

Discussion Paper Series – CRC TR 224

Discussion Paper No. 131  
Project A 03

Understanding the Sources of Earnings Losses After Job  
Displacement: A Machine-Learning Approach

Andreas Gulyas\*  
Krzysztof Pytko\*\*

October 2019

\*University of Mannheim; andreas.gulyas@uni-mannheim.de  
\*\*University of Mannheim; pytko@uni-mannheim.de

Funding by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)  
through CRC TR 224 is gratefully acknowledged.

# UNDERSTANDING THE SOURCES OF EARNINGS LOSSES AFTER JOB DISPLACEMENT: A MACHINE-LEARNING APPROACH\*

Andreas Gulyas<sup>†</sup>      Krzysztof Pytka<sup>‡</sup>

September 24, 2019

## Abstract

We document the sources behind earnings losses after job displacement adapting the generalized random forest due to [Athey \*et al.\* \(2019\)](#). Using administrative data from Austria over three decades, we show that displaced workers face large and persistent earnings losses. We identify substantial heterogeneity in losses across workers. A quarter of workers face cumulative 11-year losses higher than 2 times their pre-displacement annual income, while another quarter experiences losses less than 1.1 times their income. The most vulnerable are older high-income workers employed at well-paying firms in the manufacturing sector. Our methodology allows us to consider many competing theories of earnings losses prominently discussed in the literature. The two most important factors are the displacement firm’s wage premia and the availability of well paying jobs in the local labor market. Our overall findings provide evidence that earnings losses can be understood by mean reversion in firm rents and losses in match quality, rather than by a destruction of firm-specific human capital.

Keywords: Job displacement, Earnings losses, Causal machine learning

---

\*Funding by the German Research Foundation (DFG) through CRC TR 224 (Project A3) is gratefully acknowledged. Johannes Gönsch and Michaela Paffenholz provided excellent research assistance.

<sup>†</sup>University of Mannheim; [andreas.gulyas@uni-mannheim.de](mailto:andreas.gulyas@uni-mannheim.de)

<sup>‡</sup>University of Mannheim; [pytka@uni-mannheim.de](mailto:pytka@uni-mannheim.de)

## I. INTRODUCTION

A sizable literature documents large and long lasting consequences of job loss. Workers displaced during a mass-layoff experience significant losses in annual income lasting over 15 to 20 years.<sup>1</sup> The sources behind the persistent and long lasting earnings losses are still an open question in labor market research. Several channels have been put forward to explain those sources. For instance, two recent papers study the role of losses in employer specific wage components, and arrive at different conclusions. Schmieder *et al.* (2018) argue that they play an important role in Germany, whereas Lachowska *et al.* (2018) show that their importance is limited for the state of Washington during the financial crisis. Long tenure in prior job, industry, and occupation lead to substantially larger losses, hinting that job-specific skills are lost with job displacements.<sup>2</sup> Disentangling these channels is a premise for designing effective policy responses. A particular challenge for identifying the contribution of these channels is that specific workers tend to sort into different types of firms.<sup>3</sup> Ideally, one needs to identify how certain variables affect earnings losses while holding confounding factors constant. In addition, potential interactions between these channels should be taken into account.

To resolve these challenges, we adapt the generalized random forest method by Athey *et al.* (2019) to a difference-in-difference setup and estimate how the causal effect of job displacement varies with a large number of observables. We draw upon administrative social security for the universe of male Austrian employment spells over three decades.<sup>4</sup> There are two advantages of our research design. First, this methodology allows us to document substantial heterogeneity in the individual cost of job displacement.<sup>5</sup> While the average cumulative earnings losses over 11 years amount to 155 percent of pre-displacement earnings, the interquartile range of these losses is 95 percent. Second, contrary to traditional

---

<sup>1</sup>Jacobson *et al.* (1993), Neal (1995), Couch and Placzek (2010), Davis and Von Wachter (2011), Farber (2011), Jarosch (2014), Farber (2017), Krolikowski (2017), Jung and Kuhn (2018), among many others. Furthermore, job displacements have been shown to have detrimental effects on health (Schaller and Stevens, 2015), longevity (Sullivan and Von Wachter, 2009), and on children of displaced workers (Lindo, 2011; Rege *et al.*, 2011).

<sup>2</sup>Jacobson *et al.* (1993), Neal (1995).

<sup>3</sup>*E.g.*, Card *et al.* (2013), Song *et al.* (2018).

<sup>4</sup>Earnings losses of job loss in the context of Austria have been studied before by e.g. Schwerdt *et al.* (2010) and Ichino *et al.* (2017). Halla *et al.* (2018) show limited increases of spousal labor supply after husbands' job displacement in Austria.

<sup>5</sup>To the best of our knowledge, this is the first paper documenting the whole distribution of the causal cost of job loss. Guvenen *et al.* (2017) study the heterogeneity in long-term earnings losses from spending at least one year out of work in the US.

subgroup analysis, where subgroups differ in many characteristics, in our approach we can hold confounding factors constant. This matters for one of the most prominent theories about earnings losses, destruction in firm-specific human capital. We show after controlling for other worker and firm characteristics, high tenure workers do not face higher losses any more. Thus, in our data the differences originate from compositional differences across subgroups. Out of all considered channels, we find that the firm wage premia stand out as the most important driving force of the earnings losses.

The applied machine-learning procedure is a fully data-driven theory-agnostic way to estimate the causal cost of job loss as the function of observable variables. This is achieved by exploring possible data splits and choose those ones for which between-group differences in the cost of job losses are the highest. By recursively splitting the dataset into smaller and smaller subsamples, the algorithm builds a regression tree, which detects heterogeneity in the cost of job losses. In contrast to a more traditional approach where splits are motivated by assumed theories, our procedure lets the data speak for itself. To avoid detection of artificial heterogeneity, instead of growing a single tree we build a random forest consisting of many tree-based models. Every tree is trained using a random subset of observations and variables and then their estimates are averaged. It is known that while a single tree might suffer from overfitting, the random forest is much more robust to this concern. Furthermore, another appealing property of random forests is that we are able to quantify standard errors arising from two sources, the machine learning procedure and the estimation procedure.

With the trained random forest, we can estimate the cost of job displacement for each individual in our sample. We document a large heterogeneity in earnings losses. While median cumulative earnings losses in the 11 years following separation amount to €66,000, over 1/8 of our sample incur losses of above €100,000. Close to 10 percent of individuals do not face any losses upon job displacement. We further show that the most vulnerable individuals are older high-income workers employed in firms with high-wage premia. We follow the typical sample restrictions in the literature in order to get a clean identification of the displacement event, which involves conditioning on larger firms and long-tenured workers. Is the cost of job loss representative for whole population? We use the random forest to predict earnings losses for the population that fail to meet the sample selection criteria and show that they are surprisingly similar.

We construct 15 variables that are associated with some of the most prominent theories discussed in the earnings loss literature as plausible explanatory factors. We seek to understand how earnings losses vary with these channels and to identify which of these are the

most important ones. We include a number of socio-demographic worker variables such as age and dummies for Austrian citizenship, blue collar occupation, and manufacturing sector. These per-se might not be informative about potential channels, but we want to make sure that any of the identified effects is not originating simply through compositional differences in these factors. This ability to control for observable differences is in fact one of the advantages of our machine learning approach compared to subgroup analysis. We consider theories that tie earnings losses to losses in job-specific human capital through an inclusion of job tenure. To quantify the effects of lost firm rents we compute firm wage premia through an inclusion of firm fixed effects in a Mincer wage regression following [Abowd \*et al.\* \(1999\)](#). We also use this firm pay measure to construct a (leave-out) mean firm pay measure at the local labor market level. Another hypothesis relates earnings losses to losses of particularly good matches. We consider this channel by including income prior to displacement, which after conditioning for all aforementioned variables, proxies for match quality. Since mass-layoffs at large firms can have regional spillover effects ([Gathmann \*et al.\*, 2018](#)), we also include firm size and the Herfindahl-Hirschmann index of labor market concentration as explanatory variables. We also consider environmental variables that might affect earnings losses such as the state of the business cycle through regional and industry unemployment rates.

A contribution to the existing literature is that our methodology allows us to compare all those theories at the same time. We show that from all these variables, firm wage premia is the most important one. We arrive at this conclusion through three different analyses. First, we use a standard measure of variable importance in the machine learning literature for random forests, which essentially estimates how often a variable is used in the construction of the random forest. Not only is the pre-displacement firm wage premium by far the most important variable, but the second most important is the average firm pay premium in the region. We further conduct a variance decomposition by linearly projecting earnings losses on our variables. We show that variation in firm wage premia alone explains 40 percent of the overall variation in earnings losses across individuals.

To understand the contribution of each variable individually, we compute how earnings losses change by varying one channel at a time, holding all other variables constant at their median. This reinforces the importance of lost firm premia in understanding earnings losses. Earnings losses rise from €42,965 for workers employed at the lowest paying decile of firms to almost triple that amount for workers displaced from the highest paying decile of firms. We also show strong interactions effects with the availability of well paying jobs in the region. Workers employed at low fixed effect firms, but located in regions with high average firm

premia do not face any earnings losses. The losses in firm wage premia are simply driven by mean reversion in firm pay. Workers employed in firms with above median firm wage premia lose in terms of firm pay, whereas workers separating from below median paying firms gain in terms of firm fixed effects. Worker’s age and income prior to displacement also plays an important role. The oldest workers in our sample face losses double as high as young workers. This is not driven by differential wage changes, but because older workers face higher employment losses. Earnings losses show a steep slope in income, which after controlling for all other observable factors proxies for match quality. This channel operates through wage losses, not employment losses. This highlights that losses of particularly good matches is another important channel in understanding earnings losses. We do not find that earnings losses vary much with all other factors. Given that lost job-specific human capital is one of the most prominent theories for scarring effects of job loss, it is especially surprising that we do not observe large differences in earnings losses across job tenure. Workers with the highest job tenure are estimated to incur losses of only €2,200 more than employees with 2 years of job tenure, with these estimates being well within confidence intervals. All in all, our findings provides evidence that earnings losses can be understood by mean reversion in firm wage premia and match quality, rather than by a destruction of firm specific human capital, while earnings losses for older workers is mostly driven by employment losses.

The ability of our machine learning approach to hold confounding factors constant is very important. We show that a more traditional subgroup analysis: (1) would overestimate the contribution of firm wage premia and, worse yet, (2) would spuriously detect the impact of the job tenure. This is due to the fact that standard sample splitting ignores compositional changes with respect to other covariates. In our dataset the firm wage premia and the job tenure are positively correlated with the income level prior to the job displacement and the worker age, respectively. Sample splitting with respect to one variable leads to a comparison between subsamples that differ from each other in more than one characteristic. Our approach allows us to compute the treatment effect conditional on observables, hence, avoiding the issue of compositional differences across subgroups. Hence, we argue that the proposed theory-agnostic and completely data-driven random forest is especially well suited for this setup.

In addition to previously mentioned papers, we also contribute to the small but growing number of papers using machine learning in Economics. Examples include [Davis and Heller \(2017\)](#) who use a similar approach to ours to understand the heterogeneous effects of youth employment programs, while [Knaus \*et al.\* \(2017\)](#) use a LASSO model to study treatment

heterogeneity of job search programs. Kleinberg *et al.* (2017) study how machine-learning methods can support bail decisions made by judges in New York City. Bertrand and Kamenica (2018) analyze temporal trends in cultural distance between groups in the US defined by income using ensemble methods. Albanesi and Vamossy (2019) employ a deep-learning approach to improve predictions of consumer default.

The rest of the paper is organized as follows. The next section describes the empirical setting in Austria, as well as the sample selection. Section III presents the average cost of job displacement. Section IV describes the machine learning algorithm used to identify the driving forces behind earnings losses. Section V documents heterogeneous scarring effects of job displacement and section VI discusses the sources behind earnings losses. The last section concludes.

## II. EMPIRICAL SETTING

We use the administrative employment and unemployment records from the social security administration in Austria from 1984 - 2017. This data comprises day-to-day information on all jobs and unemployment spells covered by social security in Austria (Zweimüller *et al.*, 2009). This data contains information on yearly earnings for each worker-firm pair.<sup>6</sup> It further contains basic socio-demographic information at the worker level such as age, gender, occupation, and citizenship. Each establishment has a unique identifier, which allows us to study changes in employer specific characteristics over time. At the establishment level we have data on the geographic location and a 4-digit industry classifier.

### A. Definition of Job Displacement and Mass Layoff

To ensure comparability with the previous literature on displaced workers, we follow the typically applied definitions and sample restrictions as much as possible. A worker is considered displaced if she separated from her primary employer that experienced a mass layoff in the given year. We define a mass layoff event at the firm level in year  $t$  if it declined by more than 30 percent in size during year  $t$ . To avoid selecting volatile firms, we exclude firms that either grew rapidly the years before the mass layoff, or rebounded in size 3 years after the mass layoff event. More precisely, we exclude firm that grew by more than 30 percent in either  $t - 1$ , or  $t - 2$ , as well as firms that are larger 3 years after the event than before. To have a meaningful measure of firm growth, we only consider establishment with at least

---

<sup>6</sup>We deflate all earnings to 2017 level using the CPI index provided by the Austrian Statistical Agency.

30 employees. In addition, to avoid mis-specifying mergers, outsourcing or firm restructures as mass layoffs, we compute a cross flow matrix for all firms in each year. We exclude all firm where more than 30 percent of its workforce ends up working for the same employer in  $t + 1$ .<sup>7</sup> Thereby we exclude mass layoff firms with large worker flows to other firms. Not correcting for these potential measurement-errors might lead to a significant underestimation of earnings losses.

### *B. Construction of Sample*

We focus on male workers only. In addition to enhancing comparability to previous studies that mainly study males, we do this because of two reasons. First, the higher labor force attachment of men workers reduces selection issues in our analysis. Second, the social security data only contains information on employment status, not on hours worked. Whereas during our study period, between 27-47 percent of women were working in part-time, only between 4-11 percent of men were employed part-time.<sup>8</sup> We proceed by selecting everybody who is employed on the reference day of 1st of January each year. This results in 45,689,628 person-year observations. We follow the literature and restrict our sample to workers aged 24-50, employed at a firm larger than 30 employees, with job tenure longer than 2 years on the reference day. We remain with 8,276,796 person-year observations.

Out of these remaining observations, we define a person to be displaced in  $t$  if a worker separates from a firm experiencing a mass layoff, and the worker is not reemployed at the same firm at any point in the next 10 years. We identify 45,521 displaced (male) worker events between 1989 and 2007.

Note that we do not restrict the control group to have stayed at their employer after  $t$ . The potential comparison group consists of non-displaced workers subject to the same sample restrictions. This includes workers employed at firms without any mass layoff event during year  $t$  or worker in mass layoff firm who did not separate.

Some workers in our dataset disappear over time from our dataset. This happens on the one hand because workers might not find employment subject to social security insurance anymore and drop out of the labor force. On the other hand, this could also happen if workers move into self-employment or move abroad. We decided to use only information on workers who either have an employment spell covered by social security or a registered unemployment

---

<sup>7</sup>For an in-depth discussion see Hethey-Maier *et al.* (2013).

<sup>8</sup>Source:[https://www.statistik.at/web\\_de/statistiken/menschen\\_und\\_gesellschaft/arbeitsmarkt/arbeitszeit/teilzeitarbeit\\_teilzeitquote/062882.html](https://www.statistik.at/web_de/statistiken/menschen_und_gesellschaft/arbeitsmarkt/arbeitszeit/teilzeitarbeit_teilzeitquote/062882.html)



	Displaced	Selected Control Group	Not Selected
Age	37.86	37.84	37.82
$\log(w_{t-1})$	4.65	4.66	4.71
$\log(w_{t-2})$	4.64	4.64	4.69
Job Tenure (in days)	2560.97	2537.33	2661.38
Manufacturing	0.52	0.53	0.50
Firm Size	1656.70	1634.75	1462.52
Obs	45521	45521	8185754

Table 1: Sample characteristics of workers

spell in a given year. This likely underestimates the true costs of job-displacement, as we do not measure losses associated with dropping out of the labor force.

### *C. Propensity Score Matching*

Non displaced workers may differ in many characteristics from the displaced workers. This is also confirmed by Table 1, which shows that displaced workers have more tenure, slightly higher earnings, are more likely to work in manufacturing and are employed in larger firms compared to non-displaced workers. In order to obtain a control group that is as similar as possible to the displaced workers, we use propensity score matching for the selection of our control group. In each year, for all workers satisfying the sample restrictions, we estimate the propensity to experience a displacement event as a function of observable worker and firm characteristics. Specifically, we use worker’s log wage in year  $t - 1$  and  $t - 2$ , tenure, age, establishment size in year  $t$  as well as a dummy for working in the production sector as matching variables.<sup>9</sup> For each displaced worker in a given year, we select the non-displaced worker with the nearest propensity score without replacement. Table 1 shows that our matched control group is very similar to displaced workers in observable characteristics. The two groups are virtually indistinguishable in terms of pre-displacement evolution of earnings, days employed and log-wages, as can be seen in Figure 1, which plots raw averages across these groups over time.

---

<sup>9</sup>We also experimented with different sets of matching variables, all of which lead to similar results.

### III. THE AVERAGE COST OF JOB DISPLACEMENT

To estimate the cost of job displacement relative to the counterfactual of no displacement, we estimate the following regression model:

$$y_{it} = \sum_{j=-4}^{10} \delta_j \mathbb{1}(t = t^* + j) \times D_i + \theta D_i + \gamma_t + \epsilon_{it}, \quad (1)$$

where  $D_i$  is an indicator equal to one for a displaced persons,  $t^*$  the displacement year and  $t$  the current year. We include year fixed effects  $\gamma_t$  to control for the evolution of the control group's earning. We also control for initial differences in earnings by including a displacement dummy  $D_i$ .<sup>10</sup> Because of the propensity score matching, workers in the control group are very similar.  $\delta_j$  consequently measures the change in workers earnings relative to the baseline year  $t^* - 5$ , after controlling for differences in initial earnings between the two groups.

The main outcome variables we consider are total annual earnings, total annual days employed, and log daily wage from the employer on January 1st each year. Figure 2 shows the estimated causal effect of job-displacement for these three outcome variables. In the year after job displacement, earnings losses amount to approximately €10,000, or 26 percent of pre-displacement earnings. In the following years earnings increase, but the recovery fades out after 5-6 year, after which the losses still amount to €5,000 yearly, or 13 percent in terms of pre-displacement earnings. The lower panels of Figure 2 show that this decline in earnings both stem from employment losses and declines in log-wages. In two years after job displacement, employment losses amount to almost 50-70 days, but employment almost fully recovers after a couple of years. More strikingly, there is essentially no recovery in log-wages. Displaced workers' wages decline by 7-8 percent, with no noticeable recovery in the first 10 years after job-displacement. The evolution of earnings losses looks surprisingly similar to those in the US (see *e.g.*, [Davis and Von Wachter \(2011\)](#)), despite the institutional differences between the US and Austria.

Throughout the paper we will mostly focus on the cumulative earnings losses in the 11 years following job displacement. These can be estimated either by summing appropriately over the Event Study coefficients ( $\sum_{j=0}^{j=10} \delta_j$ ) from equation (1), or more compactly using the

---

<sup>10</sup>Instead of controlling for differences in  $t - 5$  with the group dummy  $D_i$ , we could have included worker fixed effects. The results are identical for the two cases, but because of the computational demands of our machine learning algorithm we use the dummy variable approach already here.

following difference-in-difference setup:

$$y_{it} = \tau \mathbb{1}(t \geq t^*) \times D_i + \theta D_i + \gamma_t + \epsilon_{it}. \quad (2)$$

where  $\tau$  measures the average yearly cost of job loss, or if scaled by 11, the cumulative losses over 11 years. Table 2 shows the estimates from this regression for different specifications. Column (1) reports the estimates for equation (2) without any controls, column (2) and (3) a polynomial in age and worker fixed effects are added. In all specifications, the yearly earnings losses amount to close to €5,900 per year, or close to €65,000 over 11 years. Because of the computational burden of the machine learning algorithms, we are not able to include worker fixed effects as controls as is often done in the literature. But the results from Table 2 show that because our matching procedure selected very similar workers as a control group, adding worker fixed effects or a polynomial in age does not significantly change the estimated costs of job displacement.

Table 2: DiD Regression

	<i>Dependent variable:</i>		
	Yearly Income		
	(1)	(2)	(3)
$\hat{\tau}$	−5867.4 (55.5)	−5994.3 (33.5)	−5828.6 (32.5)
Worker FE		x	x
$f(age)$			x
Observations	1378385	1378385	1378385
R <sup>2</sup>	0.04	0.7	0.7
Adjusted R <sup>2</sup>	0.04	0.7	0.7

#### IV. IDENTIFYING SOURCES – MACHINE LEARNING APPROACH

The goal of our exercise is to identify the heterogeneity in earnings losses and its sources. To this end, we employ a machine-learning procedure built on the methodology of generalized random forests, very recently developed by [Athey \*et al.\* \(2019\)](#). There are several important advantages of our approach in comparison to traditional ones, *i.e.* estimating the

homogeneous scarring effect or sample splitting to compute heterogeneous effects.

The average cumulative earnings losses after job displacement can be estimated directly from formula (2) using standard statistical methods. Nonetheless, this approach has some serious limitations. First and foremost, this type of the estimate provides the average treatment effect and does not give a gist on its underlying heterogeneity. If the earnings losses are not uniform across individuals, then the individual scarring effects may be a far cry from the estimated average. This concern is of particular importance for policy makers, because accurate predictions of individual earnings losses are a prerequisite for designing optimal labor-market policies. In addition to this, sample restrictions, traditional for the literature such as focus on mass layoffs, age of workers, firm size, and job tenure, are the objects of discussion how the results can be extrapolated outside the training dataset. Differences in the composition between the training sample and the general population may mechanically result in a situation in which the average effect identified using the training sample is not representable for the general population of the interest.

Heterogeneous treatment effects are typically identified through sample splitting and estimating the model in separate data bins. An application of this method can be illustrated by a study on heterogeneous scarring effects of one-year non-employment due to [Guvonen et al. \(2017\)](#). While this approach is absolutely correct for identifying the existence of heterogeneity, however, it does not tell us what are its driving channels. The reason for this is that the composition of individual characteristics may vary across data subsets in a similar fashion to differences between the training set and the general population, mentioned before. Consequently, it is impossible to disentangle the impact caused by the variable according to which the split has been made and the changes in the composition of the other variables.

