Universität Bern

Essays in Applied Econometrics

Marc Schranz

Inauguraldissertation zur Erlangung der Würde eines

Doctor rerum oeconomicarum

der Wirtschafts- und Sozialwissenschaftlichen Fakultät

Februar 2025

The faculty accepted this thesis on the 22. May 2025 at the request of the reviewers Prof. Dr. Costanza Naguib and Prof. Dr. Martin Huber as dissertation, without wishing to comment on the views expressed therein.

Die Fakultät hat diese Arbeit am 22.Mai 2025 auf Antrag der Gutachter Prof. Dr. Costanza Naguib und Prof. Dr. Martin Huber als Dissertation angenommen, ohne damit zu den darin ausgesprochenen Auffassungen Stellung nehmen zu wollen.

Preface

I wish to express my sincere gratitude to my supervisor, Prof. Dr. Costanza Naguib, and my second supervisor, Prof. Dr. Blaise Melly, for their invaluable support and guidance during my doctoral studies. I thank them for the fruitful discussions and the insightful feedback they have provided for me, contributing substantially to the present thesis.
Writing this thesis has been both an inspiring and challenging journey, one that would not have been possible without the support and encouragement of many individuals. First and foremost, I would like to thank my peers and good friends for their support and engaging discussions, as well as for the enjoyable breaks and memorable events that enriched my PhD journey. In particular, I thank my office colleagues, Martina Pons, Fiona Rüedi, and Alexandra Piller, for their friendship, their patience during whiteboard sessions, and the provision of life-saving snacks when my stock was empty. I also wish to deeply thank Ulrike Dowidat, Manuela Solterman, Gabriella Winistörfer, and Franz Kölliker for their invaluable support regarding organizational affairs, funding, hardware, and numerous other needs.
Finally, my most profound thanks go to my family, especially Bieula. Your time, patience, and unwavering support in both challenging and enjoyable times have been a true foundation for me, and I am grateful to have had you by my side throughout this journey.

Bern, February 2025 Marc Schranz

Contents

List of Figures

List of Tables

Introduction

As economic data becomes more abundant and diverse, and new technologies grow more powerful and capable, economic analysis increasingly relies on the application of innovative methodologies to enhance our understanding of complex questions across various economic topics. In the following three chapters, I contribute to these developments by implementing innovative approaches in the fields of policy evaluation, income mobility, and monetary policy analysis, employing advanced statistical techniques, distributional modeling, and Natural Language Processing. All three studies emphasize the importance of moving beyond traditional aggregate metrics to capture more granular distributional dynamics and to improve upon past measures.

The first chapter introduces an extension to the synthetic control method, which was originally proposed by Abadie and Gardeazabal (2003) and Abadie, Diamond, et al. (2010), to evaluate effects across the distribution. Traditional synthetic control methods focus on average treatment effects, even when the treated and control units comprise a sizable number of individual entities. This chapter proposes a distributional synthetic control utilizing information that is available on a finer granularity and capturing heterogeneous effects across different thresholds of the cumulative distribution function.

The second chapter, which is joint work with Jonas Meier, proposes a new estimator of the conditional distribution of multivariate outcomes given covariates. The estimator builds on the Local Gaussian Representation from Chernozhukov, Fernández-Val, and Luo (2023) and employs distribution regression and a conditional copula with a copula parameter that is local in the value of the outcome. The proposed method allows for flexible, semi-parametric control of covariates, enabling the analysis of multivariate counterfactual distributions.

The third chapter is joint work with Alexandra Piller and Larissa Schwaller. The chapter explores the increasingly critical role of central bank communication in monetary policy using state-of-the-art Natural Language Processing techniques. In recent years, high-frequency monetary policy surprise series have been used as external instruments to identify monetary policy effects. This chapter improves upon these surprise series by employing a Natural Language Processing model that is based on transformers, an architecture introduced in a groundbreak-

ing paper by Vaswani et al. (2017). The model is trained to isolate the component of the surprises driven solely by central bank communication. This further required the creation of a text dataset comprising statements and speeches issued by the Federal Reserve Board through web scraping.

As mentioned before, the first chapter builds on the pioneering work by Abadie and Gardeazabal (2003), Abadie, Diamond, et al. (2010), and Abadie, Diamond, et al. (2015). In their seminal papers, the synthetic control method was introduced, a straightforward and intuitive tool for policymakers to analyze policy interventions. Typically, the interventions of interest address a particular region, which are then defined as the units. With the synthetic control method, a synthetic control unit is constructed that replicates the treated unit and its behavior in the absence of treatment. Specifically, the synthetic control unit is created using a weighted average of other control units that are never treated. Comparing the synthetic control unit to the observed behavior of the treated unit yields an estimate of the average treatment effect.

Traditional applications of synthetic control methods typically focus on the aggregate unit level and its outcomes (e.g., GDP for regions). However, treated and control units potentially comprise a sizable number of individual entities, and the outcome of interest might be measured at the individual level as opposed to an aggregate, regional level (e.g., the income of households within a region). In such cases, the variable is often aggregated to employ the synthetic control method, which may obscure interesting distributional effects. The proposed method addresses this by computing effects across the entire distribution rather than relying on a single aggregate measure, thereby providing a more granular understanding of policy impacts and enabling the study of heterogeneous effects.

The suggested approach naturally extends the synthetic control method to such a distributional setting. The weights of the synthetic control unit are estimated at different thresholds of the cumulative distribution function, offering a flexible approach to replicate the distribution of the treated unit. For estimation, the existing and well-established theoretical results and estimation procedures from the vast synthetic control literature can still be relied on, making the proposed method versatile and straightforward to implement.

The proposed approach further contributes to new literature on distributional synthetic control methods, such as those proposed by Chen (2020) and Gunsilius (2023). Crucially, the estimator proposed in this chapter, like that of Chen (2020), allows the weights of the synthetic control unit to vary across different thresholds of the outcome of interest. Results from the application and a simulation study presented in the chapter provide evidence that allowing for this flexibility is essential for the synthetic unit to accurately capture the treated unit's

underlying distribution.

The proposed method is applied to a policy intervention in Switzerland, analyzing the effects on the distribution of wages. In particular, the chapter examines the introduction of the minimum wage in Neuchâtel, a canton in Switzerland. Consistent with the past findings of Berger and Lanz (2020), the analysis reveals a positive effect at the lower part of the wage distribution without notable changes in work hours or unemployment.

In the second chapter, another estimator for analyzing distributions is proposed. Specifically, a new method for the analysis of multivariate counterfactual distributions is introduced, building on the idea of the Local Gaussian Representation suggested in Chernozhukov, Fernández-Val, and Luo (2023) and extended by Fernández-Val et al. (2024). The method involves a two-step procedure. In the first step, the univariate conditional distributions of the outcomes are estimated via distribution regression. In the second step, we estimate a conditional copula of the outcomes, imposing a copula parameter that varies locally with the outcome value. Thereby, covariates can be incorporated flexibly to control for different factors, allowing for the estimation of conditional multivariate distributions.

As a result, the estimated conditional multivariate distribution becomes a valuable tool for gaining deeper insights into intergenerational income mobility, a key determinant of long-term economic equality. Many commonly used measures in the intergenerational mobility literature are a direct function of the joint distribution and, thus, can be directly computed from the estimated conditional joint income distribution. In particular, this chapter focuses on rank-based measures like rank-rank correlations, conditional expected ranks, and transition matrices.

Various socioeconomic factors influence intergenerational income mobility (see, for example, Cholli and Durlauf (2022) and Mogstad and Torsvik (2023) for a detailed discussion). Traditional approaches often face limitations when accounting for covariates because they rely on a vast number of observations or strong parametric assumptions. The proposed approach overcomes these challenges by employing a flexible, semi-parametric model to account for covariates, computing different conditional mobility measures from the conditional multivariate distributions.

The method is further applied to analyze intergenerational income mobility using Swiss and U.S. data. The analysis primarily focuses on the Swiss case due to better data availability and fewer existing studies compared to the U.S. Findings indicate that mobility differs with the share of income contributed by the father. For sons, the probability of moving upwards increases with the father's income share, but for daughters, this probability of moving upwards

tends to decrease with the father's income share. In general, we find great differences in income mobility between sons and daughters, and average hours worked play an important role in driving these results.

The third and final chapter improves upon traditional monetary policy surprise measures by extending available text data and employing Natural Language Processing methods. Since the seminal work by Gertler and Karadi (2015), high-frequency price changes within a narrow time window around monetary policy announcements, referred to as market-based surprises, have been frequently used as an instrument to identify causal effects of monetary policy. However, recent research has raised concerns about their suitability, demonstrating that these market-based surprises suffer from weak relevance and endogeneity issues. The last chapter addresses these issues, constructing an alternative surprise series.

The weak relevance is addressed by creating a comprehensive text data set that incorporates not only the Federal Open Market Committee announcements but also policy-relevant speeches by the Federal Reserve Board chair and vice chair. To identify speeches that discuss monetary policy issues and are therefore relevant, the words in each speech is analyzed using the dictionary from Gardner et al. (2022).

The endogeneity issue is tackled by employing new Natural Language Processing techniques to filter out the component of surprises driven solely by central bank communication. Specifically, a large language model is trained to predict the monetary policy surprises using the text from Federal Open Market Committee announcements and Federal Reserve speech transcripts. As mentioned earlier, the language model is based on the transformer architecture introduced in the seminal paper by Vaswani et al. (2017). Following this process, a language-driven surprise series for the analysis of monetary policy effects is constructed. The findings demonstrate that language-driven surprises mitigate endogeneity concerns and produce impulse responses that align more closely with conventional economic theories than traditional market-based measures.

In conclusion, the three studies collectively advance our understanding of complex economic debates through innovative analytical approaches. From distributional synthetic control methods and copula regression to natural language process-driven monetary policy analysis, the three chapters demonstrate the potential of these novel analytical tools in addressing challenging economic questions. By leveraging these advanced methodologies, the research bridges gaps in existing literature and offers insights for policymakers, academics, and practitioners.

Chapter 1

A Synthetic Control Method for the Analysis of Effects across the Distribution

Abstract

This paper extends the synthetic control method to evaluate distributional effects. Synthetic control methods are commonly employed for policy interventions on an aggregate unit level, where the treated and control units typically comprise a sizable number of individual entities. If interest lies in a variable measured at the individual level, there is the opportunity to analyze the effects across the distribution and uncover heterogeneous treatment effects. The proposed synthetic control method introduces a novel approach for deriving such distributional effects. Crucially, the weights of the synthetic unit depend on the position within the distribution, such that the weighted sum of control units is allowed to vary across the distribution. Furthermore, the proposed method modifies how individual values are aggregated, enabling the usage of well-established estimation procedures from the synthetic control literature. The method is applied to analyze the impact of the introduction of the minimum wage in the canton of Neuchâtel in Switzerland. Furthermore, the application and simulations compare four distributional synthetic control methods. Results show an improved fit of the targeted distributions if weights are allowed to vary across the distribution.

Acknowledgment: I thank Costanza Naguib, Blaise Melly, Martin Huber, Kaspar Wüthrich, Guillaume Pouliot, Mirco Rubin, Roger Koenker, Bo Honoré for helpful comments and suggestions. Further, I am grateful for helpful comments by seminar and conference participants at the Young Swiss Economists Meeting 2024, the 5th Workshop of the SNoPE, the SSES Annual Congress 2024, the IAAE 2024, the Cookie Seminar at the Department of Economics at the University of Fribourg and the Brown Bag Seminar at the Department of Economics at the University of Bern.

1.1 Introduction

The literature on synthetic control methods has been strongly growing since the pioneering work by Abadie and Gardeazabal (2003), Abadie, Diamond, et al. (2010), and Abadie, Diamond, et al. (2015). The aim of these methods is to estimate the average treatment effect in a setting with only a few units observed over a certain time range. Typical applications comprise a policy intervention that is implied in one region, the treated unit, but not in other regions, the control units. To find the treatment effect, the key question is what the outcome of the treated unit would have been were it not treated. Synthetic control methods answer this question by building a synthetic control unit that replicates the treated unit in the case of not being treated. This synthetic control unit is constructed as a weighted average of control units. As a result, one advantage of synthetic control methods lies in their ability to enable policy evaluation in settings where several control units are available, but there is no single comparable control unit for the treated unit.

Traditionally, the units and the outcome of interest are thought of as aggregates, e.g., regions and their GDP. However, the treated and control units potentially comprise a sizable number of individual entities, and the outcome of interest might initially be measured individually. The simplest way to conduct a synthetic control analysis in such a setting is to aggregate the variable to unit averages. However, depending on the underlying question, interest might lie in other outcomes than average treatment effects. For example, Peri and Yasenov (2019) conduct a minimum wage analysis and use values aggregated to the 15^{th} and the 20^{th} wage percentiles to analyze the impact on lower wages. Stepanyan and Salas (2020) compute the Gini-coefficient and the ratio between the top and bottom 20 percent as aggregate outcome values to analyze distributional effects. I argue that in such cases, it might be much more insightful to compute the effects over the whole distribution instead of using one single aggregate value. The proposed method captures the impact of the policy intervention across the distribution and enables us to study the heterogeneity of the effects in detail.

The novel approach proposed in this paper naturally extends the synthetic control method to a distributional setting. The new distributional synthetic control method estimates the weights of the synthetic control unit at different thresholds of the cumulative distribution function. Estimating weights such that the distribution of the synthetic control unit closely replicates the distribution of the treated unit at a specific threshold and assuming these weights to stay constant across the pre- and post-treatment periods allows us to estimate a treatment effect at a specific threshold of the cumulative distribution function. As a result, the treatment effects, as well as the weights of the synthetic control units, can change across these thresholds and thus provide a flexible approach to estimating distributional effects employing synthetic control methods. Additionally, because estimation is done at different thresholds of the cumulative

distribution function, existing and well-established theoretical results and estimation procedures from the synthetic control literature still apply, making the proposed method versatile and straightforward to implement. As later findings will reveal, allowing for varying weights is of particular importance. This feature has so far been addressed differently by other distributional synthetic control methods proposed by Chen (2020) and Gunsilius (2023). In the work by Chen (2020), varying weights are imposed but not discussed. Conversely, Gunsilius (2023) emphasizes analysis on the aggregate unit level and implicitly assumes constant weights. However, as the results in the empirical application and simulation study will show, accounting for varying weights is crucial.

To illustrate the proposed method in an empirical application, the minimum wage introduction in Neuchâtel, a canton in Switzerland, is chosen. To the best of my knowledge, the paper by Berger and Lanz (2020) is the only paper that analyzed the effect of the introduction of the minimum wage in Neuchâtel. In their work, they conducted a two-wave survey focusing on restaurants, one of the sectors especially affected by the new law, and used a Difference-in-Difference as well as a distribution-based approach. They find a significant increase in workers paid just above the minimum wage level. Additionally, wages slightly above the minimum wage level tended to increase, and in the short run, neither employment nor product pricing was used as a margin of adjustment. My application will also focus on the effect on wages, but the analysis will not be restricted to a specific sector. Consistent with their findings, I find a positive effect at the lower part of the wage distribution, but no apparent reaction is observed in work hours or unemployment.

Several other extensions to the classic synthetic control method have been made. Some extensions improve the classic model proposed by Abadie and Gardeazabal (2003) and Abadie, Diamond, et al. (2010). For example, Ben-Michael et al. (2021) and Ferman and Pinto (2021) suggested improvements regarding the perfect fit assumption that is made in the classic model. Robbins et al. (2017) or Abadie and L'Hour (2021) extend the classic model to specific settings where the outcome of interest is observed on an individual level as opposed to the aggregate treated unit. Other papers extend the idea and link it to related methods. Doudchenko and Imbens (2016) develop a general framework that nests many different methods, including synthetic control. Arkhangelsky et al. (2021) and Athey, Bayati, et al. (2021) extend the classic method by focusing on relations across time for the control units and not only on relations across groups in the post-treatment period. Other papers like, Firpo and Possebom (2018), Li (2020), Chernozhukov, Wüthrich, et al. (2021), and Cattaneo et al. (2021) proposed new approaches to improve inference, which in the classic synthetic control case is done via permutation methods. For a more in-depth recent discussion of the synthetic control method, its advantages and disadvantages, data requirements, and recent extensions, see, for example, the recent work by Abadie (2021).

Other methods to conduct distributional analysis in a setting with few units have been proposed already. Athey and Imbens (2006), Callaway and Li (2019) and Biewen et al. (2022) introduce different methods, which extend the difference-in-difference approach to a distributional setting. Compared to these approaches, embedding the synthetic control method into a distributional setting brings the main advantage of doing distributional policy evaluation even if a single similar comparison group is absent, but several other untreated units are available. To the best of my knowledge, there are only a few closely related papers that try to estimate distributional effects using a synthetic control method.

A distributional synthetic control method was first proposed by Chen (2020). He decomposes the distribution into bins and obtains weights for each unit and each bin to replicate the distribution. Additionally, he augments the factor model from Abadie, Diamond, et al. (2010) to control for the bias from poor matches of the treated and synthetic unit in pre-treatment periods. The main difference to my approach is, that he replicates the distribution via quantile function, whereas my approach uses the cumulative distribution function. One possible advantage of using the quantile function is that it does not rely on the support of the distribution when estimating weights. However, using units with much different support compared to the treated unit raises the question of whether these should be included as control units, as one might end up with a synthetic control unit that is constructed from units that are dissimilar from the treated unit. In that case, it might be advisable not to use quantile functions since it could hide these differences. Furthermore, I argue that the cumulative distribution function is a more natural approach, as it is no different from applying the synthetic control method to a specific type of aggregate of the variable of interest. Additionally, it avoids estimation problems from mass points, which are a potential issue when using quantile functions. As a result, the proposed method works without modification, even if the underlying distribution is discrete or mixed. Finally, if the outcome is continuous and the researcher is more interested in quantile treatment effects, then it is straightforward to construct these effects by inverting the cumulative distribution functions via interpolation.

A second distributional synthetic control method is proposed by Gunsilius (2023), which uses the 2-Wasserstein measure as a notion of distance between two distributions. The main method proposed in his paper relies on quantile functions as well. This results from the fact that the optimal transportation problem solved by the 2-Wasserstein distance directly relates to the distance between values of the quantile functions of two distributions. Weights are chosen such that the 2-Wasserstein distance between the synthetic control and the treated unit is as small as possible. Thereby, a single weight for each unit for the whole distribution is computed. The synthetic control unit then forms a Barycenter in the 2-Wasserstein space that preserves the underlying distributional structure of the control units chosen. Following this approach avoids dependence on a specific choice of bins and allows the focus to remain on the

aggregate level of the unit. For the case that the distribution of the treated unit is expected to be a mixture of distributions, Gunsilius (2023) suggests an alternative method using differences between values of the cumulative distribution functions. This alternative approach is closely related to the method suggested in this paper. One main difference between the two methods lies in the estimation of the optimal weights. Instead of Wasserstein distances, the proposed method relies on conventional synthetic control methods. Furthermore, the proposed method allows weights to change flexibly across the distribution and does not restrict them to staying constant, such that the underlying distribution is better captured.

A recent working paper by Kato et al. (2023) proposes another method with constant weights across the distribution. Using a GMM estimator, they replicate the density of the treated with the synthetic control unit. Their paper is based on the work by Shi et al. (2022), in which the authors motivate the underlying linear factor model commonly assumed in the synthetic control literature. They show that the average outcome for a unit and time period has a factor model structure, assuming an independent causal mechanism and a stable distribution.[1]

The remainder of the paper is organized as follows. Section 1.2 recaps the basic synthetic control theory and introduces the proposed method as well as the implementation. The application and related data are described in detail in section 1.3, and the results are presented in section 1.4, which additionally includes simulations and a comparison of different distributional synthetic control methods. Finally, section 1.5 summarizes the findings.

1.2 Methodology

This section first recaps the classic synthetic control estimator and then introduces the proposed method. Whenever possible, the setup and notation are kept similar to the work by Abadie, Diamond, et al. (2010) and Abadie (2021). Additionally, I will abstract from additional covariates. A brief discussion on this topic is given in section 1.5. Let $j \in 1, \ldots, J+1$ denote the units, where $j = 1$ is the treated unit and $j = 2, \ldots J+1$ are the control units. Further, let $t \in 1, \ldots, T$ denote the time periods, where T_0 marks the last period before treatment. Hence, $t \leq T_0$ belongs to the pre-treatment period and $t > T_0$ to the post-treatment period. Potential outcomes of a variable Y are denoted by Y^I if treated and Y^N if not treated.

1.2.1 Synthetic Control Model: The Classic Case

In the classic setting by Abadie and Gardeazabal (2003) and Abadie, Diamond, et al. (2010), the goal is to estimate the average treatment effect from a policy intervention on the treated

[1] Another paper by Zhang et al. (2023) utilize the result from Shi et al. (2022) to extend the synthetic control method to an RCT setting, where they estimate the outcome in a target population only having control group data and information from other RCTs.

unit. That is, we are interested in:

$$\tau_{1t} = Y_{1t}^I - Y_{1t}^N \tag{1.2.1}$$

i.e., the difference between the outcome of unit 1 if it is treated, Y_{1t}^I, and if it is not treated, Y_{1t}^N, where Y_{jt} denotes the outcome of interest. Naturally, the former is only observed in the post-treatment periods, while the latter is only observed in the pre-treatment periods.

$$Y_{1t} = \begin{cases} Y_{1t}^N & \text{if } t \leq T_0 \\ Y_{1t}^I & \text{if } t > T_0 \end{cases}$$

By assuming that the control units remain unaffected by the treatment, it follows that $Y_{jt} = Y_{jt}^N$ for all t and $j \neq 1$. Furthermore, estimating the effect in 1.2.1 then boils down to estimating the unobserved Y_{1t}^N for $t > T_0$.

The synthetic control method estimates Y_{1t}^N as a weighted average over several control units:

$$Y_{1t}^N = \sum_{j=2}^{J+1} w_j Y_{jt}^N \tag{1.2.2}$$

If this relation between the control units and the treated unit is assumed to remain constant across time, the pre-treatment periods can be used to estimate this relation and to estimate Y_{1t}^N. For this reason, a good fit in the pre-treatment periods between the synthetic and the treated unit is crucial for the approach to be valid. The treatment effect for $t > T_0$ is then computed as:

$$\tau_{1t} = Y_{1t}^I - Y_{1t}^N = Y_{1t} - \sum_{j=2}^{J+1} w_j Y_{jt}$$

where w_j are the weights chosen to minimize the difference between the synthetic and treated unit in the post-treatment periods.

1.2.2 Proposed Method

Relation to Conventional Methods

Typically, the outcome of interest, Y_{jt}, is an aggregated value. An easy way to think about this is to let each unit j represent a different region. Then, the outcome of interest could be an aggregate variable of the region, for example, GDP or population growth, or it could also be a variable observed only for individuals living in that region, such as income or educational attainment. In this latter case, observed individual variables are aggregated depending on the underlying research question. For example, Jardim et al. (2022) use administrative data of individual employers and take simple averages over different variables to analyze the impact of

minimum wages on the labor market in Seattle. Another study on minimum wages by Peri and Yasenov (2019) uses averaged values of the CPS data for their analysis.

Formally, let Y_{ijt} be the observed variable of interest for individual i in unit j at time t. To analyze the policy effect for the region, a researcher might use the mean as an aggregated outcome of interest, $E[Y_{jt}]$. Then, the estimated treatment effect is:

$$\tau_{1t} = E[Y_{1t}^I] - E[Y_{1t}^N] \qquad \text{and} \qquad E[Y_{1t}^N] = \sum_{j=2}^{J+1} w_j E[Y_{jt}]$$

The proposed estimator takes advantage of exactly this setting to derive distributional effects. Instead of the simple mean, I suggest using the mean of an indicator variable, $E\left[\mathbb{1}(Y_{jt} \leq y)\right]$. The variable observed on an individual level is 1 whenever an individual exhibits a value of Y_{ijt} lower than or equal to some specified value y and 0 otherwise. The aggregate variable is thus the probability that Y_{jt} takes on values below or equal to y for the unit j in the period t, which by definition is the cumulative distribution function, $F_{jt}(y)$. Then, the treatment effect at a specific point y is:

$$\tau_{1t}(y) = \mathbb{E}\left[\mathbb{1}(Y_{1t}^I \leq y)\right] - \mathbb{E}\left[\mathbb{1}(Y_{1t}^N \leq y)\right] = F_{1t}^I(y) - F_{1t}^N(y) \tag{1.2.3}$$

where the first equation follows the same argument as for the common mean case, and the second equation follows by definition.

The same idea applies to derive the treatment effect as in the classic synthetic control. For $t > T_0$, $F_{1t}^I(y)$ is observed and $F_{1t}^N(y)$ is unobserved. The unobserved outcome is then estimated using a weighted average of control units, such that for a specific point y, the synthetic unit and the estimated effect are:

$$F_{1t}^N(y) = \sum_{j=2}^{J+1} w_j(y) F_{jt}^N(y) \tag{1.2.4}$$

$$\tau_{1t}(y) = F_{1t}(y) - \sum_{j=2}^{J+1} w_j(y) F_{jt}(y) \qquad \text{for } t \geq T_0 \tag{1.2.5}$$

To estimate a distributional effect, the described procedure is repeated over different values of y. However, it is important to note that for a specific point y, the suggested method does not strongly deviate from those in the current literature. The crucial difference is to define the outcome of interest observed on the individual level in a specific way. As a result, one advantage of the suggested method is its reliance on the well-established field of synthetic control methods. Properties that hold when applying a synthetic control estimator to the simple mean of a variable extend to the mean of the indicator variable. To capture the distributional effect, the y values can be chosen as a fine grid over the support of the treated unit across all periods. In

general, the choice of the specific y values has no strong impact on the analysis, except if the y values altogether lie outside of the support of the control units, such that it is impossible to construct a synthetic control unit.

Locally Varying Weights

Similar to the alternative method suggested by Gunsilius (2023), the proposed method works with cumulative distribution functions to estimate distributional effects. One important difference is that in the paper by Gunsilius (2023), a single weight is estimated for the whole distribution, while the proposed method allows them to vary across the distribution.

It follows that the relevant setting for the proposed method is when the observed distributions, $F_{jt}(y)$, are mixtures of distributions, with mixture weights varying across the distribution. Formally, this yields the following underlying model:

$$F_{jt}(y) = \sum_{d=1}^{D} \gamma_{j,d}(y) F_{t,d}(y) \tag{1.2.6}$$

where $\gamma_{j,d}$ are the mixture weights and $F_{t,d}$ are some unobserved underlying distributions. Additionally, let $\gamma_{j,d} \geq 0$ and $\sum_{d=1}^{D} \gamma_{j,d}(y) = 1$ for all j and y. With this specification, the distribution of unit j at time t is a mix of D other distributions. The underlying distributions $F_{t,d}$ are common across all units but vary with t. Each unit exhibits distinct weights, $\gamma_{j,d}$, which remain constant across time. This specification is related to the factor model, commonly used to model the outcome of interest in the synthetic control literature. There, the loadings vary across groups but stay constant over time, and the factors vary across time but stay constant across groups.

