmatrix}$$

converges to a normal distribution with mean 0 and variance V, $N(0,V)$,

$$V=\begin{bmatrix}V_h^{(1)} & V_{c_2} & V_c\\ V_{c_2}' & V_h^{(0)} & 0\\ V_c' & 0 & V_p^{(1)}\end{bmatrix}.$$

where V_{c_2} is an $L\times L$ matrix with typical elements given by

$$\left\{q_{(l,f)}^{(1,0)}-h_l^{(0)}q_{(f,l)}^{(1)}-h_f^{(1)}q_{(l,f)}^{(0)}+h_l^{(0)}h_f^{(1)}\mu_{(l,f)}\right\}_{l,f=1}^{L}.$$

Then, by the delta method,

$$\sqrt{n}\left(\hat{\Delta}_{\mathbb{C}}^{R}-\Delta_{\mathbb{C}}^{R}\right)\xrightarrow{d}N(0,V_{\Delta^R}),$$

where

$$V_{\Delta^R}=p^{(0)\prime}\left(V_h^{(1)}+V_h^{(0)}-V_{c_2}-V_{c_2}'\right)p^{(0)}+2(h^{(1)}-h^{(0)})'V_c'p^{(0)}+\left(h^{(1)}-h^{(0)}\right)'V_{p^{(0)}}\left(h^{(1)}-h^{(0)}\right).$$

Under Assumption 2(b),

$$\sqrt{n}\begin{bmatrix}\hat{h}^{(1)}-h^{(1)}\\ \hat{h}^{(0)}-h^{(0)}\\ \hat{p}^{(0)}-p^{(0)}\end{bmatrix}\xrightarrow{d}N(0,V),$$

with

$$V=\begin{bmatrix}V_h^{(1)} & 0 & 0\\ 0 & \frac{V_h^{(0)}}{\gamma} & 0\\ 0 & 0 & \frac{V_p^{(1)}}{\gamma}\end{bmatrix}.$$

and

$$V_{\Delta^R}=p^{(0)\prime}\left(V_h^{(1)}+\frac{V_h^{(0)}}{\gamma}\right)p^{(0)}+\left(h^{(1)}-h^{(0)}\right)'\frac{V_p^{(0)}}{\gamma}\left(h^{(1)}-h^{(0)}\right).$$

■

Proof of Theorem 1 and Corollary 1. Let $\rightsquigarrow$ denote weak convergence on $l^\infty(\mathcal{C})$ (see definition 13.3 in Van Der Vaart 1996, hereafter VW), where $l^\infty(\mathcal{C})$ is the space of all real-valued functions that are uniformly bounded on $\mathcal{C}$. For any class of functions $\mathcal{F}$, denote as $\{P_nf:f\in\mathcal{F}\}$ the empirical measure indexed by $\mathcal{F}$, such that $P_nf=n^{-1}\sum f(Z_i)$; alike, Pf denotes the population measure, $Pf=\int f(Z)dP$. We say that a class of functions is: i) Glivenko-Cantelli for P (hereafter, P-GC) whenever $\sup_{f\in\mathcal{F}}|P_n-P|f=o_p(1)$; ii) P-Donsker if $\{\sqrt{n}(P_n-P)f:f\in\mathcal{F}\}$ converge in distribution to a tight random element in the space $l^\infty(\mathcal{F})$. Throughout, we refer to both classes of sets with finite VC dimension and classes of functions with finite VC subgraph dimension as VC classes. These classes, having uniformly bounded covering numbers (Theorem 2.6.7 in VW), are Glivenko-Cantelli and Donsker (see Theorem 2.4.3 and 2.5.2 in VW)

for any probability measure on the sample space, provided that they have integrable and square-integrable envelope function, respectively. Below, we provide the proof for Theorem 1 under the data structure of Assumption 2 (a). The proof for both Theorem 1 and Corollary 1 under Assumption 2 (b) are similar and not reported. In the following, $\mathbb{I}_{\mathbb{C}}(X) = (\mathbb{I}\{X \in C_1\}, ..., \mathbb{I}\{X \in C_L\})'$, denotes the vector of indicator functions over $\mathbb{C}$.