Our approach consists in estimating the version of equation (2) with heterogeneous scarring effects:

$$y_{it} = \tau(\mathbf{z}_i)\mathbb{1}(t \geq t^*) \times D_i + \theta(\mathbf{z}_i)D_i + \gamma_t(\mathbf{z}_i) + \epsilon_{it}, \quad (3)$$

where  $\mathbf{z}_i$  are the values of variables (henceforth called *partitioning variables*) for individual  $i$  and  $\tau(\mathbf{z}_i)$  is a treatment effect for individual  $i$ . The functional specification of  $\tau(\mathbf{z}_i)$  is assumed to be unknown. In theory, one can imagine computing (3) for all possible data bins, for each value of every partitioning variables interacted with other values of all other partitioning variables. That said, it would be very problematic for two reasons. First, it would mean that in our application there would be  $20.8 \times 10^{12}$  subsets. This is much more than the number of observations, which makes the whole procedure computationally infeasible. Second, even

if the computation of  $\tau(\mathbf{z}_i)$  for each data subset were theoretically possible, it would not be recommended as the size of data subsets would be very small implying potentially low accuracy of the estimates. Therefore, we decide to use an algorithmic method to learn the pattern of  $\tau(\mathbf{z}_i)$  and split the whole dataset into smaller partitions in the most efficient way. In addition to this, our partitioning variables are associated with some channels from the literature that try to account for the earnings losses. Our fully theory-agnostic, data-driven procedure can be seen as a horse race between competing theories. The algorithm we use is built on the methodology of generalized random forests, which was proposed by [Athey \*et al.\* \(2019\)](#). The description of partitioning variables and the associated channels can be found in subsection A. Details on how we adapt generalized random forests for our application can be found in subsection B.

### A. Partitioning Variables

The literature on earnings losses discusses a number of channels that could explain earnings losses (see [Carrington and Fallick \(2017\)](#) for a review). We deliberately construct 15 partitioning variables that capture some of the channels that are widely considered to be important in explaining earnings losses.

First of all, we include a number of socio-demographic factors such as worker age, a dummy for Austrian citizenship, blue collar occupation, and manufacturing sector. These per-se might not be informative about potential channels, but we want to make sure that any of the identified effects is not originating simply through compositional differences in socio-demographic factors. This ability to control for observable differences is in fact one of the advantages of our machine learning approach compared to simple subgroup analysis.<sup>11</sup>

One of the most prominent theories about earnings losses is based on the notion of specific human capital. The idea is that workers accumulate job specific human capital over the course of their tenure at an employer. Earnings losses upon job loss therefore reflect the lost job-specific human capital that was embodied in the lost job. This channel is captured by the inclusion of job tenure as a partitioning variable.

We in addition include the number of distinct previous employers (censored in 1984). Workers with higher past job mobility might be less prone to high earnings losses.

A growing literature shows that differences in firm pay is a significant driver of wage inequality ([Abowd \*et al.\*, 1999](#); [Card \*et al.\*, 2013](#)). Workers might fall of the job ladder and

---

<sup>11</sup>[Schwerdt \*et al.\* \(2010\)](#) for example shows that white collar workers face higher earnings losses in the short run.

find themselves reemployed only at firms paying lower wages (Jarosch, 2014). Two recent working papers highlight lost employer rents as an important reason for reduced earnings after job loss. To understand how much of earnings losses might be attributed to losses in rents paid by firms, we use a wage-regressions following following Abowd *et al.* (1999) (henceforth, AKM). Despite well known limitations about the structural interpretation of firm wage premia<sup>12</sup>, the AKM model has become the workhorse model for empirically estimating the firm pay component. We estimate:

$$\ln(w_{it}) = \psi_{J(i,t)} + \alpha_i + \theta_t + x_{it}\beta + \epsilon_{it}, \quad (4)$$

where  $\ln(w_{it})$  is the log daily wage of the dominant employer in period  $t$ ,<sup>13</sup>  $\psi_{J(i,t)}$  represents the establishment fixed effect of the employer of worker  $i$  in period  $t$ ,  $\alpha_i$  the worker fixed effect,  $\theta_t$  the year fixed effect, and  $x_{it}$  are time varying observables, comprising of a cubic polynomial of age. We use the universe of male private sector employment spells in Austria from 1984 - 2017 in the estimation. In order to avoid endogeneity, we disregard individuals in our earnings loss sample from the computation of firm fixed effects.<sup>14</sup> This implies that from 56,254,607 person-year observation, 2,427,901 are dropped. We cannot rule out that the masslayoff event nevertheless affected the estimation of the firm fixed effects through remaining workers who were not selected by the propensity score matching. Therefore, we will also consider a robustness exercise, in which we compute the AKM firm fixed effects only on observations before 1998, and we will restrict our earnings loss sample to years after 1998.

We further construct a number of variables that are associated with the local labor market and industry of workers. We use the NUTS-3 district (35 categories) and NACE level-1 industry classification (21 categories) of employers. The NUTS-3 classification follows European standards and is a sensible definition of a local labor market and is roughly comparable to US counties. Mass-layoffs are more common in economic downturns (Davis and Von Wachter (2011)), and might be concentrated in declining industries and regions. To understand how earnings losses vary with economic conditions, we compute for each region and industry the unemployment rate January 1st (our reference day for most of our analysis), as well as the year-to-year change in the unemployment rate.

---

<sup>12</sup>Eeckhout and Kircher (2011), Bagger and Lentz (2018), Hagedorn *et al.* (2017), Gulyas (2018).

<sup>13</sup>The dominant employer is selected based on the total earnings in calendar year  $t$ .

<sup>14</sup>Firm fixed effects are identified from wage changes of workers moving across firms. Thus, if workers experience earnings losses in mass-layoffs, the mass-layoff firm will be estimated to be a high fixed effect firm.

We are further interested how the availability of well-paying jobs affects earnings losses. We build on our firm wage premium estimates from the AKM regressions and compute the average firm premium in the region. Specifically, we compute the average firm wage premia of all jobs in a given region leaving out all jobs of the worker’s current employer. Formally for every worker  $i$  employed at firm  $J(i, t)$  we compute  $\sum_{k \notin J(i, t) \wedge k \in r(i)} \hat{\psi}_{J(k, t)} / \#(k \notin J(i, t) \wedge k \in r(i))$ , where  $r(i)$  is the region of the worker  $i$ .

Another hypothesis about earnings losses is, that displaced workers lose a particularly good match. To study this channel, we include income one year prior to separation as a partitioning variable. Conditional on all previously mentioned job characteristics, income proxies for the match quality.

Recent evidence shows that many local labor markets are highly concentrated (Azar *et al.*, 2017). Since mass layoffs are conditional on large firms, the decline of these firms might have strong spill-over effects on the whole region (Gathmann *et al.*, 2018), further depressing outcomes for displaced workers. This effect is presumably stronger in large firms and concentrated labor markets. We therefore include firm size and construct a labor market concentration measure. We use the Herfindahl-Hirschman index of labor market concentration, which is the standard concentration measure used in anti-trust cases. For each region and industry combination in each year, we compute  $H_{yri} = \sum_j s_j^2$ , where  $s_j$  is the employment share of firm  $j$  in year  $y$ , region  $r$  and industry  $i$ .

To enhance interpretability, we categorize all continuous variables into deciles according to the overall distribution of Austrian male workers on our reference day, and not only the selected displaced worker sample. This way the heterogeneity is easily interpretable in terms of the global distribution. Figure 3 shows a correlogram of all partitioning variables. Earnings losses are likely a combination of all these factors, and different channels interact with each other. In the next section we describe the applied machine learning algorithm that enables us to disentangle the contribution of all these different channels.

### *B. Bird’s-Eye View of Machine-Learning Algorithm*

For exposition purposes, the algorithm is presented in two steps. First, we show how heterogeneity is identified using a single regression tree in the spirit of Breiman *et al.* (1984) with a modified splitting criterion borrowed from Athey *et al.* (2019). Next, in order to circumvent shortcomings typical for single trees, the approach is augmented to generalized random forests in the spirit of Athey *et al.* (2019).

### B.1. Tree implementation

The tree-based procedure consists in partitioning the dataset into smaller subsamples in which individuals exhibit similar earnings losses and at the same time the differences in earnings losses between subsamples are maximized. The data fragmentation is carried out using a sequence of complementary restrictions on partitioning variables. Due to the computational complexity and the fact that the problem is NP-complete (Hyafil and Rivest, 1976), a top-down, greedy approach is traditionally used. The procedure of building a tree can be characterized in the recursive way by Algorithm (1). In each data partition (called also a node or a leaf) the scarring effect is estimated from equation (2) separately.<sup>15</sup>

---

**Algorithm 1** Tree Algorithm of Recursive Partitioning

---

- i. Start with the whole dataset and consider it as one large data partition,  $\mathcal{P}$ .
- ii. For each partitioning variable  $z_k$  and its every occurring value  $\bar{z}$ , split partition  $\mathcal{P}$  into two complementary sets of individuals  $i$  such that  $\mathcal{P}_l = \{i \in \mathcal{P} : z_{ki} \leq \bar{z}\}$  and  $\mathcal{P}_r = \mathcal{P} \setminus \mathcal{P}_l$  and estimate cumulative earnings losses  $\tau_l$  and  $\tau_r$  for both partitions by running two separate regressions of form (2) on  $\mathcal{P}_l$  and  $\mathcal{P}_r$ .
- iii. Choose the variable  $z_k$  and value  $\bar{z}$  that maximizes:

$$(\tau_l - \tau_r)^2 \frac{n_l \cdot n_r}{N^2}, \quad (5)$$

where  $n_l$  and  $n_r$  are sizes of  $\mathcal{P}_l$  and  $\mathcal{P}_r$  and  $N$  is the sample size of  $\mathcal{P}$ .

- iv. If (5) is smaller than a tolerance improvement threshold, then stop. Otherwise, go to step (ii) and repeat the splitting procedure for  $\mathcal{P}_l$  and  $\mathcal{P}_r$  separately, where  $\mathcal{P}_l$  and  $\mathcal{P}_r$  are new partitions subject to the splitting procedure,  $\mathcal{P}$ .
- 

The main difference of our procedure from the textbook one, which can be found in Breiman *et al.* (1984), is the splitting criterion. In the original approach, the algorithm aims at building a tree minimizing the squared sum of residuals.<sup>16</sup> In our setup, we are interested in growing a tree that explores the underlying heterogeneity of earnings losses between

---

<sup>15</sup>It is noteworthy that while all parameters from (2) are estimated, only the parameter of our interest, the scarring effect, is used in the splitting criterion (5).

<sup>16</sup> In regression trees, the squared sum of residuals is defined as  $\sum_j \sum_{i \in \mathcal{D}_j} (y_i - \bar{y}_{\mathcal{D}_j})^2$ , where  $\mathcal{D}_j$  is a data subset obtained through sample partitioning procedure and  $\bar{y}_{\mathcal{D}_j} = \frac{1}{|\mathcal{D}_j|} \sum_{i \in \mathcal{D}_j} y_i$  is the mean of the response variable in the specific set of data.



partitions of individuals with different characteristics. For this reason, we adapt the criterion (5) proposed by [Athey \*et al.\* \(2019\)](#). This criterion maximizes between-group differences of earnings losses,  $(\tau_l - \tau_r)^2$ , with an adjustment for more balanced splits,  $\frac{n_l \cdot n_r}{N^2}$ . It should be emphasized very clearly, while in most machine-learning algorithms the goal function is to maximize its predictive accuracy, in our analysis it is not the case. In our data partitioning procedure, in every leaf for each considered split regression (2) is estimated. Consequently, a standard procedure of causal inference can be performed in every leaf separately and our machine-learning procedure does not distort it at all.<sup>17</sup>

In applied economics, one alternative to splitting the dataset is to assume that the data-generating process is known and given and to estimate according to that process. In our application it would mean that we make an arbitrary decision upon the specification of  $\tau(\mathbf{z}_i)$ . However, our strategy stays in stark contrast with that tradition. We are very upfront about our agnosticism on  $\tau(\mathbf{z}_i)$  and by employing Algorithm (1) we try to learn the true specification. As a result, the learning procedure does three things:

- i. chooses which variables are important and contribute to accounting for the heterogeneous scarring effects and which do not;
- ii. detects non-linear relationships between  $\tau$  and  $\mathbf{z}_i$ ;
- iii. detects interactions (including interactions of higher orders) between partitioning variables.

Figure 4 depicts a tree grown using the described algorithm. Every node is labelled with the average cumulative earnings losses and the overall fraction of observations in the node. On the top there is the root node containing all observations and on the bottom there are final nodes that are not subject to further partitioning. Fractions of all final nodes (leaves) sum to 100%. In the whole dataset the average earnings loss in 11 years is equal to around €65,000. In the first iteration, the split that maximizes heterogeneity between groups goes according to the firm fixed effect. Individuals displaced from firms paying below the fourth decile face earnings losses at a lower level, around €51,000, while the scarring effect of workers fired from firms paying above the median increase to €96,000. Overall, the partitioning procedure generates 32 final nodes endogenously. The number of binary

---

<sup>17</sup> In his famous essay [Breiman \(2001b\)](#) describes two distinctive statistical cultures, one using exogenously assumed data models and the other employing algorithmic models. Back then, the former was dominating academia and the latter was more typical for business analytics. In our paper we combine both strands. Estimation of equation (2) uses causal inference borrowed from traditional econometrics, while the sample splitting is closer to the algorithmic culture.

splitting conditions used to generate each leaf varies from 3 to 7. Already at this point, we can observe that high-income individuals displaced from high fixed-effect firms face higher losses than others. Those losses can be amplified even more if, in addition, workers are older than 40, they work for large firms in industries with unemployment rate below 9th decile, in regions with high changes in unemployment as it can be seen in the second leaf from the right.

## B.2. Generalization to Random Forests

One important advantage of tree-based models is its easy and very intuitive graphical illustration. Unfortunately, in comparison to other classical machine-learning methods its prediction performance is relatively low (*cf.*, Mullainathan and Spiess, 2017) and the variance of the prediction error is quite high. In addition to this, trees can be very non-robust due to potential overfitness and the greedy algorithm choosing variable with high predictive power may upstage the importance of other relevant variables. Random forests proposed by Breiman (2001a) are a refinement of the baseline method that address the typical concerns of tree-based models. A general idea behind random forests is quite easy and relies on building many trees through bootstrapping data observations. Moreover, in each split decision a subsample of considered variables is drawn. Consequently, the ensemble of decorrelated trees is grown, which means that the trees differ from each other and are built with different variables. The final prediction of a random forest is the average of individual predictions of many trees from the forest.<sup>18</sup> This way only relationships that consistently show up in different bootstrapped samples are identified. In addition to this, each tree has been built using a so-called “honest” approach. This means that half of the bootstrapped sample was used to determine conditions which constitute data partitions, while with the other half the scarring effects were estimated in those partitions.<sup>19</sup> An example illustrating how the algorithm works on simulated data is relegated to Appendix B. More details on the methodology employed by us can be found in Athey *et al.* (2019). From the specification perspective,

---

<sup>18</sup> Due to computational complexity, we follow a suggestion of Athey *et al.* (2019) and the forest-growing procedure is approximated using gradient boosting in the spirit of Friedman (2001). To this end, we modify the `grf` library in such a way that we can handle with any arbitrary least-squared regression with multiple covariates and we can choose the subset of parameters with respect to which the splitting algorithm proceeds. In our application we maximize the heterogeneity of  $\tau$  while the remaining 17 parameters from (2),  $\theta$ , and  $\gamma_t$  (where  $t \in \{-5, \dots, 0, \dots, 10\}$ ) are nuisance parameters, which are estimated to control for other sources of variability of  $y$  but are not used in the splitting procedure explicitly.

<sup>19</sup> Thanks to this procedure we make sure we do not document spurious heterogeneity. If by any chance some splits are made due to some outliers, the estimated treatment effects are not affected by this. For more details see Athey and Imbens (2016).

our problem can be framed as a problem of uplift modelling. A simulation illustrating how generalized random forests perform in comparison to the two-models approach is presented in Appendix C.

## V. HETEROGENEOUS SCARRING EFFECTS OF JOB LOSS

In this section we use the results from the machine learning approach to answer two questions. On the one hand, we are interested in the heterogeneity behind the average costs of job displacement. Second, we leverage that we also know how the heterogeneity is related to observable worker and job characteristics. This allows us to disentangle the sources of earnings losses. In the analysis, we grew a forest consisting of 5,000 trees, with the minimum leaf size equal to 1,600 person-year observations. In each considered split we draw 6 randomly chosen partitioning variables.

### *A. Heterogeneity in Earnings Losses*

In this section we focus on documenting how the scarring effect of job displacement differs across workers. The generalized random forest provides an estimated treatment effect for each individual worker based on his characteristics. We start with plotting the distribution of cumulative earnings losses  $\tau_i$  over 11 years after the displacement event. Figure 5 and Table 3 shows that the average effect hides an enormous amount of heterogeneity among workers. Median cumulative earnings losses are €65,800, or 1.7 times the yearly annual median earnings in the year before the displacement event. But 25% of individuals face staggering earnings losses larger than €87,620, whereas 10% of individuals experience zero earnings losses. Table 3 also shows that this heterogeneity is not generated through noisy point estimates of individual earnings losses. For 7 percent of the population we can rule out (at the 95 percent confidence level) smaller than €100,000 earnings losses, whereas for almost 5 percent of workers we can rule out any detrimental effects on earnings.

Another angle of heterogeneity that we document is the distribution of earnings losses relative to income before the displacement. It is noteworthy that in the presence of dispersion in pre-displacement income, heterogeneous scarring effects in absolute terms does not translate automatically to heterogeneity relative to prior income.<sup>20</sup> However, in our study we find that individuals with higher absolute losses also suffer from higher relative losses.

---

<sup>20</sup>For instance, if every displaced worker lost exactly 10% of the income, then still there would be heterogeneity in losses in absolute terms but not heterogeneity in relative terms at all.

$x$	-125,000	-100,000	-75,000	-50,000	-25,000	0	25,000
$P(\hat{\tau}_i^{\oplus} < x)$	0.053	0.152	0.384	0.640	0.843	0.917	0.981
$P(\hat{\tau}_i^{\oplus} + 1.96*se < x)$	0.018	0.065	0.211	0.503	0.727	0.860	0.943
$P(\hat{\tau}_i^{\oplus} - 1.96*se > x)$	0.871	0.702	0.439	0.265	0.106	0.045	0.005

Table 3: Distribution of cumulative individual earnings losses, estimates from a generalized random forest. The first row shows the empirical cumulative distribution of earnings loss estimates  $\hat{\tau}_i^{\oplus}$ . Row 2 reports the fraction of workers with point estimates statistically significantly below  $x$ , whereas row 3 shows the fraction of workers with point estimates statistically significantly above  $x$  (at the 95 % confidence level).

This is shown by the high correlation between both measures being equal to 0.84. Row 1 of Table 4 shows that also the relative heterogeneity is substantial, with interquartile range amounting to more than one.

Given the heterogeneity we document, the questions arises whether the usually applied sample restrictions select individuals with particularly high earnings losses. In fact, a long standing concern in the earnings loss literature is about generalizability of the results to the whole population. With the use of our random forest we are able to address this question by predicting earnings losses for a random subset of 1 million individuals not satisfying the sample restrictions on firm size or tenure restriction. Table 4 shows that the distributions of in-sample and out-of-sample predictions are surprisingly similar. If at all, it shows that the low-tenure workers or employees at small firms exhibit worker and job characteristics that imply slightly higher and more dispersed earnings losses.

To study difference in the evolution of earnings losses, as well as employment and wage losses we proceed by binning workers into quartiles according to their estimated earnings losses. Figure 6 plots the evolution of earnings losses, employment losses and log wage losses by estimating the Event-study specification from equation (1) separately for every quartile of predicted earnings losses, alongside with the conventional 95% confidence intervals. The figure reconfirms that workers face significantly different scarring effects of job loss. First of all, it is clearly visible that the heterogeneity does not only originate from different short term evolution of earnings, but are persistent throughout the 10 year window after job displacement. The group of workers with the lowest predicted losses (Q1) gain in terms of yearly income from job displacement. They only face short employment losses of about one month in the year of displacement, but recover the employment losses almost completely afterwards. The income gains result from an increase in wages by 3-4 percent. This is in

	Percentile	P10	P25	P50	P75	P90
Displaced Worker Sample		-2.597	-2.077	-1.578	-1.127	-0.197
Out-of-sample Population		-3.022	-2.252	-1.660	-0.965	0.020

Table 4: Distribution of earnings losses relative to prior income in the displaced worker sample and in the population either not satisfying firm size or tenure restriction (1 million random subsample).

stark contrast to the other 3 quarters of workers, who all face long-term earnings losses. The quartile of workers with the highest predicted earnings losses experience employment losses of one and a half years and conditional on reemployment, wages losses of 20 log points. Workers with high absolute earnings losses therefore also face high wage losses in relative terms. This group of workers also face significant employment losses. Even 10 years after separation, days employed is still depressed by over one month, while employment losses for all other groups recovered almost fully recovered. Visually, it appears that heterogeneity in earnings losses arise mostly through differences in wage changes. Next, to address this question more precisely, we decompose earnings losses into employment and wage losses more formally.

### *B. Decomposing Earnings Losses into Employment and Wage Losses*

Losses in earnings are a resultant of losses in their two margins, wages and employment. For this reason, we quantify which fraction of earnings losses originate from declines in employment and wage losses, and how this decomposition differs across workers. To this end, we grew two additional forests maximizing heterogeneity in the treatment effect  $\tau$  with changing the dependent variable  $y$  in equation (3) to wages and employment. Figure 7 depicts the joint distribution of the estimated losses in wages in log points and employment in days through 11 years after the job displacement. Both estimates are positively correlated with each other with the correlation coefficient amounting to 0.462. The wage of the displaced workers after the layoff event is on average 6% lower than their counterparts from the control group. That said 18% of the treatment group do not suffer from any wage losses. The average employment loss over 11 years after the displacement is equal to 345 days and for 91% workers the loss is longer than half a year.