Using equation 1.2.4, we obtain the following relation for the treated group:

$$\begin{aligned} F_{1t}(y) &= \sum_{j=2}^{J+1} w_j(y) \sum_{d=1}^{D} \gamma_{j,d}(y) F_{t,d}(y) \\ &= \sum_{d=1}^{D} F_{t,d}(y) \sum_{j=2}^{J+1} w_j(y) \gamma_{j,d}(y) \end{aligned}$$

where the first equation follows from equation 1.2.6. This result shows that a solution exists if the last part replicates the weights of the treated unit, i.e., $\gamma_{1,d}(y) = \sum_{j=2}^{J+1} w_j(y)\gamma_{j,d}(y)$. Given that $\gamma_{j,d}$ could be equal to zero, this condition requires that for a specific point y, the treated unit and some control units are related to the same underlying distributions. Another interesting implication is that the weights $\gamma_{j,d}$ cannot change arbitrarily. The size in which

weights can change across y is linked to the properties of the underlying distributions.[2]

Intuitively, allowing the weights to change across the distribution relaxes the assumption of them being constant. A heuristic argument for this point can be formulated by considering the related papers by Athey and Imbens (2006) and Gunsilius (2023). These papers motivate their model by specifying a function $h(.)$ that maps an unobserved variable u into an observed variable y. The function is also allowed to change across t, i.e., $y = h(u, t)$. For simplicity of the argument, assume that u stays constant across t.[3] Importantly, $h(.)$ applies to all values of u, which results in the weights of the synthetic control unit being constant across the distribution as well.

Consider as an example $h(u, t) = a_t + b_t u$ with $b_t > 0$, i.e. it is strictly increasing in u. Then equation 1.2.4 can be rewritten as:

$$\mathbb{P}\left(Y_{1t} \leq y\right) = \sum_{j=2}^{J+1} w_j(y) \mathbb{P}\left(Y_{jt} \leq y\right)$$

$$\mathbb{P}\left(U_1 \leq (y - a_t)/b_t\right) = \sum_{j=2}^{J+1} w_j(y) \mathbb{P}\left(U_j \leq (y - a_t)/b_t\right)$$

where the second equation follows from plugging in $y = h(u, t)$ and rearranging. Other than $b_t \geq 0$, there are no restrictions on the behavior of a_t and b_t. Thus, the weights $w_j(y)$ have to hold across the whole distribution of U and hence must be constant across different values of y.[4] So, if the underlying model is described by $y = h(u, t)$, then the weights are constant across the distribution. However, my method should exhibit constant weights across y in such a case.

1.2.3 Implementation

As described earlier, the basic idea of the suggested method is to define a grid of different y values and to iteratively apply a well-established synthetic control method using an indicator variable as an outcome of interest. In general, there is no restriction on the specific synthetic control method adopted. In the following, I will again focus on the classic case by Abadie and Gardeazabal (2003) and Abadie, Diamond, et al. (2010).

For the classic synthetic control method to be feasible, the weights must be chosen such that the synthetic unit replicates closely the behavior of the treated unit. In the classic setting,

[2] Further remarks are provided in appendix 1.C.

[3] As described in Gunsilius (2023) one can also relax from the assumption that $u_t = u \; \forall t$. However, u_t needs to be a linear function of its past, and hence the argument is straightforward to extend to that case.

[4] The same result can be derived for $b_t < 0$. The only difference is an additional step that uses the property $\mathbb{P}(X > x) = 1 - \mathbb{P}(X \leq x)$ and the fact that the weights sum up to one.

the weights are chosen such that the sum of the squared distances between the treated and synthetic unit across different predictors is minimized. In the case without covariates, the predictors correspond to the outcome of interest from different pre-treatment periods, and the weights are chosen to solve the following minimization problem:

$$\hat{\boldsymbol{W}} = \min_{\boldsymbol{W}} \left(\sum_{t \in \mathcal{T}_0} (Y_{1t} - w_2 Y_{2t} - \ldots - w_{J+1} Y_{J+1t})^2 \right)^{1/2} \tag{1.2.7}$$

where $\mathcal{T}_0 \subset \{1, \ldots, T_0\}$ and $\boldsymbol{W} = (w_2, \ldots w_{J+1})$.[5] Additionally, the weights are assumed to be non-negative, $w_j \geq 0$ for all j and sum up to one, $\sum_{j=2}^{J+1} w_j = 1$.

For the proposed method, the first step is to estimate the aggregate variable of interest, $E\left[\mathbb{1}(Y_{jt} \leq y)\right]$. The corresponding sample analog is $N_{jt}^{-1} \sum_{i=1}^{N_{jt}} \mathbb{1}(y_{it} \leq y)$, where N_{jt} is the total number of observations in unit j at time t.

Then, applying the normal synthetic control method once to the mean of an indicator variable allows us to derive the treatment effect for one point in the distribution. Relying on the classic approach, the weights for the specific point at y are chosen to solve the following minimization problem:

$$\hat{\boldsymbol{W}}_{\boldsymbol{y}} = \min_{\boldsymbol{W}} \left(\sum_{t \in \mathcal{T}_0} (F_{1t}(y) - w_2(y) F_{2t}(y) - \ldots - w_{J+1}(y) F_{J+1t}(y))^2 \right)^{1/2} \tag{1.2.8}$$

where y is a specific value from the grid and hence $\boldsymbol{W}_{\boldsymbol{y}} = (w_2(y), \ldots w_{J+1}(y))$. The weights are also assumed to be non-negative and sum up to one.

The above procedure is applied to multiple parts of the distribution to derive a distributional effect. Thereby, the weights are allowed to change flexibly across the distribution. Hence, weights are chosen by minimizing:

$$\hat{\boldsymbol{W}} = \min_{\boldsymbol{W}} \left(\sum_{t \in \mathcal{T}_0} (F_{1t}(\boldsymbol{y}) - w_2(\boldsymbol{y}) F_{2t}(\boldsymbol{y}) - \ldots - w_{J+1}(\boldsymbol{y}) F_{J+1t}(\boldsymbol{y}))^2 \right)^{1/2} \tag{1.2.9}$$

where $\boldsymbol{y}$ represent different grid-points and $\boldsymbol{W}$ can now be thought of as a matrix with a different $\boldsymbol{W}_{\boldsymbol{y}}$ in each row. That is, $\boldsymbol{W}$ is a matrix with J columns and a number of rows equal to the number of grid points chosen.

Importantly, by specifying the weights in each row to be non-negative and sum up to one, the support of F_{1t} must lie within the joint support of the control units. Otherwise, the distribution

[5] In the classic setting, an additional weighting vector, $\boldsymbol{v}$ is included in the objective function. These weights capture the importance of the single predictors and impact the resulting synthetic control unit. Methods in choosing these $\boldsymbol{v}$ are discussed in Abadie and Gardeazabal (2003), Abadie, Diamond, et al. (2010), and Abadie, Diamond, et al. (2015) For the case without covariates, the predictors relate to each period and $\boldsymbol{v} = (v_1, \ldots, v_{T_0})$.

of the treated unit cannot be replicated at these points. To avoid this issue, one can trim the y-grid in the exceptional case, where y-values are defined, which only lie once or never in the support of the control distributions in the pre-treatment period.

1.2.4 Inference and Evaluation

To test the statistical significance of the estimated treatment effect, I adjust the placebo permutation method suggested by Abadie, Diamond, et al. (2010), following the notation by Abadie (2021).[6] They suggest assigning the treatment separately for each unit and estimating the placebo treatment effect by applying the synthetic control method. The estimated effect of the treated unit is then compared with the effects from the placebo runs. For this comparison, it is crucial to take into account the overall performance of the synthetic control unit for each run. If the synthetic control unit produces large differences in the pre-treatment period, then a large effect in the post-treatment is likely not related to the treatment. The estimated effect is considered significant when the size of the effect of the treated unit is among the most extreme values relative to the placebo runs.

For each unit to which the treatment is assigned, i, the resulting squared distance between the treated and synthetic unit for each period t and each grid point y can be computed. That is:

$$SE_{it}^{y} = \left(F_{it}(y) - \hat{F}_{it}(y)\right)^2, \text{ where } \hat{F}_{it}(y) = \sum_{j \neq i} \hat{w}_j(y) F_{jt}(y)$$

For the synthetic control methods to be valid, a small SE_{it}^{y} is required for the pre-treatment periods $t \leq T_0$ and each point y. Large differences indicate that the synthetic unit cannot replicate the potential outcome well. For the post-treatment periods $t > T_0$, a larger difference indicates the presence of a treatment effect, especially when the value is much larger compared to pre-treatment values. To control for the overall performance of the synthetic control unit, the difference in the post-treatment period is set in relation to the difference in the pre-treatment period. For our case, this ratio is computed for each point of the distribution:

$$r_i^y = \frac{R_i^y(T_0+1, T)}{R_i^y(1, T_0)}, \text{ where } R_i^y(t_1, t_2) = \left(\frac{1}{t_2 - t_1 + 1} \sum_{t=t_1}^{t_2} SE_{it}^{y}\right)^{1/2} \tag{1.2.10}$$

As a last step, the results of the treated unit are compared to the placebo runs. Furthermore, we can compute a p-value, $p_1^y = 1 - \mathcal{R}_1^y / J + 1$, where $\mathcal{R}_1^y$ is the rank of r_1^y in all runs r_i^y. Please

[6] Despite the various improvements in inference in recent years, I refrain from following a new approach, as it is above this project's scope and the approach by Abadie, Diamond, et al. (2010) remains popular. The same holds true for the fact that the results depend on an assumed underlying distribution of treatment assignment. Here, the assignment is assumed to be uniform across units, i.e., each unit has the same probability of having the treatment assigned. This could be questioned in practice, and there are extensions to address this issue.

note that the metric is referred to as p-value out of convenience. It is not the same metric as the p-value normally used for significance testing. The metric only refers to the same idea, which is to measure how likely we were to observe a similar effect if there was no effect at all.

The above results are straightforward to calculate for the quantile values via interpolation, assuming that the underlying variable is continuous. Chen (2020) outlines the same idea using the quantile function. The approach by Gunsilius (2023) is motivated by the same idea, but he compares the results of the treated unit to the placebo runs for each period and thereby focuses on the overall difference in distributions. In contrast, the metric above is intentionally chosen to depend on a specific point in the distribution. The reason is that the size of the effect potentially differs across the distribution, while the performance of the synthetic control unit also does. As we will see later, the results suggest that, especially in the tails of the distribution, the synthetic control unit does a poorer job of replicating the treated unit well. Consequently, a metric capturing the overall difference in distributions might not properly indicate the presence of a treatment effect. Large variations in the tails might hide a sizable treatment effect located away from the tails of the distribution. The difference from the treatment effect appearing in the post-treatment period becomes negligible compared to the overall variation across periods. While the above metric provides the benefit of capturing a significant effect at a specific threshold, the increased likelihood of finding a significant effect due to repeated testing is not accounted for. Hence, a careful interpretation of the results is essential.

For the comparison of methods, however, an aggregate metric is employed. In that setting, our focus lies less on evaluating the significance of an effect but on assessing the overall precision with which the unobserved distribution as a whole is replicated. Additionally, the comparability of methods is improved since the estimator by Gunsilius (2023) minimizes the Wasserstein distance between distributions to keep the analysis on the aggregate unit level. Therefore, I will use:

$$RMSE_t = \left(\frac{1}{G} \sum_{g=1}^{G} \left(F_{1t}^{-1}(\tau_g) - \hat{F}_{1t}^{-1}(\tau_g) \right)^2 \right)^{1/2}$$

where $\tau_g \in (0,1)$ represents a grid point, $F_{1t}^{-1}(\tau_g)$ is the quantile value of interest at τ_g and $\hat{F}_{1t}^{-1}(\tau_g)$ is the quantile value estimated by the synthetic control unit. Note that the finer the grid values taken, the closer the performance measure is to the 2-Wasserstein distance between the synthetic and treated unit.

1.3 Application and Data Description

This section introduces the application and describes the variables used. As an application, I analyze the impact of the minimum wage introduction in Neuchâtel, a canton in Switzerland. The data used for the analysis is from the Swiss Earnings Structure Survey (ESS) provided by the Federal Statistical Office in Switzerland. The ESS provides repeated cross-sectional data on firms and their employees, where the variables contain mostly but not exclusively information about the employees. The survey has been carried out every second year since 1994, and I have access until 2018. Unfortunately, the ESS is not constructed to be representative on a cantonal level. Hence, the results in this and the following section have to be interpreted with caution.

1.3.1 Minimum Wages in Neuchâtel, Switzerland

Neuchâtel was the first canton in Switzerland to introduce a minimum wage in 2017. Other cantons followed in the years after. The canton of Jura in 2018, Geneva in 2020, and Ticino and Basel-city in 2021. For the analysis, the canton of Jura is therefore excluded from the pool of control units. In 2011, the people living in the canton of Neuchâtel voted for the introduction of a minimum wage. The cantonal government passed the corresponding law in 2014, setting the minimum wage to 20 CHF, pegged to inflation with a basis in August. The case of Neuchâtel is of special interest, given the enforcement. When the law was passed in May 2014, firms had time to adjust their wages until the end of the year. However, several entities, e.g., industry associations, companies, and private individuals, appealed against the new law to the Federal Supreme Court. The law then immediately came into force on the 4th of August 2017, when the Federal Supreme Court rejected the appeal. Given the uncertainty on how the Federal Supreme Court will decide, the time of treatment can be pinned down closely.

1.3.2 Variables

Units of Interest

To tackle how the minimum wage introduction affected people in Neuchâtel, a potential first approach is to identify all individuals affected by the minimum wage law and aggregate them as the treated unit. However, note that one political motivation for the introduction of the minimum wage was to prevent wage dumping, which is a highly debated topic in border cantons with many guest workers. Hence, the minimum wage law was specified to affect all employment relationships, where employees commonly work within the canton of Neuchâtel. Additionally, the introduced law allows for exceptions. To name a few, employees in training or on holiday contracts are excluded from the law, or employees in the agriculture, viticulture, and horticulture sectors face a lower minimum wage. As a result, it is impossible to identify exactly

which individuals are affected by the introduced law and which are not. Therefore, I refrain from focusing on individuals directly affected by the MW and turn my interest to individuals generally being employed in the canton of Neuchâtel. Thus, the units $j = 1, \dots J+1$ represent the cantons where individuals work.

This implies that my results could also be interpreted as a lower bound of the treatment effect on the employees targeted by the minimum wage. That is, people employed in other cantons, especially cantons closer to Neuchâtel, still might generally work within the canton of Neuchâtel and hence would be directly affected by the introduction of the minimum wage. However, a significant amount of individuals generally working within the canton of Neuchâtel and being employed in another canton are needed to distort the estimation of the distributional effects. Additionally, if the results are affected by the described spillover effect, then they are likely to be downward biased; that is, the synthetic control suggests wages that are too high.

The ESS comprises different regional variables of which two are important to identify the canton individuals are employed in. The first variable assigns all employees of a private firm to the canton in which most employees of the firm are working. The second assigns employees to regions in which their workplace is located. Unfortunately, these regions exhibit a finer granularity than cantonal borders, and sometimes these regional and cantonal borders do not coincide. Of the three regions covering Neuchâtel, one also covers the canton of Bern. For the analysis, a new cantonal variable is created. Employees with a workplace in a region that consistently lies within a canton are also assigned to this canton. An employee with a workplace in a region that overlaps multiple cantons is only assigned to a specific canton if the firm in which the employee is working has most of the employees working in that canton. If the firm has most employees working in a canton not covered by the region, I exclude the observation from the analysis. This new variable gives the best approximation for the canton an individual is actually employed in and avoids losing too many observations from regions not consistently lying within a canton.

I will only keep control units large enough to estimate distributional effects for the analysis. Cantons with, on average, less than 10′000 individuals per year in the post-treatment periods are excluded, leading to $J = 17$. Individuals above the age of 65 and below the age of 25,[7] as well as individuals employed in the public sector, are dropped. Lastly, unrealistically high or low values for wages and working hours are excluded by trimming the lowest and highest values.

[7] A robustness check in which individuals with an age of 18 or older are kept in the sample is presented in appendix 1.B.

Outcome of Interest

The focus of the analysis lies on the impact on hourly wages, which are conducted by combining working hours and wages. Typically, the ESS takes the month of October as a reference point when measuring these variables. For working hours, two variables are available, measured monthly or weekly, depending on the employee's contract. To construct a single monthly variable, weekly working hours are multiplied by 52/12. The reason is that while monthly work hours are recorded for individuals employed on an hourly basis, weekly work hours are recorded for individuals with a regular contract and fixed income for a year. The ESS further comprises three different variables for wages. A standardized wage, a gross wage, and a net wage. However, each has been altered differently and includes different additional payments. For example, the standardized wages additionally include social payments, or gross wages don't consider the 13th monthly salary. The analysis will focus on net wages as it also takes into account additional payments like the 13th monthly salary or Sunday supplements and hence displays the available money to the individual. Further, payments for overtime and other special payments are excluded, and wages are deflated using the consumer price index provided by the Federal Statistical Office.

1.3.3 Descriptives

The total number of observations used in the analysis is around 11.4 million and 13 waves. In earlier years, fewer observations were available. Until the year 2000, each wave covered approximately between 400 and 450 thousand observations, in the years after, each wave had between 0.9 and 1.3 million observations.[8]

From tables 1.1 and 1.2, we see that changes over time are relatively similar for Neuchâtel and Switzerland. One exception is in 2000, where wages for Neuchâtel are slightly higher across the distribution. Additionally, it looks as if in 2018 wages also tend to decrease a bit less for

Table 1.1: Hourly Wages for Neuchâtel

	min	25%	med.	mean	75%	max	obs.
1994	4.82	18.56	23.22	25.78	29.90	106.44	9357
2000	6.82	20.35	25.75	28.27	32.78	109.84	6512
2006	4.41	19.92	24.12	26.79	30.55	108.87	25003
2012	3.51	18.04	22.00	24.70	28.26	108.94	22373
2018	3.49	21.78	26.43	29.30	33.50	109.83	26590
total	3.49	19.84	24.28	26.94	30.80	109.84	264150

[8] See appendix 1.A.1 for more detailed description tables across all years.

Neuchâtel compared to Switzerland. The results also indicate that the wage distribution of Neuchâtel is comparable to that of Switzerland, such that it seems feasible to find weights that allow the synthetic control unit to replicate the distribution of Neuchâtel closely.

Table 1.2: Hourly Wages for Switzerland

	min	25%	med.	mean	75%	max	obs.
1994	3.58	20.49	26.07	28.58	33.22	109.99	394959
2000	3.70	20.12	25.56	28.32	33.23	109.90	449940
2006	3.50	20.97	26.06	28.95	33.52	109.99	1038001
2012	3.48	19.29	24.18	26.78	31.11	109.99	1046126
2018	3.48	22.23	27.75	30.63	35.76	109.95	1249239
total	3.48	20.93	26.21	28.98	33.75	110.00	11386093

1.4 Results

1.4.1 Application: Neuchâtel

In this section, the proposed method is applied to the minimum wage introduction in the canton of Neuchâtel. The minimum wage policy was introduced in 2017 when the Federal Supreme Court rejected the appeal. Given the coverage of our data, the post-treatment period ranges up to $T_0 = 12$, which is the wave of 2016. The only post-treatment period, $T = 13$, is the wave of 2018.

Figure 1.1 plots the CDF of the treated unit, Neuchâtel, and of the synthetic control unit. Results indicate a small improvement for individuals with a wage of around 18 to 25 CHF per hour and 35 CHF per hour. Otherwise, the introduction of the minimum wage has no visible impact on the distribution of hourly wages. The difference at the bottom of the distribution is as expected and is consistent with the results from Berger and Lanz (2020). Introducing a minimum wage for employees who generally work in Neuchâtel should also increase wages for the lower part of individuals employed in Neuchâtel. The difference in the upper half of the distribution is less expected; however, it seems smaller in absolute terms and relative to the hourly wages. Hence, the difference potentially captures an inaccuracy resulting from the synthetic unit being unable to replicate the treated unit well.

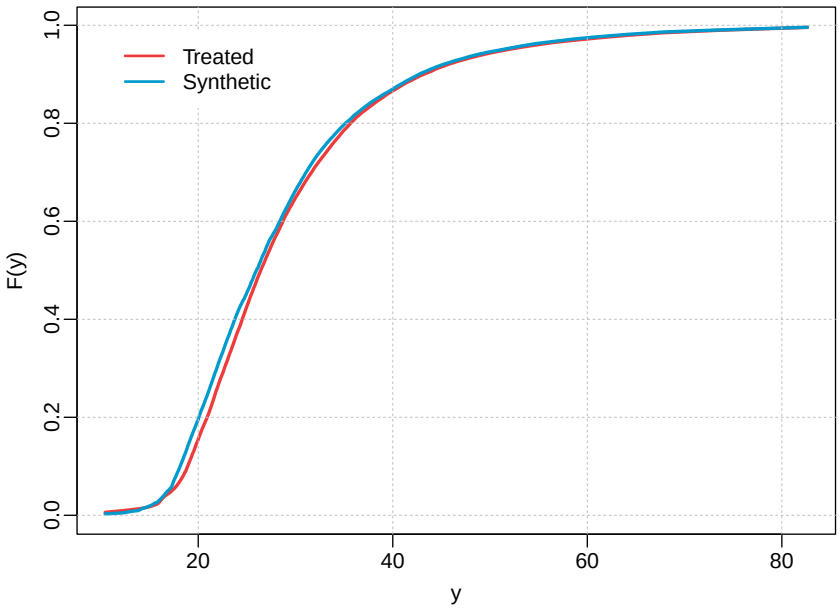
Additionally, some individuals earn an hourly wage less than the minimum wage. This has several reasons. First, as described in section 1.3, we look at individuals working in Neuchâtel, which does not necessarily mean that most of their work is done in Neuchâtel and, hence, that they are affected by the minimum wage law. Second, the minimum wage law allows for

exceptions from the law. Thus, some individuals are only affected by a lower minimum wage or not at all.

Figure 1.2 gives insights into the economic significance of the effect as well as the performance of the synthetic control unit. To improve readability, the more common quantile treatment effect (QTE) is chosen to display distributional effects rather than the difference between the cumulative distribution functions. Computing the QTE is straightforward by first inverting the cumulative distribution function to get the corresponding quantile functions and then taking the difference across different τ.

The gray lines in figure 1.2 are the computed differences in the quantile functions between the treated and the synthetic unit for pre-treatment periods. Hence, they show how well the synthetic control unit was able to replicate the distribution of the treated unit in periods before the treatment. With the exception of two years, the synthetic control unit is able to replicate the distribution of Neuchâtel closely. The first exceptional year is 2000 when the synthetic unit generates wages that are too low to replicate the true distribution well. Tables 1.6 and 1.7 from the appendix 1.A.1 reveal that in the year 2000, wages in Neuchâtel are higher relative to other years compared to wages in Switzerland, which is why the synthetic control underestimates wages in that year. The second exception is 2014, for which the QTE starts to drop around $\tau = 0.6$. This observed difference in 2014 is highly unlikely to be a reaction to the minimum

Figure 1.1: CDFs of the Treated and Synthetic Unit

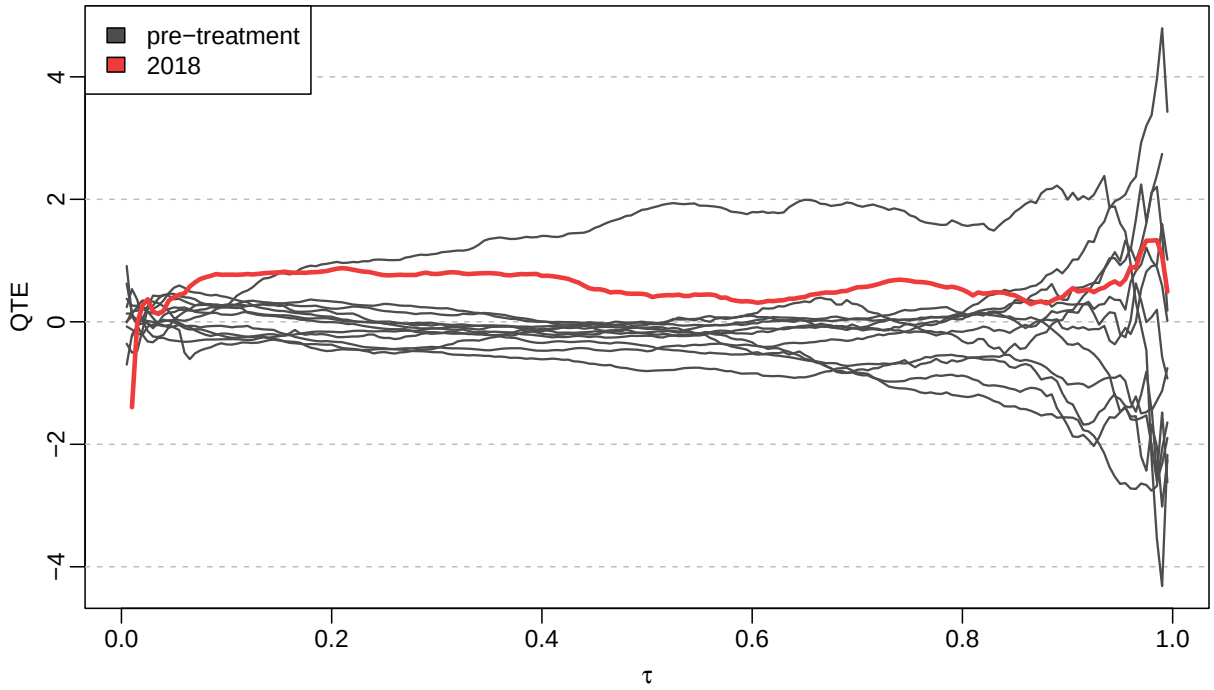


Notes: The lines display the CDFs of Neuchâtel (red) and of the synthetic control unit (blue) for the year 2018. Both are close to each other, with small differences in hourly wages between 20 and 32.

wage. The minimum wage was communicated at the end of May 2014, and wages in the ESS are collected for the month of October. Within less than five months, a significant share of firms would have needed to react by lowering wages. However, lowering wages is mostly possible when hiring new employees; otherwise, it is more difficult. Hence, it is more likely that the observed effect results from having an exceptional year for which the synthetic unit cannot capture the treated unit well for the upper half of the distribution. Again, consulting tables 1.6 and 1.7 in the appendix 1.A.1 support this argument. In the year 2014, the decrease in wages in the upper part of the distribution compared to other years was relatively stronger in Neuchâtel than in the rest of Switzerland. From figure 1.2, we further see that in the tails and notably in the upper part of the distribution, volatility in estimated QTEs becomes larger. This results from the wage distribution being left skewed. For the upper part of the distribution, the numeric range of hourly wages becomes larger, moving from one percentile to another.

For the post-treatment period, displayed as a red line in figure 1.2, the QTE stays overall positive, which means that at each point of the distribution, the hourly wage of the synthetic Neuchâtel is lower. As seen in the figure before, there are two visible differences in the lower and upper part of the distribution, whereas this effect seems to be a bit higher for the lower part of the distribution.