Let $\hat{\varepsilon}(\mathbb{C}) = (\hat{\varepsilon}'_q(\mathbb{C}), \hat{\varepsilon}'_{\mu_1}(\mathbb{C}), \hat{\varepsilon}'_{p_0}(\mathbb{C}))'$, where $\hat{\varepsilon}_q(\mathbb{C}) = (P_n - P)g(y)\mathbb{I}\{j = 1\}\mathbb{I}_{\mathbb{C}}(x)$, $\hat{\varepsilon}_{p_1}(\mathbb{C}) = (P_n - P)\mathbb{I}\{j = 1\}\mathbb{I}_{\mathbb{C}}(x)$, and $\hat{\varepsilon}_{p_0}(\mathbb{C}) = (P_n - P)\mathbb{I}\{j = 0\}\mathbb{I}_{\mathbb{C}}(x)$ are empirical processes indexed by partitions in $\mathcal{C}$. By the CLT, the finite-dimensional distributions of $\sqrt{n}\hat{\varepsilon}(\mathbb{C})$ are the same of $\sqrt{n}(\hat{q}^{(1)} - q^{(1)}, \hat{p}^{(1)} - p^{(1)}, \hat{p}^{(0)} - p^{(0)})$. Also, by Lemma 2.6.17 and 2.6.18 in VW, and Assumption 4, $\mathcal{F}_q = \{y\mathbb{I}\{j = 1\}\mathbb{I}_{\mathbb{C}}(x) : \mathbb{C} \in \mathcal{C}\}$, $\mathcal{F}_{p_1} = \{\mathbb{I}\{j = 1\}\mathbb{I}_{\mathbb{C}}(x) : \mathbb{C} \in \mathcal{C}\}$, and $\mathcal{F}_{p_0} = \{\mathbb{I}\{j = 0\}\mathbb{I}_{\mathbb{C}}(x) : \mathbb{C} \in \mathcal{C}\}$ are VC classes with square integrable envelope, $F_q = |y|$, $F_{p_1} = |1|$, and $F_{p_0} = |1|$, respectively. It follows that, $\mathcal{F} = \mathcal{F}_q \times \mathcal{F}_{p_1} \times \mathcal{F}_{p_0} = \{(y\mathbb{I}_{\mathbb{C}}(x), \mathbb{I}\{j = 1\}\mathbb{I}_{\mathbb{C}}(x), \mathbb{I}\{j = 0\}\mathbb{I}_{\mathbb{C}}(x))' : \mathbb{C} \in \mathcal{C}\}$ is also VC and P-Donsker. Thus,

$$\sqrt{n}\hat{\varepsilon}(\cdot) \rightsquigarrow \varepsilon_0(\cdot) \text{ as a process on } l^\infty(\mathcal{C}),$$

where $\varepsilon_0(\cdot)$ is an $\mathbb{R}^{3L}$-valued Gaussian process with zero mean vector and covariance structure given by,

$$\mathbb{E}\left[H(Y,j)\mathbb{I}_{\mathbb{C}_1}(X)\mathbb{I}_{\mathbb{C}_2}(X)'H(Y,j)\right] - \mathbb{E}\left[H(Y,j)\mathbb{I}_{\mathbb{C}_2}(X)\right]\mathbb{E}\left[H(Y,j)\mathbb{I}_{\mathbb{C}_2}(X)\right]' \quad \forall \mathbb{C}_1, \mathbb{C}_2 \in \mathcal{C},$$

with $H(y,j) = (g(y), \mathbb{I}\{j = 1\}, \mathbb{I}\{j = 0\})' \otimes I_L$ and $\otimes$ being the Kronecker product.