More formally, we can write annual earnings as the product of the number of days employed multiplied by the daily wage in a that year, *i.e.*  $y = N_d w$ . We follow Schmieder *et al.* (2018), and decompose earnings losses into losses stemming from working fewer days

and losses in daily wages. The wage gap between displaced workers and their control group  $\Delta = y^C - y^D$  can be decomposed into three terms the following way:

$$\begin{aligned}
\mathbb{E}[\Delta] &= \mathbb{E}[y^C] - \mathbb{E}[y^D] = \mathbb{E}[N_d^C w^C] - \mathbb{E}[N_d^D w^D] \\
&= \mathbb{E}[N_d^C] \mathbb{E}[w^C] - \mathbb{E}[N_d^D] \mathbb{E}[w^D] + \text{Cov}(N_d^C, w^C) - \text{Cov}(N_d^D, w^D) \\
&= (\mathbb{E}[N_d^C] - \mathbb{E}[N_d^D]) \mathbb{E}[w^C] + \mathbb{E}[N_d^D] (\mathbb{E}[w^C] - \mathbb{E}[w^D]) + \Delta \text{Cov}(N_d, w) \\
&= \Delta \mathbb{E}[N_d] \mathbb{E}[w^C] + \mathbb{E}[N_d^D] \Delta \mathbb{E}[w] + \Delta \text{Cov}(N_d, w).
\end{aligned} \tag{6}$$

Figure 8 shows this decomposition by quartile of predicted treatment effect. Overall, losses in days employed contribute a significant fraction in the short run to overall earnings losses, but the long run persistent losses are almost entirely driven by changes in wages. In the first year after separation, employment losses explain approximately one third of earnings losses. But the contribution of employment losses fades away quickly over time as workers transition back to work. In the long run, losses in days employed only contribute around 10 percent, and earnings losses are almost entirely driven by losses in wages. The change in the covariance term counteracts earnings losses. This implies, that shortly after displacement, workers with higher wages are employed more days per week compared to the control group. The decomposition results are very similar across the different treatment effect groups except the group with the lowest earnings losses. For this group, short term losses are entirely driven by employment losses, whereas they even experience wage gains in the long run.

### *C. Who Losses More?*

Which group of workers face larger than average earnings losses, and which workers are unscarred by job displacement? To address this question, Table 5 reports descriptive statistics broken down by quartile of predicted earnings losses. The workers with the highest predicted earnings losses have above average tenure and income, are employed at better paying firms, and are more likely to work in the manufacturing sector and have a white collar occupation, and have worked for fewer firms over their careers. It is notable how different the average firm pay is across these four groups. While workers facing the highest losses work on average for firms that are paying above the 8 deciles, the ones with the lowest losses are employed on average in the 3rd firm pay decile. These two groups also differ significantly in their age. Workers with the highest losses are on average 5 years older than workers with the lowest losses.

While it is interesting to understand the composition of workers with high earnings losses,

these documented differences still do not address which of the factors are the driving forces behind earnings losses. Many of these variables are correlated with each other (see Figure 3), so it is hard to draw definite conclusions from these compositional differences. Before this question will be tackled in section VI, we evaluate how accurately our random forest estimates the heterogeneity of losses.

Table 5

	Q1	Q2	Q3	Q4
Blue Collar	0.442	0.487	0.638	0.523
Manufacturing	0.615	0.645	0.556	0.271
Austrian	0.786	0.809	0.771	0.790
Firm FE	8.224	7.476	5.872	3.106
Firm Size	6.741	6.412	6.252	6.560
Income t-1	6.860	5.734	3.787	3.023
Job Tenure	8.001	7.357	7.202	7.261
Herfindahl Index	5.282	5.115	5.440	5.974
Avg. Firm FE	6.146	6.522	6.593	6.926
Industry UE-Rate	5.461	6.067	5.863	4.736
Change Ind. UE-Rate	6.059	5.635	5.622	5.447
Regional UE-Rate	5.977	5.513	5.500	6.066
Change Regional UE-Rate	6.400	5.895	5.807	6.024
Number of Firms	2.059	2.451	2.517	2.525
Age	41.832	37.534	36.508	35.409

Notes: Table shows mean baseline characteristics for each quartile of predicted treatment effects. Predictions from a causal forest

#### *D. Accuracy of Prediction*

How accurate are the estimates made by the random forest? Evaluating the accuracy of the random forest is not a straightforward task. We are not estimating an observed outcome, but a treatment effect. Thus, there is no ground truth which we can use to evaluate the predictions. To nevertheless provide an estimate of the accuracy, we execute the following exercise. First, using the estimates by our random forest, we bin individuals into 50 groups based on their predicted earnings losses. For each of these groups, we separately estimate equation (1) using OLS. We then compare the OLS earnings losses estimates with the estimates of our random forest. The left panel of Figure 25 compares the rank correlation between the

two approaches. Both, the OLS estimates and the random forest rank the groups almost in the same way, the rank correlation being 0.98. The right panel plots the OLS estimates against the earnings loss estimates from the random forest. The correlation is with 0.96 equally high. A closer inspection reveals that the earnings loss estimates from the random forest are somewhat regularized, meaning that the OLS estimates suggest a higher level of heterogeneity. We think if this a feature, rather than a shortcoming. OLS is going to overfit towards outliers, whereas the bootstrapping estimation procedure of the random forest is only picking up heterogeneity that consistently occurs across the bootstrapped sample (bags). Put differently, the random forest only identifies predictable heterogeneity.

## VI. SOURCES OF EARNINGS LOSSES

The previous section documented the heterogeneous scarring effect of job displacement. The results were already indicative of which channels are important to explain earnings losses. Workers with the highest earnings losses have higher firm fixed effects, higher income and higher job tenure for example. But this information is still not enough to identify which factors are causing higher earnings losses. First, it just documents how workers with higher earnings losses differ from workers with little losses and not which factor is leading to higher losses. Second, as workers with high losses differ in many characteristics from workers with little earnings losses, so it is hard to single out the most important factors. One of the succinct ways to understand which factors are the most important is the variable importance measure obtained from the random forest, to which we turn next.

### *A. Which channels are the most important?*

A compact way to assess which factors are the most important in explaining differences in earnings losses is the occurrence frequency of variables in the splitting criteria. Variables chosen more frequently have higher contribution in explaining the heterogeneity of scarring effects. Figure 10 shows that the variable used by far the most often was the firm-fixed effects, which has been used in 36% of all splits of the depth level lower than 4. That said, this raw statistics might be misleading. Due to a top-down greedy approach, variables chosen first tend to be more important. The importance of variables used at a lower deepness level are underestimated.<sup>21</sup> For this reason, it is a common practise to compute a depth-adjusted

---

<sup>21</sup> To understand this property, imagine two binary partitioning variables,  $x_1$  and  $x_2$ , where  $x_1$  is more important than  $x_2$ . The tree-building algorithm will always choose  $x_1$  first and  $x_2$  will be chosen later and will be conditioned on two values of  $x_1$ . As a result, the more important variable  $x_1$  will occur in the splitting



variable frequency, which puts higher weights for splits of lower levels. Figure 11 shows the adjusted frequency with a decay exponent equal to  $-2$ .<sup>22</sup> For this measure, the firm fixed effect is even more important relatively. This result is quite striking. Due to variable randomization, random forests tend to spread the importance out across many variables.<sup>23</sup> However, in our exercise, the firm fixed effect continues being the dominating variable. The regional average firm fixed effect is the second most important variable and the worker's income before displacement is the third one.

Another way to judge the statistical importance of the different channels is to estimate for how much of the total variance each individual factor can explain. In order to estimate this, we linearly project the 3 most important variables, *viz.*: firm wage premia  $\psi_i$ , regional average wage premia  $\psi_i^R$ , and income prior to the job displacement  $inc_i$ , onto the individual earnings losses  $\hat{\tau}_i^{\oplus}$  estimated using our random forest and perform a variance decomposition. In practise, we estimate the following model:

$$\hat{\tau}_i^{\oplus} = \beta_0 + \beta_1\psi_i + \beta_2\psi_i^R + \beta_3inc_i + \epsilon_i. \quad (7)$$

Table 6 depicts the variance decomposition of the estimated model. Here again, the firm wage premia have the biggest contribution and its variance alone accounts for almost 40% of variability in  $\hat{\tau}_i^{\oplus}$ . Then additional 10% is contributed by the interaction with the income prior to the displacement, exhibited by their covariance term.

Share of total variance	
$\text{Var}(\hat{\tau}_i^{\oplus})$	1
$\text{Var}(\beta_1\psi_i)$	0.382
$\text{Var}(\beta_2\psi_i^R)$	0.040
$\text{Var}(\beta_3inc_i)$	0.094
$2\text{Cov}(\beta_1\psi_i, \beta_2\psi_i^R)$	-0.028
$2\text{Cov}(\beta_1\psi_i, \beta_3inc_i)$	0.104
$2\text{Cov}(\beta_2\psi_i^R, \beta_3inc_i)$	-0.007
$\text{Var}(\epsilon_i)$	0.416

Table 6: Variance decomposition. Calculations based on regression model (7)

It is noteworthy that two measures used so far evaluate the variable importance in the criteria only once, while a variable of the second order,  $x_2$ , will be used twice.

<sup>22</sup>It means that split frequencies for nodes of depth  $k$  is 50% less important than split frequencies for nodes of depth  $k - 1$ .

<sup>23</sup>For more details see Efron and Hastie (2016, pp. 331–332).

training set. Nevertheless, in our analysis we are able to go even further and we investigate the counterfactual changes in the value of one variable while keeping all others at their empirical level. For this purpose, we employ partial dependence plots, which are presented in subsection C.

### *B. Conditional Average Treatment Effects*

How do earnings losses vary by changing one factor at a time? Do these variables affect earnings losses through declines in wages, or through losses in days employed, or both? To answer this question, we take advantage of our generalized random forests where we can compute losses conditional on particular realizations of the partitioning variables  $\tau(z)$ . We compute how earnings, employment, wage and firm wage premia losses vary, by changing the realization of one factor at a time, while holding all other variables fixed at their median (or mode for variables with two categories). This way we measure the impact of changing one channel at the time, while holding confounding factors constant. In addition, by comparing the outcomes for earnings, employment and wage losses, we can study whether the channel affects earnings losses through employment or wage losses.

Figure 12 and 13 report how earnings losses change with the 15 different partitioning variables considered, sorted by the variable importance measure. We categorize all continuous variables in deciles according to the overall distribution of workers employed on the reference day. Since displaced workers and the selected control group might differ from the general population, we also include a boxplot of the distribution of our sample over the variable of interest on top of every plot. Alongside the point estimate, the figures also contain 95% confidence intervals that account for the uncertainty arising from both the machine learning procedure and the estimation procedure.<sup>24</sup> We scale our results by 11 so that the results can be interpreted as 11 year cumulative earnings losses. To enhance comparability, we are holding the  $y$ -scale constant across figures. This way it is easy to see which variables are more important drivers of earnings losses.

#### *B.1. Firm Wage Premia*

Figure 12 confirms the finding from the variable importance measure, that the displacement firm's wage premium is key to understand the cost of job loss. Of all variables considered, earnings losses vary the most with firm wage premia. A worker separating firms paying in

---

<sup>24</sup>See [Athey \*et al.\* \(2019\)](#) and [Sexton and Laake \(2009\)](#) for a detailed description behind the estimation of standard errors.

the bottom 10 percent of the firm pay distribution face earnings losses of €42,965, whereas workers formerly employed in the top decile paying jobs forgo almost 3 times as much, with earnings losses amounting to almost €118,658. These heterogeneous effects are also very precisely estimated. With the exception of the lowest paying jobs, where there are fewer observations (see box-plot over the plot), the confidence intervals are not wider than  $\pm\text{€}10,000$ .

The importance of lost firm rents in explaining earnings losses is further confirmed by studying how losses change by the availability well paying jobs in the region. Moving a worker with median characteristics from a region in the bottom decile of the average firm pay distribution to the highest reduces the estimated earnings losses by over €12,000. Interestingly, compared to other variables, this slope seems low, given that it is the second most important factor in explaining the heterogeneity in earnings losses. This suggests that there are strong interactions effects with other variables, which is confirmed in Figure 14. It shows the effect of firm fixed effects by the average regional firm pay premium. It shows that workers employed in low fixed effect firms, but located in high firm wage premium regions do not experience any income losses.

Holding all confounding variables fixed, worker's separating from high fixed effect firms face higher losses. To understand whether this is purely coming from losses in employer specific wage components, or also through declines in employment, we study how firm fixed effects affect losses in employment, log-wages and firm fixed effects. Figure 16 plots all these outcomes alongside with earnings losses. It is visible, that the differences in earnings losses arise through wage losses, and not through employment losses. The differences across firm FE deciles are small and statistically not significant. In contrast, the slope in wage losses mirrors the slope in earnings losses. Panel (d) reveals further striking results: First, losses in firm fixed effects, which are measured in log-wages, only explain part of wage losses. Wages across all firm fixed effect deciles decline by about 5 log-points more than what can be explained by changes in firm fixed effects. But the *differences* in wage losses across firm fixed effect deciles are entirely explained by differences in lost firm wage premia. Second, changes in firm fixed effects show a mean reversion pattern. Workers employed in firms with above median firm pay face losses in firm wage premia, whereas workers employed in below median paying firms gain in terms of firm pay.

We also consider a robustness exercise, where we compute the firm fixed effects on observations before 1998, and reestimate the random forest on earnings loss observations after 1998. This rules out that the mass layoff event affected our measure of earnings losses.

Appendix D shows that the main findings are essentially unchanged.

## B.2. Worker’s Job Tenure , Age, Income

Perhaps the most prominent theory about the sources of earnings losses is that workers lose job specific human capital. Several papers have shown that workers with higher tenure experience higher losses (Topel, 1990; Jacobson *et al.*, 1993). This is in stark contrast to our findings. The bottom panel in Figure 12 shows that earnings losses only marginally increase from €72,000 for a low tenure worker with otherwise median characteristics to €75,000 for high tenure, but otherwise observational identical worker. In subsection D we show that it is crucial to hold confounding factors constant. A traditional sub-group analysis would yield higher losses for high tenure workers in our sample, but this because high tenure workers differ in many other characteristics from low tenure workers.

In contrast to losses in firm specific human capital, we find evidence that part of the earnings losses arise from losing particularly good matches. Figure 18 displays how losses in earnings, employment, wage and firm premia are affected by the predisplacement income. We interpret income as a proxy for the match specific wage component. This is motivated by a AKM wage specification  $\ln w_{it} = \alpha_i + \psi_{J(it)} + \beta x_{it} + \epsilon_{it}$ , where log wages are a function of a fixed worker, firm components  $\alpha_i$  and  $\psi_{J(i,t)}$ , time-varying worker characteristics such as age and job-tenure  $x_{it}$  and a residual  $\epsilon_{i,t}$  that captures a match specific term. Since we vary income by conditioning on firm fixed effects and time varying worker characteristics, we effectively move the fixed worker component and match specific productivity. Since the former is by definition fixed over time, only match specific productivity can affect wage losses. Panel (a) shows that high income workers face higher earnings losses compared to observationally similar workers with lower income. Although high income workers face lower employment losses, they face higher wage losses. The shape of wage losses closely mirrors the shape of earnings losses. Panel (d) shows that varying income while holding other factors constant does not affect losses in firm wage premia. This is evidence that workers in particularly good matches fall off the match quality ladder as in Jovanovic (1979).

Workers’ age also plays an important role in understanding earnings losses. Workers aged 30 and below face losses of about €50,000, whereas otherwise identical older worker experience double the losses. Thus older workers face higher losses not only because they differ in terms of (observable) characteristics. Figure 17 shows that age affects earnings losses mostly through employment losses. Workers aged 50 experience cumulative employment losses of 22 months, compared with less than 9 months for young workers. The difference

is wage losses is with 3 percent relatively small. These differences cannot be explained by changes in firm wage premia. Panel (d) clearly shows, that losses in firm rents are not influenced by worker’s age.

### B.3. Other Factors

All other factors do not seem to play an important role, varying these variables while holding other factors constant barely changes the estimated earnings losses. For some of this variables this is nevertheless noteworthy, as they have been shown in the prior literature to imply heterogeneous earnings losses by following a traditional sample splitting approach, where the treatment effect is computed on subsamples. For example, [Schwerdt \*et al.\* \(2010\)](#) show that white collar workers face higher losses using the same dataset as we do, although for a different time period with slightly different definitions.

Earnings losses have been shown to substantially vary over the business cycle ([Davis and Von Wachter, 2011](#); [Schmieder \*et al.\*, 2018](#)). Interestingly, for all four variables that capture the business cycle state of the local labor market and the industry, earnings losses do not vary much. It has been shown that the composition of unemployed workers change over the business cycle ([Mueller, 2017](#)). The difference to the prior findings could arise because of our ability to control for compositional differences in workers across business cycle states. But we cannot rule out another explanation. Austria exhibited only very mild business cycles throughout our study period. The national unemployment rates fluctuated between 3.6 and 5.8 percent, which is considerably less compared to Germany or the United States, which were studied in ([Davis and Von Wachter, 2011](#); [Schmieder \*et al.\*, 2018](#)).

### C. Partial Dependence Plots

One criticism for comparative statics from subsection B is that an individual, exhibiting all variables at their median values, might not be representative for the whole population or even may not exist at all. Moreover, focus in the analysis on such an artificial unit might generate misleading results due to plausible non-linearities and interactions between the partitioning variables which occur far from the median. The computed measures may be relevant only for the neighborhood of the median and may disregard other strong effects in the whole sample. For this reason, as a sanity test, we use partial dependence plots proposed by [Friedman \(2001\)](#) to better understand how *on average* a single variable affects the earnings losses in the whole data set. This approach consists in predicting the average earnings losses  $\tau$  where a value of the variable of our interest is kept fixed  $z^k = \bar{z}$  and values of other variables  $\mathbf{z}^{-k}$

are taken from the global distribution with a cdf denoted  $F(\mathbf{z}^{-k})$  :

$$\mathbb{E}_{\mathbf{z}^{-k}} \hat{\tau}(z_k = \bar{z}; \mathbf{z}_{-k}) = \int \hat{\tau}(z^k = \bar{z}; \mathbf{z}^{-k}) dF(\mathbf{z}^{-k}), \quad (8)$$

which in our application can be estimated on our training set:  $\frac{1}{N} \sum_{i=1}^N \hat{\tau}(z^k = \bar{z}; \mathbf{z}_i^{-k})$ .

Figures 19 and 20 depict partial dependence plots for different deciles of all partitioning variables.<sup>25</sup> All the main findings from the previous exercise preserve, with the difference, that the level of effects from partial dependence plots is slightly shifted upwards for some variables.

#### *D. Comparison to Sample Splitting Approach*

The most conventional method of understanding heterogeneity in treatment effects is sample splitting, where the treatment effect is computed separately on sub-groups. In Figure 21 we compare our results to the sample splitting approach. For this, we re-estimate equation (2) on the sample of workers with the specific realization of the variable of interest. The boxplots at the top of the plots indicate the sample distribution over the variable of interest and allows a judgment of the sample size in the respective group.

The comparison reveals that in general, that the subgroup analysis using sample splitting yields much noisier estimates than the results from the generalized random forest. This is to be expected, OLS tends to overfit the data, since it is sensitive to outliers. This is unsurprisingly more likely to occur in regions with fewer observations, as can be seen from the boxplots at the top of the plots which depict the sample distribution of the variable of interest. The machine learning approach is less prone to over-fitting, since it is estimating the heterogeneity by averaging over thousands of trees, each of which is only using a bootstrapped sample of the whole data.

In addition to this regularization advantage of the generalized random forest, there is a more substantial difference to the sample splitting approach. We can change one variable at the time, while holding all other confounding factors constant. This is in contrast to the sample splitting approach. If one splits the sample according to some characteristics in two subgroups, they are likely be different in many dimensions, not only in the variable of interest. Although it might be of interest how the cost of job loss differs across different sub populations, it is hard to draw conclusions about the driving factors behind it as the

---

<sup>25</sup> In this exercise due to a very high memory consumption, we needed to decrease the number of trees to 2,000 instead of initial 5,000.

two groups differ in many characteristics. Figure 21 shows that this difference is not only a theoretical concern, but leads to different results. The sample splitting approach leads to an overestimation of the importance of each factor. This is because these factors are correlated with each other. High income workers tend to work for firms with high wage premia. At the same time, job tenure tends to be higher for older workers with higher income as can be seen in the correlogram (Figure 3).<sup>26</sup> Therefore, from the sample splitting with respect to one variable it is difficult to infer what characteristics account for differences in earnings losses.

Next we show that differences between subgroup analysis and the random forest outcomes are truly driven by our ability to control for confounding factors. Once we further condition on more similar workers in terms of age and income in the traditional subsample splitting, we also conclude that job tenure is not an important driver for heterogeneity anymore. Figure 22 presents earnings losses for each subsample in a hierarchical structure, quite analogous to a tree-based model. One can see that the split according to 8<sup>th</sup> job tenure decile generates a between-group difference in scarring effects amounting to almost 11,000 Euros ( $-69,276 + 58,491$ ). However, if for both samples we conduct a further split for individuals younger and not younger than 39, then the impact of higher job tenure conditional on age will dwindle to less than 5,000 Euros for younger workers ( $-56,366 + 51,653$ ) and less than 8,000 Euros ( $-79,667 + 71,803$ ) for older workers. The subsequent split according to 6<sup>th</sup> income decile will reduce the impact of job tenure even more. For around 57% of the population the impact of job tenure is smaller than 1,000 Euros (for young low-income workers and old high-income workers; for the latter group it is even negative), while for around 18% of the population the impact of job-tenure is around 2,500 Euros. The remaining 25% exhibit the impact on the treatment effect that amounts to less than 8,500 Euros. Consequently, we see the effect of job tenure conditional on other confounding factors is lower than suggested by a standard subgroup analysis. While in standard subgroup analysis different compositions of all characteristics cannot be ruled out as the driver for changes in treatment effects, in tree-based models we are always able to find two subsamples with all characteristics but one at the same level. Furthermore, a random forest being an ensemble of many tree-based models allows us to evaluate the heterogeneity in scarring effects in a far richer and complete way than in the presented example.

---

<sup>26</sup> Admittedly, there are other correlated variables, but in our illustration we chose those which exhibit a relatively high importance in our forest (Figures 10 and 11).

## VII. CONCLUSIONS

We adapt the generalized random forest approach by [Athey \*et al.\* \(2019\)](#) to a difference-in-difference setting to study the sources of earnings losses of displaced workers. This methodology allows us to make two important empirical contributions to the existing literature. First, we document the heterogeneity in the scarring effect of job loss across individuals. Using the universe of Austrian social security records from 1984-2017, first we show that there is substantial heterogeneity in earnings losses across individuals. While the average cumulative earnings losses over 11 years after job displacement amount to €62,000, the interquartile range of individual losses is almost €52,000. We also show that almost 10 percent of workers do not face any earnings losses. Second, the machine learning procedure allows us to conduct a horse race between many competing theories about earnings losses. We show that the pre-displacement firm wage premium is by far the most important factor. Holding all other channels fixed, the counterfactual earnings losses rise from €43,000 for workers employed at the lowest paying decile of firms to almost triple that amount for workers displaced from the highest paying decile of firms. The ability to hold other characteristics constant is important.