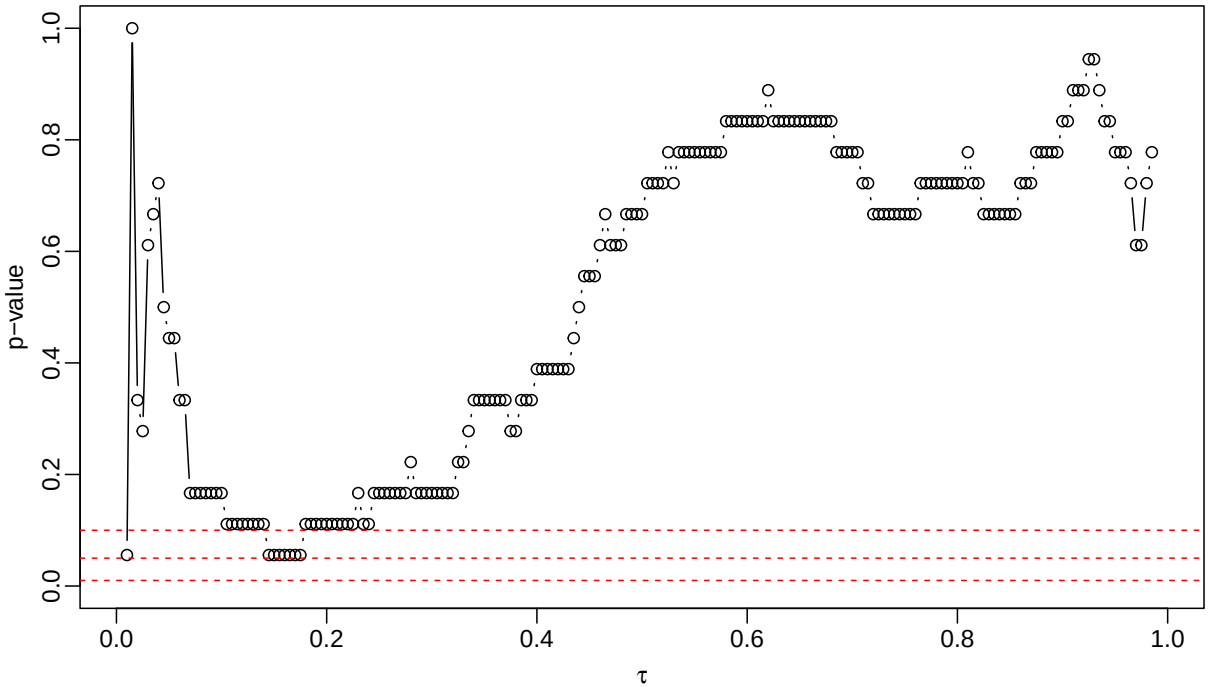
Figure 1.2: QTE Compared to Pre-Treatment Periods



Notes: The figure displays the QTE for Neuchâtel across different years. The colored line displays the QTE for the post-treatment years. Lines in gray display the QTE for the pre-treatment years.

To gain insights into whether the effect is statistically significant, I follow the placebo permutation test outlined in section 1.2. To preserve consistency with the results above, differences between the treated and synthetic units are computed between quantile values. That is, my suggested method is applied to each control unit, and the resulting QTE is computed. These effects are then utilized as placebo treatment effects. Figure 1.6a in appendix 1.A.2 displays the QTE for the post-treatment period for units that were actually not treated. For some placebo runs, the estimated effect in the post-treatment is sizable. However, those effects are likely to be of similar size in the pre-treatment periods, i.e., we cannot construct a valid synthetic control unit for these placebo runs. Compared to units with more reasonable placebo effects, the positive QTE in the lower part of the distribution potentially indicates a weak significant effect. For a more formal comparison, figure 1.3 displays the p-values as described earlier. Taking into account the performance of the synthetic control unit in each run, each p-value indicates how exceptional the treatment effect at point τ for Neuchâtel is compared to the placebo runs. The results indicate a weakly significant effect at the lower part of the distribution. Some p-values of 1/9 are barely above 10%. For our case with $J = 17$, a value of 1/9 implies that next to Neuchâtel, there was one other control canton with a higher ratio between post- and pre-treatment differences at the bottom of the distribution. Further, the results clearly rule out the

Figure 1.3: Placebo Permutation Results



Notes: The figure displays in black the p-values across the different values of τ, derived via placebo permutation as described in section 1.2. The red dotted lines mark the 10%, 5%, and 1% values.

smaller difference in the upper part of the distribution to be a significant effect. Thus, evidence suggests a weakly positive effect at the bottom of the distribution, which is also expected for the introduction of the minimum wage. It follows that for individuals with a lower wage who are employed in Neuchâtel, the introduction of the minimum wage seems to have increased the hourly wages. The effect lies mostly between 0.5 and 1 CHF per hour, a small but notable change for individuals with an already low hourly wage.[9]

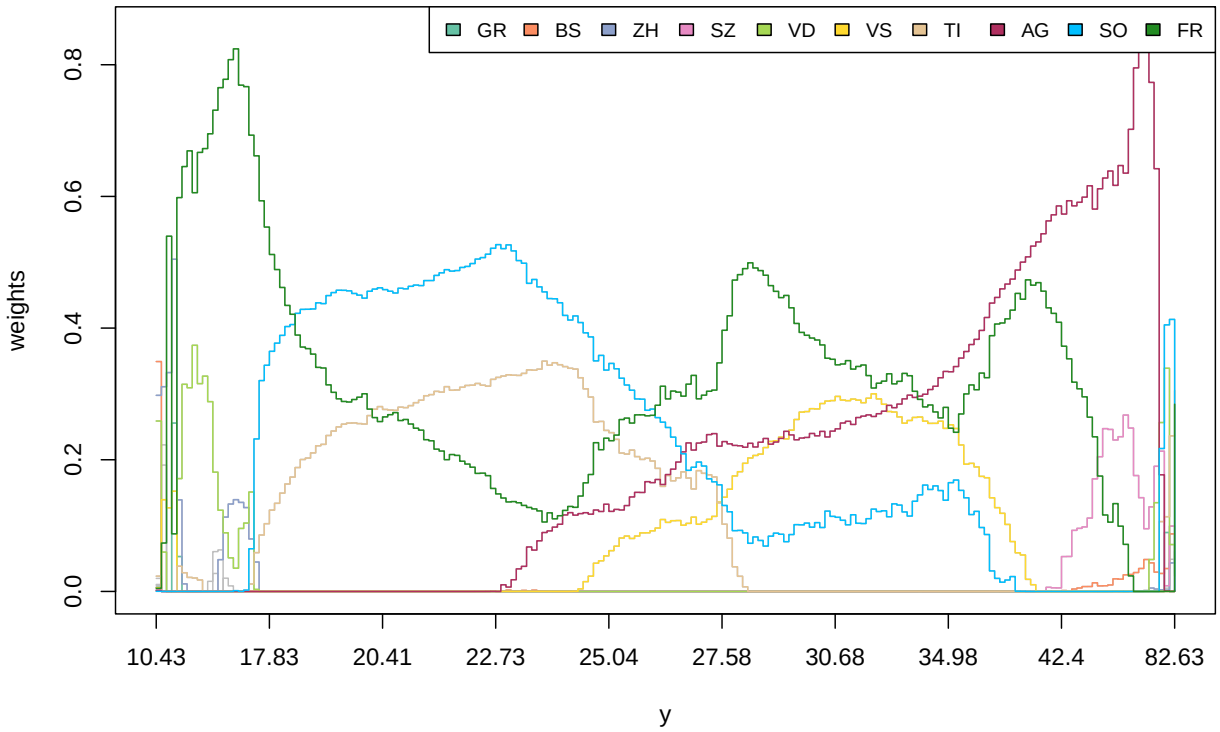
It remains to point out that the small effect at the bottom extends across a broader range of the distribution. This supports the findings by Berger and Lanz (2020), indicating that not only individuals with an income below the minimum wage level benefit from the new policy but also individuals with an income slightly above. Furthermore, Berger and Lanz (2020) find no significant effect on employment or product pricing. Hence, employers do not use these channels to adjust for the higher wages. Following this idea, I further analyze the potential impact of the minimum wage intervention on various outcomes (see appendix 1.A.3 for further details). First, the proposed method is applied to work hours, which are also available in the ESS, to analyze if the number of hours decreases following the minimum wage intervention. The findings suggest no strong reaction in work hours. If anything, there would be a slight shift from part-time to full-time employment. The conventional synthetic control method is applied without covariates for the remaining additional variables. Consistent with the findings in Berger and Lanz (2020), results do not suggest a significant effect on unemployment rates. Additionally, there seems to be no impact on the fraction of closing firms. For the number of workers from abroad as well as the number of registered open job positions, a significant difference between the treated and synthetic units in post-treatment periods is apparent. While the number of workers from abroad decreases, the number of open job positions increases. The introduction of the minimum wage could explain these findings to some degree. However, the observed effects are very likely influenced by another national policy intervention occurring around the same time. The policy was initially aimed at reducing the number of workers from abroad, and when implemented, it certainly increased the number of registered job positions. Hence, it remains uncertain how much of the observed differences in the post-treatment periods for the two variables can be attributed to the minimum wage policy.

Finally, figure 1.4 shows the weights used for the synthetic Neuchâtel. Results suggest that allowing weights to change across the distribution is important. The canton of Fribourg is the only control unit that persistently contributes to the synthetic unit with a positive weight. Other cantons like Ticino contribute only to the lower part of the distribution, while other cantons like Aargau contribute to the upper part. Furthermore, the weights are relatively smooth along y for the large part of the distribution. Large jumps are only visible in the

[9] Additional robustness checks regarding anticipation effects, the inclusion of young employees, as well as the choice of pre-treatment periods are reported in appendix 1.B. Results point towards the same finding as in the main analysis.

tails, especially in the lower tail. This could be related to the multiple solutions issue. For many units, the values in the tails of the CDF are close to either 0 or 1, i.e., differences across units are rather small. As a result, the potential for many different combinations of weights to yield a good fit becomes larger. That is, the solution is no longer unique, which leads to the jumping results from one y-value to another. Furthermore, stronger changes in the lower tail are theoretically feasible since the cumulative distribution function goes to 0. Given the unstable choice of weights in the tails, a potential disadvantage of the method is that the treated unit cannot be matched well in the tails of the distribution. Especially for the case where the true weights are constant across the whole distribution, it is likely that my method fails to estimate the weights in the tails well, while another method that assumes constant weights across the distribution yields better results in the tails. This imposes some sensibility on the specific choice of bins, which could be addressed by smoothing the weights or assuming some functional form. In appendix 1.B an analysis using smoothed weights is conducted.

Figure 1.4: Weights Across the Distribution



Notes: Values of the weights are displayed on the y-axis in percentage points. For readability, the x-axis is stretched to represent quantile values from Neuchâtel over all years.

1.4.2 Comparison to Other Approaches

In the following, the performance of four different approaches is compared. The first approach will be denoted as *CSC*, which is the suggested method in this paper. The second approach will be denoted as *QSC*. It is the quantile equivalent to the suggested method, which is also mentioned in Chen (2020).[10] The last two approaches are both methods suggested in Gunsilius (2023) and are based on optimal transportation theory. The main method, which relies on the quantile function, will be denoted as *QOT*, and the alternative method, which relies on the cumulative distribution function, will be denoted as *COT*.

For the approaches using quantile values, a grid with values between zero and one is required. For the approaches using the cumulative distribution functions, a y-grid is chosen to cover the support of the distributions of the treated unit over all periods. Note that all four approaches discussed base their estimation of the synthetic control weights on several grid points that span across the distribution. My approach and the quantile equivalent use a fixed pre-specified grid, while the approaches by Gunsilius (2023) randomly draw points. As a result, the computation of the weights is not based on the exact same positions in the distribution. However, a fine grid for estimation is chosen, and the distribution of the treated and the synthetic units are compared at the same thresholds. Thus, employing estimation at slightly varying grid points should not change overall results. This is especially the case if one uses a performance measure, like the $RMSE_t$ introduced earlier in section 1.2, that takes an average over the whole distribution.

Accordingly, I will use the $RMSE_t$ as a simple measure of how well the synthetic control unit replicates the distribution of the treated unit. For comparability, the performance measure is computed in terms of quantile values over the same grid. Hence, the results from the approaches using cumulative distribution functions are inverted. This ensures the same scale in the outcome, i.e., in each case, we compare differences in the value y at specific percentiles.

Application Results

In the following, my proposed method and the main method suggested by Gunsilius (2023) are applied for the analysis of the minimum wage in Neuchâtel.

Figure 1.5 contains the same results as in figure 1.2, but also adds the results when using the estimator by Gunsilius (2023) in blue and brown. As expected, the latter exhibits a less flexible behavior. This is of advantage, e.g., when looking at the pre-treatment period from 2014. Instead of a strong drop between $\tau \in (0.6, 0.7)$, the QTE exhibits a rather constant, slightly negative slope across the distribution. However, as seen for 2018, the estimator also reduces differences that are likely to be local effects.

Table 1.3 compares the 2-Wasserstein distance between the synthetic and the treated unit for

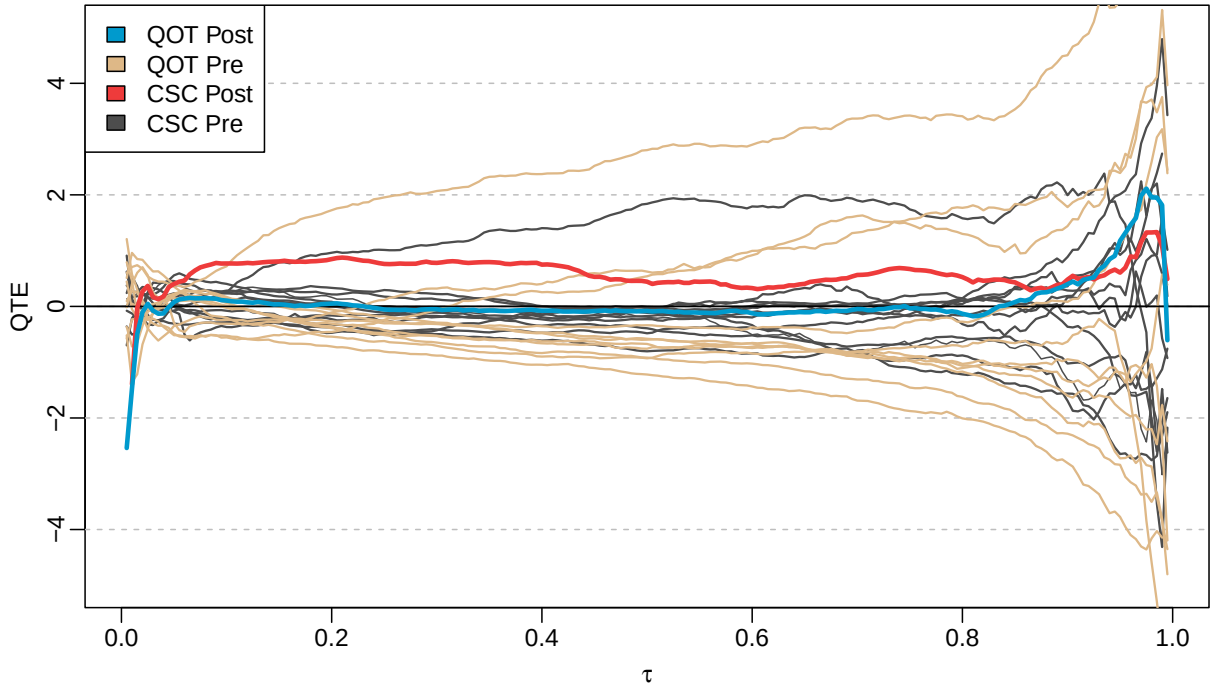
[10] see equation 35

Table 1.3: Comparison of 2-Wasserstein Distances

	1994	1996	1998	2000	2002	2004	2006	2008	2010	2012	2014	2016	2018
CSC	0.70	0.47	0.22	1.63	1.03	0.24	0.47	0.54	0.26	0.70	0.76	0.42	0.65
QOT	0.39	1.27	1.20	3.08	1.68	0.72	1.14	1.39	0.82	1.22	0.91	0.65	0.46

each year and both methods. The 2-Wasserstein distance is an integral over squared differences in quantile functions. However, neither method imposes a functional form on the distribution, and hence, the integral is computed using the $RMSE_t$ with F_{1t}^{-1} being equal to the observed pre-treatment quantile values and a fine grid for G. With the exception of the year 1994, the synthetic unit from my approach performs better in replicating the treated unit. Additionally, it provides evidence for the issue discussed above, that the standard permutation methods are unsuitable when the treatment only affects one part of the distribution. The table shows that the computed 2-Wasserstein distance for the post-treatment period is still very small, even with a larger difference in one part of the distribution.

Figure 1.5: Comparison of QTE Compared to Pre-Treatment Periods



Notes: The colored lines display the QTE for the post-treatment years. Lines in gray and brown display the QTE for the pre-treatment years.

Simulation Results

For the simulation part, different combinations of $T_0 = \{5, 10, 50, 100\}$ and $J = \{5, 10, 50\}$ with one treated unit and one post-treatment period are considered. For each combination, I repeatedly draw a new sample of $N_{jt} = 4000$ observations from a data generating process (DGP) and estimate the resulting $RMSE_t$ between the estimated and true quantile values if not treated for the post-treatment period. The number of total repetitions is 1000. Furthermore, I use two different DGPs, and the treatment effect is specified to negatively affect the lower tail of the distribution and to fade off towards the upper tail.

The first DGP is based on my approach and proceeds in three steps: Following equation 1.2.6, in the first step, pools of underlying distributions and mixture weights are constructed. Ten underlying gamma distributions with different combinations of shape and scale parameters are used for the distributions. For the mixture weights, five different patterns across y are defined. Then in the second step, the underlying distributions and mixture weights are randomly sampled. For each period, three distributions, $F_{t,d}$, are drawn from the pool, i.e., $D = 3$. For each unit, one to three weighting patterns, $\gamma_{j,d}(y)$ are randomly drawn and assigned to a distribution. Hence, $\gamma_{j,d}(y)$ is allowed to be zero for some d, but at least for one distribution, we have non-zero values. The weights are then rescaled to sum up to 1. Additionally, a check is run to ensure that the resulting CDFs are increasing. If the check fails, the mixture weights are smoothed. In the last step, we compute the distribution, $F_{jt}(y)$ for each unit and period according to equation 1.2.6 and generate random draws from it. Furthermore, the treatment effect is added for $t > T_0$.

The second DGP is based on the approach by Gunsilius (2023). Here, the relation between the underlying unobserved variable and the observed variables, as well as future unobserved variables, is assumed to be linear. I will further assume the underlying unobserved variable is normally distributed, which allows me to compute the resulting normal distribution of the observed variable directly. The DGP proceeds in three steps. First, the underlying distribution of the unobserved variables is defined. For the variance-covariance matrix, I specify two different settings. In the first, small correlation values between the groups are randomly drawn. In the second, we don't have any correlation between the groups, and the relations among them result from the linear functions that depend on time and are shared across groups. In the second step, we iterate through time and compute the resulting unobserved and observed variables. In the final step, observations are drawn from these distributions, and a treatment effect is added for $t > T_0$.

Table 1.4 displays results from the first DGP for different combinations of the number of pre-treatment periods, T_0, and control groups, J. The results show that the methods using varying weights perform the best, with the proposed method achieving the best pre-treatment fit between synthetic and treated units. The two approaches working with Wasserstein distances

and constant weights achieve worse results. This is as expected, given the underlying DGP. Additionally, the methods using cumulative distribution functions perform better than their counterparts using quantile functions. This is also as expected since the underlying mixture model imposes the relations between the treated and control unit depending on y and not the percentiles. Overall, the results improve as the number of control groups and the number of pre-treatment periods increases. The longer the pre-treatment period, the better the synthetic control captures the underlying structure, and the more control units are drawn, the more likely some of the control units are close to the treated unit, such that a better synthetic unit can be constructed.

So far, results have confirmed the importance of the proposed method. To check the performance of the proposed method in more detail, the simulation study is repeated with the second DGP that is closer to the approach by Gunsilius (2023). Table 1.5 displays the post-treatment $RMSE_t$ for different combinations of T_0 and J. Overall, results indicate that the four methods perform similarly well. Hence, even if the DGP imposes constant weights across the distribution, the methods using varying weights capture the underlying structure well. The results also show that the methods using quantile functions perform slightly better than those using cumulative distribution functions. Overall, these findings further strengthen the importance of varying weights.

Table 1.4: Average Performance for the Post-Treatment Periods

T0	J	CSC	COT	QSC	QOT
5	5	2.1245	2.5618	2.4881	2.7212
5	10	1.0914	1.7679	1.5586	1.9986
5	50	0.3813	0.9274	0.8442	1.2766
10	5	2.1933	2.631	2.4443	2.7635
10	10	1.0254	1.6671	1.3966	1.9075
10	50	0.2766	0.8274	0.5645	1.2105
50	5	2.1329	2.481	2.3645	2.6596
50	10	0.862	1.2888	1.1644	1.6485
50	50	0.1864	0.6032	0.3532	1.0068
100	10	0.8841	1.2875	1.1753	1.6413
100	50	0.1625	0.5718	0.3159	0.9822

Notes: Results display the $RMSE_t$ averaged over all $t > T_1$ for different combinations of the number of control units, J, and pre-treatment periods, T_0. The abbreviations describe the underlying methods. The first letter indicates whether the cumulative distribution function (C) or the quantile function (Q) is used, and the last two letters indicate whether the traditional synthetic approach (SC) or the 2-Wasserstein distance (OT) is implemented.

Table 1.5: Results with Alternative DGP

		Uncorrelated				Correlated			
T0	J	CSC	COT	QSC	QOT	CSC	COT	QSC	QOT
5	5	0.111	0.143	0.059	0.118	0.112	0.143	0.059	0.117
5	10	0.056	0.06	0.034	0.046	0.056	0.06	0.033	0.045
5	50	0.04	0.03	0.03	0.027	0.044	0.033	0.035	0.031
10	5	0.068	0.092	0.037	0.076	0.067	0.091	0.035	0.073
10	10	0.031	0.037	0.02	0.029	0.03	0.036	0.019	0.027
10	50	0.021	0.018	0.018	0.016	0.023	0.02	0.02	0.019
50	5	0.008	0.007	0.003	0.006	0.008	0.007	0.003	0.006
50	10	0.003	0.003	0.002	0.002	0.003	0.003	0.001	0.002
50	50	0.002	0.002	0.001	0.001	0.002	0.002	0.002	0.002
100	10	0.019	0.018	0.006	0.009	0.018	0.017	0.005	0.009
100	50	0.013	0.012	0.005	0.005	0.011	0.011	0.006	0.005

Notes: Results display the $RMSE_t$ averaged over all $t > T_1$ for different combinations of the number of control units, J, and pre-treatment periods, T_0. 'Uncorrelated' refers to the case where the underlying unobserved variables from each unit are uncorrelated. 'Correlated' refers to the opposite case where they are correlated. The abbreviations describe the underlying methods. The first letter indicates whether the cumulative distribution function (C) or the quantile function (Q) is used and the last two letters indicate whether the traditional synthetic approach (SC) or the 2-Wasserstein distance (OT) is implemented.

1.5 Conclusion

This paper contributes to the growing literature on synthetic control methods by formalizing an intuitive extension that allows for the estimation of effects across the distribution. In a setting where the variable of interest is observed at a lower level than the aggregate unit, the researcher decides how to aggregate the observed values to analyze the effect on the treated unit. The proposed method takes advantage of this setting by aggregating individual values as the mean over an indicator function, which represents a specific point in the cumulative distribution function. The effect across the distribution is then derived by repeatedly applying a synthetic control estimator while varying the indicator function specification. Thereby, the method relies on well-established synthetic control methods. Compared to the existing distributional synthetic control methods, the proposed extension is not only natural but avoids other issues like mass points and allows weights to vary across the distribution.

The suggested method is applied to the introduction of the minimum wage in Neuchâtel, Switzerland. Crucially, the estimated weights of the synthetic control unit provide evidence that allowing these weights to change across the distribution is essential to replicate the distribution of the treated unit. Results further support the finding that for individuals employed in Neuchâtel, there is a positive effect on labor income in the lower part of the distribution,

and the results from the permutation test indicate that the estimated effect is weakly significantly different from zero. These findings, together with further results on various outcomes, like unemployment rates, are in line with past findings by Berger and Lanz (2020). Still, it remains to point out that given the limitations of the available survey, it would be interesting and important to conduct further analysis on that matter.

To compare the various distributional synthetic control methods, the application is repeated using a second method, and the simulation study compares four different methods. Findings suggest that the proposed method and its quantile equivalent, which both use varying weights, yield the best results, as their synthetic control unit replicates the target distribution of the treated unit in the post-treatment periods the closest. Hence, these results further strengthen the finding that allowing for varying weights across the distribution is important.

For simplicity, I refrained from including covariates in the model. However, other variables might be of importance as well to find a synthetic control unit that matches the characteristics of the treated unit well. For the proposed method, including covariates is possible in different ways. On the one hand, when a synthetic control method is iteratively applied over different points in the distribution, additional covariates can be added to each estimation. However, the choice of these covariates for different thresholds of the distribution has to be discussed. On the other hand, a two-step procedure can be used. In the first step, the cumulative distribution functions conditional on covariates are estimated using distribution regression, and counterfactual cumulative distribution functions conditional on the characteristics of the treated unit are constructed for each unit (see Chernozhukov, Fernández-Val, and Melly (2013) for details).[11] In the second step, the suggested method is applied using the counterfactual instead of the observed cumulative distribution functions.

Further, this paper strongly focuses on the relationship between the proposed method and the classic synthetic control method. However, other extensions of the synthetic control method could be used as well for estimation. For future research, it would be interesting to assess the relation to other extensions theoretically and empirically. Of special interest are extensions that address the poor-matching problem in pre-treatment periods; see, for example, Doudchenko and Imbens (2016), Ferman and Pinto (2021), and Ben-Michael et al. (2021). In the classic case, Abadie, Diamond, et al. (2010) and Abadie and Gardeazabal (2003) advice against using the synthetic control method if the pre-treatment fit between the synthetic and the treated unit is poor. Neither this paper nor the paper by Gunsilius (2023) discuss this issue. However, Chen (2020) augments his model and uses estimated residuals, which allows him to address the

[11] It is important to think carefully about the relations between covariates and outcome of interest and how the distribution regression should be modeled. The point in using different covariates for the outcome of interest, Y, is that one suspects these to have an impact on Y and that this impact is the same across units, but not necessarily across time. Hence, one probably would want the coefficients for different covariates from the distribution regression to be the same across units but to be able to vary across periods.

poor-matching problem using observable heterogeneity in the pre-treatment periods. Hence, it might be interesting to see whether combining the proposed method with one of the extensions addresses the issue properly.

Finally, I followed standard permutation methods to do inference in this paper. As discussed in section 1.2 there are likely to be disadvantages if a single average measure of the distance between distributions is used. A local treatment effect could be missed. The approach to account for this issue suggested in the paper is very heuristic, and therefore, it would be interesting to improve inference to convincingly identify significant treatment effects along the distribution.

1.A Appendix I - Supplementary Insights and Results

1.A.1 Detailed Descriptives

In this section, the tables 1.1 and 1.2 are displayed in more detail. The columns for the 1^{st}, 5^{th}, and 10^{th} percentile at the bottom and top of the wage distribution are added as well as all available years. Overall, wages for individuals employed in the canton of Neuchâtel are not very different compared to Switzerland. However, as described in section 1.4, we observe small differences that potentially explain some of the observed results. First of all, we observe a less strong decrease in wages for the year 2018 for Neuchâtel compared to Switzerland. Additionally, we see that for the year 2000, wages in Neuchâtel are higher relative to other years compared to wages in Switzerland. Finally, for the year 2014, wages in Neuchâtel from the 75^{th} percentile upwards are lower relative to other years compared to the same part of the distribution of Switzerland.