Furthermore, by Assumption 4, both $\mathcal{C}\Delta\mathcal{C} = \{\mathbb{C}_1\Delta\mathbb{C}_2 : \mathbb{C}_1, \mathbb{C}_2 \in \mathcal{C}\}$ and $(\mathcal{C}\Delta\mathcal{C})^{(1)} = \{\mathbb{I}\{j = 1\}\mathbb{I}_{\tilde{C}}(x) : \tilde{C} \in \mathcal{C}\Delta\mathcal{C}\}$ are VC classes, where Δ is the symmetric difference operator. Therefore, by Assumption 3,

$$|\hat{p}^{(1)}(\hat{\mathbb{C}}) - \hat{p}^{(1)}(\mathbb{C}) \leq \sup_{\tilde{C} \in \mathcal{C}\Delta\mathcal{C}} |P_n - P|\mathbb{I}_{\tilde{C}}(x)\mathbb{I}\{j = 1\} + \mathbb{E}\left[\mathbb{I}\{j = 1\}\mathbb{I}_{\hat{\mathbb{C}}\Delta\mathbb{C}}(X)\right] = o_p(1),$$

where $\hat{p}^{(1)}(\mathbb{C}) = P_n\mathbb{I}\{j = 1\}\mathbb{I}_{\mathbb{C}}(x)$ highlights the dependence of the empirical measure on the partition.

Finally, by Assumptions 3, 4, and the uniform continuity of the sample paths of $\varepsilon_0(\cdot)$,

$$\sqrt{n}\left(\hat{\mu}^{(1,0)}_{Y,\hat{\mathbb{C}}} - \mu^{(1,0)}_{Y,\mathbb{C}}\right) = \begin{bmatrix} d(\hat{p}^{(1)}(\hat{\mathbb{C}})) & -d(\hat{p}^{(1)}(\hat{\mathbb{C}}))^{-1}d(h^{(1)}(\hat{\mathbb{C}})) & 0 \\ 0 & 0 & I_L \end{bmatrix} \hat{\varepsilon}(\hat{\mathbb{C}})$$
$$\stackrel{d}{=} \begin{bmatrix} d(\hat{\mu}^{(1)}(\mathbb{C})) & -d(\hat{\mu}^{(1)}(\mathbb{C}))^{-1}d(h^{(1)}(\mathbb{C})) & 0 \\ 0 & 0 & I_L \end{bmatrix} \varepsilon_0(\mathbb{C}) + o_p(1).$$

■

C Codification of the Poverty Risk Factors

Nat & Cb

The variable *Nat* indicates the nationality of the individual, while *Cb* the country of birth. They take the following values: Spain (1), rest of European Union (EU28) (2), rest of the world (3).

Educ

The variable *Educ* denotes the education level: Less than primary (0), primary education (1), first-stage secondary education (2), second-stage secondary education (3), post-secondary education (4), higher education (5).

Ws

The variable *Ws* denotes the working status of the individual: employed (full-time) (1), employed (part-time) (2), inactive (3), student or in formation (4), retired, early retired or have closed down a business (5), permanently unable to work (6), compulsory military service or substitute social service (7), dedicated to housework, care of children or other persons (8), other class of economic inactivity (9).

HHtype

The variable *HHtype* denotes the type of household where the individual lives: 1 Adult: Male $<$ 30 years old (1), 1 Adult: 30 $\geq$ Male $\leq$ 64 years old (2), 1 Adult: Male $\geq$ 65 years old (3), 1 Adult: Female $<$ 30 years old (4), 1 Adult: 30 $\leq$ Female $\leq$ 64 years old (5), 1 Adult: Female $\geq$ 65 years old (6), 2 Adults without financially dependent children[2],

[2]Note, the classification of financially dependent children includes all those under the age of 18 and those who are 18 and older but under 25 and economically inactive.

at least one person above 65 years of age (7), 2 Adults without financially dependent children, both below 65 years of age (8), other type of household without financially dependent children (9), 1 Adult with at least a dependent child (10), 2 Adults with a dependent child (11), 2 Adults with 2 dependent children (12), 2 Adults with 3 or more dependent children (13), other type of household with dependent children (14).

HHsyze

The variable *HHsyze* denotes the size of the household. It can take values from 1 to 14.

HHhage

The variable *HHsyze* denotes the age of the household head. It can take value from 25 to 99.