We illustrate this by a comparison of our results to a traditional subgroup analysis via sample splitting. Perhaps the most striking difference is regarding the lost job-specific human capital channel. A sample splitting approach would suggest this as an important mechanism, while we show that once holding other confounding factors constant, we do not find any impact of job tenure on the level of earnings losses.

Further important factors are the availability of well paying jobs in the local labor market, income prior to displacement and worker’s age. These findings are consistent with a job-ladder search models, where separating workers fall off the firm and match quality ladder. Younger workers as well as workers located in regions with many well paying jobs find it easier to climb back, while older workers face higher earnings losses.



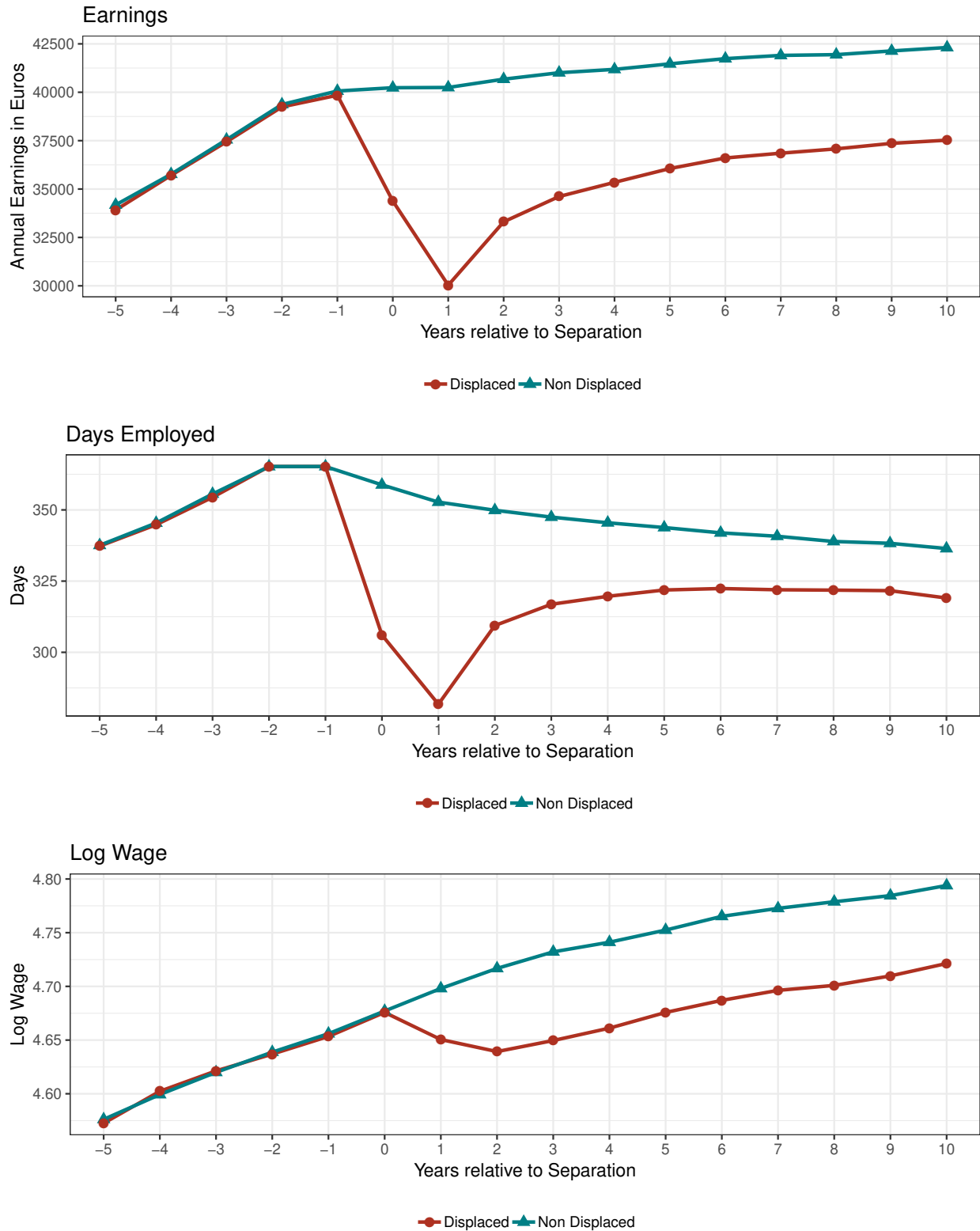


Figure 1: Earnings evolution and days employed over time for workers displaced and control group. Control group was select via propensity score matching

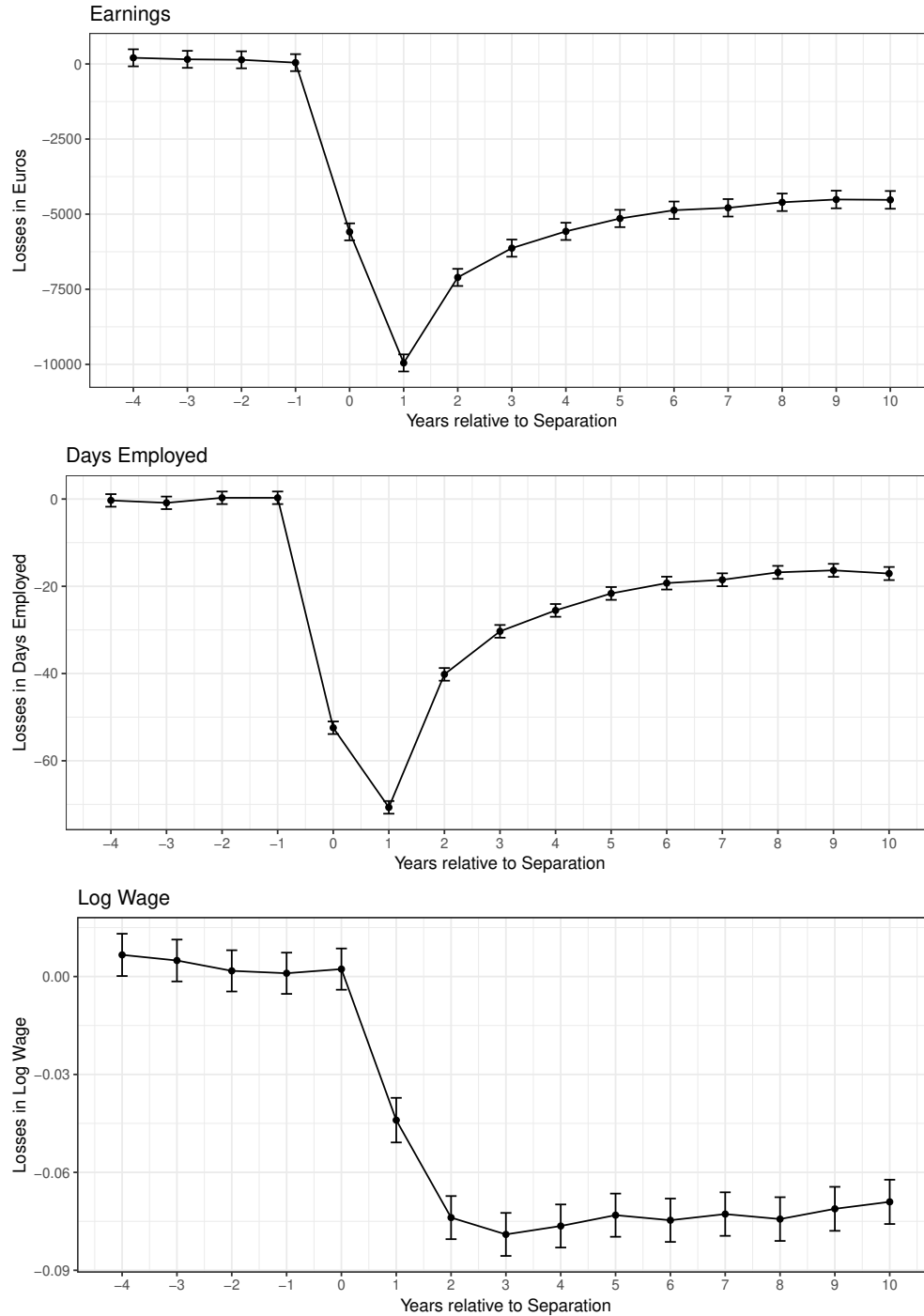


Figure 2: Earnings Losses of displaced workers - Eventstudy regression estimates, average treatment effects. The estimates show the average treatment effects of job displacements. Period 0 corresponds to the separation year. Results from pooling workers displaced between 1989-2007, while outcome data spans 1984-2017. Earnings and days employed are computed for the whole year, log wages are computed as the log average daily wage from the employer on 1st January. Control group is selected via propensity score matching.

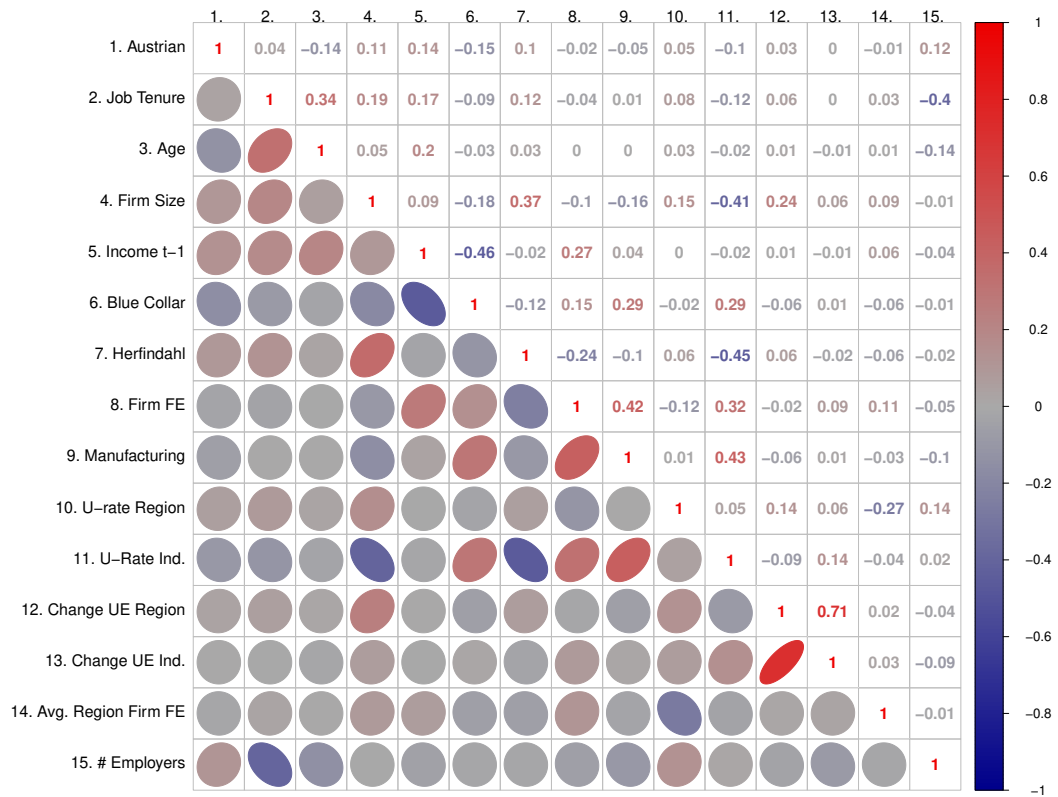


Figure 3: Correlogram of partitioning variables

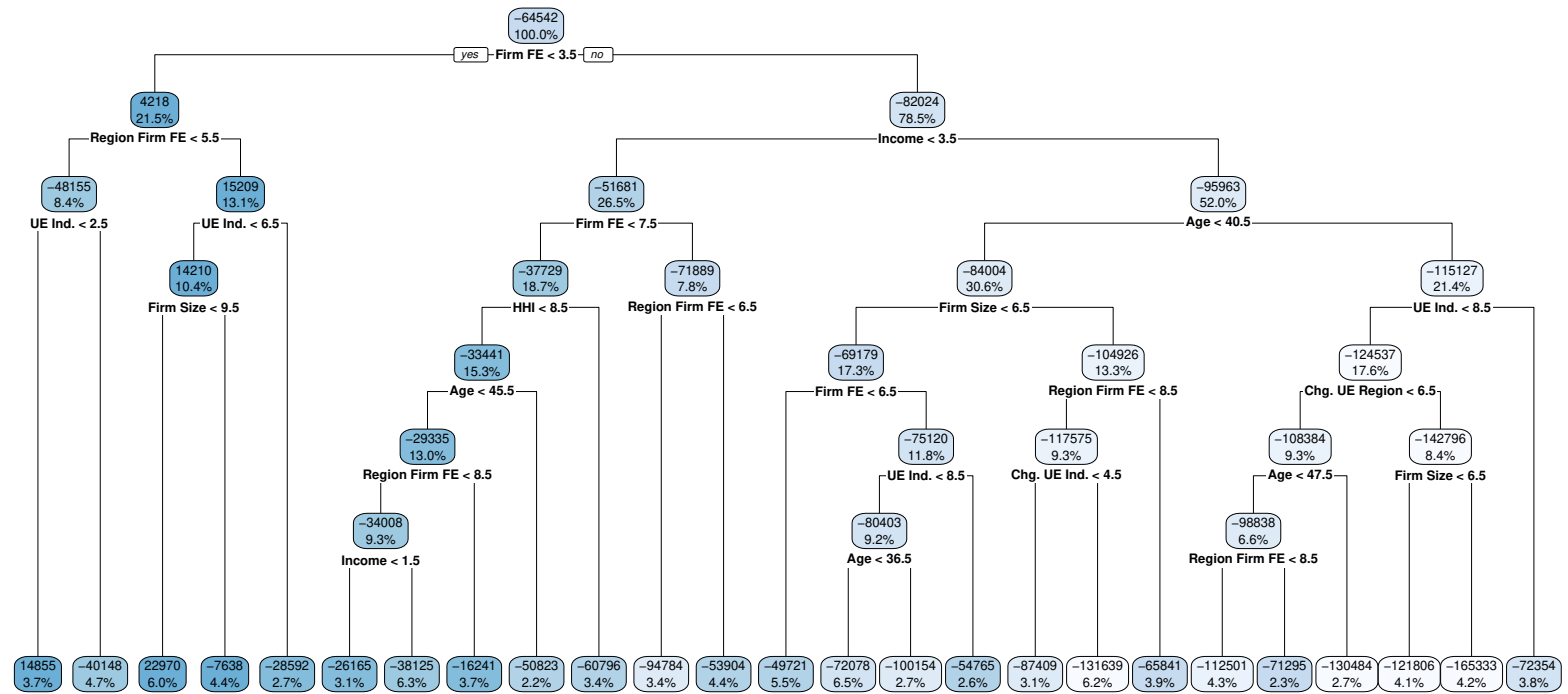


Figure 4: Heterogeneous treatment effect Job Displacement. Regression tree was built with a CART algorithm using the relabeling strategy as described in the text. Min node size of 30000.

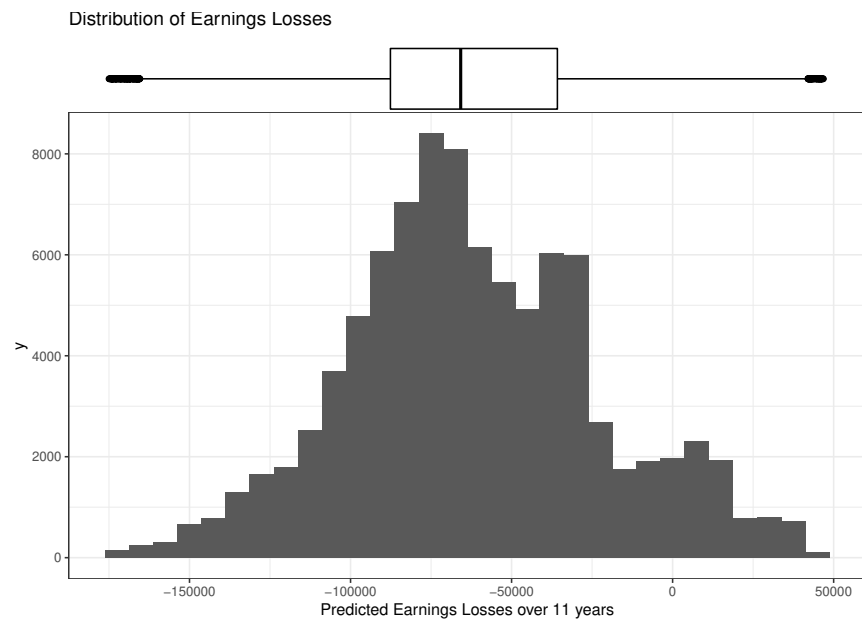


Figure 5: Distribution of predicted cumulative earnings losses over the first 11 years after job displacement. Predictions from a generalized random forest.

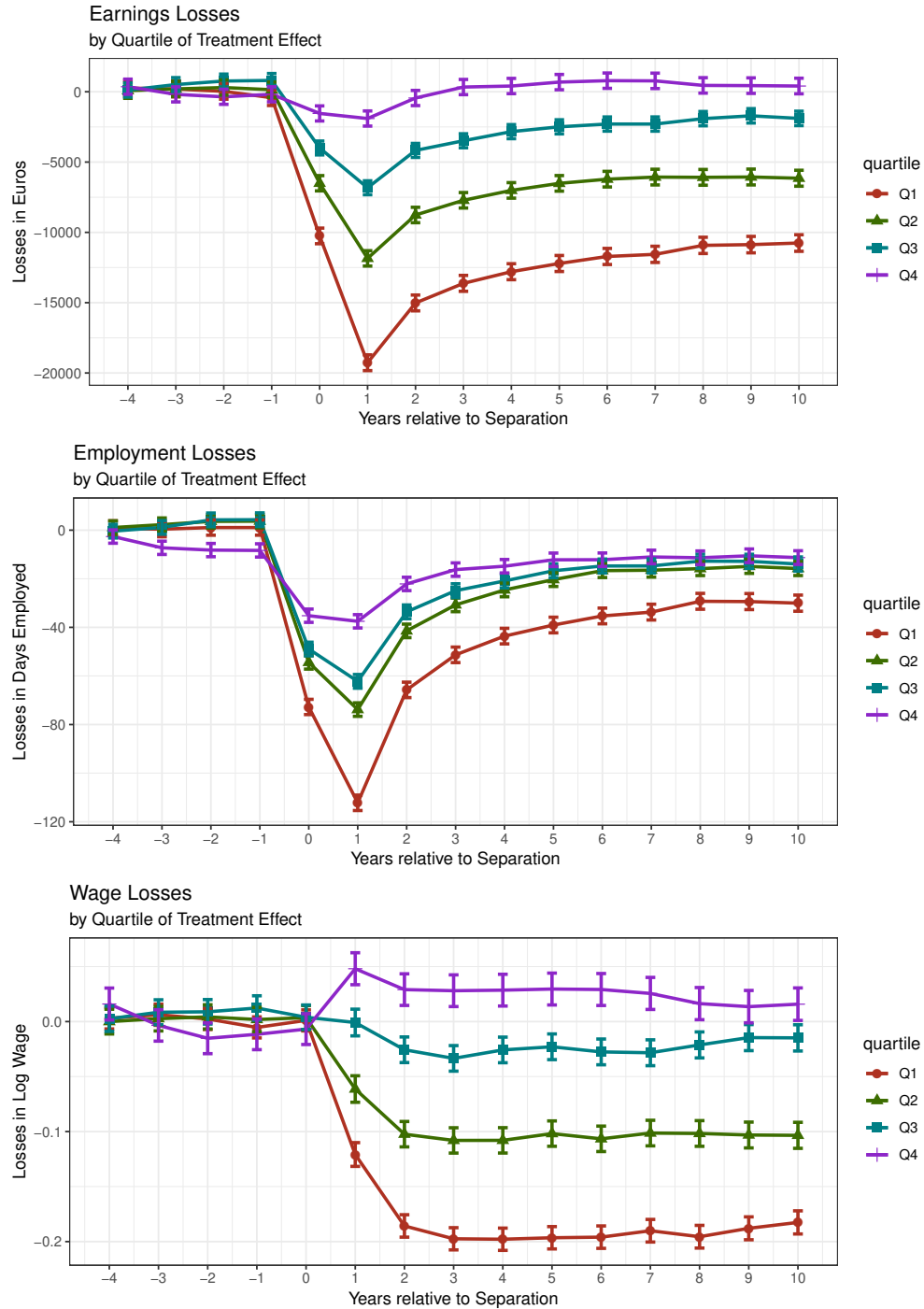


Figure 6: Earnings, employment and wage losses of displaced workers - Eventstudy regression estimates, by quartile of predicted earnings losses over first 11 years after displacement, predictions from a random forest. Period 0 corresponds to the separation year. Results from pooling workers displaced between 1989-2007, while outcome data spans 1984-2017. Earnings and days employed are computed for the whole year, log wages are computed as the log average daily wage from the employer on 1st January. Control group is selected via propensity score matching.