Table 1.6: Hourly Wages over Years for Neuchâtel

	min	1%	5%	10%	25%	med.	mean	75%	90%	95%	99%	max	obs.
1994	4.82	12.49	15.03	16.07	18.56	23.22	25.78	29.90	38.24	45.59	65.70	106.44	9357
1996	6.60	12.30	13.97	15.10	17.91	23.26	25.69	30.18	38.29	46.01	67.67	109.50	8963
1998	6.26	12.98	15.00	15.83	18.87	24.26	26.62	31.25	39.42	47.17	68.60	104.71	8318
2000	6.82	13.12	15.06	16.54	20.35	25.75	28.27	32.78	42.43	51.24	73.50	109.84	6512
2002	4.57	12.64	14.89	16.31	19.19	23.33	25.78	29.13	37.50	45.17	68.50	109.57	29363
2004	6.74	13.56	15.84	17.20	20.06	24.34	27.09	30.71	39.69	48.29	71.90	109.41	24867
2006	4.41	12.08	15.81	17.19	19.92	24.12	26.79	30.55	38.96	47.05	68.46	108.87	25003
2008	5.08	12.97	15.79	17.12	19.86	24.07	26.66	30.33	38.88	46.64	68.63	109.73	27028
2010	4.45	12.87	15.88	17.56	20.55	24.98	27.82	31.71	41.47	49.74	71.70	107.23	28550
2012	3.51	6.55	13.37	15.35	18.04	22.00	24.70	28.26	37.22	45.43	66.97	108.94	22373
2014	3.56	11.52	15.07	16.49	18.76	22.42	24.67	27.81	35.00	41.82	62.26	109.42	22675
2016	4.54	14.68	17.56	19.11	22.13	26.79	29.80	33.94	44.00	53.05	74.71	108.81	24551
2018	3.49	12.53	17.28	18.91	21.78	26.43	29.30	33.50	43.22	51.74	73.80	109.83	26590
total	3.49	12.35	15.35	16.87	19.84	24.28	26.94	30.80	39.87	47.89	70.10	109.84	264150

Table 1.7: Hourly Wages over Years for Switzerland

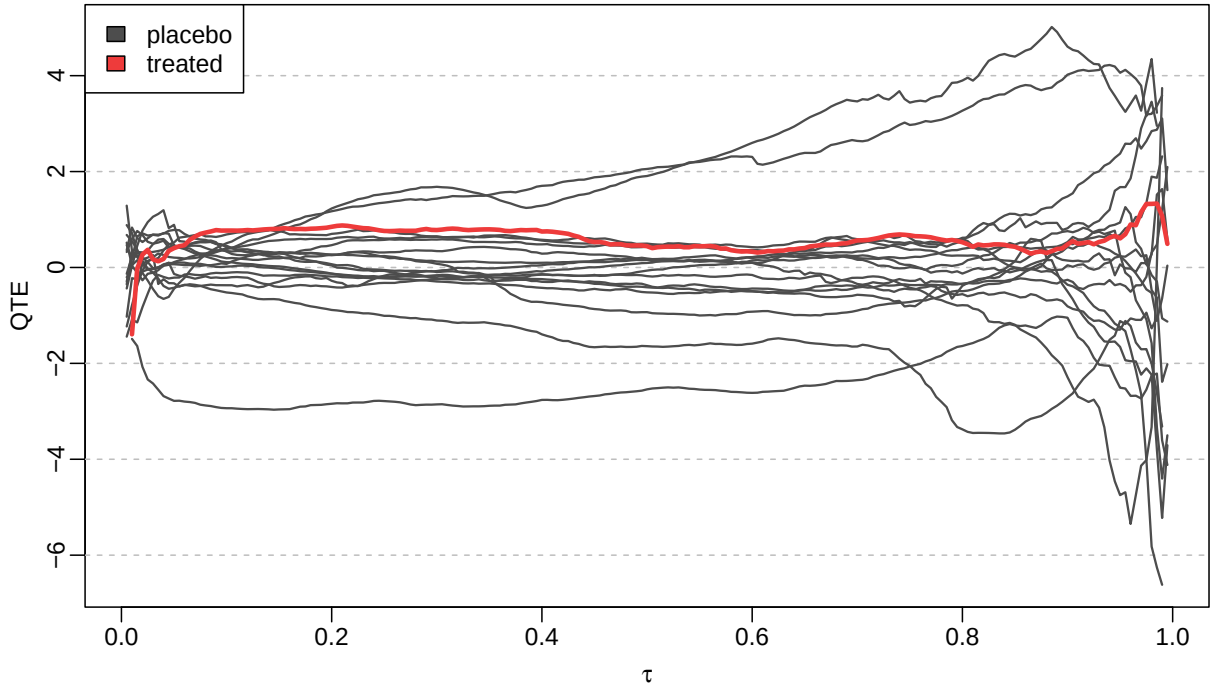
	min	1%	5%	10%	25%	med.	mean	75%	90%	95%	99%	max	obs.
1994	3.58	12.13	15.17	16.72	20.49	26.07	28.58	33.22	43.16	51.56	73.01	109.99	394959
1996	4.17	12.62	15.31	16.61	19.92	25.30	27.73	32.17	41.62	49.44	69.75	109.92	414349
1998	4.67	12.68	15.58	16.95	20.64	26.17	28.64	33.44	42.75	50.67	71.95	109.92	382800
2000	3.70	12.15	15.15	16.69	20.12	25.56	28.32	33.23	43.19	51.23	72.41	109.90	449940
2002	3.65	13.12	15.63	17.28	20.89	26.06	28.82	33.41	43.46	51.66	73.02	109.98	928850
2004	3.48	13.50	15.96	17.54	21.04	26.19	29.04	33.63	43.82	52.14	73.58	110.00	1010126
2006	3.50	12.80	15.83	17.50	20.97	26.06	28.95	33.52	43.89	52.20	73.57	109.99	1038001
2008	3.48	12.71	15.86	17.49	20.88	26.09	28.92	33.61	43.85	51.91	72.70	109.97	1094337
2010	3.49	12.54	15.79	17.55	21.11	26.58	29.41	34.37	44.73	52.85	73.87	109.99	1227273
2012	3.48	6.13	13.95	15.98	19.29	24.18	26.78	31.11	40.49	48.49	74.03	109.99	1046126
2014	3.48	11.49	15.32	16.91	19.99	24.68	27.17	31.49	40.14	46.95	68.39	109.98	1022063
2016	3.71	14.42	17.69	19.37	23.01	28.65	31.51	36.91	47.19	55.06	75.00	109.98	1128030
2018	3.48	13.77	17.19	18.81	22.23	27.75	30.63	35.76	45.93	53.94	74.99	109.95	1249239
total	3.48	12.28	15.75	17.41	20.93	26.21	28.98	33.75	43.83	51.87	73.26	110.00	11386093

1.A.2 Details Related to the Main Analysis

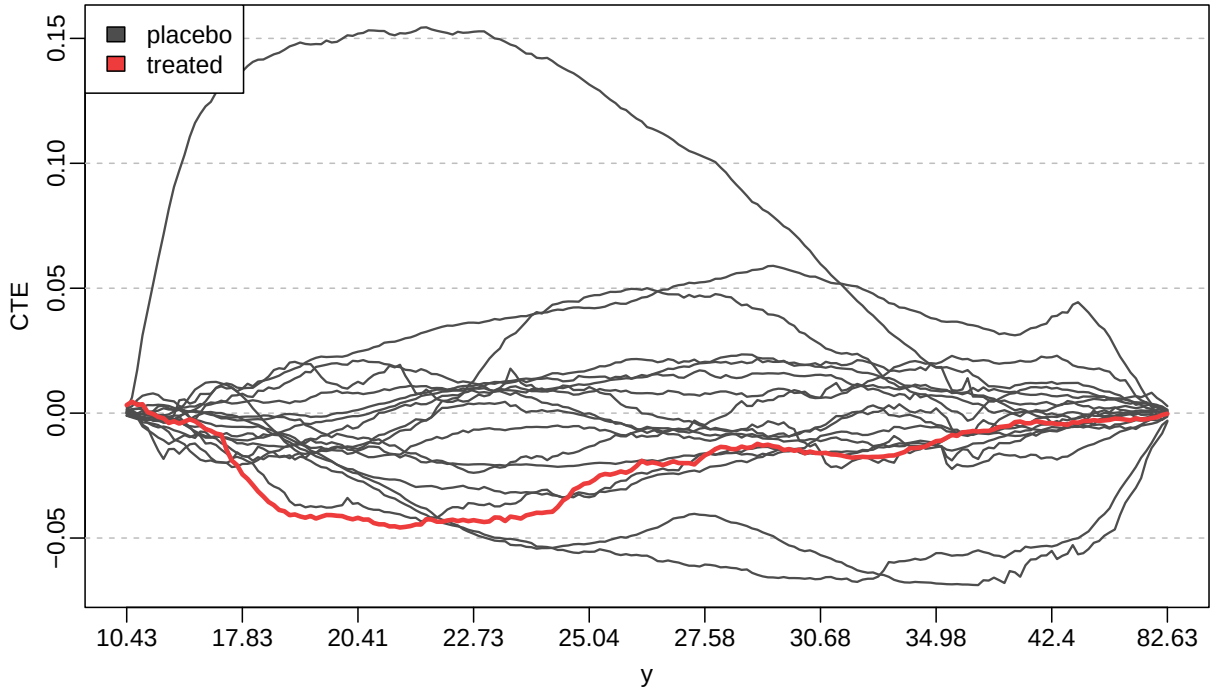
Figure 1.6 captures the result from the placebo permutation test outlined in section 1.2. The gray lines in figure 1.6a display the QTE for the post-treatment period for units that were actually not treated, i.e., the placebo QTE. Some placebo units exhibit differences far away from zero. However, this is not because we expect there to be an effect but simply because these units don't have a fitting synthetic counterpart. The two cantons with very high QTE are Basel and Zug, which have exceptionally high GDP per capita compared to all other cantons. Consequently, the two cantons exhibit higher wages compared to other control units, and, combined with the restrictions on the weights, the constructed synthetic units are unable to replicate the two placebo-treated units well. The canton with a very low QTE is Ticino, which is a result of a similar reason as before. Ticino exhibits very low wages compared to other control units and, hence, cannot be replicated well by its synthetic counterpart. Compared to units with more reasonable placebo effects, the positive QTE in the lower part of the distribution for Neuchâtel, displayed by the red line, is just above other placebo effects. Also, compared to negative placebo effects in the lower part, the QTE for Neuchâtel seems slightly larger in absolute size. This is not the case for the smaller effect in the upper part of the distribution. Hence, there is some evidence that the effect on lower wages is significant, but it is rather weak. Importantly, figure 1.6a does not provide any information regarding the pre-treatment fit of the placebo runs. The lines in figure 1.6b display the difference between CDF values across different points of y. The interpretation of the results is the same as for figure 1.6a.

Figure 1.6: Placebo Permutation Results

(a) Quantile Treatment Effects



(b) CDF Treatment Effects



Notes: The figures display the QTE and the percentile treatment effect for different cantons in 2018. The colored line displays the effect for the treated unit, Neuchâtel, in the post-treatment year. Lines in gray display the effect of the post-treatment year if we apply the synthetic control method to other units.

1.A.3 Minimum Wage Impact on Various Outcomes

In addition to the hourly wages, this section analyses the impact of the minimum wage introduction in Neuchâtel on the following outcomes: Work hours, unemployment, exiting and entering behavior of firms, number of workers from abroad, and registered available job positions. For work hours, the distributional synthetic control method, as suggested in the paper, is applied. For the other variables, the conventional synthetic control method is used. The information on work hours is available in the ESS used for the main analysis. Information on the remaining outcomes is publicly accessible on the website of the Federal Statistical Office or on Amstat.ch, a website provided by the State Secretariat of Economic Affairs.

Work Hours

The variable work hours is clearly discrete since most people are employed in a contract with fixed work hours. The standard working time in Switzerland is around 40 to 42 hours per week, and a large share of individuals are employed full-time. As a result, the variable is not continuous with sizable mass points, especially around the standardized work time for full-time employment. Nevertheless, with the proposed method, it is straightforward to implement a distributional synthetic control analysis for work hours.

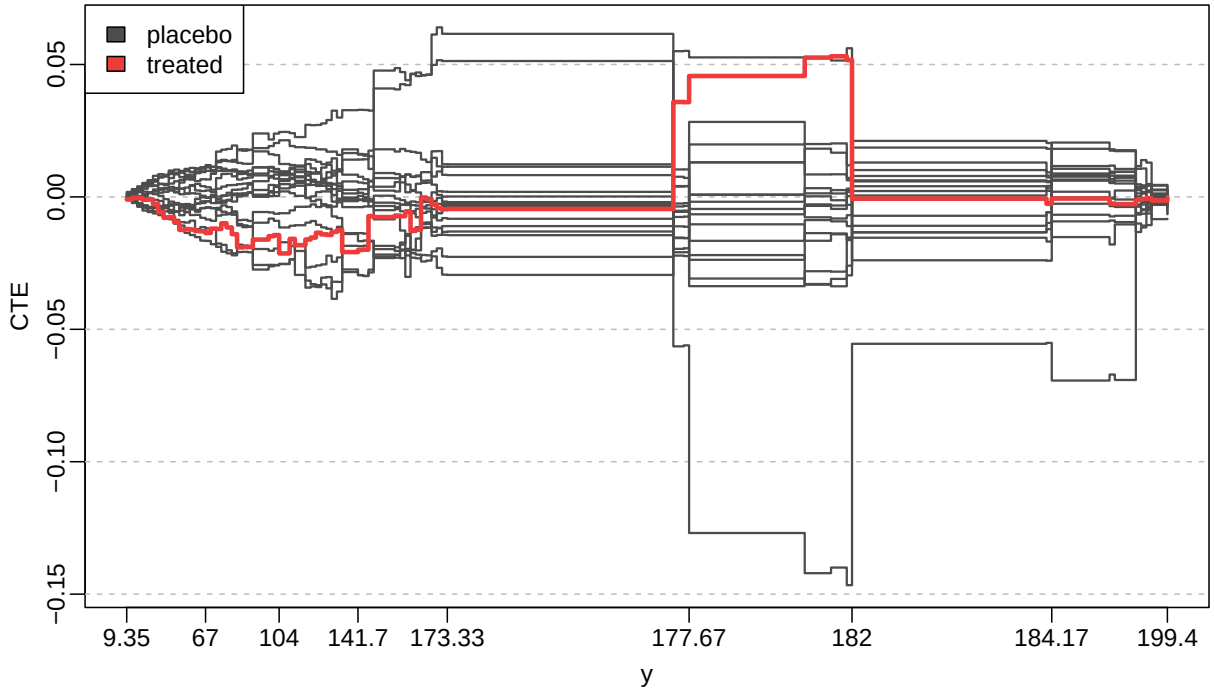
Figure 1.7 captures the results from the permutation analysis. In subfigure 1.7a, the red line displays the estimated effect of the minimum wage introduction on work hours. Results suggest that there could be a small shift in the number of part-time workers, visible by a decrease in probabilities in the lower part of the distribution, towards a higher number of full-time workers, visible through the higher probabilities in the upper part of the distribution at the mass point. It could be argued that with the increase in the minimum wage, there might be an incentive for individuals in lower-paid jobs who profited from the policy intervention to increase their workforce participation. However, compared to other placebo runs, the effect estimated for Neuchâtel does not seem to be among the most extreme cases. The corresponding p-values displayed in subfigure 1.7b are also further away even from 10% for the most part of the distribution. There are two exceptions, but given the overall picture, they seem more likely to result from repeated testing and not from a significant effect. Hence, results point more towards no effect on work hours due to the minimum wage introduction.

Unemployment Rates

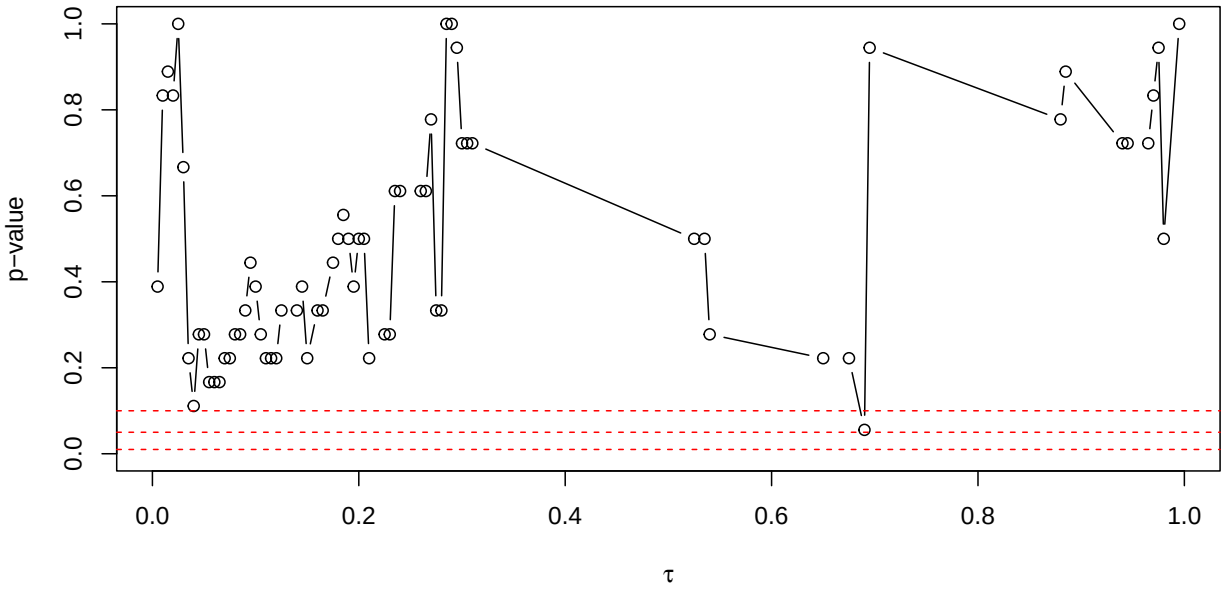
The variable captures the number of registered unemployed for each canton. Data is available monthly since 1993, and all periods are used up to the year 2020. I refrain from including the years from 2020 and onwards due to the Covid-19 global pandemic. This holds true for unemployment rates as well as all other subsequent variables. The data further allows for

Figure 1.7: Placebo Permutation Results for Work Hours

(a) Quantile Treatment Effects



(b) P-Values



Notes: The results are constructed by applying the proposed distributional synthetic control method on work hours. In figure 1.7a, the red line displays the difference between CDF values of work hours of Neuchâtel, the treated unit, and the synthetic control unit. In gray, the same differences for each of the placebo runs are displayed. In figure 1.7b, the p-values related to the placebo runs are displayed.

distinguishing between youth unemployment, including individuals between the ages of 15 and 24, and regular unemployment, including individuals with an age of 25 or older. For the analysis, the monthly percentage changes are computed, and an average for each year is constructed. The last pre-treatment period is then $T_0 = 2016$.

The results of the conventional synthetic control method are displayed in subfigures 1.8a and 1.8b. The synthetic control unit does manage to replicate the treated unit fairly well in the pre-treatment period. The positive difference in the post-treatment period observed for both variables seems not to be larger than observed differences in the pre-treatment periods; hence, a treatment effect is not apparent.

Exiting and Entering Behavior of Firms

The variable captures the number of active, closing, and newly founded firms for each canton. Data is available yearly since 2013, and all periods are used up to the year 2020. For the analysis, the fraction of closing firms relative to the number of active firms for each canton is constructed. The same fraction is constructed for the newly founded firms. The last pre-treatment period is set to $T_0 = 2016$.

The resulting findings when applying the conventional synthetic control method are displayed in the subfigures 1.8c and 1.8d. Note that we only have four pre-treatment periods; hence, interpretation should be conducted with care. For the fraction of closing firms, the synthetic control method is able to closely replicate Neuchâtel in the pre-treatment period. However, there seems to be no clear treatment effect in the post-treatment periods. For the fraction of new firms, the method fails to provide a close match and, hence, cannot be used to get reliable estimates.

Workers from Abroad

The variable captures the number of workers with foreign residences coming from abroad to work within the canton. These individuals need to have permanent residence in a country neighboring Switzerland, within a closer region to the canton in which they are working. Furthermore, they typically return to their residence every day and are obliged to do so by law at least once per week. Data is available quarterly since 1999, and all periods are used up to the year 2020. For the analysis, yearly sums are computed to get rid of seasonalities. The last pre-treatment period is again $T_0 = 2016$.

The results of the conventional synthetic control method are displayed in subfigure 1.8e. Throughout the pre-treatment period, the synthetic control unit replicates Neuchâtel very well. Additionally, we see a clear negative difference for the post-treatment period. Strikingly, this difference already appears before T_0, namely around 2014. There are two possible explanations for this observation. First the minimum wage law was first presented in May 2014, and the

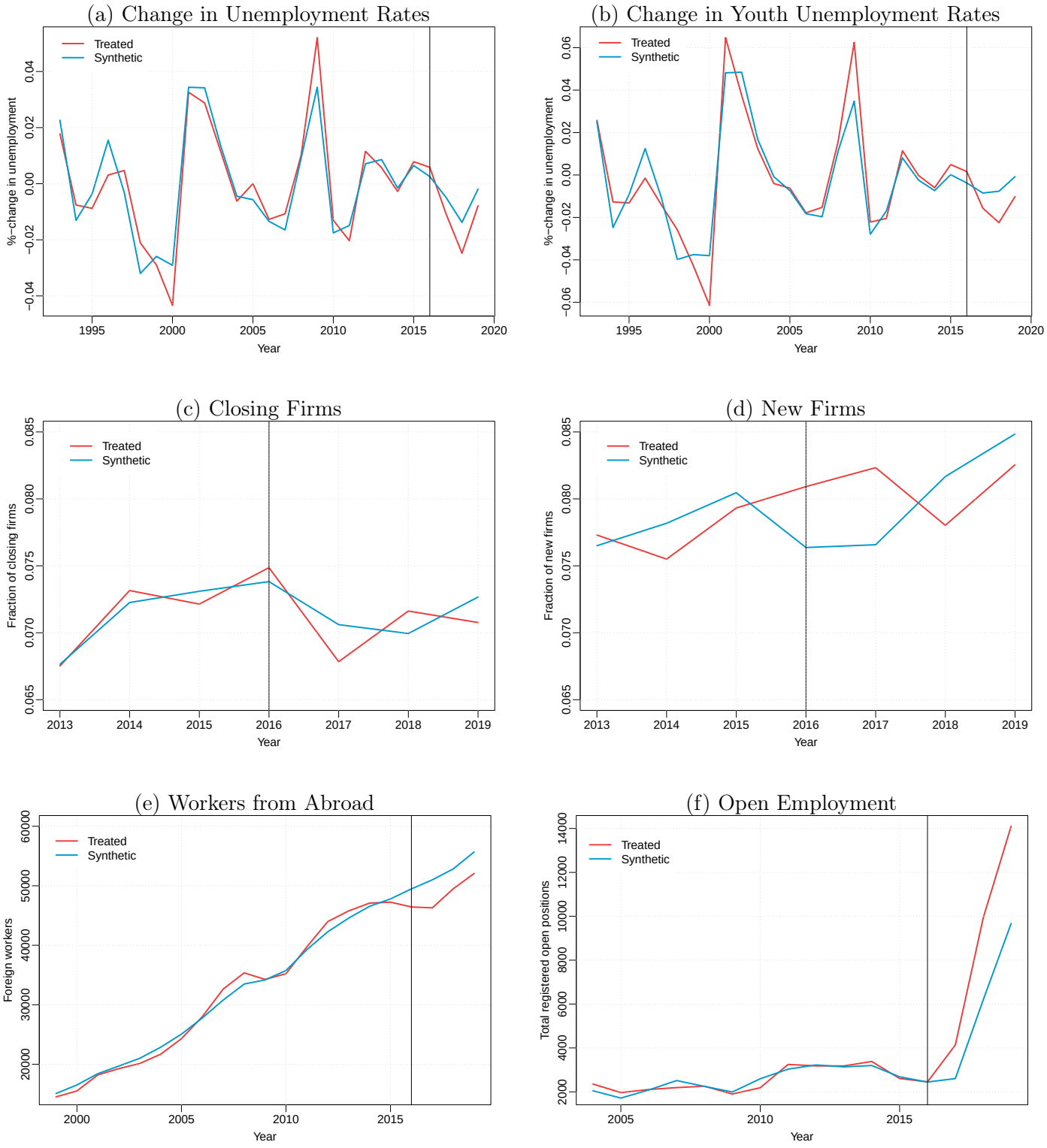
enforcement was set to January 2015. Hence, the result could be explained by anticipatory reactions from either employers or workers abroad. Second, in February 2014, the initiative "against mass immigration" (Eidgenössische Volksinitiative "Gegen Masseneinwanderung" in German) was accepted, which, among other things, aimed to reduce the number of workers from abroad. Even though it takes time for new laws to come into force (until July 2018 for the specific case), employers and workers could have again reacted earlier. Conclusively, even though there is a negative impact on the number of workers from abroad, it is hard to disentangle how much of the effect can be attributed to the new minimum wage policy.

Registered Available Job Positions

The variable measures the number of registered open positions at the public job placement centers. Regional assignment to specific cantons is done depending on the location of the firm with the open position rather than the center. Data is available monthly since 2004, and all periods are used up to the year 2020. Furthermore, information on full- and part-time employment is available separately. For the analysis, full- and part-time employment is combined and summed up over the year to smooth the series. The last pre-treatment period is $T_0 = 2016$.

The results after applying the conventional synthetic control method are displayed in subfigure 1.8f. Again, the synthetic control unit replicated the treated unit well in the pre-treatment periods. After the intervention, there is a very strong increase in the number of registered open positions for both the treated and the synthetic control unit, whereas, for the treated unit, the effect is stronger. Especially for the year 2017, Neuchâtel already exhibits a strong increase, while the synthetic control unit remains at the same level. It could be that the positive effect is driven to some extent by the new minimum wage law. However, similar to the number of workers from abroad, interpreting these results must account for the initiative "against mass immigration." The federal law implementing the initiative was designed to prioritize national workers, which was a softer intervention than demanded by the initiating party. The new law required employers to register open job positions in certain fields with higher unemployment rates. The Swiss parliament agreed upon the new law in December 2016. The final law came into force in July 2018. Thus, the large increase starting in the post-treatment periods for both the treated and synthetic control unit is almost surely due to the new law. It is unclear to which degree the increase for Neuchâtel for the year 2017 is driven by the minimum wage policy or the fact that employers in Neuchâtel reacted earlier to the new law compared to other cantons in Switzerland.

Figure 1.8: Synthetic Control Estimation Results for Various Outcomes



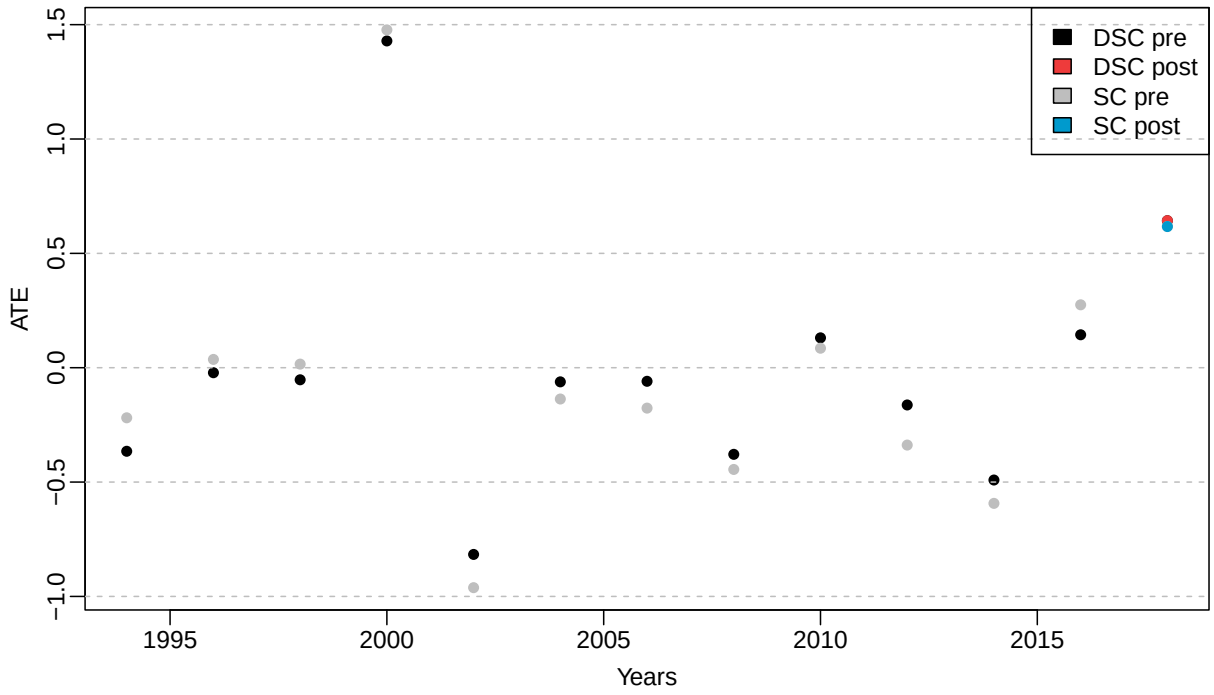
Notes: Each figure displays the results from applying the conventional synthetic control method to different outcome variables. The red line displays the observed values of the treated unit, and the blue line displays the estimated values of the synthetic control unit. The black vertical line marks the last pre-treatment period, $T_0 = 2016$.