Sex

The variable *Sex* denotes the gender of the individual: Male =1; Female=2.

D CART

The algorithm consists of two parts, initial tree building and cross-validation phase. In the initial tree building phase, the algorithm recursively partitions the data according to a in-sample goodness-of-fit measure and a penalty term, which regulates the depth of the tree. In the cross-validation phase, the algorithm determines the optimal penalty term by repeatedly splitting the sample into a training sample and a cross-validation sample. The training sample is used to generate a partition and to estimate the conditional mean function; the cross-validation sample is used to evaluate the estimates based on the partition generated by the training sample. The optimal penalty term maximizes a goodness-of-fit criterion in cross-validation samples. In the grouping procedure we use the sample of population (1) to generate the splitting rules, which are then applied to split the sample of both populations.

Initial Tree-Building Phase

Let denote a partition as $\mathbb{C}_{L_t}^{CART}$, where the sub-script L_t indicates the dependency of the number of set in each partition from the t-th iteration of the algorithm. Fix a given iteration t and partition $\mathbb{C}_{L_t}^{CART}$, then for any $C_l \in \mathbb{C}_{L_t}^{CART}$ the algorithm selects a threshold c and one of the covariates X_{ik} to split C_l into two smaller sets by minimizing the following

objective function:

$$Q_{C_l}(k,c) \text{ subject to } Q_{C_l} \geq \lambda L_t + Q_{C_l}(k,c)$$

Where Q_{C_l} and $Q_{C_l}(k,c)$ denote the in-sample goodness-of-fit before and after the split respectively, such that:

$$Q_{C_l} = \sum_{i:X_i^{(1)} \in C_l} \left(Y_i^{(1)} - \hat{h}^{(1)}(C_l)\right)^2$$

And

$$Q_{C_l}(k,c) = \sum_{i:X_i^{(1)} \in C_l, X_{ik}^{(1)} \leq c} \left(Y_i^{(1)} - \hat{h}_l^{(1)}(C_l,c)\right)^2 + \sum_{i:X_i^{(1)} \in C_l, X_{ik}^{(1)} > c} \left(Y_i^{(1)} - \hat{h}_r^{(1)}(C_l,c)\right)^2$$

Where $\hat{h}_r^{(1)}(C_l,c) = \frac{\sum_{i=1}^{n_1} g(Y_i^{(1)}) \mathbb{I}\{X_i^{(1)} \in C_l\}}{\sum_{i=1}^{n_1} \mathbb{I}\{X_i^{(j)} \in C_l : X_{ik}^{(1)} \leq c\}}$ and $\hat{h}_r^{(1)}(C_l,c) = \frac{\sum_{i=1}^{n_1} g(Y_i^{(1)}) \mathbb{I}\{X_i^{(1)} \in C_l\}}{\sum_{i=1}^{n_1} \mathbb{I}\{X_i^{(j)} \in C_l : X_{ik}^{(1)} > c\}}$.

The penalty term λL_t punishes the partitions with too many classes. For a split to be done, the splitting point c on the covariate X_{ik} should improve the in-sample goodness-of-fit criterion by at least λL_t. When X_{ik} is a factor variable (non-ordered or non-numerical), the split is done by selecting two subsets of categories of X_{ik}, rather than dividing above and below the threshold c.

Cross-Validation Phase

Let $S^{(1)}$ be the sample of population (1), and denote the partition obtained in the initial tree building phase as $\mathbb{C}^{\text{CART}}(S^{(1)}, \lambda)$. The next step is to search over a grid of values $\Lambda_M = \{\lambda_1, ..., \lambda_M\}$ the λ that maximizes the goodness-of-fit criterion in cross-validation samples. In this step, we drop the dependence of L_t from the number of iteration since it is not relevant for the discussion and consider it as a function of the sample and the penalty term; that is, $L = L(S^{(1)}, \lambda)$. Common cross-validation practices include the k-fold cross-validation and the leave-one-out cross-validation. Without loss of generality, we discuss the latter. For each $\lambda \in \Lambda_M$ and $i \in \{1, 2, ..., n_1\}$ we split $S^{(1)}$ into a training sample $S_{-i}^{(1)}$ including all the observations in $S^{(1)}$ but i and a cross-validation sample $\{i\}$. For