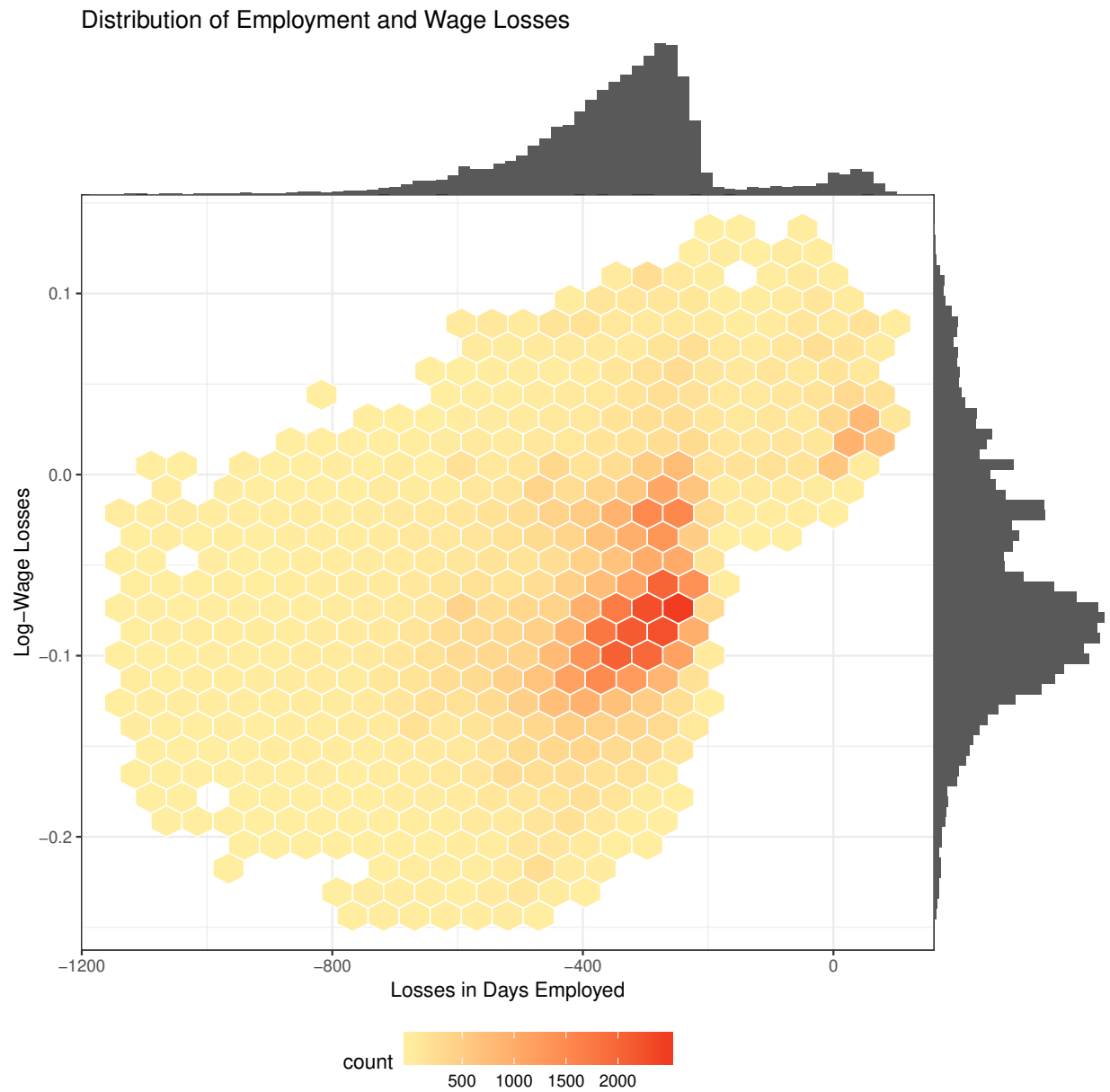


Figure 7: Joint distribution of wage and employment losses generated by two independent random forests.

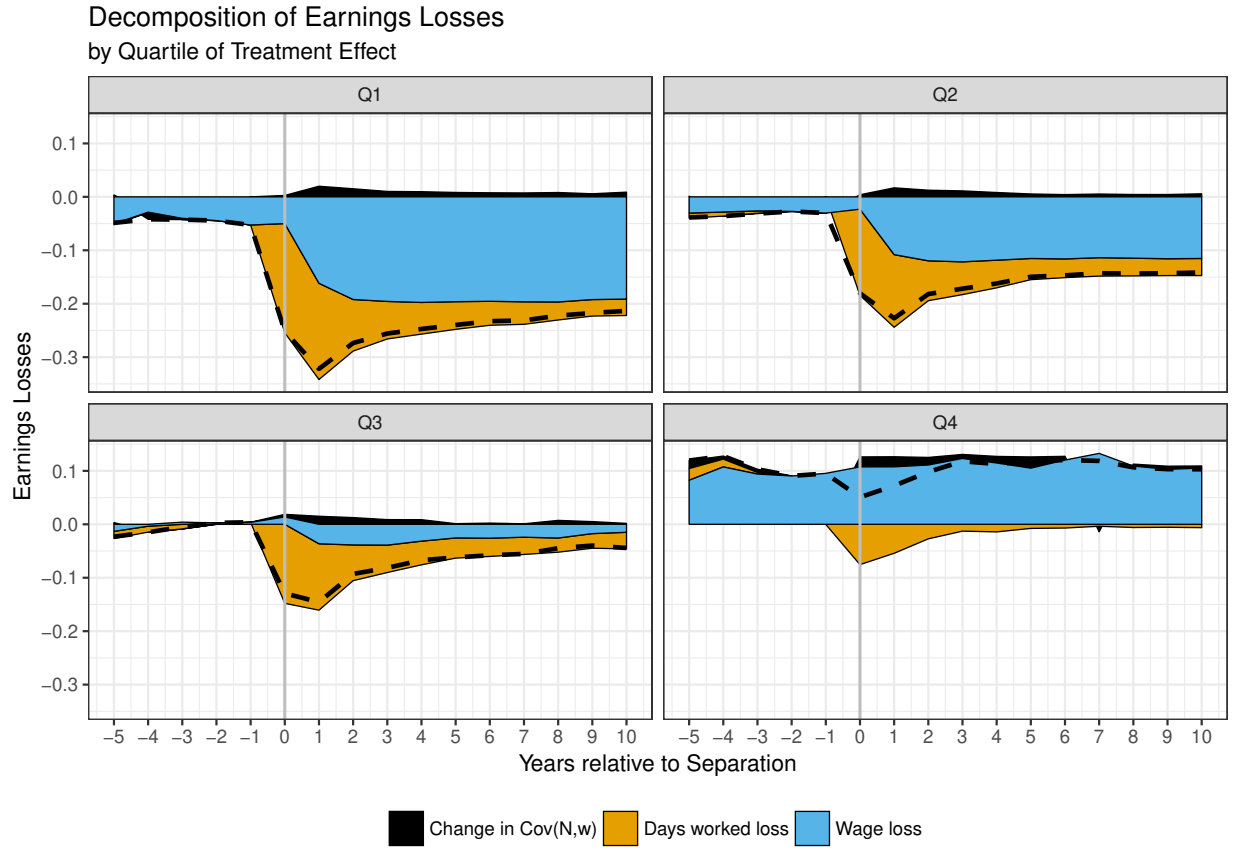
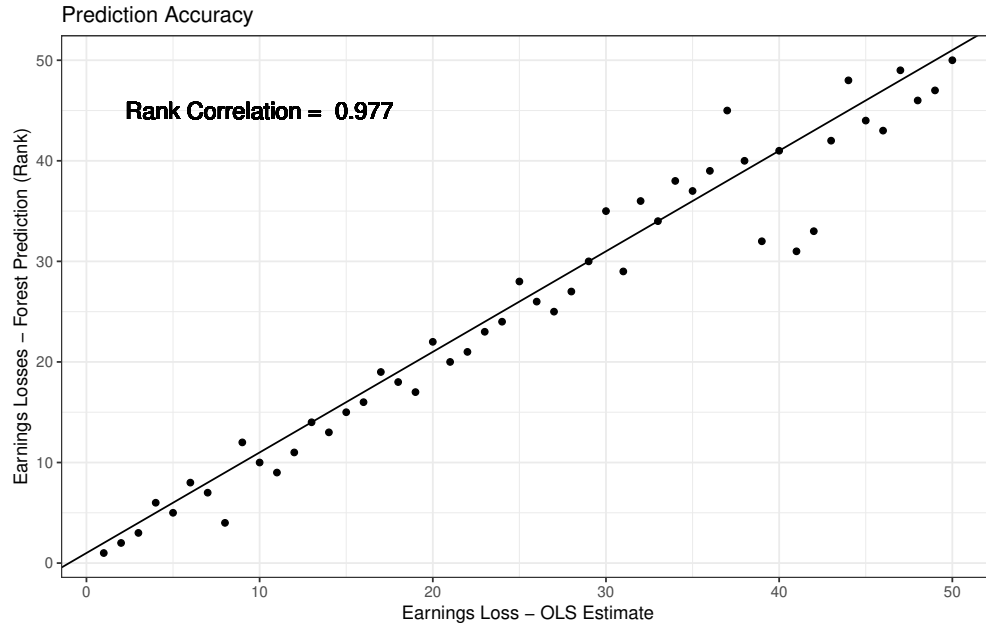
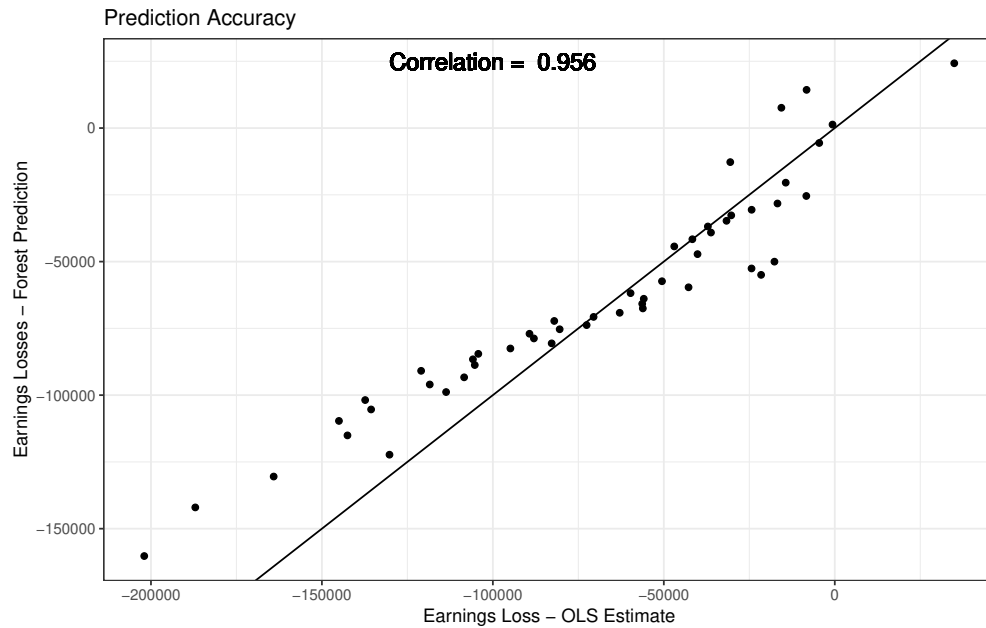


Figure 8: Decomposition of earnings losses by quartile of predicted treatment effect using equation (6). Predictions from a generalized random forest. Broken line indicates total earnings losses.





(a) Rank scatter plot of forest prediction against OLS estimate



(b) Bin scatter plot of forest prediction against OLS estimate

Figure 9: Bin scatter plot of prediction accuracy. We bin all individuals by their predicted treatment effects into 50 bins. For these 50 subgroups, we compute the OLS regression and plot the estimated cost of job displacement against the average forest prediction

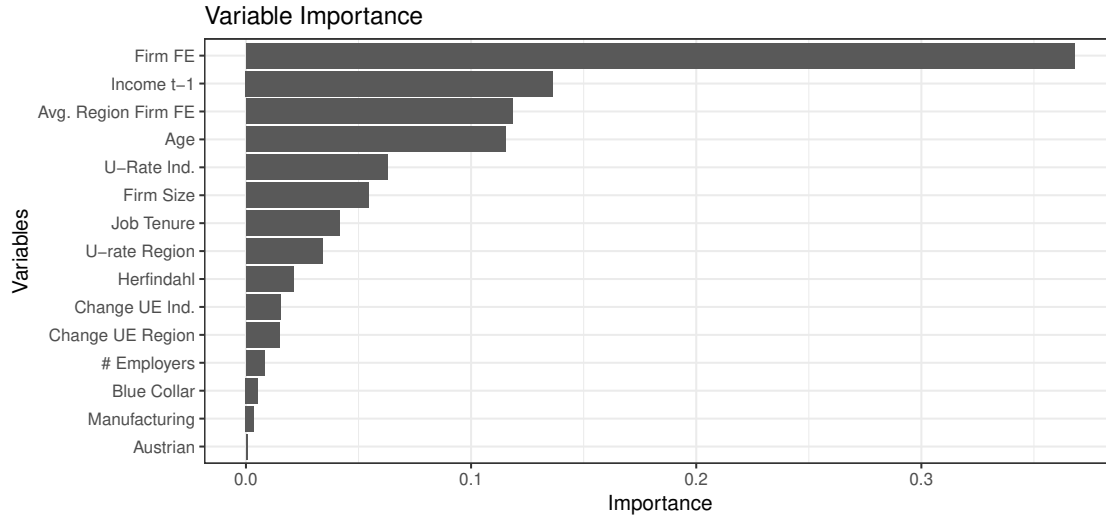


Figure 10: Variable frequency in splits in the GRF and the maximum depth level of nodes equal to 4. All values sum to 1.

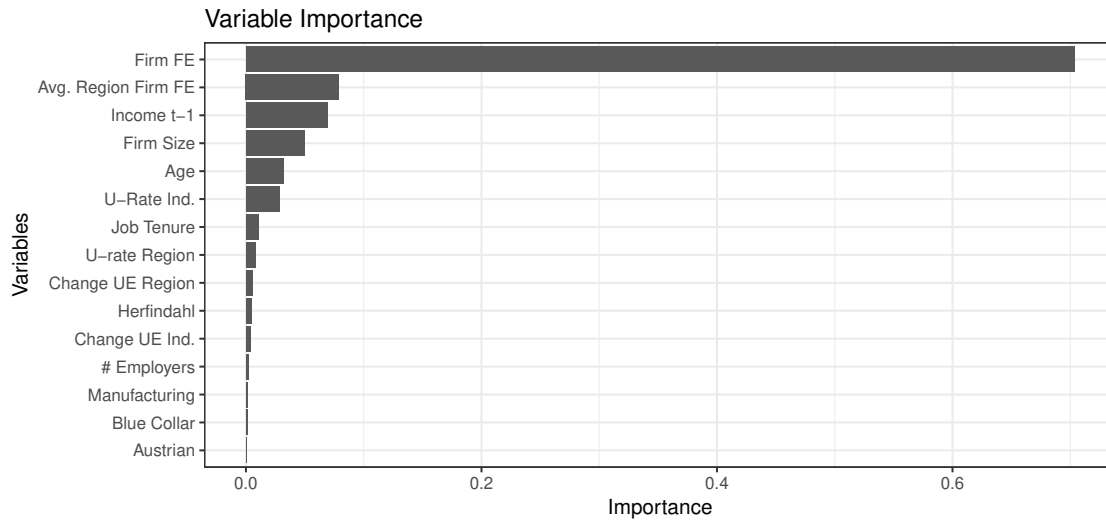


Figure 11: Depth-adjusted variable frequency in splits in the GRF with a decay exponent equal to  $-2$  and the maximum depth level of nodes equal to 4. All values sum to 1.

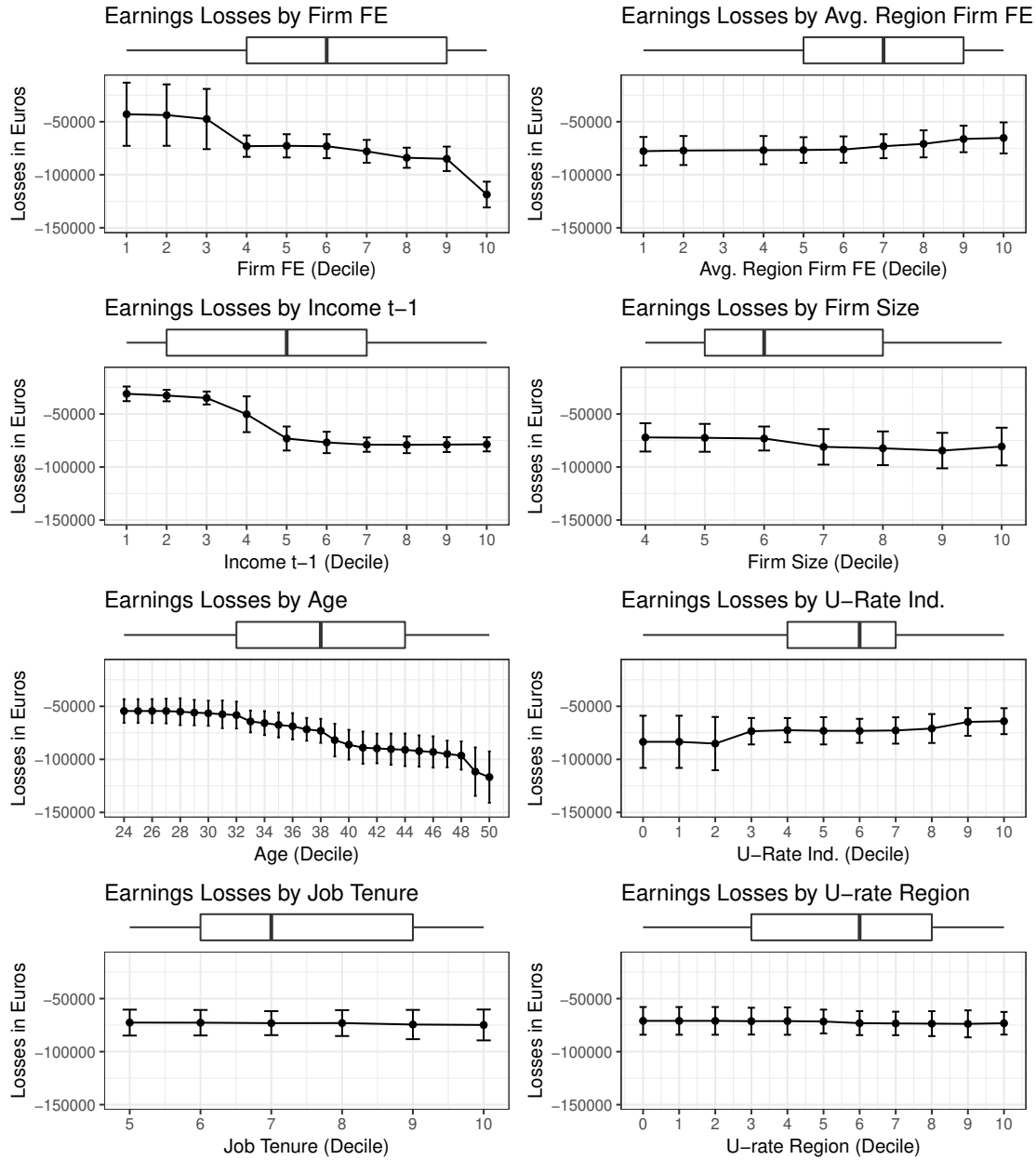


Figure 12: GRF estimates of the cumulative 11-year earnings losses with 95% confidence intervals (part I). All variables that are not used in a given plot are set to their mean values. The  $y$ -axis indicates cumulative 11-year earnings losses, while the  $x$ -axes depict global deciles of the unrestricted dataset (except for the age, number of employees, blue collar, manufacturing, and Austrian where the actual values are used). The boxplots above figures represent the distribution of the partitioning variable in the restricted dataset (for the details see Subsection B of Section II).

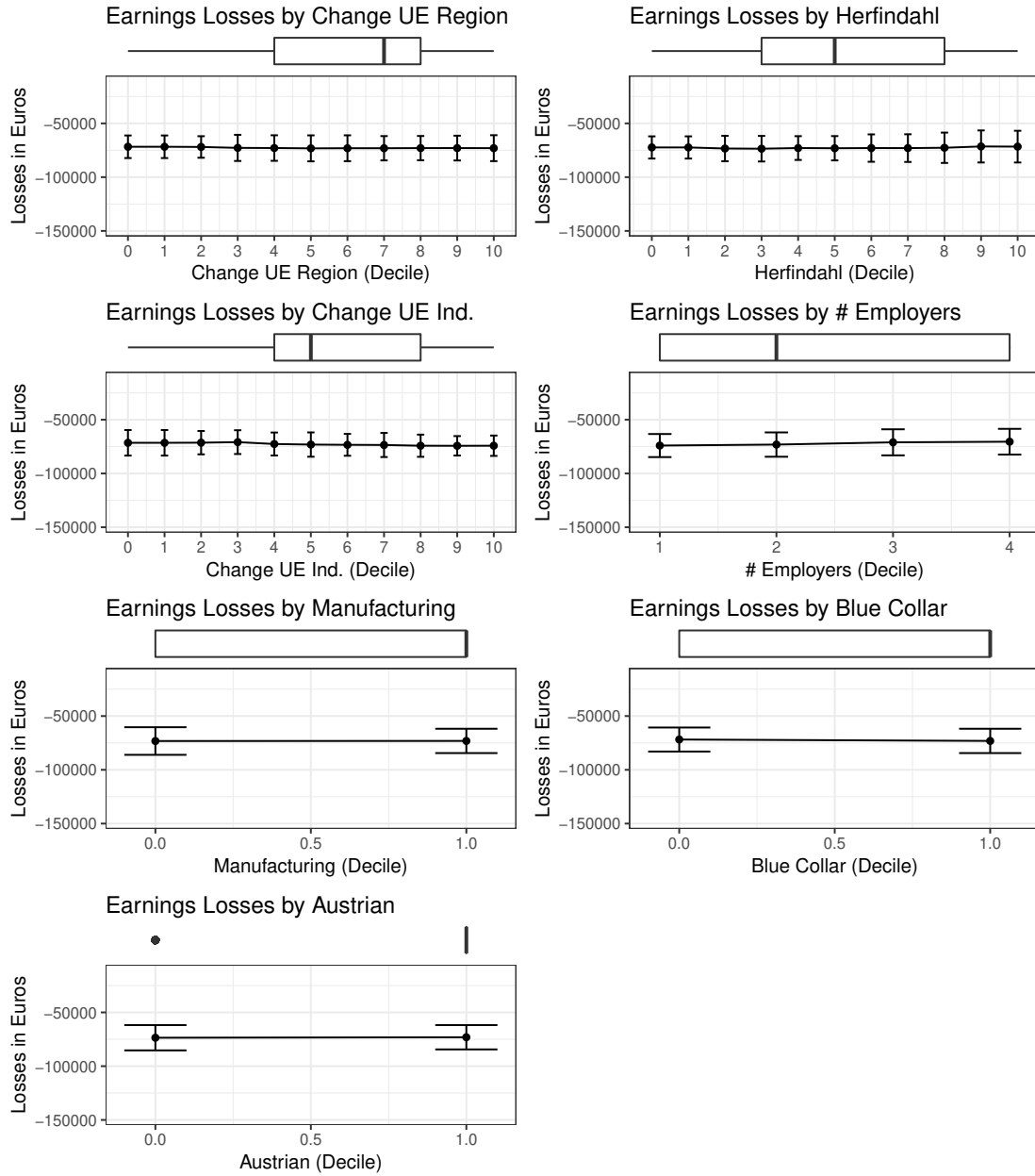


Figure 13: GRF estimates of the cumulative 11-year earnings losses with 95% confidence intervals (part II). All variables that are not used in a given plot are set to their mean values. The  $y$ -axis indicates cumulative 11-year earnings losses, while the  $x$ -axes depict global deciles of the unrestricted dataset (except for the age, number of employees, blue collar, manufacturing, and Austrian, where the actual values are used). The boxplots above figures represent the distribution of the partitioning variable in the restricted dataset (for the details see Subsection B of Section II).