1.A.4 Comparison to the Standard Synthetic Control Method

In this section, the suggested method is compared to the standard Synthetic Control Method by comparing the average treatment effect (ATE) and the control units used to build the synthetic control unit. Note that for this section, the term ATE refers to the difference between the treated and the synthetic unit unrelated, regardless of whether we are examining the pre- or post-treatment periods. Overall, the two methods yield similar results, which speaks in favor of the proposed method.

To derive the ATE from the suggested method, the observed empirical CDF of the treated unit and the estimated CDF of the synthetic control unit are used. For each year, the average income is approximated by treating income as a discrete and truncated variable, which takes values lying in the middle of the available CDF grid values. Figure 1.9 comprises the estimated ATE for the post- and pre-treatment periods of both methods. In general, the two methods exhibit very similar results. For each pre-treatment year the average income of the treated and the synthetic control unit are, with the exception of some points, close to each other, that is, the ATE is close to 0. Further, both methods suggest an ATE of around 0.6 CHF per hour following the minimum wage introduction.

Figure 1.9: Comparison of the ATE



Notes: Each dot represents the difference between the treated and the synthetic control unit. DSC denotes the suggested method and SC denotes the classic Synthetic control methods.

To compare the control units used in the synthetic control unit for each method, a single weight for each control unit from the suggested method is created by averaging the weights across the distribution. In table 1.8, the weights of all control units with a nonzero weight in one of two methods are displayed. For both methods, almost the same control units are used. The weights from the proposed method are less scarce, with smaller weights being put on additional units.

Table 1.8: Weights

	ZH	SZ	FR	SO	AG	TI	VD	VS
DSC	0.01	0.01	0.34	0.24	0.20	0.13	0.02	0.08
SC	0.00	0.00	0.21	0.31	0.31	0.17	0.00	0.00

1.A.5 Inclusion of an Intercept

In this section, the proposed method is extended to include an intercept to replicate the treated unit following the discussion in Doudchenko and Imbens (2016). The advantage of including an intercept is that the treated unit can be closely matched, even if the CDF of the treated unit does not lie within the convex hull of the CDF values of the control units at a specific point of the distribution.

Overall, the results point towards the same direction as the main analysis. However, as visible in subfigure 1.11a, the estimated effect remains relatively constant across the distribution. Additionally, the estimated effect is less exceptional compared to other placebo runs. Subfigure 1.11b exhibits p-values, which are mostly over 20%, with the lowest values reaching 16.7%.

Consulting subfigure 1.11c together with figure 1.10 provides some further insights. Note that even though Neuchâtel belongs to one of the poorer cantons, it is not the poorest. Hence, it is possible to find weights for the synthetic control unit, which span the unit simplex and replicate the treated unit. However, the estimated intercept is clearly nonzero, reaching values higher than 8%. The estimated weights indicate that, in general, there is a dominant control unit with a high weight of around $0.7 - 0.8$. For the bottom of the distribution, the dominant control unit is Fribourg, followed by the cantons Solothurn and Aargau. Hence, by including an intercept, the method seems to pick a control unit that behaves similarly to the treated unit at a specific point of the distribution and adds the difference between this control unit and the treated unit as a coefficient. For example, Fribourg is a canton with low incomes, i.e., very high CDF values. To match the treated unit, Neuchâtel, at the bottom of the distribution, a lot of weight is assigned to Fribourg, and a negative coefficient is added.

Results suggest that extending the proposed method might be critical and interesting. To assess whether or not this behavior is desired to estimate effects across the distribution, a more

in-depth analysis would be necessary, but this is beyond the scope of this paper.

Figure 1.10: Estimated Intercept

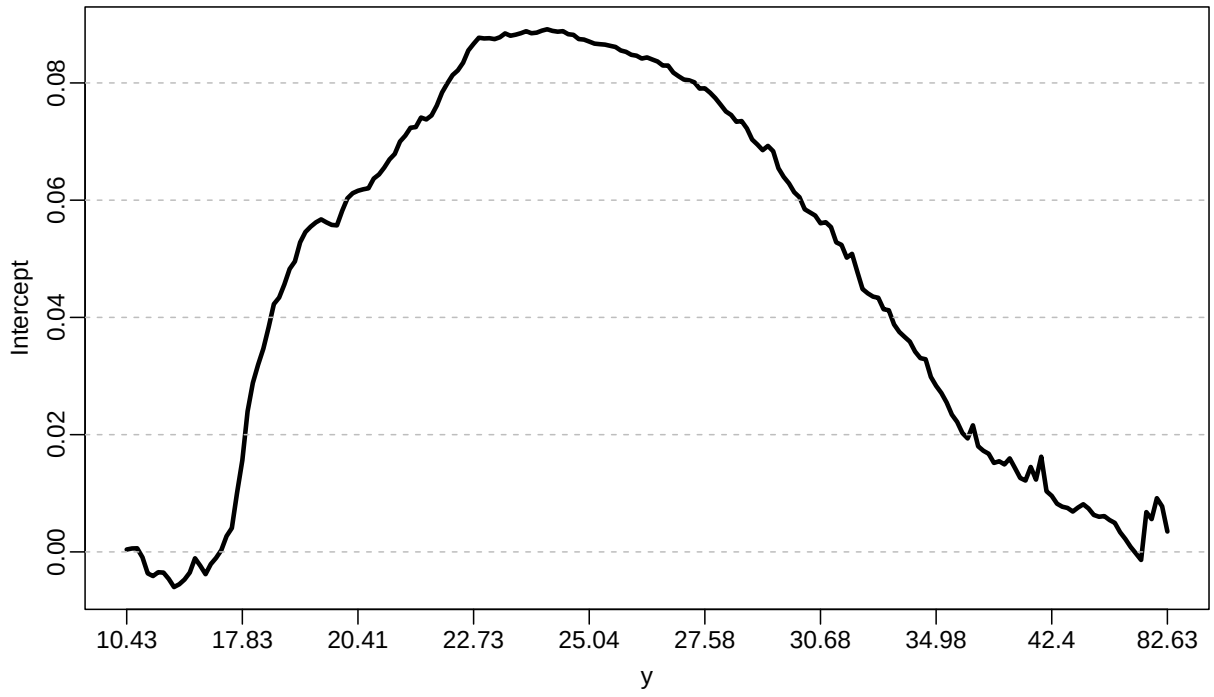
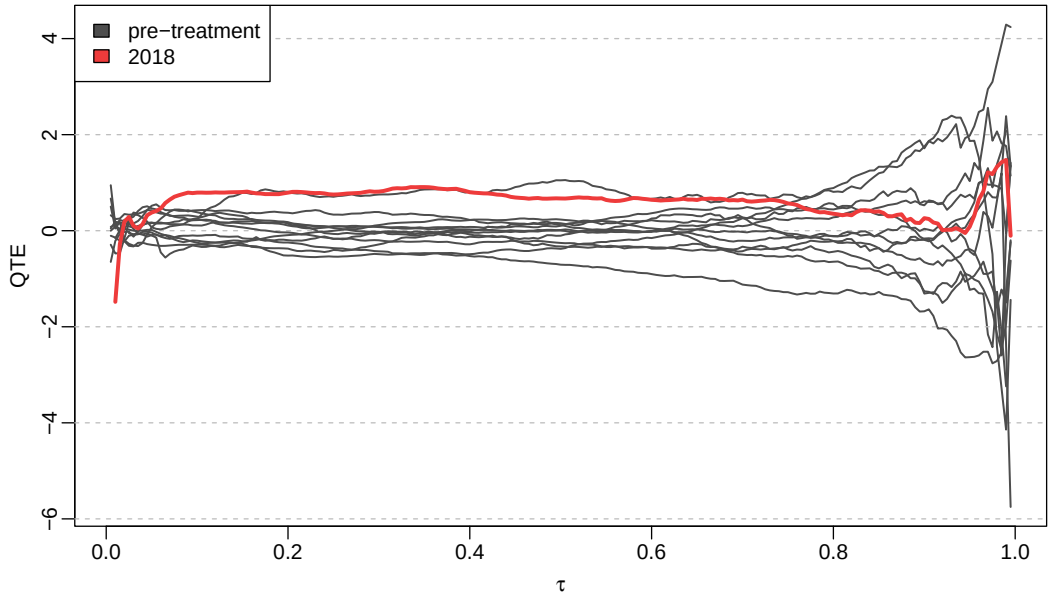
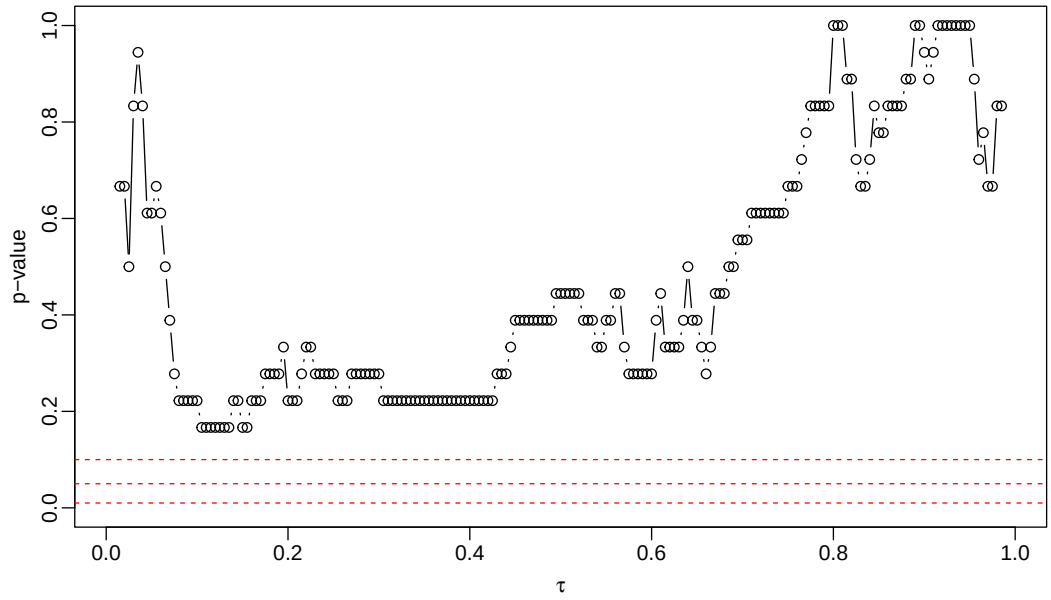


Figure 1.11: Results Including an Intercept

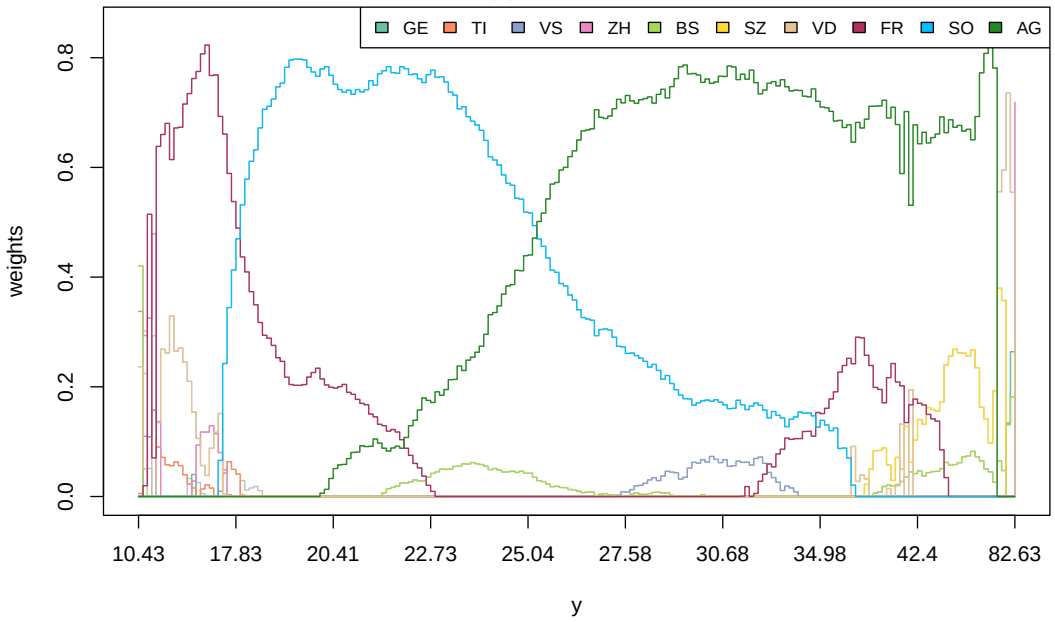
(a) QTE Compared to Pre-Treatment Periods



(b) P-Values



(c) Weights

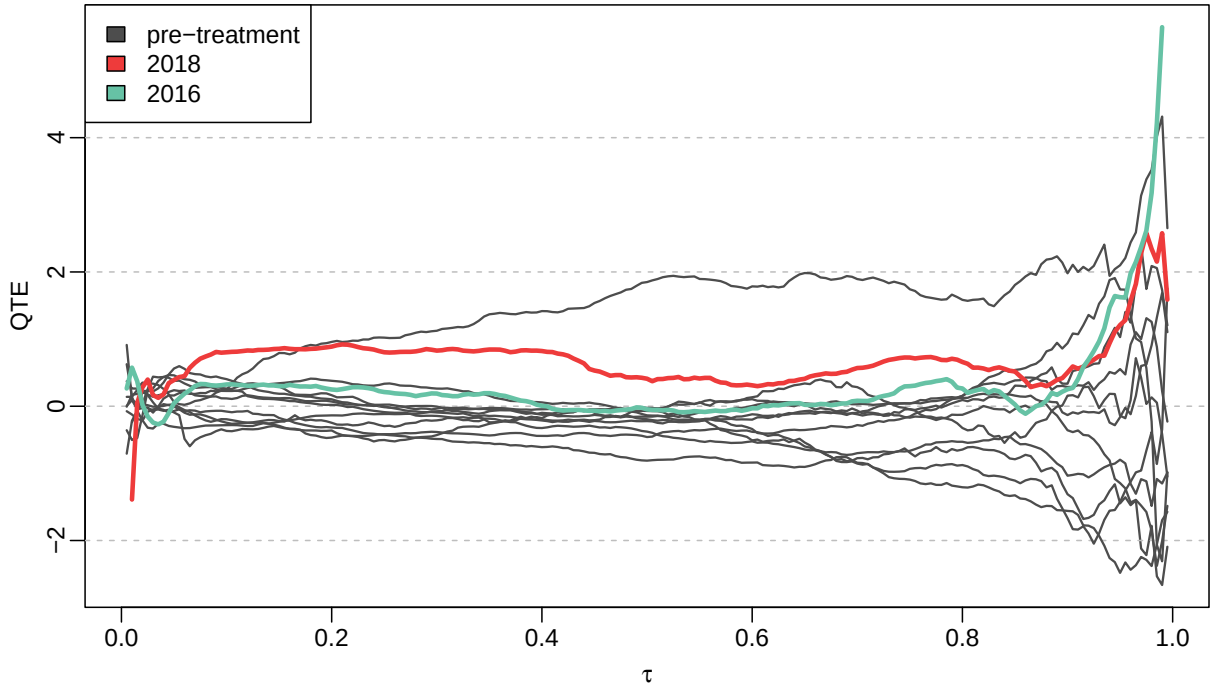


1.B Appendix II - Robustness Checks

Anticipation Effects

As described in section 1.3.1, the level of the minimum wage was communicated by the cantonal government in 2014, hence some firms could have already adopted the new policy in 2016 even if it was not into force yet. Therefore, the analysis is repeated with T_0 set to 2014. Results displayed in figure 1.12 are reassuring that no such early adoption took place. The green line showing the QTE for the year 2016 is close to zero except for the tails, where volatility in general is higher. Additionally, the positive effect for the year 2018 remains unchanged.

Figure 1.12: Magnitude of the QTE Compared to Pre-Treatment Periods



Notes: The colored line displays the QTE for the post-treatment year. Lines in gray display the QTE for the pre-treatment years.

Including Young Workers

Employees at a young age have sometimes not yet fully entered the labor market. The transition from education to employment is a lengthy process, especially for individuals with a tertiary education. Therefore, individuals below the age of 25 were excluded from the main analysis. As an additional robustness check, the analysis is repeated, including individuals aged between 18 and 25. Overall, the results displayed in figure 1.13 for the QTE, the significance across the distribution, and the weights of the synthetic control unit do not change heavily. However, the

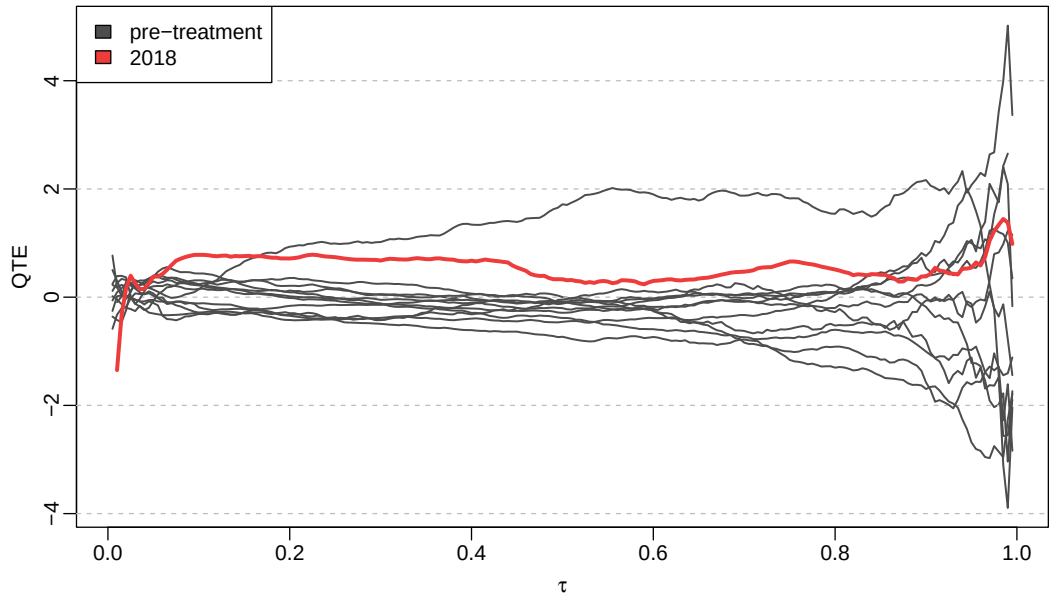
p-values never drop 10% across the whole distribution but stay slightly above it. The reason is likely related to how the unit of interest is defined. We can expect that some individuals in the lower part of the distribution employed in Neuchâtel will also commonly work within Neuchâtel and hence are directly affected by the minimum wage. By including individuals under the age of 25, the sample contains more individuals who are not yet fully integrated into the labor market and who are likely employed in settings excluded from the minimum wage law. Hence, it seems that for this sample, not enough people employed in Neuchâtel are impacted by the minimum wage law. Therefore, the positive effect at the bottom is still apparent, but the number of people affected is not large enough for the effect to be significant.

Exclude Early Pre-Treatment Years

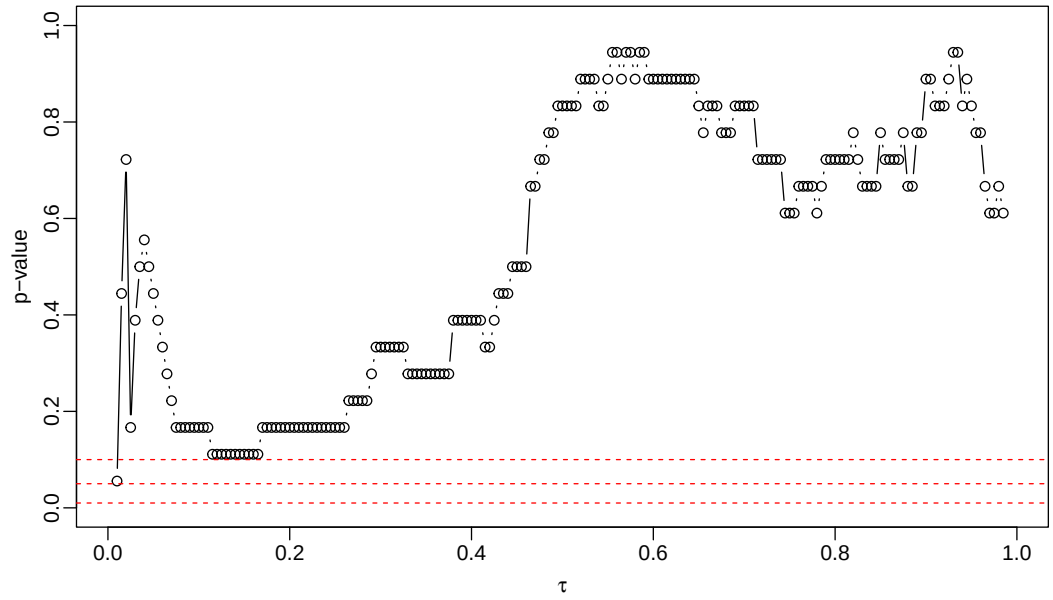
For the years from 1994 to 2000, the waves are much smaller compared to the years after. Therefore, an additional robustness check is employed to make sure that the results are not affected by structural changes in the underlying cross-sectional data. Overall, results in figure 1.14 point towards the same findings as in the main analysis. Nevertheless, changes are stronger compared to the other robustness checks. The QTE exhibits a sharp drop after the 0.5-quantile, and the p-values change correspondingly. This result is driven by an increase in the weights for Ticino around the median, followed by a sharp drop to zero, slightly above the median. In general, the weights seem to be more volatile in the upper half of the distribution, suggesting that reducing the number of pre-treatment periods also weakens the performance of the synthetic control unit.

Figure 1.13: Results Including Employees Below the Age of 25

(a) QTE Compared to Pre-Treatment Periods



(b) P-Values



(c) Weights

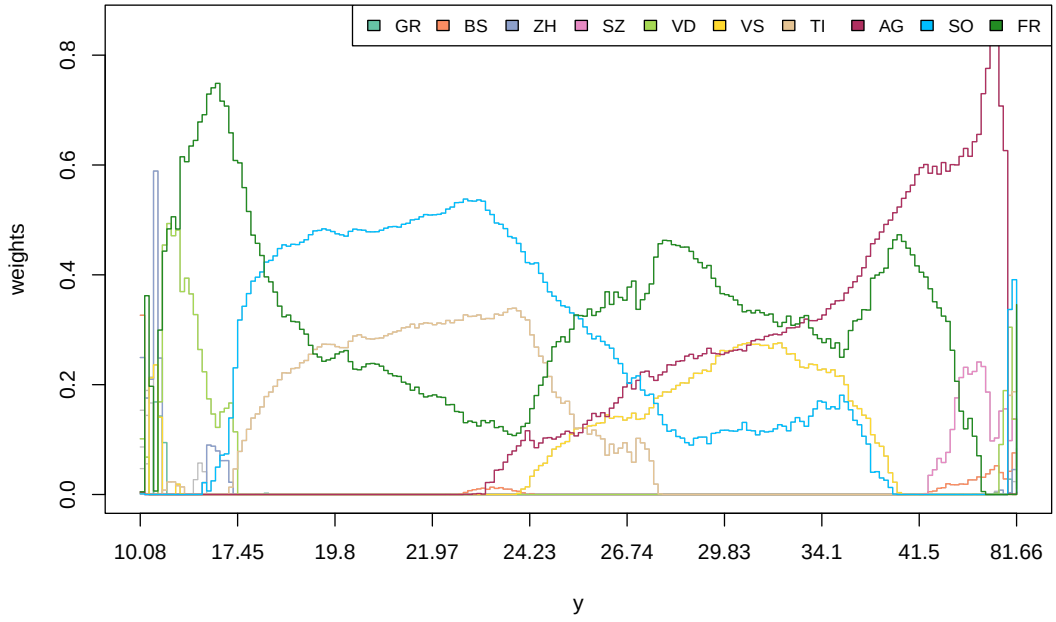
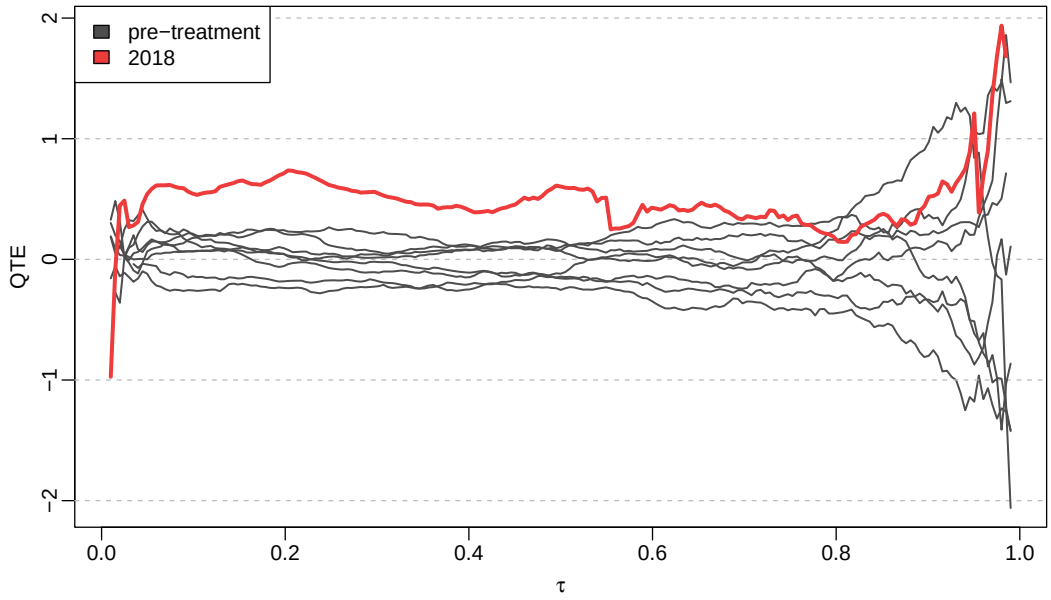
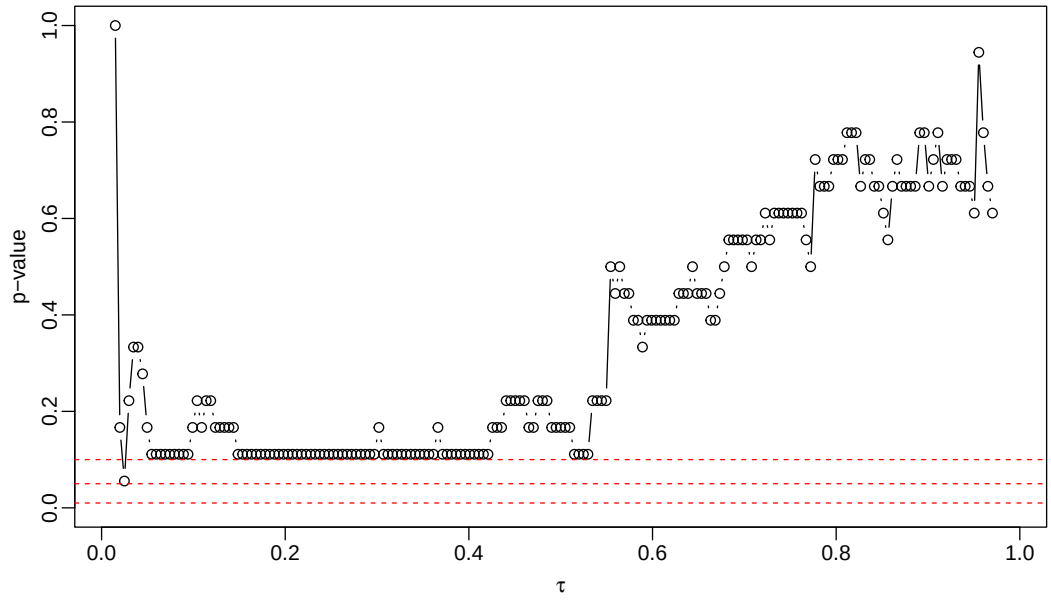