each training sample $S_{-i}^{(1)}$ we generate a partition $\mathbb{C}^{\text{CART}}(S_{-i}^{(1)}, \lambda) = \{C_l(S_{-i}^{(1)}, \lambda)\}_{l=1}^{L(S_{-i}^{(1)}, \lambda)}$ and estimate the conditional mean in the leaf (set) of $\mathbb{C}^{\text{CART}}(S_{-i}^{(1)}, \lambda)$ where the cross-validation sample lies. That is:

$$\hat{h}^{(1)}\left(X_i^{(1)}, \mathbb{C}(S_{-i}^{(1)}, \lambda)\right) = \sum_{l=1}^{L(S_{-i}^{(1)}, \lambda)} \frac{\sum_{j \neq i} Y_i^{(1)} \mathbb{I}\{X_j^{(1)} \in C_l(S_{-i}^{(1)}, \lambda)\}}{\sum_{j \neq i} \mathbb{I}\{X_j^{(1)} \in C_l(S_{-i}^{(1)}, \lambda)\}} \mathbb{I}\{X_i^{(1)} \in C_l(S_{-i}^{(1)}, \lambda)\}$$

We calculate the error on observation i: $e_i^{(1)}(\lambda) = Y_i^{(1)} - \hat{h}^{(1)}\left(X_i^{(1)}, \mathbb{C}^{\text{CART}}(S_{-i}^{(1)}, \lambda)\right)$ and choose the λ^* that minimize the average mean square error over all n_1 observations:

$$\lambda^* = \underset{\lambda \in \Lambda_M}{\operatorname{argmin}} \sum_{i=1}^{n_1} \left[e_i^{(1)}(\lambda)\right]^2$$

Finally, we get our chosen partition $\mathbb{C}_L^{\text{CART}} = \mathbb{C}^{\text{CART}}(S^{(1)}, \lambda^*)$ and use (12) to estimate the counterfactual mean. All the above procedure using CART can be easily implemented using the R-package *Rpart*.

References

Andrews, D. W. (1988): "Chi-square diagnostic tests for econometric models: theory," *Econometrica: Journal of the Econometric Society*, pp. 1419–1453.

Athey, S., and S. Wager (2021): "Policy learning with observational data," *Econometrica*, 89(1), 133–161.

Ayala, L., A. Jurado, and J. Pérez-Mayo (2011): "Income poverty and multidimensional deprivation: Lessons from cross-regional analysis," *Review of income and wealth*, 57(1), 40–60.

——— (2021): "Multidimensional deprivation in heterogeneous rural areas: Spain after the economic crisis," *Regional Studies*, 55(5), 883–893.

Barsky, R., J. Bound, K. K. Charles, and J. P. Lupton (2002): "Accounting for the black–white wealth gap: a nonparametric approach," *Journal of the American statistical Association*, 97(459), 663–673.

Bauer, T., S. Göhlmann, and M. Sinning (2007): "Gender differences in smoking behavior," *Health Economics*, 16(9), 895–909.

Biewen, M., and S. P. Jenkins (2005): "A framework for the decomposition of poverty differences with an application to poverty differences between countries," *Empirical Economics*, 30(2), 331–358.

Black, D. A., A. M. Haviland, S. G. Sanders, and L. J. Taylor (2008): "Gender wage disparities among the highly educated," *Journal of human resources*, 43(3), 630–659.

Blinder, A. S. (1973): "Wage discrimination: reduced form and structural estimates," *Journal of Human resources*, pp. 436–455.

Bourguignon, F., and F. Ferreira (2005): "Decomposing changes in the distribution of household incomes: methodological aspects," *The microeconomics of income distribution dynamics in East Asia and Latin America*, pp. 17–46.

Bourguignon, F., F. H. Ferreira, and P. G. Leite (2008): "Beyond Oaxaca–Blinder: Accounting for differences in household income distributions," *The Journal of Economic Inequality*, 6(2), 117–148.