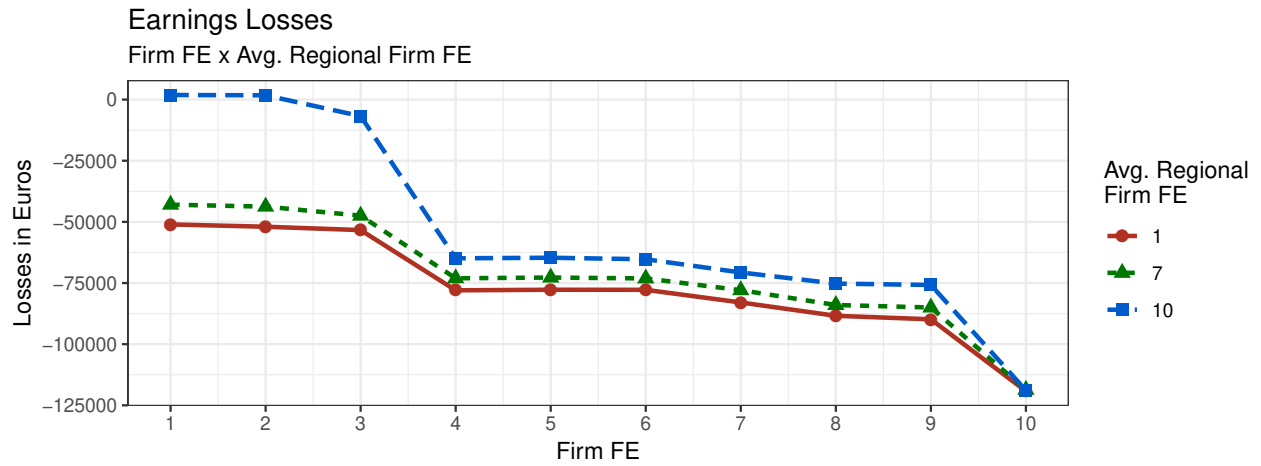


Figure 14: GRF estimates of the cumulative 11-year earnings losses for different values of the firm fixed effect and the regional average firm fixed effect. All other variables are set to their mean values.

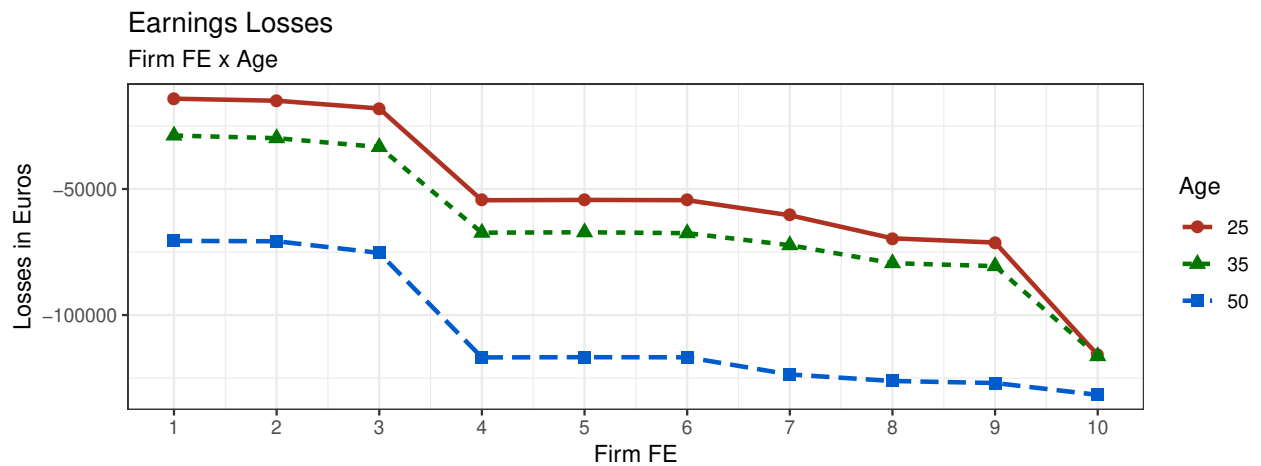


Figure 15: GRF estimates of the cumulative 11-year earnings losses for different values of the firm fixed effect and worker age. All other variables are set to their mean values.

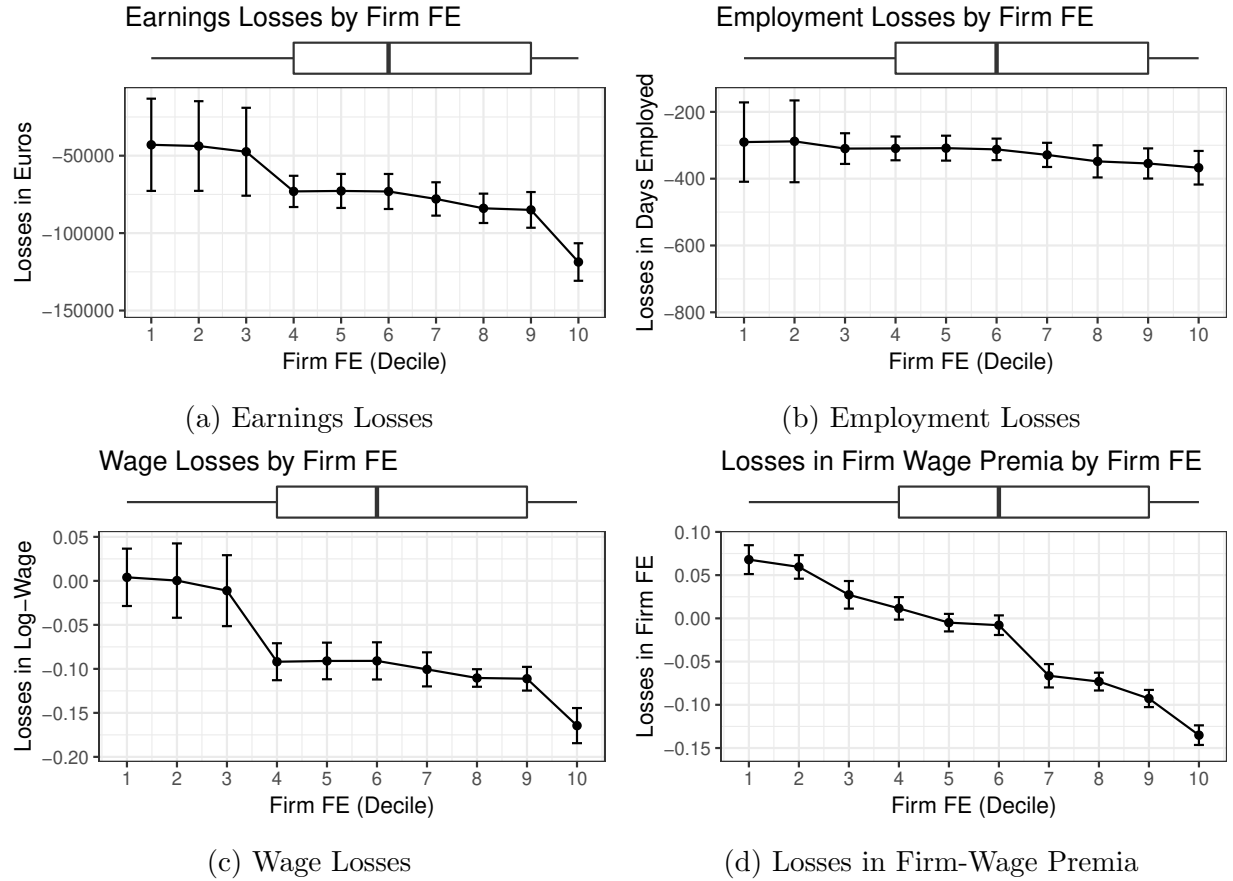


Figure 16: GRF estimates with 95% confidence intervals of losses in earnings, employment, wages, and firm premia for different values of the firm fixed effect. All other variables are set to their mean values. The boxplots above figures represent the distribution of the partitioning variable in the restricted dataset (for the details see Subsection B of Section II).

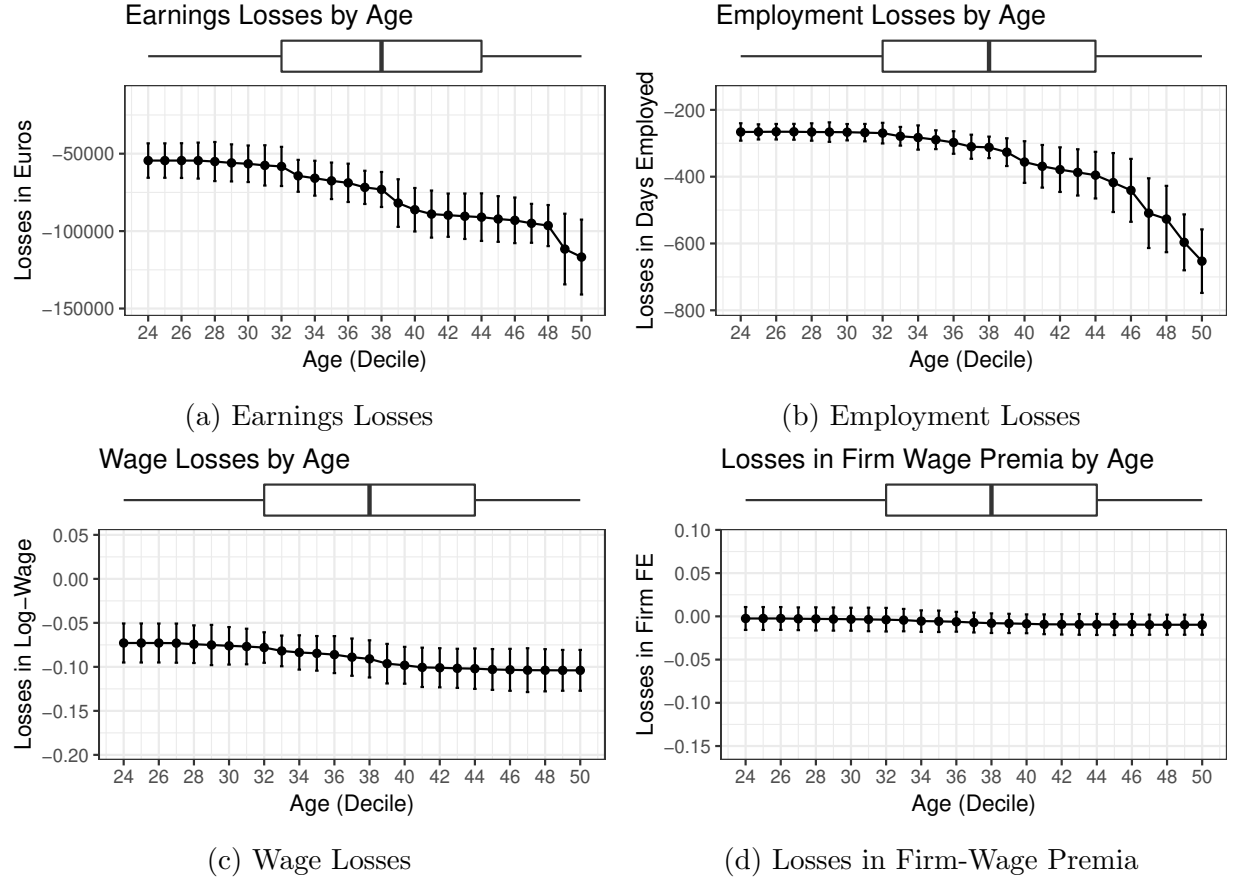


Figure 17: GRF estimates (with 95% confidence intervals) of losses in earnings, employment, wages, and firm premia for different values of worker age. All other variables are set to their mean values. The boxplots above figures represent the distribution of the partitioning variable in the restricted dataset (for the details see Subsection B of Section II).

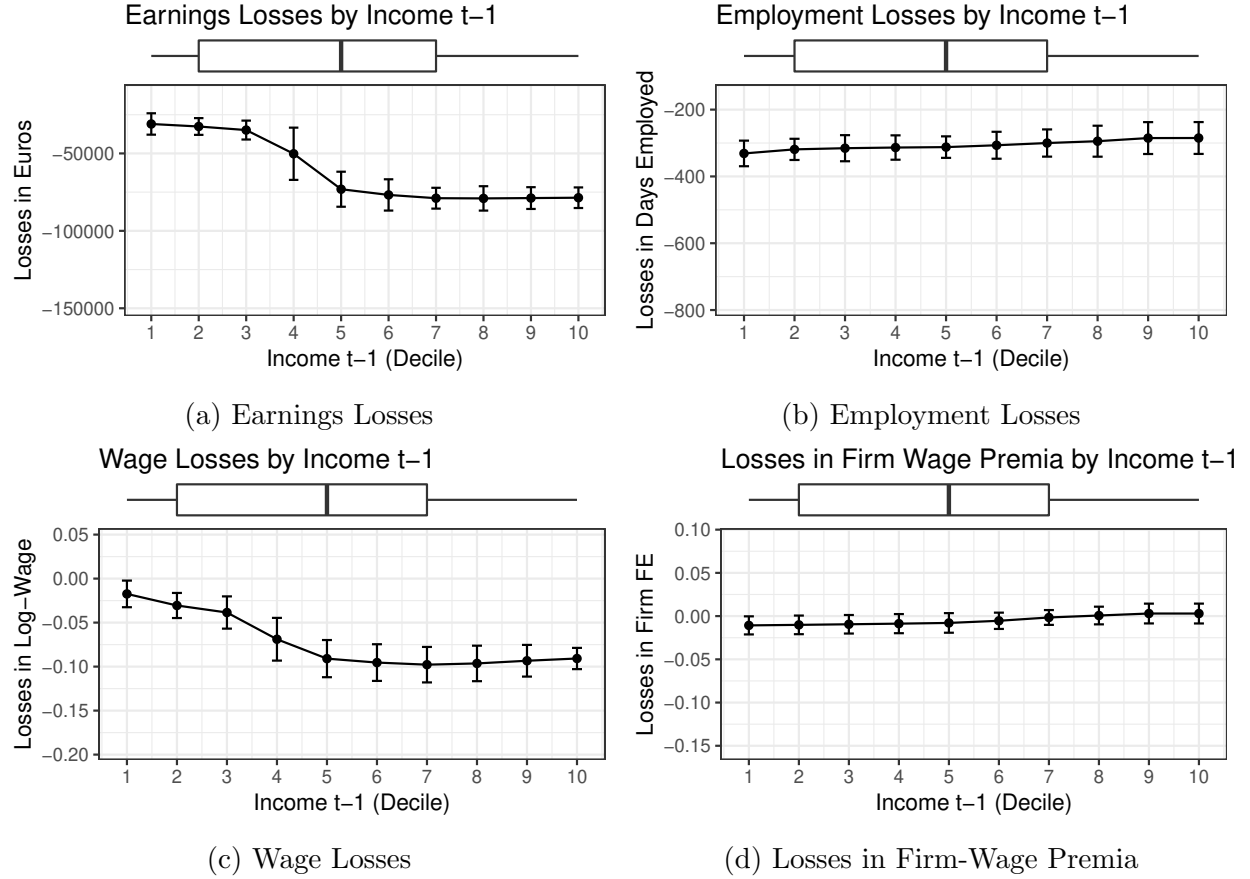


Figure 18: GRF estimates (with 95% confidence intervals) of losses in earnings, employment, wages, and firm premia for different values of income prior to the displacement. All other variables are set to their mean values. The boxplots above figures represent the distribution of the partitioning variable in the restricted dataset (for the details see Subsection B of Section II).



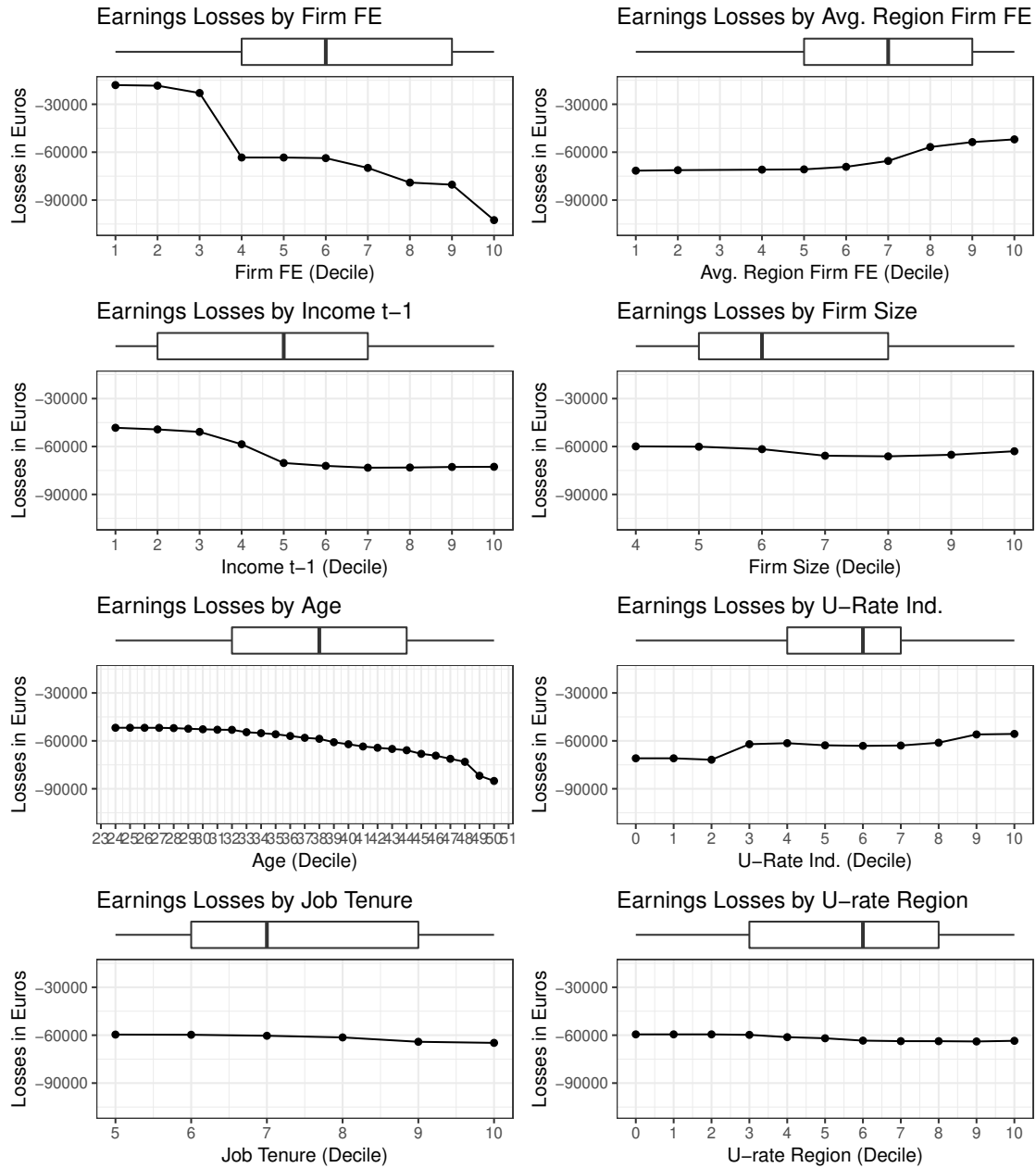


Figure 19: The figure shows partial dependence plots (for details on computations see Subsection C of Section VI) for different variables (part I). The  $y$ -axis indicates cumulative 11-year earnings losses, while the  $x$ -axes depict global deciles of the unrestricted dataset (except for the age, where the actual values are used). The boxplots above figures represent the distribution of the partitioning variable in the restricted dataset (for the details see B).