Figure 1.14: Results using a Reduced Pre-Treatment Period from 2002-2016

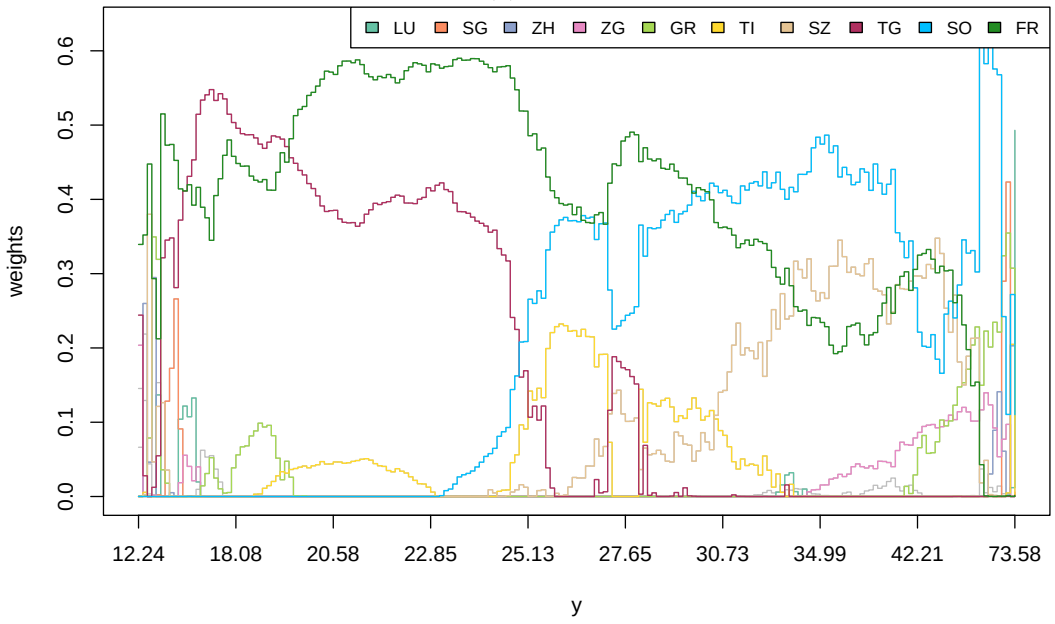
(a) QTE Compared to Pre-Treatment Periods



(b) P-Values



(c) Weights



Smoothed Weights

The smoothed weights $\tilde{w}_j(y)$ are a weighted average of weights positioned before and after y, mathematically: $\tilde{w}_j(y) = \sum_{i=-k}^{k} m_i w_j(y+i)$, where m_i are the weights in position i. In other words, $\tilde{w}_j(y)$ is a rolling average. In the underlying case, I choose k such that around 5% of the number of weights of a unit are used. Furthermore, weights are chosen to create a triangular shape and sum up to one. Using this specification and the fact that the total number of weights used for the rolling average is always odd (which results directly from the rolling average covering weights at points $y-k$ up to $y+k$), one obtains $s = [(k+1)/2]^{-2}$ as slope coefficient. It follows that:

$$m_i = \begin{cases} (k+1+i)s & \text{for } i \leq 0 \\ (k+1-i)s & \text{for } i > 0 \end{cases}$$

Following this basic approach leads to the loss of observations in the tail, namely the lowest and highest 2.5%. However, as seen in section 1.4 estimated weights in the tails are quite volatile. To deal with this case, it might be advisable to assume some parametric form, which would be an interesting question for future research.

Figure 1.15 displays the result if we smooth the weights following the approach outlined above. Figure 1.16 displays the QTE when using smoothed weights. Overall, results remain very similar to the ones shown in section 1.4.

Figure 1.15: Smoothed Weights Across the Distribution

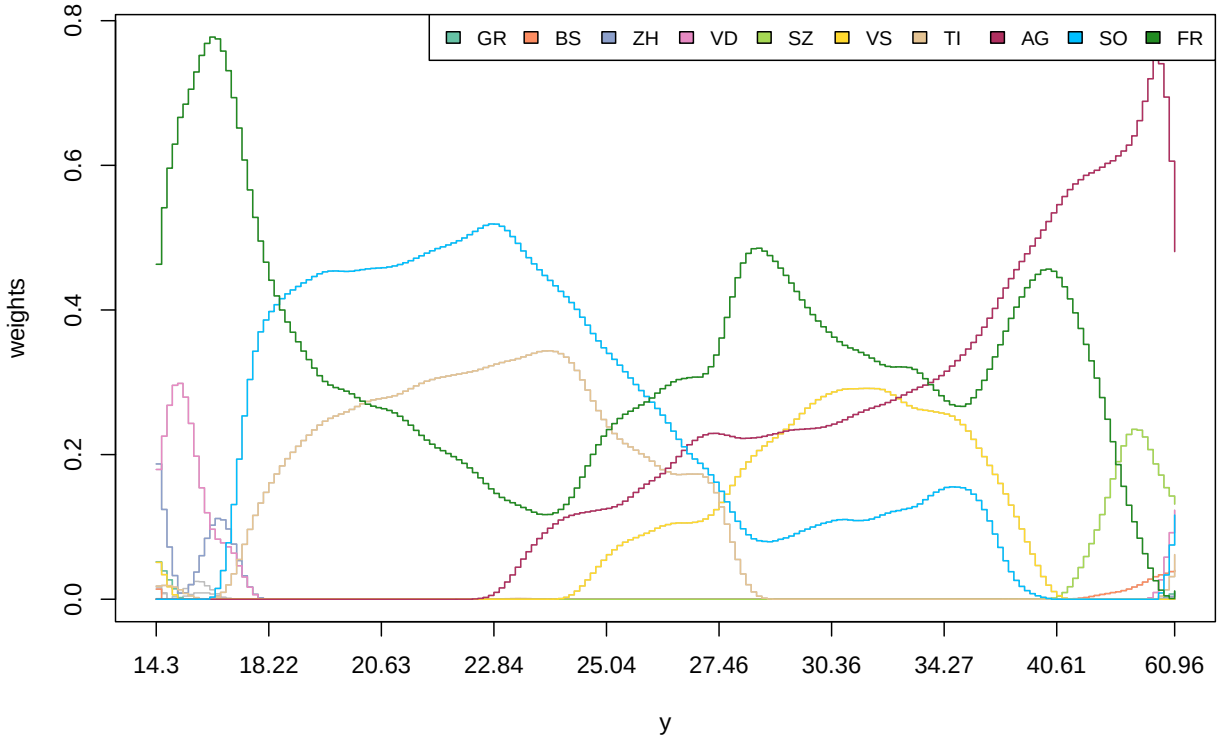
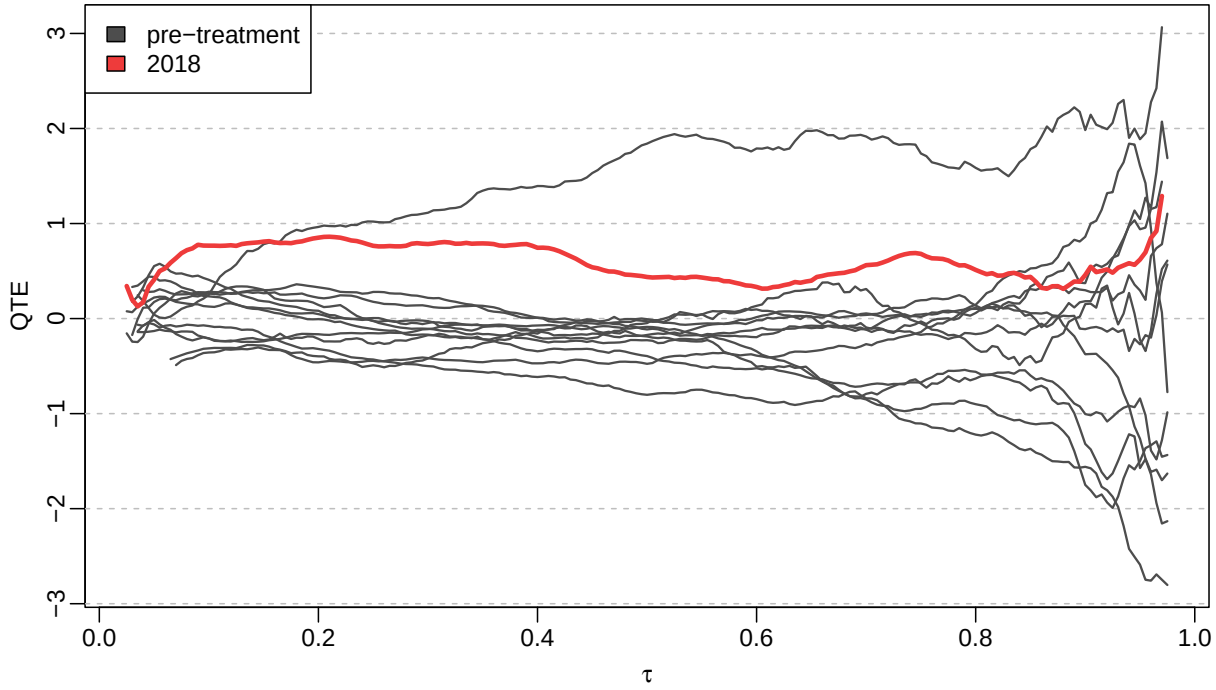


Figure 1.16: QTE Compared to Pre-Treatment Periods



1.C Appendix III - Remarks on Mixture Weights

The size in which mixture weights can change across y is linked to the properties of the underlying distributions. Taking derivatives with respect to y, equation 1.2.6 can be written as:

$$\sum_{d=1}^{D} \gamma_d'(y)F_d(y) + \gamma_d(y)F_d'(y) \geq 0 \tag{1.C.1}$$

where subscripts j and t are dropped for visibility. Since the weights are restricted to sum up to one, any change in mixture weights imposes $\gamma_d'(y) < 0$ for at least one distribution. Let the distribution for which weights are decreased be denoted as d^*, and let d^{-*} be any distribution that has positive weights at y but is not d^*. Further, we can think of three cases. First, if the decrease occurs for a distribution with lower CDF values, i.e., $F_{d^*}(y) < F_{d^{-*}}(y)$ for all d^{-*}, then the inequality 1.C.1 is always satisfied. The decrease in weights must be matched by other weights with higher CDF values. As a result, the first term of inequality 1.C.1 must be positive, and thus, the weights can change strongly. However, note that a strong decrease implies a strong increase in weights of distribution(s) with higher CDF values. As a result, for values of y larger than the current point, the weights are restricted from reverting to d^*. The reason is the following second case. If the decrease occurs for a distribution with higher CDF values, i.e., $F_{d^*}(y) > F_{d^{-*}}(y)$ for all d^{-*}, then the first term of the above inequality 1.C.1 becomes negative. To avoid a violation of the inequality above, the second term has to compensate. The higher the density, i.e., the steeper the CDFs of the distributions with positive weights, the more likely the negative term can be compensated. Hence, if the decrease happens for the distribution with higher CDF values, then the change in weights becomes more strongly bounded as the CDFs become flatter. Importantly, for y values in which the distributions are equal, that is $F_{d^*}(y) = F_{d^{-*}}(y)$, the first term of the above inequality is always zero, and hence weights can change arbitrarily. It follows that the weights, $\gamma_d(y)$, are to some degree bounded by the underlying properties of the distributions that have positive weights at y. There exist cases for which stronger changes in weights could occur. However, if the underlying distributions are not identical, we should not expect to observe these strong changes across all values of y.

Chapter 2

Intergenerational Income Mobility: A Copula Regression Approach

Abstract

We propose a new estimator of the conditional distribution of multivariate outcomes given covariates. In the first step, the univariate conditional distributions of the outcomes are estimated via distribution regression. In the second step, we estimate a conditional copula of the outcomes, imposing a copula parameter that is local in the value of the outcome. Without covariates, the estimator reduces to the empirical distribution function. We apply the estimator to study intergenerational income mobility in Switzerland and the U.S. We estimate the joint cumulative distribution of children's and parents' income, controlling for different explanatory variables. Derivation of specific mobility measures from the joint distribution is straightforward, and taking covariates into account during the estimation of the joint distribution enables us to gain deeper insights into drivers of intergenerational income mobility. Further, this innovative approach allows us to decompose structural from compositional differences. Results focus on mobility differences between sons and daughters. Specifically, we find that a higher father's income share for sons is related to higher upward mobility. However, this relationship does not hold for daughters. We further analyze differences in mobility outcomes between sons and daughters and conduct a decomposition analysis. The decomposition shows that average hours worked are crucial in explaining the differences between sons and daughters. Nevertheless, a large part of these observed differences remains unexplained.

Acknowledgment: Joint work with Jonas Meier. We thank Costanza Naguib, Blaise Melly, Bo Honoré, and Matias Cortes for helpful comments and suggestions. Further, we are grateful for helpful comments by seminar and conference participants at the Young Swiss Economists Meeting 2023, the IAAE 2023, and the Brown Bag Seminar at the Department of Economics at the University of Bern.

2.1 Introduction

This paper introduces a new estimator for the analysis of multivariate counterfactual distributions, providing deeper insights into intergenerational income mobility. Income mobility is an essential economic concept since it is strictly connected with the level of long-term inequality (see, e.g., Arellano et al., 2017, Bonhomme and Robin, 2009) and is a crucial issue in both academic research and the current political debate. Our results have clear implications for policymaking: If the degree of association between parent and child income is low in the bottom part of the income distribution, then individuals can climb the income ladder from one generation to the next. Hence, income inequality would be a transitory phenomenon, which reduces itself over time and disappears eventually. Other recent developments in intergenerational mobility studies include the analysis of the intergenerational transmission of wealth (Adermon et al., 2018), consumption (Waldkirch et al., 2004), and occupational choice (Boar and Lashkari, 2021). Specific aspects of intergenerational income mobility, such as differences between family structures (Björklund and Chadwick, 2003) or father-daughter degree of income persistence (Chadwick and Solon, 2002) have also been put under scrutiny.

Intergenerational income mobility is affected through different channels. The early model by Becker and Tomes (1979) and Becker and Tomes (1986) suggests that labor market outcomes are directly linked to human capital, e.g., knowledge, skills, and attitudes, valued by the labor market. As past research has shown, the human capital a child brings to the labor market results from a complex process involving endowments and investments into human capital. A typical example of an investment is education provided by the parents or the government.[1] Past research has emphasized various factors to be of importance. For example, Björklund, Roine, et al. (2012) and Blanden et al. (2007) attributed a significant role in determining the degree of father-son income persistence to education. Bhattacharya and Mazumder (2011) as well as Chetty, Hendren, Jones, et al. (2020) find that race determines intergenerational income transmission and Abramitzky et al. (2021) show that children of immigrants have higher rates of upward mobility. Carneiro et al. (2021) find that investments when children are between 6 and 11 years old are relatively less important compared to investments made during the early and later stages of childhood. Additionally, studies like Chetty, Hendren, Kline, et al. (2014), Chetty and Hendren (2018a), Chetty and Hendren (2018b), and Corak (2020) find significant regional differences within a country underscoring the influence of neighborhood effects. Our approach allows us to explicitly account for such factors when estimating the joint distribution of children's and parent's income. Differences in mobility across characteristics were traditionally addressed mainly by estimating different mobility measures for different groups, e.g., high- or low-educated individuals. However, such an approach either requires a very large dataset or

[1] For a more in-depth discussion, see appendix 2.C.

limits the analysis to a smaller set of variables. Additionally, categorizing a continuous variable typically involves defining arbitrary thresholds to group the data. Our approach simplifies controlling for multiple variables, even when they are continuous.

The proposed method incorporates covariates in a semi-parametric way during the estimation of the joint income distribution. From there, multiple mobility measures are computable. As a result, different conditional measures can be computed even with a smaller data set. Therefore, the crucial first step in our approach relies on estimating the joint income distribution. There is some related research on this topic. For example, Bhattacharya and Mazumder (2011) analyze differences in intergenerational income transmission for race in the U.S. However, they adopt a fully nonparametric model that is hence prone to the well-known problems of a slow rate of convergence, the curse of dimensionality, and bandwidth choice, to name only a few. Additionally, they consider a single parameter, whereas we provide a method that works with all functionals of the joint distribution of parent's and children's income. Similar to our paper, Richey and Rosburg (2018) suggest a decomposition of the transition matrix. They use the fact that transition matrices are built up by the children's distribution conditional on the parents' rank and a number of covariates. However, for their estimation, they rely on parametric copulas to estimate the distributional structure of the covariates. With our method, we avoid these strong parametric assumptions. Instead, we employ a copula with a copula parameter that is modeled via a parametric link function (e.g., a logistic function) and local in the value of the outcome. This allows for a flexible specification of the multivariate distribution while ensuring efficient estimation.

In our paper, we will analyze Swiss data from the Economic well-being of the working- and retirement-age population (WiSiER) as well as U.S. data from the Panel Study of Income Dynamics (PSID). Given the limited number of studies on intergenerational mobility in Switzerland and the comparatively richer dataset available compared to the PSID, this analysis will primarily focus on Switzerland. Previous research on intergenerational mobility in Switzerland examined the mobility of education. Bauer and Riphahn (2006) use data from the Swiss population census to study intergenerational mobility of education. Jann and Seiler (2014) use a combination of different surveys to study intergenerational mobility of education and social class. Favre et al. (2018) analyze intergenerational mobility in occupations for the City of Zurich in the 19th century. Recently, with the new WiSiER database available, other aspects of mobility have been analyzed. Chuard and Grassi (2020) are the first to analyze intergenerational income mobility in Switzerland. A working paper by Kalambaden and Martınez (2021) studies intergenerational mobility of income, wealth, education, and occupation status. Both papers investigate intergenerational income mobility in Switzerland but use traditional approaches. The explanatory variables are included by dividing the data into subgroups, thereby estimating the measures for each subgroup separately. Our paper contributes to this literature by applying

a more sophisticated method, which avoids subsampling and allows us to gain deeper insights into the driving factors of mobility. We find that mobility differs with the share of income contributed by the father. For sons, the probability of moving upwards increases with the father's income share, but for daughters, this probability of moving upwards tends to decrease with the father's income share. In general, we find great differences in income mobility between sons and daughters and that average hours worked play an important role in driving these results.

2.1.1 Motivational Example

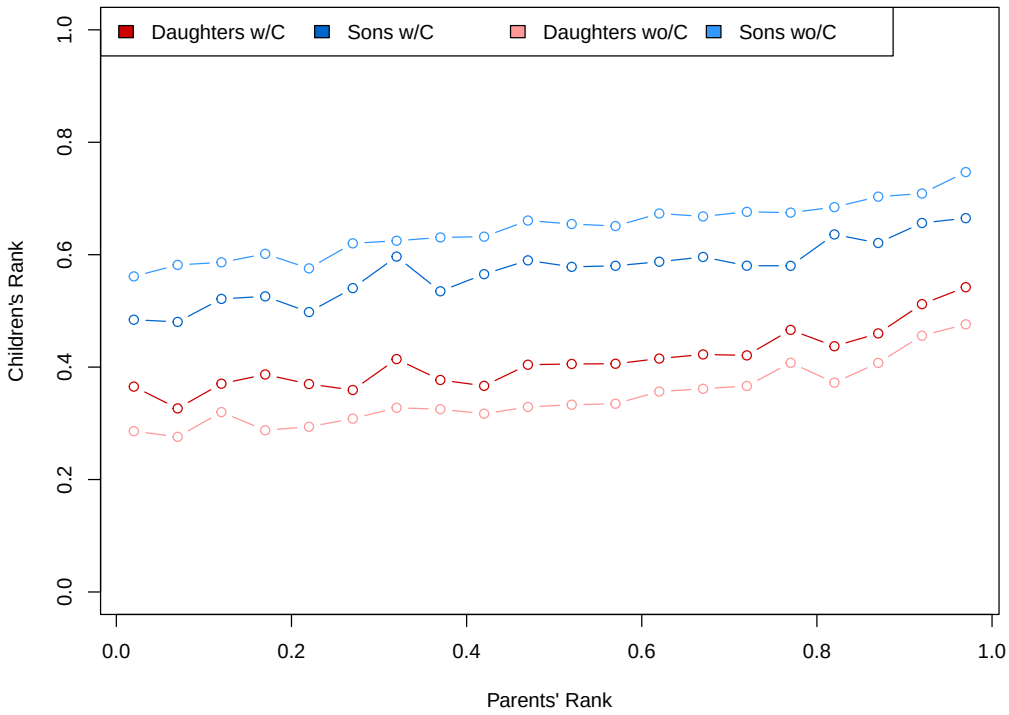
Our approach combines two very convenient features. First, we estimate the whole joint distribution of the outcomes of interest, which allows us to easily construct different types of intergenerational mobility parameters. Second, when estimating the joint distribution, we can account for other variables that influence the outcomes of interest and construct a counterfactual distribution.

In figure 2.1, we plot 25 points of the conditional expected ranks – one of several possible mobility measures – for sons and daughters. The light-colored lines result from our approach when only the sex of children is included as an explanatory variable. In line with previous findings, the relationship between the conditional expected rank of children and parents is rather linear. The relationship seems a bit steeper for children with parents in the highest ranks. As previously found in papers by Chuard and Grassi (2020) and Kalambaden and Martınez (2021), the relationship for Switzerland is relatively flat compared to the U.S. For example, if the slope of the relationship is modeled as linear, Chetty, Hendren, Kline, et al. (2014) find the slope to be around 0.3 for the U.S., while we estimate a value of 0.158 for Switzerland. There are apparent differences between daughters and sons. On average, sons have an income rank of around 0.55, and daughters' rank is around 0.4. This result is driven by the fact that, in general, daughters are on lower ranks in the child distribution. However, we want to investigate how much of this difference can be explained by observable characteristics.

To do so, we estimate conditional expected ranks of sons and daughters, conditioning for other observed factors. The blue dotted lines in figure 2.1 are estimated, including several covariates in addition to the sex of a child. In this case, the gap between the two lines narrows. The variable with the strongest impact is likely to be working hours due to the fact that in Switzerland, women more often work part-time compared to males. Controlling for these covariates reduces the influence of daughters being on lower ranks in the overall distribution. However, a large part of the gap remains unexplained.[2]

[2]Note that working hours itself is potentially endogenous and part of the relation between parents' and children's incomes. Hence, it is debatable whether it should be included as a control or not.

Figure 2.1: Conditional Expected Ranks of Sons and Daughters



Notes: Each dot displays the expected income rank of a child, conditional on the parent's income rank, estimated as outlined in section 2.2.2. Lines in lighter colors display results without controlling for covariates (wo/C). The ones in darker colors display results if we control for covariates (w/C).

The outline of the paper is as follows: In section 2.2, we introduce our method and link it to some parameters of interest from the intergenerational mobility literature. In section 2.3, we present the data and variables used for the analysis. Section 2.4 presents the results, and section 2.5 concludes.

2.2 Model and Estimator

2.2.1 Copula Regression

In this section, we propose an extension to the local Gaussian representation (LGR), which was introduced in Lemma 2.1 of Chernozhukov, Fernández-Val, and Luo (2023). They show that for a bivariate case with variables Y_1 and Y_2, the joint distribution, F_{Y_1,Y_2} can be represented via a Gaussian copula with local correlation parameter that depends on (y_1, y_2). The LGR has also been utilized as a variation by Fernández-Val et al. (2024), who use a general link function to model the marginal distributions rather than the Gaussian.

We propose a generalization of the LGR that allows for different copula families and improves estimation efficiency. The new estimator is a semi-parametric method that aims to weaken the assumptions compared to fully parametric copula-type approaches while maintaining accuracy. At the same time, it reduces the risk of dimensionality issues compared to fully nonparametric approaches. Additionally, by moving beyond the Gaussian specification used in the LGR, significant computational speed improvements are achieved.

For simplification and given the focus of the application on joint income distributions, we focus on the bivariate case. However, the extension to the multivariate case is conceptually straightforward. The starting point is the estimation of the joint conditional distribution of the variables Y_1 and Y_2 given a vector of covariates X using a two-step procedure. By Sklar's Theorem (see also Theorem 2.3.3. in Nelsen (2006)):

$$F_{Y|X}(y_1, y_2|x) = C_{U|X}\left(F_{Y_1|X}(y_1|x), F_{Y_2|X}(y_2|x)|x\right) \tag{2.2.1}$$

where $F_{Y|X}$ is the conditional joint CDF of $Y = (Y_1, Y_2)$ given X. Similarly $C_{U|X}$ is the conditional copula function of $U = (U_1, U_2)$ given X, where, U_1 and U_2 correspond to the conditional marginal distributions $F_{Y_1|X}$ and $F_{Y_2|X}$ of Y_1 and Y_2 given X.

In the first step, the conditional marginal distribution functions $F_{Y_1|X}(y_1|x)$ and $F_{Y_2|X}(y_2|x)$ are estimated using distribution regression. For instance, by estimating a probit regression of $1(Y_1 \leq y_1)$ on X and another probit regression of $1(Y_2 \leq y_2)$ on X.[3]

In the second step, the joint distribution $F_{Y|X}(y_1, y_2|x)$ is estimated using the estimated values for $\hat{F}_{Y_1|X}(y_1|x)$ and $\hat{F}_{Y_2|X}(y_2|x)$ from the first step. To do so, a copula family is chosen to model the conditional copula, where the copula parameter depends on thresholds of the distribution and X. The conditional copula is then written as:

$$C_{U|X}\left(u_1, u_2|x\right) = C_{U|X}\left(u_1, u_2, \theta(y_1, y_2, x)\right) \tag{2.2.2}$$

where the θ is the local copula parameter and $C_{U|X}$ denotes a parametric specification for the conditional copula depending on the chosen copula family. Note that given the local copula parameter, the specification remains flexible despite choosing a specific copula family. Since $u_d = F_{Y_d|X}(y_d|x)$ for $d \in (1, 2)$, we could equivalently let the correlation coefficient be a function of u_1 and u_2 instead of y_1 and y_2. For practical and computational reasons, we assume that the copula parameter is a known transformation of a linear function of X:

$$\theta(y_1, y_2, x) = \Lambda(x'\beta(y_1, y_2)) \tag{2.2.3}$$

The function Λ is useful to impose the support constraints on the parameter θ. For example,

[3] X may be a transformation of the original variables.

in the Gaussian case, θ corresponds to the correlation coefficient and is restricted to lie within $[-1, 1]$. For example, Chernozhukov, Fernández-Val, and Luo (2023) set $\Lambda(x) = tanh(x) = (e^x - e^{-x})/(e^x + e^{-x})$ to impose this support. For the Frank copula, we can use the identity function because the parameter is unbounded, or for the Plackett copula, we can use the exponential function.