Breiman, L., J. Friedman, R. Olshen, and C. Stone (1984): "Cart," *Classification and Regression Trees; Wadsworth and Brooks/Cole: Monterey, CA, USA*.

Brunori, P., V. Peragine, and L. Serlenga (2019): "Upward and downward bias when measuring inequality of opportunity," *Social Choice and Welfare*, 52(4), 635–661.

Cattaneo, M. D., and M. H. Farrell (2013): "Optimal convergence rates, Bahadur representation, and asymptotic normality of partitioning estimators," *Journal of Econometrics*, 174(2), 127–143.

Chernozhukov, V., I. Fernández-Val, and B. Melly (2013): "Inference on counterfactual distributions," *Econometrica*, 81(6), 2205–2268.

Chiquiar, D., and G. H. Hanson (2005): "International migration, self-selection, and the distribution of wages: Evidence from Mexico and the United States," *Journal of political Economy*, 113(2), 239–281.

DiNardo, J., N. M. Fortin, and T. Lemieux (1996): "Labor market institutions and the distribution of wages, 1973-1992: A semiparametric approach," *Econometrica: Journal of the Econometric Society*, pp. 1001–1044.

Fortin, N., T. Lemieux, and S. Firpo (2011): "Decomposition methods in economics," in *Handbook of labor economics*, vol. 4, pp. 1–102. Elsevier.

Gessaman, M. (1970): "A consistent nonparametric multivariate density estimator based on statistically equivalent blocks," *The Annals of Mathematical Statistics*, 41(4), 1344–1346.

Gradín, C. (2012): "Poverty among minorities in the United States: Explaining the racial poverty gap for Blacks and Latinos," *Applied Economics*, 44(29), 3793–3804.

Hastie, T., R. Tibshirani, and J. Friedman (2009): "The elements of statistical learnin," *Cited on*, p. 33.

Kitagawa, E. M. (1955): "Components of a difference between two rates," *Journal of the american statistical association*, 50(272), 1168–1194.

Machado, J. A., and J. Mata (2005): "Counterfactual decomposition of changes in wage distributions using quantile regression," *Journal of applied Econometrics*, 20(4), 445–465.

MacQueen, J., et al. (1967): "Some methods for classification and analysis of multivariate observations," in *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, vol. 1, pp. 281–297. Oakland, CA, USA.

Melly, B. (2005): "Decomposition of differences in distribution using quantile regression," *Labour economics*, 12(4), 577–590.

Neison, F. (1844): "On a method recently proposed for conducting inquiries into the comparative sanatory condition of various districts, with illustrations, derived from numerous places in Great Britain at the period of the last census," *Journal of the Statistical Society of London*, 7(1), 40–68.

Ñopo, H. (2008): "Matching as a tool to decompose wage gaps," *The review of economics and statistics*, 90(2), 290–299.

Oaxaca, R. (1973): "Male-female wage differentials in urban labor markets," *International economic review*, pp. 693–709.

Oaxaca, R. L., and M. R. Ransom (1999): "Identification in detailed wage decompositions," *Review of Economics and Statistics*, 81(1), 154–157.

Pollard, D. (1979): "General chi-square goodness-of-fit tests with data-dependent cells," *Zeitschrift fur Wahrscheinlichkeitstheorie und verwandte Gebiete*, 50(3), 317–331.

——— (1984): *Convergence of stochastic processes*. Springer Science & Business Media.

——— (1990): "Empirical processes: theory and applications," Ims.

Racine, J., and Q. Li (2004): "Nonparametric estimation of regression functions with both categorical and continuous data," *Journal of Econometrics*, 119(1), 99–130.

Rothe, C. (2010): "Nonparametric estimation of distributional policy effects," *Journal of Econometrics*, 155(1), 56–70.

Stock, J. H. (1989): "Nonparametric policy analysis," *Journal of the American Statistical Association*, 84(406), 567–575.

Van Der Vaart, J. W. (1996): *Weak convergence*. Springer.