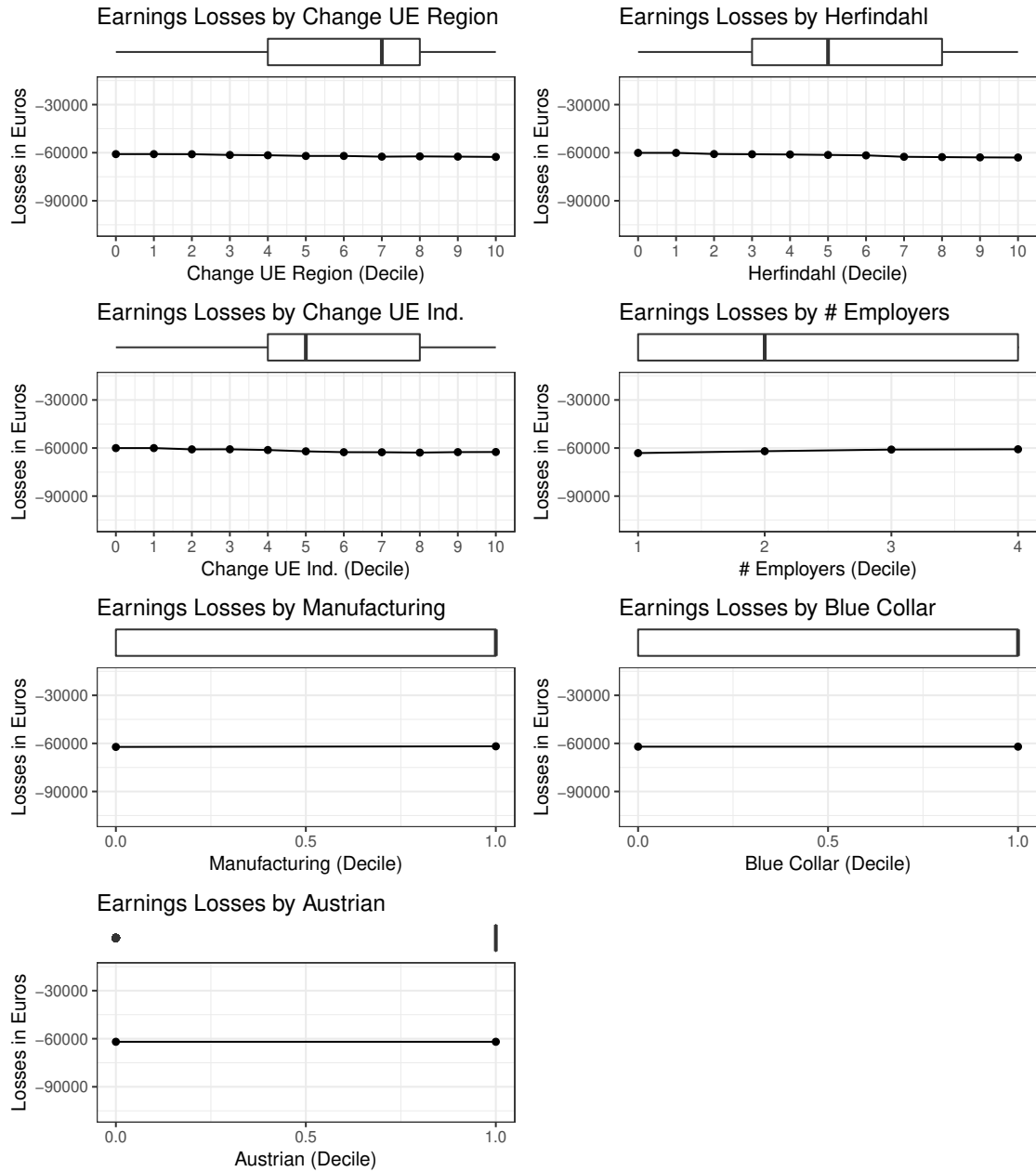
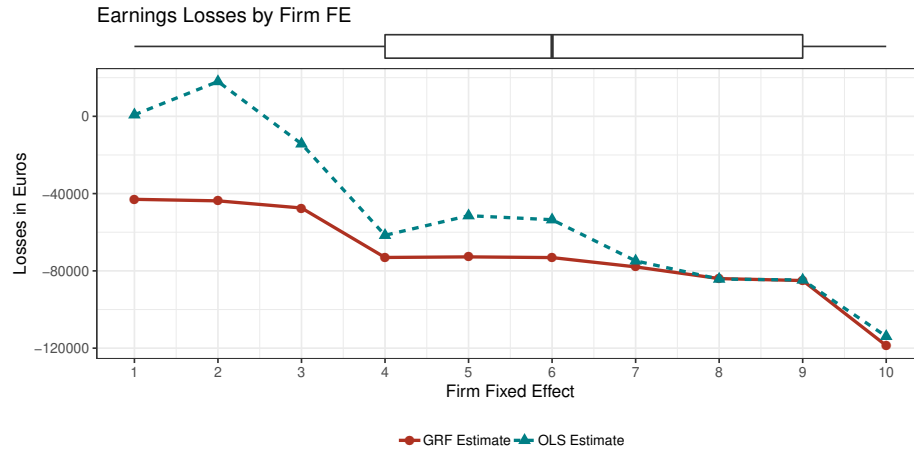
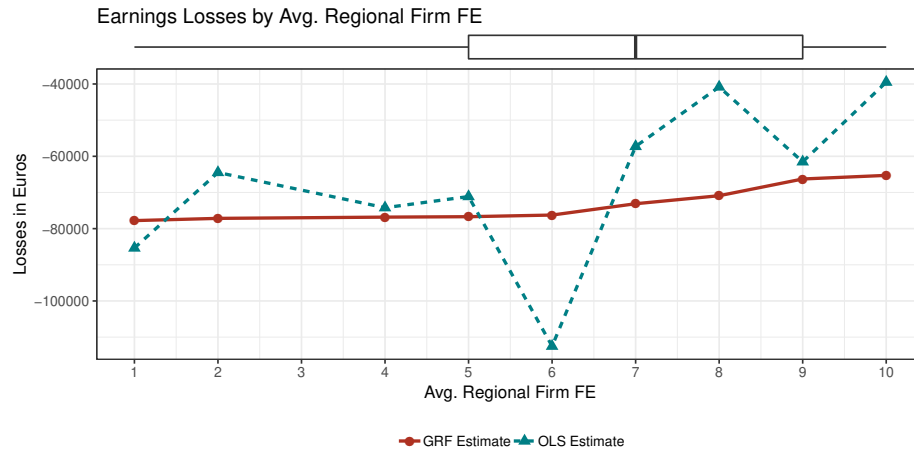


Figure 20: The figure shows partial dependence plots (for details on computations see Subsection C of Section VI) for different variables (part II). The  $y$ -axis indicates cumulative 11-year earnings losses, while the  $x$ -axes depict global deciles of the unrestricted dataset (except for the age, where the actual values are used). The boxplots above figures represent the distribution of the partitioning variable in the restricted dataset (for the details see B).



(a) Earnings Losses by Firm FE



(b) Earnings Losses by Income



(c) Earnings Losses by Job Tenure

Figure 21: Comparison between Sample Splitting and GRF.

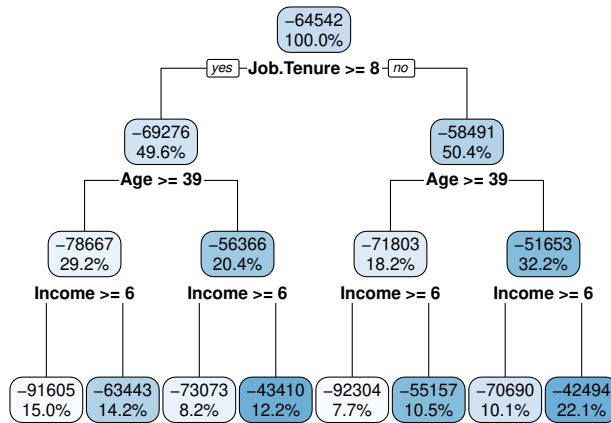


Figure 22: Hierarchical subgroup analysis of the scarring effect with respect to some splits in job tenure, age, and income prior to displacement. Subgroup estimates obtained with OLS.

## REFERENCES

- ABOWD, J. M., KRAMARZ, F. and MARGOLIS, D. N. (1999). High wage workers and high wage firms. *Econometrica*, **67** (2), 251–333.
- ALBANESI, S. and VAMOSSY, D. F. (2019). *Predicting Consumer Default: A Deep Learning Approach*. mimeo.
- ATHEY, S. and IMBENS, G. (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, **113** (27), 7353–7360.
- , TIBSHIRANI, J., WAGER, S. *et al.* (2019). Generalized random forests. *The Annals of Statistics*, **47** (2), 1148–1178.
- AZAR, J., MARINESCU, I. and STEINBAUM, M. I. (2017). *Labor Market Concentration*. Tech. rep., National Bureau of Economic Research.
- BAGGER, J. and LENTZ, R. (2018). An empirical model of wage dispersion with sorting. *The Review of Economic Studies*, **86** (1), 153–190.
- BERTRAND, M. and KAMENICA, E. (2018). *Coming Apart? Cultural Distances in the United States over Time*. Working Paper 24771, National Bureau of Economic Research.
- BREIMAN, L. (2001a). Random forests. *Machine Learning*, **45** (1), 5–32.
- (2001b). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statist. Sci.*, **16** (3), 199–231.
- , FRIEDMAN, J. H., OLSHEN, R. A. and STONE, C. J. (1984). *Classification and Regression Trees*. Statistics/Probability Series, Belmont, California, U.S.A.: Wadsworth Publishing Company.
- CARD, D., HEINING, J. and KLINE, P. (2013). Workplace heterogeneity and the rise of west german wage inequality\*. *The Quarterly Journal of Economics*, **128** (3), 967–1015.
- CARRINGTON, W. J. and FALLICK, B. (2017). Why do earnings fall with job displacement? *Industrial Relations: A Journal of Economy and Society*, **56** (4), 688–722.
- COUCH, K. A. and PLACZEK, D. W. (2010). Earnings losses of displaced workers revisited. *The American Economic Review*, pp. 572–589.

- DAVIS, J. and HELLER, S. B. (2017). *Rethinking the benefits of youth employment programs: The heterogeneous effects of summer jobs*. Tech. rep., National Bureau of Economic Research.
- DAVIS, S. J. and VON WACHTER, T. (2011). Recessions and the costs of job loss. *Brookings Papers on Economic Activity*, **43** (2 (Fall)), 1–72.
- ECKHOUT, J. and KIRCHER, P. (2011). Identifying sorting in theory. *The Review of Economic Studies*, p. rdq034.
- EFRON, B. and HASTIE, T. (2016). *Computer Age Statistical Inference: Algorithms, Evidence, and Data Science*. New York, NY, USA: Cambridge University Press, 1st edn.
- FARBER, H. S. (2011). *Job loss in the Great Recession: Historical perspective from the displaced workers survey, 1984-2010*. Tech. rep., National Bureau of Economic Research.
- (2017). Employment, hours, and earnings consequences of job loss: Us evidence from the displaced workers survey. *Journal of Labor Economics*, **35** (S1), S235–S272.
- FRIEDMAN, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, **29** (5), 1189–1232.
- GATHMANN, C., HELM, I. and SCHÖNBERG, U. (2018). Spillover Effects of Mass Layoffs. *Journal of the European Economic Association*.
- GULYAS, A. (2018). Firm dynamics and labor market sorting.
- GUVENEN, F., KARAHAN, F., OZKAN, S. and SONG, J. (2017). Heterogeneous scarring effects of full-year nonemployment. *American Economic Review*, **107** (5), 369–73.
- HAGEDORN, M., LAW, T. H. and MANOVSKII, I. (2017). Identifying equilibrium models of labor market sorting. *Econometrica*, **85** (1), 29–65.
- HALLA, M., SCHMIEDER, J. and WEBER, A. (2018). Job displacement, family dynamic, and spousal labor supply.
- HANSOTIA, B. and RUKSTALES, B. (2002). Incremental value modeling. *Journal of Interactive Marketing*, **16** (3), 35 – 46.

- HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. H. (2017). *The elements of statistical learning: data mining, inference, and prediction, 2nd Edition*. Springer series in statistics, Springer.
- HETHEY-MAIER, T., SCHMIEDER, J. F. *et al.* (2013). Does the use of worker flows improve the analysis of establishment turnover? evidence from german administrative data. *Schmollers Jahrbuch: Journal of Applied Social Science Studies/Zeitschrift für Wirtschafts-und Sozialwissenschaften*, **133** (4), 477–510.
- HYAFIL, L. and RIVEST, R. L. (1976). Constructing optimal binary decision trees is np-complete. *Information Processing Letters*, **5** (1), 15 – 17.
- ICHINO, A., SCHWERDT, G., WINTER-EBMER, R. and ZWEIMÜLLER, J. (2017). Too old to work, too young to retire? *The Journal of the Economics of Ageing*, **9**, 14–29.
- JACOBSON, L., LALONDE, R. and SULLIVAN, D. (1993). Earnings losses of displaced workers. *The American Economic Review*, pp. 685–709.
- JAROSCH, G. (2014). Searching for job security and the consequences of job loss.
- JOVANOVIC, B. (1979). Job matching and the theory of turnover. *Journal of Political Economy*, **87** (5), 972–990.
- JUNG, P. and KUHN, M. (2018). Earnings losses and labor mobility over the life cycle. *Journal of the European Economic Association*, **17** (3), 678–724.
- KLEINBERG, J., LAKKARAJU, H., LESKOVEC, J., LUDWIG, J. and MULLAINATHAN, S. (2017). Human Decisions and Machine Predictions. *The Quarterly Journal of Economics*, **133** (1), 237–293.
- KNAUS, M., LECHNER, M. and STRITTMATTER, A. (2017). Heterogeneous employment effects of job search programmes: A machine learning approach.
- KROLIKOWSKI, P. (2017). Job ladders and earnings of displaced workers. *American Economic Journal: Macroeconomics*, **9** (2), 1–31.
- LACHOWSKA, M., MAS, A. and WOODBURY, S. A. (2018). *Sources of displaced workers' long-term earnings losses*. Tech. rep., National Bureau of Economic Research.

- LINDO, J. M. (2011). Parental job loss and infant health. *Journal of Health Economics*, **30** (5), 869–879.
- MUELLER, A. I. (2017). Separations, sorting, and cyclical unemployment. *American Economic Review*, **107** (7), 2081–2107.
- MULLAINATHAN, S. and SPIESS, J. (2017). Machine learning: An applied econometric approach. *Journal of Economic Perspectives*, **31** (2), 87–106.
- NASSIF, H., KUUSISTO, F., BURNSIDE, E. and SHAVLIK, J. (2014). Uplift modeling with roc: An srl case study. *CEUR Workshop Proceedings*, **1187**, 40–45.
- NEAL, D. (1995). Industry-specific human capital: Evidence from displaced workers. *Journal of Labor Economics*, **13** (4), pp. 653–677.
- RADCLIFFE, N. J. (2007). Using control groups to target on predicted lift: Building and assessing uplift model.
- REGE, M., TELLE, K. and VOTRUBA, M. (2011). Parental job loss and children’s school performance. *The Review of Economic Studies*, **78** (4), 1462–1489.
- SCHALLER, J. and STEVENS, A. H. (2015). Short-run effects of job loss on health conditions, health insurance, and health care utilization. *Journal of health economics*, **43**, 190–203.
- SCHMIEDER, J. F., VON WACHTER, T. and BENDER, S. (2018). The Costs of Job Displacement over the Business Cycle and Its Sources: Evidence from Germany.
- SCHWERDT, G., ICHINO, A., RUF, O., WINTER-EBMER, R. and ZWEIMÜLLER, J. (2010). Does the color of the collar matter? employment and earnings after plant closure. *Economics Letters*, **108** (2), 137–140.
- SEXTON, J. and LAAKE, P. (2009). Standard errors for bagged and random forest estimators. *Computational Statistics and Data Analysis*, **53** (3), 801 – 811, computational Statistics within Clinical Research.
- SONG, J., PRICE, D. J., GUVENEN, F., BLOOM, N. and VON WACHTER, T. (2018). Firming up inequality. *The Quarterly Journal of Economics*, **134** (1), 1–50.
- SULLIVAN, D. and VON WACHTER, T. (2009). Job displacement and mortality: An analysis using administrative data. *The Quarterly Journal of Economics*, **124** (3), 1265–1306.



TOPEL, R. (1990). Specific capital and unemployment: Measuring the costs and consequences of job loss. In *Carnegie-Rochester Conference Series on Public Policy*, Elsevier, vol. 33, pp. 181–214.

ZWEIMÜLLER, J., WINTER-EBMER, R., LALIVE, R., KUHN, A., WUELLRICH, J.-P., RUF, O. and BÜCHI, S. (2009). Austrian social security database. *Available at SSRN 1399350*.

## A. PROPENSITY SCORE MATCHING

<i>group</i>	<i>age</i>	<i>logwmin1</i>	<i>logwmin2</i>	<i>jobTenure</i>	<i>prodsector</i>	<i>fsize</i>	<i>propensity</i>	<i>n</i>
<b>Year = 1989</b>								
1	37.72	4.63	4.62	1586.22	0.62	217.95	50.15	1300
2	37.76	4.63	4.62	1586.24	0.60	210.33	50.15	1300
3	37.91	4.61	4.59	1643.67	0.48	1566.20	50.08	383017
<b>Year = 1990</b>								
1	37.36	4.57	4.56	1771.02	0.56	116.87	50.20	1410
2	37.33	4.57	4.56	1806.26	0.57	114.40	50.20	1410
3	37.81	4.63	4.60	1870.38	0.49	1541.54	50.09	374545
<b>Year = 1991</b>								
1	37.00	4.57	4.54	1970.06	0.54	156.46	50.15	1236
2	36.60	4.56	4.53	1937.70	0.55	159.43	50.15	1236
3	37.70	4.65	4.62	2027.62	0.50	1672.56	50.08	383144
<b>Year = 1992</b>								
1	37.60	4.64	4.60	2072.85	0.42	302.98	50.24	2875
2	37.32	4.65	4.61	2087.12	0.48	265.11	50.24	2875
3	37.50	4.67	4.63	2167.48	0.51	1640.96	50.18	387729
<b>Year = 1993</b>								
1	37.60	4.67	4.66	2235.55	0.70	649.40	50.32	4231
2	37.88	4.66	4.65	2177.72	0.68	582.87	50.32	4231

(continued)

<i>group</i>	<i>age</i>	<i>logwmin1</i>	<i>logwmin2</i>	<i>jobTenure</i>	<i>prodsector</i>	<i>fsize</i>	<i>propensity</i>	<i>n</i>
3	37.34	4.67	4.65	2290.31	0.51	1628.39	50.26	395203
<b>Year = 1994</b>								
1	37.28	4.62	4.61	2097.54	0.53	75.99	50.29	1561
2	37.24	4.63	4.62	2122.31	0.53	75.75	50.29	1561
3	37.17	4.67	4.65	2377.48	0.51	1577.82	50.09	405626
<b>Year = 1995</b>								
1	37.74	4.64	4.61	2266.26	0.50	244.13	50.27	3134
2	37.50	4.64	4.61	2247.47	0.51	241.16	50.27	3134
3	37.23	4.69	4.66	2492.18	0.50	1552.11	50.19	406903
<b>Year = 1996</b>								
1	37.78	4.64	4.62	2355.85	0.70	192.62	50.39	3390
2	37.67	4.64	4.62	2362.40	0.73	198.06	50.39	3390
3	37.47	4.71	4.68	2631.92	0.50	1526.64	50.20	404581
<b>Year = 1997</b>								
1	37.76	4.64	4.63	2407.20	0.61	90.21	50.21	1353
2	37.51	4.63	4.62	2486.15	0.60	87.93	50.21	1353
3	37.52	4.71	4.70	2713.32	0.50	1462.91	50.08	410672
<b>Year = 1998</b>								
1	37.53	4.69	4.67	2345.60	0.45	194.64	50.16	1689
2	37.65	4.67	4.64	2390.67	0.44	186.12	50.16	1689
3	37.54	4.71	4.69	2807.54	0.50	1452.41	50.10	412663
<b>Year = 1999</b>								
1	37.70	4.66	4.64	2709.91	0.56	355.54	50.24	3149
2	37.56	4.65	4.63	2673.30	0.57	367.26	50.24	3149
3	37.62	4.73	4.70	2887.17	0.50	1409.91	50.19	408815
<b>Year = 2000</b>								
1	37.84	4.70	4.68	2382.79	0.66	156.46	50.23	1959

(continued)

<i>group</i>	<i>age</i>	<i>logwmin1</i>	<i>logwmin2</i>	<i>jobTenure</i>	<i>prodsector</i>	<i>fsize</i>	<i>propensity</i>	<i>n</i>
2	37.82	4.70	4.68	2362.34	0.66	158.08	50.23	1959
3	37.73	4.74	4.72	2947.87	0.49	1412.60	50.12	409693
<b>Year = 2001</b>								
1	37.43	4.63	4.63	2439.30	0.60	116.62	50.33	2389
2	37.44	4.63	4.62	2431.16	0.62	119.04	50.33	2389
3	37.89	4.74	4.73	3011.00	0.49	1393.34	50.14	409842
<b>Year = 2002</b>								
1	37.88	4.72	4.70	3000.88	0.68	573.29	50.23	3241
2	37.81	4.74	4.72	3097.73	0.70	558.25	50.23	3241
3	38.02	4.75	4.72	3042.32	0.49	1315.66	50.19	411583
<b>Year = 2003</b>								
1	38.85	4.60	4.59	3367.89	0.25	10152.64	54.40	6237
2	38.75	4.63	4.62	3078.66	0.22	10054.63	54.00	6237
3	38.07	4.75	4.73	3054.22	0.49	1111.77	50.25	409574
<b>Year = 2004</b>								
1	37.13	4.70	4.68	2479.83	0.30	193.34	50.18	1792
2	37.33	4.71	4.68	2542.28	0.34	172.37	50.18	1792
3	38.15	4.76	4.74	3082.23	0.50	1041.74	50.11	402703
<b>Year = 2005</b>								
1	38.01	4.62	4.63	2608.31	0.62	88.96	50.26	1391
2	38.01	4.63	4.64	2550.76	0.63	88.08	50.26	1391
3	38.21	4.76	4.75	3063.74	0.49	1083.65	50.08	407225
<b>Year = 2006</b>								
1	39.28	4.85	4.82	3892.02	0.30	769.82	50.20	2130
2	39.67	4.84	4.82	4111.85	0.33	797.69	50.20	2130
3	38.28	4.76	4.74	2994.88	0.47	1199.27	50.12	425399
<b>Year = 2007</b>								

(continued)

<i>group</i>	<i>age</i>	<i>logwmin1</i>	<i>logwmin2</i>	<i>jobTenure</i>	<i>prodsector</i>	<i>fsize</i>	<i>propensity</i>	<i>n</i>
1	37.40	4.61	4.59	2335.67	0.59	163.39	50.14	1054
2	37.84	4.60	4.59	2427.92	0.59	166.56	50.13	1054
3	38.38	4.78	4.75	2983.83	0.47	1230.65	50.06	439162

## B. THE ALGORITHM AT PLAY.

To illustrate how the algorithm works, we simulate a dataset of 10,000 observations, which is then used for building a random forest. Then the dataset is used for training a random forest of 1,000 trees. Suppose that data-generating process is of the following form, which is *unknown* to the researcher:

$$y_i = \alpha_1(\mathbf{z}_i)x_{1i} + \alpha_2(\mathbf{z}_i)x_{2i} + \varepsilon_i, \quad (9)$$

such that:

$$\alpha_1(\mathbf{z}_i) = \begin{cases} 0, & \text{if } z_{1i} < \frac{1}{2}, \\ 100(z_{1i} - \frac{1}{2})^2, & \text{if } z_{1i} \geq \frac{1}{2}, \end{cases} \quad (10)$$

$$\alpha_2(\mathbf{z}_i) = 10|z_{2i} - z_{3i}|, \quad (11)$$

$$(12)$$

where 5 partitioning variables are drawn from the uniform distribution,  $\mathbf{z} \sim \mathcal{U}[0, 1]^5$  and the noise term is normally distributed,  $\varepsilon_i \sim \mathcal{N}(0, 1)$ .