For estimation, note that $F_{Y|X}(y_1, y_2|x) = P(Y_1 \leq y_1 \cap Y_2 \leq y_2|x)$. It follows that equations (2.2.1), (2.2.2), and (2.2.3) imply a binary choice model for the variable $\tilde{Y}_i(y_1, y_2) = 1(Y_{1i} \leq y_1) \cdot 1(Y_{2i} \leq y_2)$ with

$$P(\tilde{Y}_i(y_1, y_2) = 1|X = x) = C_{U|X}(F_{Y_1|X}(y_1|x), F_{Y_2|X}(y_2|x), \Lambda(x'\beta(y_1, y_2)))$$

When using the estimates from the first step, we obtain

$$P_i(y_1, y_2, \beta(y_1, y_2)) = C_{U|X}(\hat{F}_{Y_1|X}(y_1|x), \hat{F}_{Y_2|X}(y_2|x), \Lambda(x'\beta(y_1, y_2)))$$

The MLE is

$$\hat{\beta}(y_1, y_2) = \underset{\beta \in \mathcal{R}^K}{\arg\max} \sum_{i=1}^{n} \tilde{Y}_i(y_1, y_2) \ln(P_i(y_1, y_2, \beta)) + (1 - \tilde{Y}_i(y_1, y_2)) \ln(1 - P_i(y_1, y_2, \beta))$$

The estimator is implemented in R using the method by Nelder and Mead (1965). The method is neither based on first nor second-order derivatives and is hence suitable for the different copula families implemented.

2.2.2 Parameters of Interest

Intergenerational mobility is a versatile concept that cannot be grasped with a single measure. Today's researchers are equipped with a large toolbox to measure intergenerational mobility (see Deutscher and Mazumder, 2023 for an in-depth analysis of different mobility measures), and most of these measures are functions of the joint distribution. An advantage of the proposed approach is that it delivers direct results for all these functions of the joint distribution. In general, predominant measures are the so-called intergenerational elasticity (IGE) and intergenerational correlation (ICE), the rank-rank slope (RRS), and transition matrices. The IGE, ICE, or RRS capture mobility in a single parameter that results from a simple linear regression. While these measures are easy to interpret and simplify comparison, e.g., cross-country comparisons, they bear the risk of missing crucial aspects of mobility. Transition matrices provide a more detailed picture of mobility but are harder to interpret. The cells of the matrix display the probability for the child to transit in a children's income bracket, given the parents' income bracket. Transition matrices are commonly displayed with a restricted number of cells

or even single cells of particular interest (for example, the "rags-to-riches" measure) to simplify interpretation. While very insightful, the results depend on the number of brackets specified by the researchers, which is somewhat arbitrary.

Further note that while the IGE and ICE are log-income-based measures, the other measures presented above are rank-based. Nybom and Stuhler (2017) find the attenuation bias from using annual income instead of true lifetime income to be weaker for rank-based measures. Additionally, the life-cycle bias, which arises when lifetime income is measured using income at a young age, is smaller for rank-based measures. However, in the tails of the distribution, rank-based measures can be inaccurate due to these biases. As a general pattern, transition probabilities are found to be understated along the diagonal and overstated off the diagonal.

In this paper, we resort to rank-based measures to capture different aspects of mobility. In particular, we will derive the following measures from the estimated joint distribution: The Rank correlations, transition matrices, and the conditional expected ranks. Additional measures can be derived if needed, though they are not addressed in the following. For simplicity, all measures are presented without considering covariates. Once the conditional distributions are estimated, incorporating covariates for the mobility measures is straightforward. However, it remains important to point out that the specific approach taken affects the interpretation of the computed values.

Rank Correlations (Rank-Rank Slope)

The RRS is a very popular measure, especially since Chetty, Hendren, Kline, et al. (2014) favored the RRS over the more traditional IGE, as the RRS proved to be more robust. The measure is the slope coefficient obtained from regressing the children's ranks on the parent's rank.

The RRS is based on the famous Spearman's rho or Spearman's correlation. When measuring the dependency between two variables, measures like Spearman's rho or Kendall's tau are often preferred over Pearson's correlation. Unlike Pearson's correlation, which captures only linear dependencies, Spearman's rho can capture a broader range of dependencies and is robust to nonlinear transformations of the underlying variable. Spearman's rho is defined as Pearson's correlation of the ranks of the underlying variable.

$$\rho_S = \frac{Cov(R_{Y_1}, R_{Y_2})}{sd(R_{Y_1})sd(R_{Y_2})} \tag{2.2.4}$$

where R_{Y_d} are the ranks of the outcome Y_d for $d \in (1, 2)$. Note that if ranks are normalized to lie within 0 and 1, an individual's rank corresponds to the marginal CDF value of that individual. Consequently, the ranks follow a uniform distribution over the interval $[0, 1]$, and the standard deviation for both ranks is the same. As a result, the denominator equals the variance, $\rho_S =$

$Cov(R_{Y_1}, R_{Y_2})/Var(R_{Y_2})$, which is the slope coefficient of a simple OLS regression of children's ranks on parent's ranks, i.e., the RRS. Given the uniformity of the ranks, $Var(R_{Y_2}) = 1/12$, and Spearman's rho can equivalently be rewritten as:

$$\rho_S = 12E[(R_{Y_1} - 0.5)(R_{Y_2} - 0.5)] \tag{2.2.5}$$

or alternatively as

$$RRS = 12 \int_0^1 \int_0^1 R_{Y_1} R_{Y_2} dC(R_{Y_1}, R_{Y_2}) - 3 \tag{2.2.6}$$

where $C_R(\cdot)$ represents the copula describing the joint distribution of the ranks.

It is crucial to emphasize that the equivalence between the RRS and Spearman's rho holds only as long as ranks are uniformly distributed, meaning they are derived from a well-defined univariate distribution. As pointed out by Deutscher and Mazumder (2023), describing the measure as a slope is, in many applications, more accurate than a correlation. Once the ranks are computed based on an external benchmark rather than the underlying population for which the measure is computed, the measure no longer exactly corresponds to Spearman's Rho. An in-depth discussion is also provided in Chernozhukov, Fernández-Val, Meier, et al. (2024). For the analysis below, we will focus on the correlation measure and the estimation procedure as described in Chernozhukov, Fernández-Val, Meier, et al. (2024). Their approach requires only the estimation of univariate distributions, which corresponds to the first step of our two-step procedure.

Transition Matrices

Transition matrices are one of the most commonly used methods to assess the degree of income mobility within an economy (see, e.g., Shorrocks, 1978, Fields and Ok, 1999, Dickens, 2000, Bonhomme and Robin, 2009). The transition matrix captures the transition probabilities of children conditional on parents' rank. More precisely, it is the probability of having an income rank within some interval, conditional on having parents with an income rank within some interval. Using again the marginal distribution to represent the ranks, a cell in the transition matrix is:

$$TP(u, v) = P(u_1 \leq F_{Y_1}(y_1) \leq u_2 | v_1 \leq F_{Y_2}(y_2) \leq v_2) \tag{2.2.7}$$

where $u = (u_1, u_2), v = (v_1, v_2)$ are grid points between 0 and 1. F_{Y_1} and F_{Y_2} denotes the rank of the children and parents. Equivalently, the transition probability can be expressed as:

$$\begin{aligned} TP(t, s) &= P(t_1 \leq y_1 \leq t_2 \,|\, s_1 \leq y_2 \leq s_2) \\ &= \frac{F_Y(t_2, s_2) - F_Y(t_2, s_1) - F_Y(t_1, s_2) + F_Y(t_1, s_1)}{F_{Y_2}(s_2) - F_{Y_2}(s_1)} \end{aligned}$$

where $t = F_{Y_1}^{-1}(u)$ and $s = F_{Y_2}^{-1}(v)$. All terms in the expression are directly estimated using the new approach. Consequently, the transition probability can be readily obtained by plugging these estimates into the formula.[4]

Typically, the intervals are defined by quantiles, e.g., quintiles. Combining these transition probabilities into a single matrix yields the transition matrix. In this paper, rows capture children's ranks, and columns parents' ranks. Stayers, i.e., individuals who occupy the same part of the distribution as their parents, are displayed along the main diagonal of the transition matrix. Upward movers, i.e., individuals who occupy a higher part of the distribution than their parents, are in the lower triangle, and downward movers are in the upper triangle of the transition matrix. For example, the last cell of the first column of a quintile transition matrix contains the probability of a child reaching the highest quintile of the children's distribution, conditional on having parents in the lowest quintile of the parents' distribution. Symmetric transition matrices that use quantiles and ranks of the underlying distribution as intervals have rows and columns that sum up to one. This no longer holds for the rows whenever ranks are computed using another distribution than the underlying one. For example, a transition matrix for sons that uses the whole population, i.e., sons and daughters, to compute ranks will not necessarily have rows that sum up to one. A limitation of the transition matrix approach is that all intra-quintile transitions are not shown.

Conditional Expected Rank

Another insightful measure is the expected rank of a child, given the rank of the parent. Again, by defining ranks to take values between 0 and 1, the conditional expected rank can be written as:

$$CER(v) = \mathbb{E}\left[F_{Y_1}(y_1) \,|\, v_1 \leq F_{Y_2}(y_2) \leq v_2)\right] \tag{2.2.8}$$

Note that this measure is closely related to the RRS. Using the slope coefficient together with the coefficient of the constant yields an estimate for the same conditional average rank. However, this assumes the conditional average rank to be linear, while with our approach, this assumption is relaxed. Furthermore, instead of the average rank of the child, it is straightforward to derive measures of the conditional distribution of the children's ranks given the rank of the parents, e.g. to estimate the median rank of a child given the rank of the parent (or at any desired quantile).

[4]Note that $F_{Y_2}(y_2) = \lim_{y_1 \to \infty} F_Y(y_1, y_2)$, hence all terms can be derived from the joint distribution.

2.2.3 Decomposition

The proposed estimator further allows for a decomposition analysis by constructing counterfactual distributions. Extending the idea of Chernozhukov, Fernández-Val, and Melly (2013) to our case, we can define the counterfactual joint distribution as:

$$F_{Y\langle 1|0\rangle}(y_1, y_2) = \int_{\mathcal{X}_0} F_{Y_1|X_1}(y_1, y_2|x) dF_{X_0}(x) \tag{2.2.9}$$

where 0 and 1 represent different groups of individuals, e.g., educated and uneducated. Thus, equation 2.2.9 represents the joint distribution for individuals with characteristics of group 0 if they faced the distributional structure of group 1. The subscript $\langle\cdot|\cdot\rangle$ is used to indicate to which group the structure and characteristics correspond to. Using the counterfactual joint distributions allows us to compute the counterfactual for mobility measures explained earlier. In the following, we will focus on the transition probabilities as an example, as they provide a very detailed picture of mobility.[5]

To compare the results from two transition matrices from different subgroups, the difference between each cell, i.e., *total difference* can be computed:

$$TE(t, s) = TP_{\langle 1|1\rangle}(t, s) - TP_{\langle 0|0\rangle}(t, s)$$

Similar to Richey and Rosburg (2018), this total difference can be decomposed into two parts: a compositional and a structural difference. Since the transition matrix results directly from the joint distribution, we can decompose the total difference using the counterfactual joint distribution $F_{Y\langle 0|1\rangle}(y_1, y_2)$. Then, $TP_{\langle 1|0\rangle}(t, s)$ is a transition matrix for individuals with characteristics of group 0 if their conditional joint distribution had the structure of that from group 1. The *compositional difference* can then be defined as:

$$CE(t, s) = TP_{\langle 1|1\rangle}(t, s) - TP_{\langle 1|0\rangle}(t, s)$$

The compositional difference captures how much of the total difference between groups 0 and 1 results from differences in observed characteristics. The *structural difference* is:

$$SE(t, s) = TP_{\langle 1|0\rangle}(t, s) - TP_{\langle 0|0\rangle}(t, s)$$

and captures how much of the total difference between groups 0 and 1 results from differences in the structure of the joint distribution.

[5] Once the counterfactual distribution is estimated, the decomposition of other measures is straightforward.

2.3 Data

We use data from two different sources. The first is the WiSiER data from Switzerland, and the second is the well-known PSID from the United States. Both sets of data allow us to capture the family relations of individuals and provide important control variables. For the analysis, we focus on labor income as our outcome variable. However, it is beneficial to have these control variables at hand since our approach allows us to control for additional variables when estimating the multivariate distribution of income.

2.3.1 WiSiER

The WiSiER is a combination of many different sets of data capturing different characteristics of the individuals. Our data includes individuals from eight Cantons from 1982 to 2016. The main outcome variable is insured labor income, which is aggregated on a yearly basis. Basic covariates are available for all years (e.g., age, gender, marital status, number of children, place of birth, years worked), whereas others are only available from 2011-2016 (e.g., education, company size, company sector, hours worked, nationality). Based on a unique ID, parents and children can be matched, and some of the covariates are imputed backward. For the analysis, each observation corresponds to a child-parent combination with corresponding values averaged over the observed years relevant to the analysis. We restrict the sample to children born in 1968 or later to approximate work experience as described below and drop observations for which the father or mother was below 14 when the child was born.

Variable of Interest

Labor income is measured as yearly income insured by the mandatory old-age and survivors' insurance in Switzerland. Following the literature, income is measured over several years for individuals between 30 and 50 years old to reduce potential attenuation and life-cycle biases. We deflate income using the Consumer Price Index provided by the Federal Statistical Office. We also drop unrealistic values by trimming the data, dropping the highest 1%, and excluding the few negative values[6]. Thereby, only the information for the specific year is removed and not all years for which the individual is observed. The final measure of income is a simple average over all years observed, filtering out noise from single years. For the parent's labor income, the sum of the father's and mother's average labor income is taken.

[6] About 0.05% of the observed years contained negative values. Also, note that trimming the data is not necessary for the underlying estimator. However, it excludes unrealistic high values.

Explanatory Variables

Table 2.1 gives an overview of the control variables included in the analysis. We will especially focus on two crucial variables: the father's income share and the sex of the child. Income share captures the share of income contributed by the father. The father's income share captures several aspects impacting income mobility. For example, the father's income share is likely to be linked to the decision of whether one parent works part-time and thus to the time parents dedicate to their children. Additionally, a higher father's income share is associated with traditional family settings and hence captures family culture and role model effects. For the analysis, the variable is modeled as a polynomial of order two to allow for more flexibility. Furthermore, distinguishing the sex of the children allows us to analyze differences in mobility between sons and daughters.

For individuals born in Switzerland, the place of birth is known on a communal level. For individuals born outside of Switzerland, the place of birth is known at the country level. Given the eight cantons participating, the number of observations varies across regions. For the children born in Switzerland, we include the place of birth on a cantonal level, as cantons tend to be an important factor in Switzerland. For children born abroad, we create six broad regions. For the parents, the mother's and father's place of birth is included separately, but it is only distinguished if they are born abroad or in Switzerland.

Table 2.1: Control Variables WiSiER

Variable	Child	Parents	Type
Income share		x	continuous
Sex	x		binary
Place of birth (CH)		x	binary
Place of birth (regions)	x		categorical
Age	x	x	continuous
Age at measurement	x		continuous
Number of children		x	categorical
Education	x	x	categorical
Wealth		x	continuous
ISEI		x	continuous
Hours worked per week	x		categorical
Years worked	x		categorical

Notes: The columns child and parents indicate for who the control variable is typically included. The column type shows how we included the variable in the analysis.

We additionally control for the average age of children and parents by taking averages over the years when income is observed. The age of the parents is included as an average of the mother's and father's age. Since the timing of investments is essential, we additionally control for the age of the children when parents' income is measured. All variables related to age are modeled as a polynomial of order two to allow for more flexibility.

The number of children of both children and parents is also included, which might capture the available resources of the parents for one child.

Further, the WiSiER contains a categorical variable on the highest education completed. We construct five categories, low education, vocational training and education, high school degree, degree from a higher specialization school (HF), or some higher education (FH, PH, Uni), to control for education. Parents' education is defined as the highest observed education of the father and mother to capture the highest education available within the family.

We also include the net wealth of the parents measured in thousands of CHF. The variable is trimmed on both ends, such that observations that belong to the lowest or highest 1% are dropped. As for income, net wealth is deflated using the Consumer Price Index. Parents' wealth is defined as the sum of the net wealth of both parents. Observations for which the wealth of only one parent is observed are dropped.

The ISCO numbers of the occupations of parents and children are available between 2012 and 2016. The occupation most frequently observed within these years is defined as occupation. Our interest lies in the status attached to certain occupations, which might impact children's income. Thus, we construct the international socio-economic index of occupational status (ISEI) from the ISCO numbers using the iscogen package from Jann (2019). For the analysis, the parent's ISEI is defined as the highest observed ISEI of the father and mother.

Another important variable is the average working hours per week for children. Again, the average is taken over the years when income is observed. The variable is categorized into eight categories (between 0 to 59.5 in steps of 8.5 and one category for values > 59.5).

Finally, we also include years worked as the total number of years for which an income has been observed since 1982. To capture all years worked, we restrict the sample to children born after 1968. It is noteworthy that we still observe an income during absences like maternity leave or military service due to the insurance system. Furthermore, the variable does not differentiate years of full-time or part-time work, and we can only capture years worked in Switzerland. Hence, the variable presents the number of years an individual has been in the Swiss labor market and provides a proxy for work experience. In the analysis, we include years worked for children, averaged over the years when income is observed. We include years worked as a categorical variable with six categories $([0, 5], (5, 7.5], (7.5, 10], (10, 12.5], (12.5, 15], (15, \infty))$.

Descriptive Statistics

The data used for the analysis includes 12,641 pairs of children and parents with complete information on all variables.[7] Table 2.2 displays some descriptive statistics for these observations. We observe that mothers and fathers are recorded mostly between the years 1988 and 2001, while children are observed mostly from 2008 to 2015. The count variables display the number of years the children or parents are observed. Parents are observed for 16 years on average, which is most of the time span between 30 and 50. Unlike their parents, we observe each child for about 8.19 years, with the median being also at 8. This is a result of excluding years when the child is below 30. Furthermore, note that the average age of children when their income is measured is mostly between 30 and 38. The average age of parents when their income is measured is between 40 and 45. A majority of the children are between the ages of 10 and 20 when the income of the parents is measured. The median and mean are around 15, which, as described above, is a crucial stage of childhood. Finally, in general, the fathers provide most of the parent's income. The median for the father's income share is at 1, and a value below 0.4 is rarely observed.[8]

2.3.2 PSID

The PSID is a longitudinal household survey dataset that started in 1968 and was conducted every year until 1997 and biannually afterward. The basic survey captures many different aspects of individuals and families, and the same data has been used previously to study intergenerational mobility by Callaway and Huang (2018). The PSID collects information for individual units as well as the family unit. As a result, variables collected for the family unit and the related roles like 'head' and 'spouse' are subject to change if the family composition changes. In these cases, the observed individual's information in a specific year depends on the family role of that individual for the given year, which must be taken into account. Basic covariates like age, gender, marital status, and more are available for all years. Other variables like employment status, occupation, and education are not covered for all years. Moreover, the survey is continuously being updated such that variables and the information they capture are added, removed, or adapted across the years.[9]

In the following, an observation is equivalent to a child-parents relationship. Parents exclusively refer to biological mothers and fathers, and we only keep observations if each parent is

[7] Table 2.6 in the appendix displays the same statistics for all observations. The results are very similar. Hence, the observations used in the main analysis are similar to those of the people we are interested in. The only exception is the place of birth, for which the composition changes, as shown in table 2.7 in the appendix. Variables only observed from 2011 to 2016 (e.g., education, occupation, working hours) lead to this quite reduced number of observations.

[8] Details are displayed in figure 2.8 in the appendix.

[9] Whenever possible, adjustments were made guided by the codebook to allow for consistency across years.

Table 2.2: Descriptive Statistics WiSiER

	Mean	SD	Median	Q5	Q95
Child's inocme	45,356.67	21,670.40	45,184	10,399.13	82367.92
Parents' income	77,170.27	30,940.47	71,368	39,865.00	134753.81
Income share	0.75	0.16	1	0.48	0.97
Son	0.51	0.50	1	0.00	1.00
Child's age	33.68	2.03	34	30.50	37.50
Parent's age	41.76	1.91	41	39.76	45.40
Income measurement age	14.76	3.04	15	9.95	20.00
Parents no. of children	2.57	0.85	2	1.00	4.00
Child's education	8.24	2.54	8	6.00	12.00
Parents education	6.72	2.57	6	3.00	12.00
Parents' wealth	598.07	798.62	343	0.00	2192.75
Parents' ISEI	45.49	17.35	42	18.00	71.00
Child's work-hours	35.79	12.59	41	9.00	50.00
Years worked	14.62	2.85	15	9.50	19.17
Child year	2,012.31	2.00	2,013	2,008.50	2015.50
Parents year	1,993.71	3.98	1,993	1,988.00	2001.25
Child count	8.19	3.92	8	2.00	15.00
Parents count	16.40	3.84	17	9.00	21.00

observed at least once.

In the following, the variables and the most important preprocessing steps are briefly described. A more detailed description of the variables and the construction of the data is given in Appendix 2.A.2. In general, variables are constructed as averages across the observed years.

Variable of Interest

To compare the results to the WiSiER data, we will focus on total yearly labor income, which is available starting from 1968. Following the same argument as before, we consider only income earned between the ages of 30 and 50 years old. For the head of the family unit, total labor income, comprised of wage income plus others like farming, business, bonuses, and overtime, is available for all years. Unfortunately, a similar labor income variable for the spouse was only recorded until 1993. After 1993, the spouse's labor income is measured excluding farm income and the labor portion of business income. While business income remains easily accessible after 1993, farming income does not. To ensure consistency across years, all families with income from farming are excluded. Finally, we deflate labor income using the Consumer Price Index

provided by the U.S. Bureau of Labor Statistics, trim the highest and lowest yearly values, and take the average over all observed years. The parent's labor income is then constructed as a sum over the father's and mother's labor income.

Explanatory Variables

The number of available control variables is smaller for the PSID compared to the WiSiER. However, since the number of observations is also smaller for the PSID, including a smaller number of control variables is desirable. Table 2.3 displays an overview of the control variables. The variables income share and sex capture the same information as for the WiSiER case. Most variables are available starting from 1968. The only exceptions are college degree, available only since 1975, and urbanization of birthplace for the spouse, which is available from 2009 on, and the years 1976 and 1985. Time-varying variables are only considered for years in which the individual is between 30 and 50 years old, ensuring that only the relevant data from those years, in which income is also measured, are captured.

The race of the children is included as a control variable. We work with the race information, which was first mentioned by the individuals and carried forward in time by the PSID. The majority of children belong either to the group "white" or "black". Accordingly, a categorical variable with three categories "white", "black" and "others" is created for the analysis.

Furthermore, urbanization captures the size of the area in which an individual grew up. As shown in past research, the location in which a child grew up impacts mobility (see for example Chetty, Hendren, and Katz (2016), Chetty and Hendren (2018a), and Chetty and Hendren

Table 2.3: Control Variables PSID

Variable	Child	Parent	Type
Income share		x	continuous
Sex	x		binary
Race	x		categorical
Urbanization	x		categorical
Age	x	x	continuous
Age at measurement	x		continuous
Number of children		x	continuous
College degree		x	binary
Hours worked		x	continuous, mothers

Notes: The columns child and parents indicate for who the control variable is typically included. The column type shows how we included the variable in the analysis.

(2018b)). The variable available comprises the categories "rural area", "small town, suburb", "large city" and "other".

The variables age and age at measurement capture the same information as for the WiSiER case.

The number of children captures the number of individuals below 18 living in the parents' family unit averaged over the years when income is observed. Note that these must not necessarily be the actual children. Since this number potentially varies across the years in which income is measured, the variable is continuous. We model the relationship to be linear.

Further, we include college degree as a binary variable, indicating if the parents completed college when we measure their income. To construct the parents variable, the maximum of the father and the mother is taken, capturing if at least one of the parents obtained a college degree to focus on the highest available education within the family.

The PSID also includes information on the total annual working hours on all jobs for money. For the analysis, only the working hours of the mothers are included as control. For fathers, a large majority works 1600 hours or more per year, which is close to full-time, leaving little variation for the analysis. The impact of the variable is approximated using a linear specification.

Descriptive Statistics

Table 2.4 displays some descriptive statistics for the observations with no missing values, which are subsequently used in the analysis.[10] In total, 5,777 observations are available. Some important observations should be noted. First, parents' information, in general, was captured in earlier years, around 1984, while children's information was captured in more recent years, around 2006. Additionally, parents are observed later in their lives, mostly between the ages of 36 and 48, while children are observed earlier between the ages of 30 and 40. Given the strong relationship between age and labor income, controlling for these differences is crucial for the analysis. The age at measurement exhibits the age of the child when the income of the parents is observed. The average age at measurement is 16, and a majority of the children are between the ages of 8 and 23 when the income of the parents is measured. The count variables display the number of years an individual was observed in the PSID. In general, parents are observed for longer periods, around 14 years, and children are observed for shorter periods. On average, children are observed for 7.5 years in a right-skewed distribution, whereas 556 children are only observed for one year. As for the WiSiER data, this is a result of excluding years when the child is below 30. We further see that for a majority, the father contributes the main part to the total labor income of the parents.

[10] Table 2.9 in the appendix displays the same statistics for all observations.

Table 2.4: Descriptive Statistics PSID

	Mean	SD	Median	Q5	Q95
Child's Income	16,306.92	11,944.17	14,242	333.42	40394.15
Parents' Income	28,926.52	14,544.21	27,553	7,516.67	55522.46
Income share	0.75	0.21	1	0.38	1.00
Daughter	0.53	0.50	1	0.00	1.00
Race	1.44	0.88	1	1.00	2.00
Child Urbanization	2.36	0.65	2	1.00	3.00
Child's age	35.55	3.40	36	30.33	40.05
Parent's age	41.10	3.53	40	36.14	47.50
Income measurement age	15.80	4.58	16	8.00	23.00
Parents' no. of children	2.44	1.56	2	0.66	5.76
Parent's college degree	0.25	0.42	0	0.00	1.00
Mother's work-hours	1,030.64	706.70	1,042	0.00	2053.67
Child year	2,005.53	10.54	2,009	1,988.55	2019.00
Parents' year	1,983.96	10.24	1,984	1,970.50	2000.62
Child count	7.55	5.19	6	1.00	17.00
Parents' count	13.51	5.03	14	4.50	21.00

2.4 Results

In the following sections, we will demonstrate the capabilities of our approach through some illustrative examples. We first compare mobility measures constructed using the proposed estimator to those constructed via traditional estimation. Afterward, we examine how mobility is related to a continuous variable, the father's income share, which is straightforward to analyze using our approach. In the last part, we focus on differences in mobility between sons and daughters, thereby showing the potential of decomposition.