*Non-linearities and variable selection.* In our example, the only variable that affects  $\alpha_1(\mathbf{z}_i)$  is  $z_1$ . That said, this impact is non-linear and conditioned on the value of  $z_1$  greater than  $\frac{1}{2}$ . In this sense, the identification of this relationship by the algorithm can be used as a benchmark for both the selection of variables and detection of the functional relationship. Figure 23 shows the estimates of  $\alpha_1$  conditioned on all considered variables,  $\{z_1, z_2, z_3, z_4, z_5\}$ . As can be seen, the algorithm is succesful in selecting partitiong covariates accounting for the value of  $\alpha_1(z)$ . The estimated impact of all variables but  $z_1$  are very close to zero. The

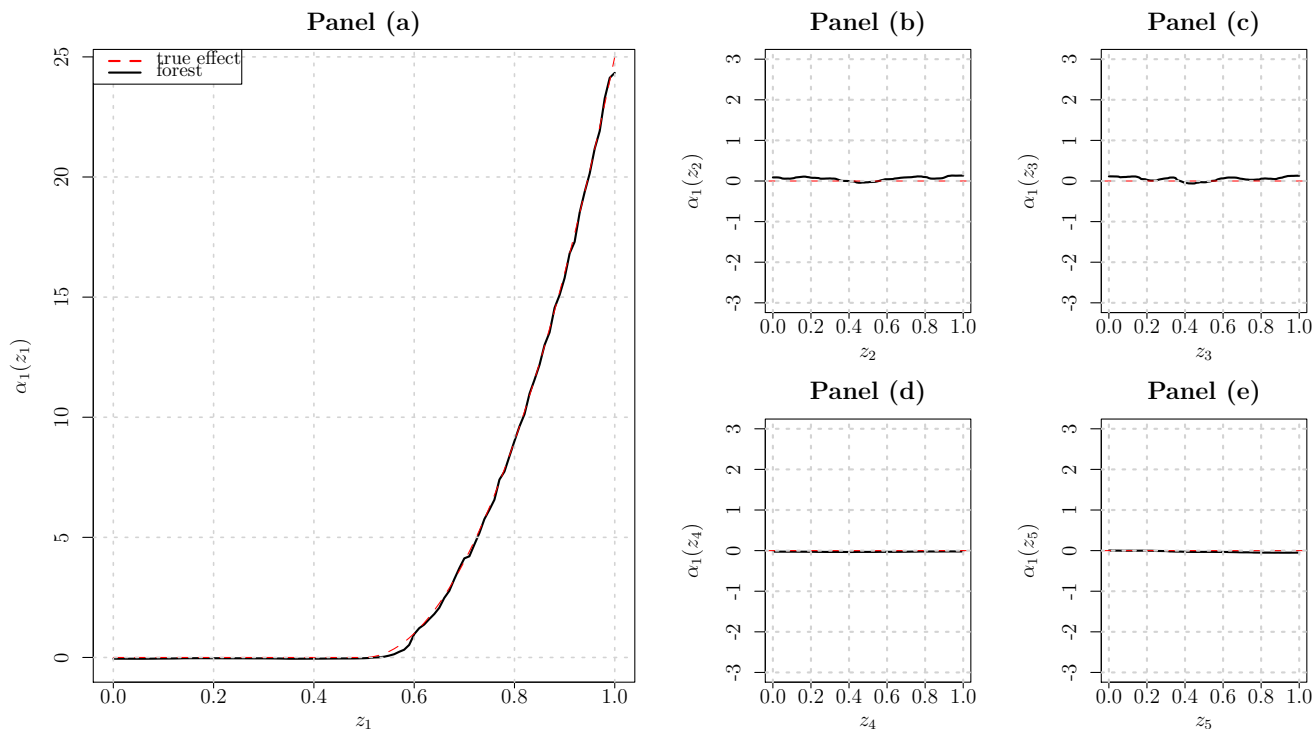


Figure 23: Estimated and true value of  $\alpha_1$  given by (10) for all partitioning variables. For each panel, all variables which are not included in a figure are set to .5. The random-forest algorithm applied to simulated data selects the correct variable as the driving force. Besides, it detects the non-linear functional form.

noise is a bit higher for covariates  $z_2$  and  $z_3$  than for  $z_4$  and  $z_5$ . The reason for this is that the former ones influences  $y_i$  through  $\alpha_2$ , while the latter have no impact on the regressant at all. In addition to this, we can see that the algorithm detects the non-linear relationship between  $\alpha_1$  and  $z_1$  correctly (see Panel (a) of Figure 23). The generated prediction is very smooth and this is thanks to the use of random forests instead of a single tree<sup>27</sup>.

*Interactions.* The next challenge that arises during the model selection stage is the nature of interactions between variables. In our simulation we test this by assuming that the value  $\alpha_2(\mathbf{z}_i)$  depends on the distance between  $z_2$  and  $z_3$ . Figure 24 illustrates heatmaps for true values of  $\alpha_2(\mathbf{z}_i)$  and the estimated one obtained with the random forest. Our exercise shows that the method captures the interaction in the correct way not only qualitatively but also in terms of the magnitude.

<sup>27</sup> Prediction surface made by single trees is a step function and as a result it lacks smoothness in comparison random forests averaging many step functions (Hastie *et al.*, 2017).

Forest predictions,  $\hat{\alpha}_2(z_2, z_3)$       True values,  $\alpha_2(z_2, z_3)$

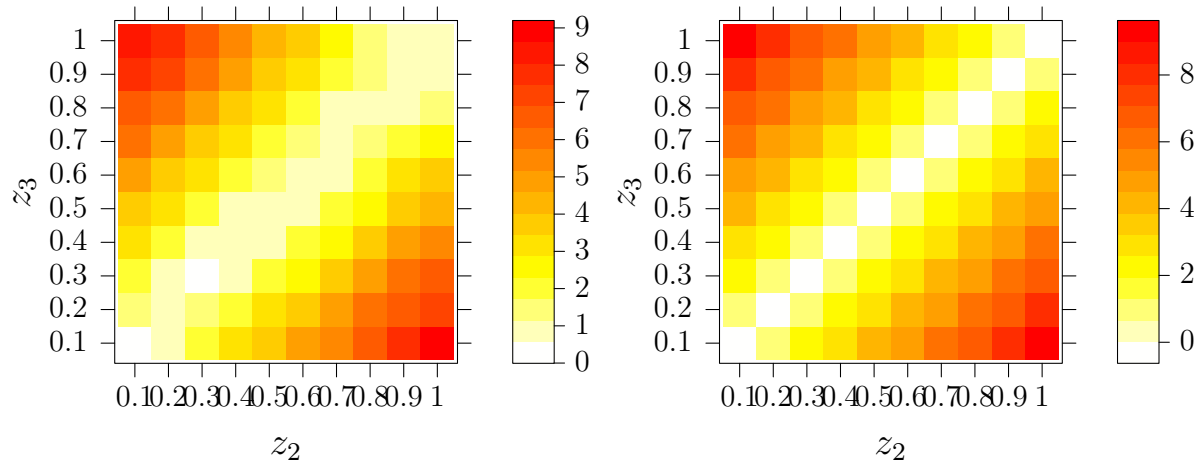


Figure 24: True vs. estimated value of  $\alpha_2(\mathbf{z})$  in  $(z_2, z_3)$ -plane. All other partitioning variables are set to .5 ( $\forall_{j \in \overline{1,6} \setminus \{2,3\}} z_j = \frac{1}{2}$ ). The random-forest algorithm applied to simulated data captures the interaction  $10|z_{2i} - z_{3i}|$  in the correct way not only qualitatively but also in terms of the magnitude.

## C. RELATIONS TO UPLIFT MODELLING

From the specification perspective, our problem can be framed as a problem of uplift modelling. In techniques of this type, the main goal is to model the incremental impact of a treatment. In business analytics, one of the most popular methods to model the uplift is a two-model approach (*e.g.*, [Hansotia and Rukstales, 2002](#); [Radcliffe, 2007](#); [Nassif \*et al.\*, 2014](#)). The idea behind it is to build two predictive models independently, one for treated objects,  $F(\mathbf{x}|\text{treatment})$ , and one for the control group,  $F(\mathbf{x}|\text{control})$ . Then the treatment effect for an individual with characteristics  $\mathbf{x}$  is pinned down as a difference between two predictions,  $\tau = F(\mathbf{x}|\text{treatment}) - F(\mathbf{x}|\text{control})$ . In this setup the treatment effect is residual value of two models, which were built to maximize their prediction accuracy of the dependent variable. However, this does not necessarily translate to the prediction accuracy of the treatment effect.

An alternative approach for modelling the incremental impact is the two-model approach. To illustrate how two algorithms, generalized random forest and two-model, work we consider two scenarios and for each we simulate a dataset of 10,000 observations, which is then used for building models. In the first scenario we simulate the model of 10,000 observations where values of  $y$  do not change rapidly in covariates. Each observation is characterized with 10 independent variables  $x_1 \dots x_{10} \sim \mathcal{U}[-2, 2]$  and random binary treatment  $W \in \{0, 1\}$ . The dependent variable  $y$  is generated by:

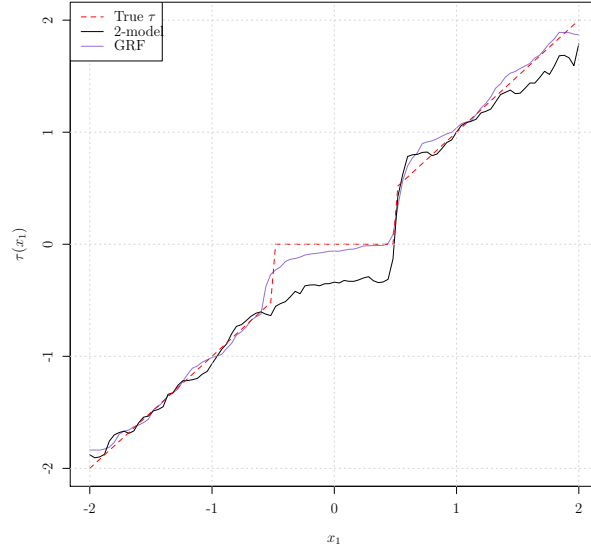
$$y = W \cdot \underbrace{\mathbb{1}(|x_1| > 0.5) \cdot x_1}_{\tau(x_1)} + x_2 + \varepsilon, \quad (13)$$

where  $\varepsilon \sim \mathcal{N}(0, 1)$  and  $\mathbb{1}$  is an indicator function.

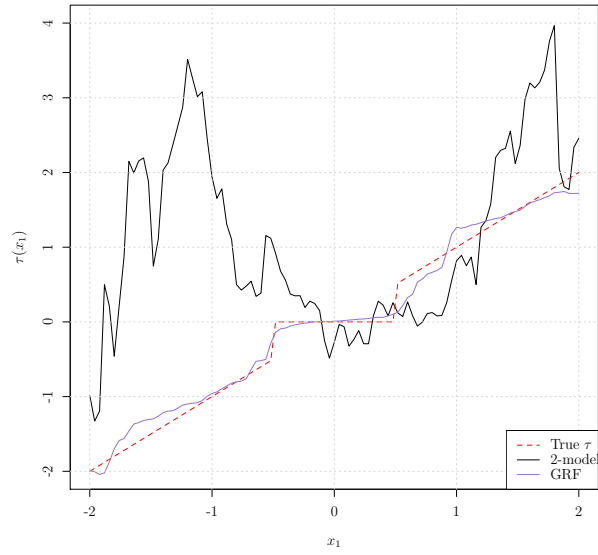
In the second scenario (“hard” one), we introduce additional rapid fluctuations in  $x_3$  :

$$y = W \cdot \underbrace{\mathbb{1}(|x_1| > 0.5) \cdot x_1}_{\tau(x_1)} + x_2 + 50 \cdot \sin(10x_3) \cdot x_3 + \varepsilon. \quad (14)$$

The main predictor of  $y$  is  $x_3$ , but it has no impact on treatment effect at all. Figures 13 and 14 show how the considered methods work in both scenarios. We see that in the “easy” scenario both methods provide similar estimates, while in the “hard” one the GRF is substantially better. The reason for this is that the treatment effect in the hard scenario does not contribute much in the overall variance of  $y$ .



(a) “Easy” scenario (13)



(b) “Hard” scenario (14)

Figure 25: Treatment effects (true and estimated).



## D. ROBUSTNESS: FIRM FIXED EFFECT

We perform the following robustness exercise. In the baseline specification, we computed the firm fixed effect measure using the universe of all male observations except the observations from our earnings loss sample. This means that however the masslayoff affected labor market outcomes of the individuals in our sample, it cannot have an effect on the measurement of the firm fixed effect, as these individuals were excluded in the AKM regression. But Post-masslayoff of those workers that stayed at mass layoff firms, but were not selected by the propensity score matching could influence the estimation of firm fixed effects and lead to a mis-measurement of the firm pay premia at the time of separation.

To rule out that any of our results are driven by this, we re-estimate the AKM regression equation (4) only on observations before 1998. We then re-estimate our forest with workers after 1998. Figure 26 shows that the estimated importance of firm fixed effects is essentially unchanged compared to our baseline specification. Furthermore we recompute the marginal effects for all partitioning variables. Figures 27 and 28 show that these are also very similar to our baseline results. Some changes are to be expected, as we also consider only a subperiod. The main result is essentially unchanged, workers from low fixed effect firms still face considerably lower earnings losses than workers employed at high fixed effect firms. It is still the variables that shows the highest slope in earnings losses.

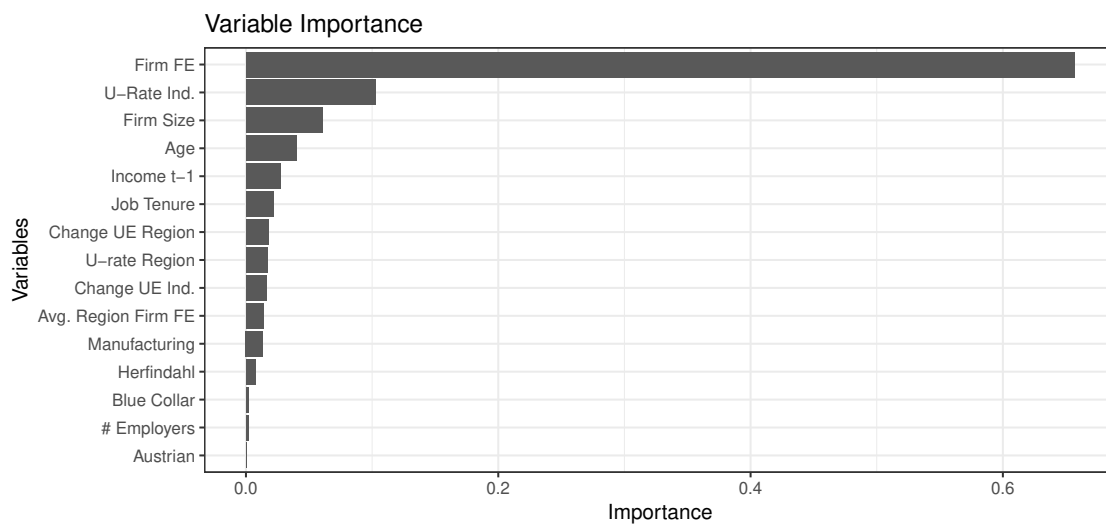


Figure 26: Variable importance using only workers from 1999 or later, while firm fixed effect is computed on observations before 1998. Depth-adjusted variable frequency in splits in the GRF with a decay exponent equal to  $-2$  and the maximum depth level of nodes equal to 4. All values sum to 1.

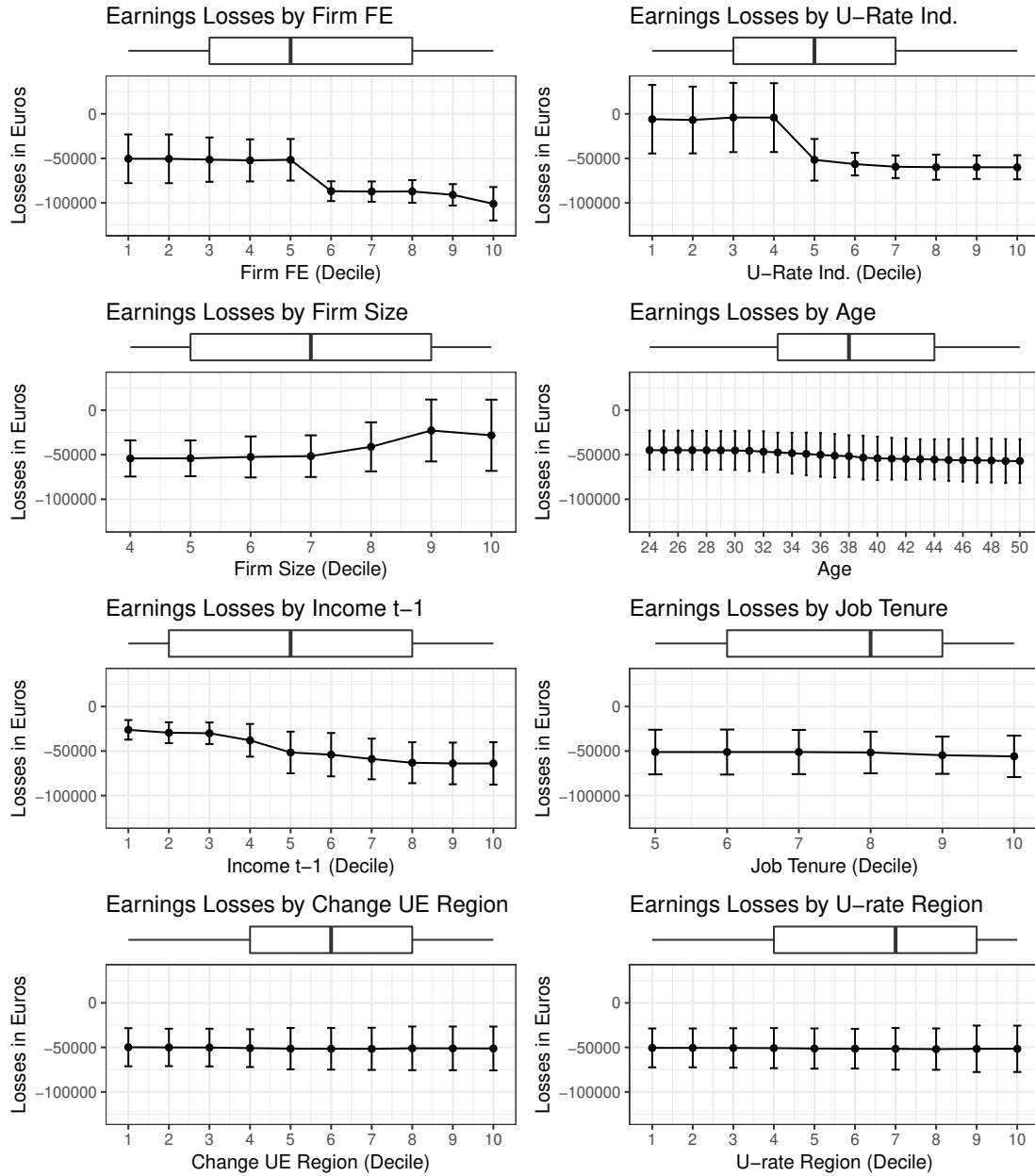


Figure 27: GRF estimates of the cumulative 11-year earnings losses with 95% confidence intervals (part I), only workers from 1999 onwards. Firm fixed effects are computed on observations before 1998. All variables that are not used in a given plot are set to their mean values. The  $y$ -axis indicates cumulative 11-year earnings losses, while the  $x$ -axes depict global deciles of the unrestricted dataset (except for the age, number of employees, blue collar, manufacturing, and Austrian where the actual values are used). The boxplots above figures represent the distribution of the partitioning variable in the restricted dataset (for the details see Subsection B of Section II).

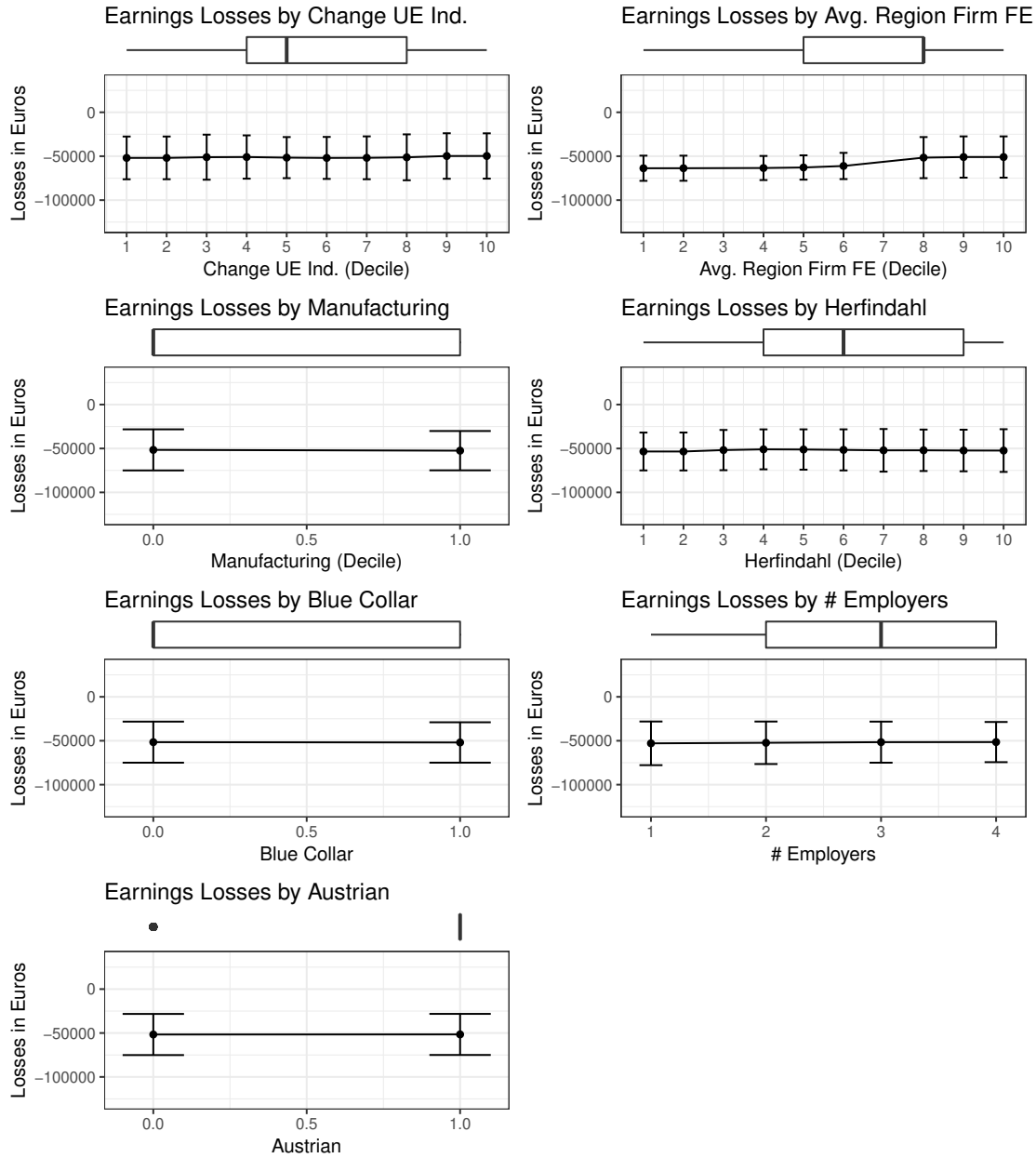


Figure 28: GRF estimates of the cumulative 11-year earnings losses with 95% confidence intervals (part II), only workers from 1999 onwards. Firm fixed effects are computed on observations before 1998. All variables that are not used in a given plot are set to their mean values. The  $y$ -axis indicates cumulative 11-year earnings losses, while the  $x$ -axes depict global deciles of the unrestricted dataset (except for the age, number of employees, blue collar, manufacturing, and Austrian, where the actual values are used). The boxplots above figures represent the distribution of the partitioning variable in the restricted dataset (for the details see Subsection B of Section II).