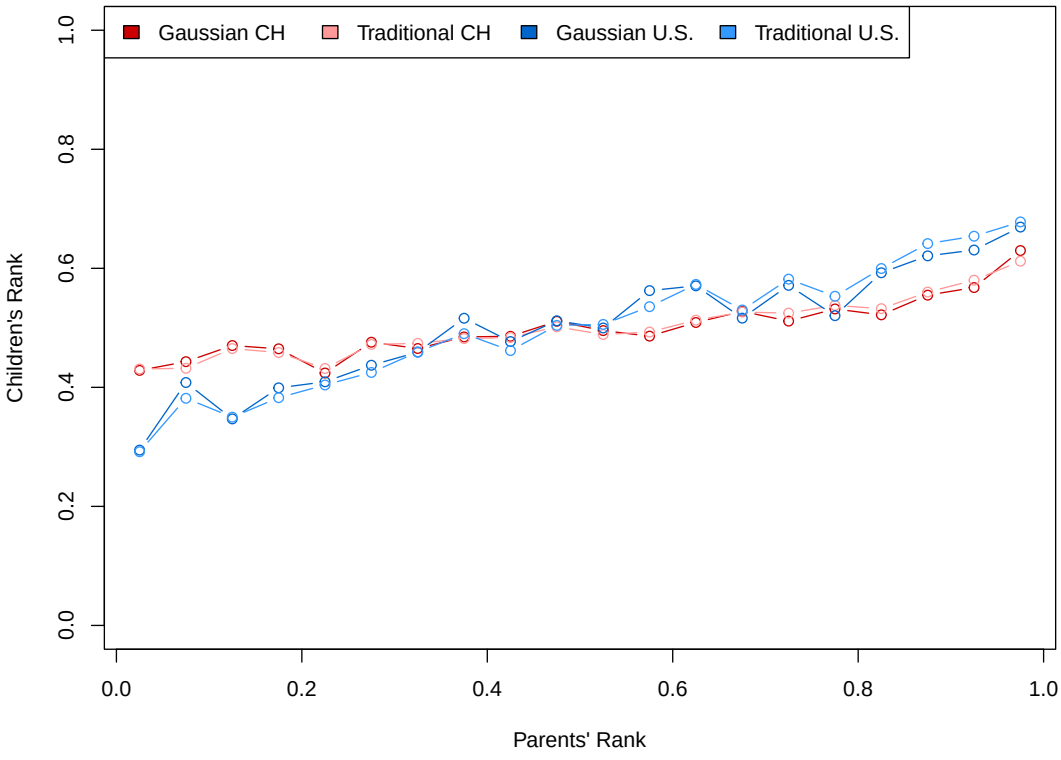
If not stated differently, we specify the link function for the first step in the estimation procedure as a Probit. For the second step, the copula is specified to be Gaussian, with the local copula parameter being $\theta(x) = \Lambda(x) = (e^x - 1)/(e^x + 1)$.

Figure 2.2 compares the CER computed by traditional estimation to the CER computed using our new approach and including all covariates. The results are computed for Switzerland, in red, and the U.S., in blue. As expected, estimated values from the suggested approach are very much the same as if we would take a more traditional approach. Furthermore, the estimates for the U.S. are more volatile compared to the ones for Switzerland, which is likely due to the smaller sample size. For Switzerland, the CER shows a slightly increasing relation

to parents' rank. A child with parents in the lowest income rank has an expected rank of a bit more than 0.4. A child with parents in the highest income rank, on the other hand, has an expected rank of 0.6. The results are also similar to the ones of Chuard and Grassi (2020). Small differences likely arise from different sample definitions. For the U.S., the slope of the estimated CER is steeper compared to Switzerland, which indicates higher persistence. A child with parents in the lowest income rank has an expected rank of around 0.3, while a child with parents in the highest rank has an expected rank of a bit more than 0.6.

Figure 2.3 shows four decile transition matrices. For the first transition matrix, (a), transition probabilities for Switzerland are estimated as the fraction of children in the m^{th} decile of the children's income distribution, conditional on having parents in the n^{th} decile of the parents' income distribution. The second transition matrix, (b), is estimated for Switzerland using our new approach and a Gaussian copula. The third transition matrix, (c), is equivalent to (b), except that a Gumbel copula was specified. The last transition matrix, (d), is again estimated using a Gaussian copula but with data from the U.S. For the estimation, only a constant is included. Hence, we don't expect the results from the different estimation procedures to differ. Figure 2.3 underlines this result. All three matrices displaying results for Switzerland are al-

Figure 2.2: Sample Conditional Expected Ranks



Notes: The figure displays 20 estimated CER for Switzerland (red) and the U.S. (blue). Each dot represents the average of children's rank conditional on parents' rank. The lighter dots are estimated using the traditional approach. The darker dots are estimated using the suggested approach as outlined in section 2.2.2.

most identical. Since we split the matrix into deciles, each cell of the matrices should contain transition probabilities of 10% if the parents' rank has no relation to the child's rank. Overall, the estimated transition probabilities for Switzerland are not far off from 10%. They are a bit higher around the diagonal, especially for the tails. That is, for children with parents in a high (low) income rank, the probability of ending up in a high (low) income rank is clearly larger than 10%. This effect is more pronounced for the upper tail, where a probability of 22% is observed. Hence, there is some persistence in income mobility. For the U.S., a similar pattern is observed, but as already observed for the CER, persistence tends to be higher compared to Switzerland. Transition probabilities along and close to the diagonal are generally higher than 10% and become much smaller moving off the diagonal. Furthermore, values at and around the tails increase to around 15%.

Figure 2.3: Transition Matrices

(a) Traditional CH

Child brackets \ Parents brackets	0–0.1	0.1–0.2	0.2–0.3	0.3–0.4	0.4–0.5	0.5–0.6	0.6–0.7	0.7–0.8	0.8–0.9	0.9–1
0–0.1	14%	11.6%	12.3%	10.4%	10.1%	10.4%	8.7%	8.5%	7.8%	6.2%
0.1–0.2	10.4%	11.3%	10.7%	9.6%	9.6%	11.7%	9.7%	9.2%	9.7%	8.1%
0.2–0.3	13%	8.8%	11.8%	10.3%	10.4%	9.3%	10.3%	8.9%	9.7%	7.6%
0.3–0.4	11.6%	10.9%	10.6%	11.6%	10.2%	9.9%	9.6%	9.2%	8.3%	8.1%
0.4–0.5	11.1%	11.8%	10.8%	10.9%	10.1%	10%	8.6%	10.2%	9%	7.5%
0.5–0.6	10.7%	11.7%	9.9%	10.8%	11.2%	8.5%	11%	9.2%	8.1%	8.8%
0.6–0.7	8.6%	10.4%	10.4%	11.4%	9.5%	11.2%	9.7%	11.3%	9.2%	8.5%
0.7–0.8	7.9%	10.5%	10.9%	9.1%	11.3%	10.2%	8.5%	9.6%	11%	10.9%
0.8–0.9	8.1%	7.6%	8.4%	8.5%	10.5%	9.8%	11.9%	11.5%	11.3%	12.4%
0.9–1	4.7%	5.4%	4.3%	7.4%	7.1%	9.2%	11.9%	12.3%	15.8%	21.9%

(b) Gaussian CH

Child brackets \ Parents brackets	0–0.1	0.1–0.2	0.2–0.3	0.3–0.4	0.4–0.5	0.5–0.6	0.6–0.7	0.7–0.8	0.8–0.9	0.9–1
0–0.1	14%	11.6%	12.3%	10.4%	10.1%	10.4%	8.7%	8.5%	7.9%	6.3%
0.1–0.2	10.4%	11.3%	10.7%	9.7%	9.6%	11.7%	9.7%	9.3%	9.6%	8%
0.2–0.3	12.9%	8.8%	11.8%	10.3%	10.3%	9.2%	10.4%	9%	9.7%	7.6%
0.3–0.4	11.6%	10.9%	10.5%	11.7%	10.3%	9.9%	9.5%	9.1%	8.1%	8.3%
0.4–0.5	11.1%	11.8%	10.9%	10.8%	10.2%	9.9%	8.6%	10.3%	9.2%	7.3%
0.5–0.6	10.8%	11.7%	9.8%	10.9%	11.2%	8.5%	11.1%	9%	8.3%	8.7%
0.6–0.7	8.6%	10.4%	10.5%	11.3%	9.5%	11.4%	9.4%	11.5%	9.1%	8.5%
0.7–0.8	7.9%	10.4%	10.9%	9.1%	11.3%	10.1%	8.7%	9.6%	11%	10.9%
0.8–0.9	8.1%	7.6%	8.3%	8.5%	10.6%	9.7%	12.1%	11.5%	11.3%	12.4%
0.9–1	4.6%	5.4%	4.3%	7.3%	7.2%	9.2%	11.9%	12.3%	15.8%	21.9%

(c) Gumbel CH

Child brackets \ Parents brackets	0–0.1	0.1–0.2	0.2–0.3	0.3–0.4	0.4–0.5	0.5–0.6	0.6–0.7	0.7–0.8	0.8–0.9	0.9–1
0–0.1	14%	11.6%	12.3%	10.4%	10.1%	10.3%	8.7%	8.5%	7.8%	6.3%
0.1–0.2	10.4%	11.4%	10.6%	9.7%	9.5%	11.8%	9.7%	9%	9.8%	8.1%
0.2–0.3	13%	8.7%	11.9%	10.2%	10.5%	9.2%	10.3%	9.1%	9.5%	7.6%
0.3–0.4	11.6%	10.9%	10.6%	11.6%	10.2%	9.9%	9.6%	9.1%	8.4%	8%
0.4–0.5	11.1%	11.8%	10.8%	11%	10%	10.1%	8.5%	10.3%	9%	7.5%
0.5–0.6	10.7%	11.7%	10%	10.8%	11.2%	8.5%	11%	9.2%	8%	8.9%
0.6–0.7	8.6%	10.4%	10.3%	11.4%	9.5%	11%	9.7%	11.3%	9.2%	8.4%
0.7–0.8	7.9%	10.5%	10.9%	9.1%	11.3%	10.2%	8.6%	9.8%	10.8%	11%
0.8–0.9	8.1%	7.6%	8.4%	8.4%	10.7%	9.8%	11.8%	11.4%	11.5%	12.4%
0.9–1	4.7%	5.4%	4.2%	7.3%	7.1%	9.2%	12.1%	12.3%	15.8%	21.9%

(d) Gaussian U.S.

Child brackets \ Parents brackets	0–0.1	0.1–0.2	0.2–0.3	0.3–0.4	0.4–0.5	0.5–0.6	0.6–0.7	0.7–0.8	0.8–0.9	0.9–1
0–0.1	20.6%	16.1%	12.3%	9.2%	8.3%	7.7%	7.7%	6.6%	5.6%	6%
0.1–0.2	14.5%	13.8%	13.3%	11.8%	10.5%	10.3%	7.3%	5.7%	6.4%	6.3%
0.2–0.3	14.8%	15.4%	12.5%	11.1%	10.9%	8.6%	9%	7.7%	5.4%	4.4%
0.3–0.4	16.1%	14.8%	12.6%	9%	9.7%	7.9%	7.4%	10.2%	7.4%	4.9%
0.4–0.5	10.1%	10.7%	11.8%	11.6%	11.7%	11.5%	8.9%	11.4%	7.4%	5%
0.5–0.6	7.4%	11.1%	10.8%	9.9%	12.6%	10.6%	11.2%	11.4%	9%	5.9%
0.6–0.7	5.6%	6.6%	10.6%	11.9%	12.3%	12.8%	11.3%	9.4%	9.8%	9.8%
0.7–0.8	4.9%	4.7%	7.3%	11.6%	11.2%	11.1%	12.6%	9.7%	14.1%	12.8%
0.8–0.9	4.1%	3.9%	5.9%	9.4%	6.9%	10.8%	13.9%	12.3%	15.5%	17.3%
0.9–1	1.9%	2.9%	2.8%	4.6%	6%	8.7%	10.7%	15.7%	19.2%	27.5%

Notes: Each cell, (m, n), contains a transition probability, which is the probability of children to end up in the m^{th} decile of the children's income distribution, conditional on having parents in the n^{th} decile of the parents' income distribution. The transition matrix (a) is estimated via traditional methods. All other matrices are estimated using the new approach as outlined in section 2.2.2.

Father's Income Share

The father's income share, which is the fraction of the total income contributed by the father, has various potential channels to impact the labor market outcomes of the child. Firstly, it is more likely that the mother will work part-time if the father's income share is higher, which leaves more time to invest in the child.[11] Hence, we expect higher upward mobility for children, with fathers contributing a higher share of total parental income. Furthermore, the main earner in families at the top of the parental income distribution must have an exceptionally high income. The position related to such a high income is likely to be linked with better labor market information and connections, which is expected to further enhance upward mobility. Secondly, a higher father's income share may also capture the influence of social norms and role models. Families where the father is the main earner embody more traditional gender roles, whereby fathers are responsible for providing for the family, and mothers look after children and the home. If this family culture is adopted by the children, we may expect the trend in upward mobility to be more pronounced among sons than daughters.

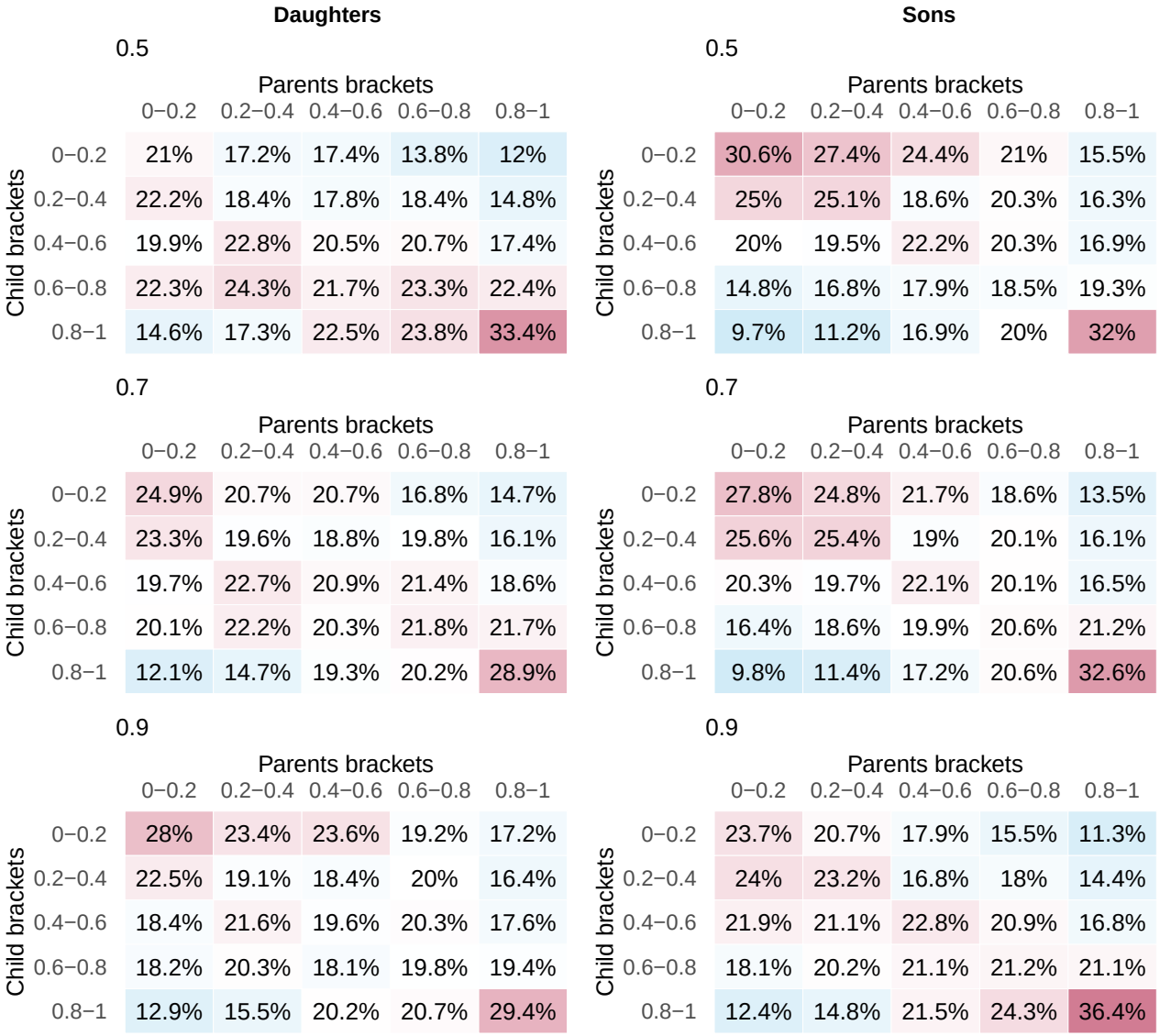
Traditional methods involve categorizing the continuous variable into subgroups to analyze differences in intergenerational mobility across the father's income share. The suggested estimator enables us to directly include the father's income share as a continuous variable and estimate mobility at any desired point along the father's income share. To analyze the effect of the father's income share on mobility for sons and daughters, we specify a model that excludes the children's education, average working hours, and the years worked. These variables are likely endogenous, as they are all chosen by the children and hence likely influenced by the parents' income or the father's income share. Figure 2.4 presents the results, where the transition matrices display the mobility of daughters and sons within their respective groups. The daughter's (sons') ranks are computed relative to the daughter's (sons') income distribution.[12] This allows us to look past the great differences in mobility between sons and daughters, as analyzed below, and enables a clearer assessment of mobility differences with respect to the father's income share.

Mobility conditional on fathers' income share clearly differs between sons and daughters. For daughters, values in the lower triangle and the lower tail tend to increase, and values in the upper triangle and the upper tail tend to decrease as the father's income share grows. These results indicate that for daughters, upward mobility decreases, and downward mobility increases as the fathers' income share grows. This effect is especially pronounced in the tails. For daughters with parents in the lowest income quintile, the probability of ending up in the lowest quintile of the daughter's income distribution increases from 21% to 28%. Conversely,

[11] Some descriptive statistics supporting this relation can be found in appendix 2.A.1.

[12] Additional transition matrices for other values of the father's income share are presented in appendix 2.B.1. Results for variations in the included covariates and transition matrix brackets are discussed in appendix 2.B.2.

Figure 2.4: Mobility Conditional on Father's Income Share using the WiSiER Data



Notes: The left column displays results for daughters, and the right column displays results for sons. The ranks are computed by the respective subsample, i.e., the brackets from the transition matrices in the left column result from the daughter's income distribution. Each row displays results for different father's income share. For example, the first row displays results if the father's income share is 50%.

for the upper tail, the probability decreases from 33.4% to 29.4%. For sons, we observe the opposite effect. Values in the lower triangle and the lower tail tend to decrease, while values in the upper triangle and the upper tail tend to increase as the father's income share grows. Thus, upward mobility increases and downward mobility decreases as the father's income share grows. Hence, even with fathers as the main earners and potentially more time and other resources available, daughters are more likely to move downward in their respective income distributions. A potential driver of this pattern is social influence, i.e., the second channel described earlier. While daughters from households where the father is the main earner might choose to live in a household where the partner is the main earner, daughters from households where the mother contributed a significant share of the household income might aim to contribute a significant share to their own household. For sons, we don't have this counteracting effect but an enforcing one. While for daughters, the positive effect of a higher father's income share on upward mobility is counteracted and actually reversed due to the second channel, the positive effect is enhanced due to the second channel for sons.

Figure 2.5 displays a summary of the results repeating the analysis of the father's income share for the U.S. PSID data.[13] Transition matrices are reduced to 4×4 given the smaller sample size. The model includes only the dummy for sons, a polynomial of order two for the father's income share, and the interaction terms of the two. We further include controls for the age of parents and sons to reduce life-cycle biases. Further, variables are not included, given the smaller sample size.[14] Overall, we see that the effect goes in a similar direction for the U.S. compared to Switzerland. An interesting difference is that the transition probabilities in the very tails for sons and daughters in their respective income distribution are similar if the father's income share is high, at 90%. Otherwise, the results are very similar. For daughters, probabilities in the upper triangle tend to increase, i.e., downward mobility increases, and the lower triangle tends to decrease, i.e., upward mobility decreases, as the father's income share increases. For sons, we observe the opposite effect. To give some examples, for daughters, the probability of reaching the highest quartile of the daughter's income distribution decreases from 46.8% to 42.8%, while for sons, the probability increases from 36.1% to 50.5%. For the lower tail, the probability increases from 30.2% to 35.7% for daughters, while for sons, it decreases from 50.3% to 40.2%. Consequently, the interpretation remains the same as that of Switzerland.

[13] Additional transition matrices for other values of the father's income share are presented in appendix 2.B.1

[14] Appendix 2.B.2 contains the results from the analysis, including all covariates.

Figure 2.5: Mobility Conditional on Father's Income Share using PSID Data

Notes: The left column displays results for daughters, and the right column displays results for sons. The ranks are computed by the respective subsample, i.e., the brackets from the transition matrices in the left column result from the daughter's income distribution. Each row displays results for different father's income share. For example, the first row displays results if the father's income share is 50%.

Daughters and Sons

In section 2.1.1, we observed that daughters tend to have lower average ranks than sons in the overall children's income distribution. Additionally, the CER for daughters appeared to increase slightly with higher parents' rank. This pattern is further reflected in the higher estimated RRS for daughters compared to sons, as shown in table 2.5. The table presents the RRS in the first row, computed using rank-rank regression with ranks derived from the children's income distribution. The second row displays the results of the conditional rank-rank regression approach presented in section 2.2.2. Estimating the RRS for the entire population yields an estimate of 0.158. Given normalized ranks ranging from 0 to 100, this implies that an increase of ten ranks for the parents relates to an average increase of 1.58 ranks for the child. For daughters, the RRS is 0.172, while for sons, it is 0.16, indicating a slightly higher persistence for daughters. The findings are in line with previous research for Switzerland. Kalambaden and Martınez (2021) find an RRS of 0.138 for the whole population, 0.161 for daughters, and 0.119 for sons. Chuard and Grassi (2020) find an RRS of 0.151 for the whole population and an RRS of 0.152 for sons and daughters when using the rank of family income. The average conditional correlation for both sons and daughters is slightly lower. As explained in Chernozhukov, Fernández-Val, Meier, et al. (2024), these values can be interpreted as within-group persistence, that is, the mobility persistence of sons and daughters within their respective groups. The correlation is slightly lower for daughters at 0.061 compared to the sons' correlation at 0.102.

The transition matrices allow for an even closer look into the mobility of daughters and sons. Figure 2.6 shows two transition matrices that capture transition probabilities for different quintiles. The first displays results for sons and the second results for daughters. Matching previous findings, transition probabilities in the lowest two quintiles are very high for daughters. For the sons, it is the opposite, as they have higher probabilities in the last two quintiles. Clearly, this results from the fact that, in general, sons obtain higher ranks in the children's income distribution compared to daughters. For example, sons with parents in the lowest bracket have an estimated probability of 6.1% to end up in the lowest bracket as well. For daughters, this probability is 41.4%. On the other end of the distribution, Sons with parents in the highest income bracket have an estimated transition probability of 45% to end up in the highest bracket. For daughters, the probability is 12.6%. It is interesting to notice that these results even hold

Table 2.5: Rank-Rank Relations by Sex

	Daughters	Sons
Rank-Rank Slope	0.172	0.160
Conditional Correlation	0.061	0.102

for extreme movements. For example, daughters with parents in the highest income quintile still face a probability of 28.1% to end up in the lowest income quintile. For sons, the same probability is at 5.5%. Hence, even if transition probabilities for daughters in the lowest two quintiles tend to decrease as parents' rank increases, it still remains quite high.

To analyze the main drivers of these results, we follow the decomposition outlined in section 2.2.3. Figure 2.7 shows the compositional, the structural, and the total difference. The total difference is largest in the first and last quintile of the children's income distribution. The compositional difference displays the difference between sons with their characteristics and sons if they had the same characteristics as daughters, and the structural difference shows the difference between sons if they had the same characteristics as daughters and daughters. Hence, it gives an idea about the difference in transition probabilities due to unobserved characteristics or the part that remains unexplained. We observe that part of the total differences between sons and daughters in transition probabilities are explained by the given characteristics of the two groups. However, a part of the difference seems to result from differences in the structure of the joint income distribution of children and parents, so even if a daughter exhibited the same observed characteristics as sons, they would find themselves on lower ranks with a higher probability.

Clearly, these results strongly depend on the explanatory variables included in the analysis. Depending on the observed characteristics, the compositional and structural differences change. As described in section 2.3, income is not restricted to the result of a specific type of job, such as a full-time job. As a result, income depends very strongly on the average working hours. Furthermore, females in Switzerland tend to work part-time. Therefore, a large part of the differences between sons and daughters is expected to result from the fact that daughters very often work less than sons.

Figure 2.6: Transition Matrices by Sex

Sons

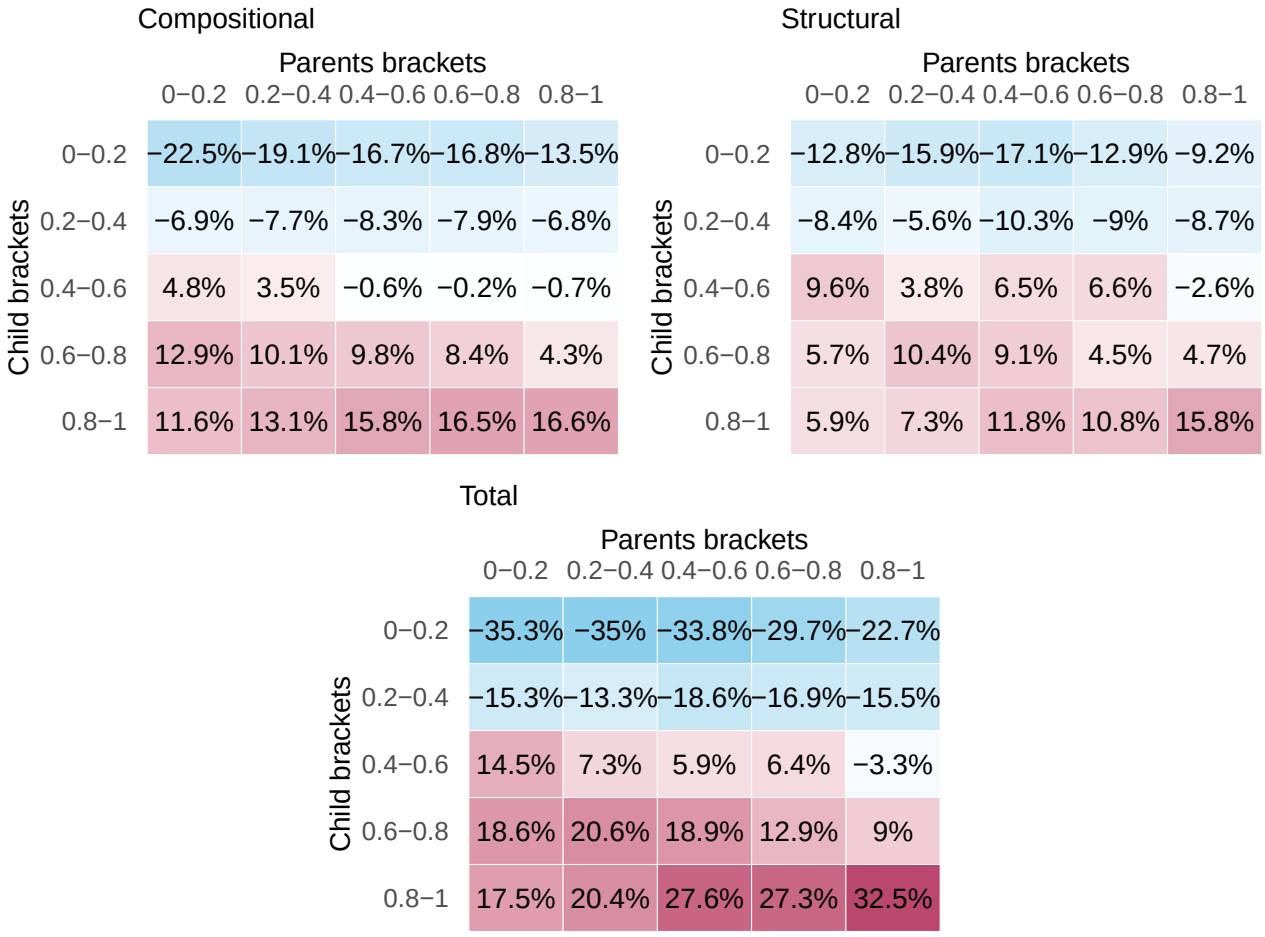
Child brackets \ Parents brackets	0–0.2	0.2–0.4	0.4–0.6	0.6–0.8	0.8–1
0–0.2	6.1%	5%	4%	5%	5.5%
0.2–0.4	14.7%	15.9%	8.9%	10.4%	10.4%
0.4–0.6	28.6%	23%	24.5%	23.1%	15.2%
0.6–0.8	29.7%	30%	30.3%	26.1%	23.9%
0.8–1	20.8%	26.2%	32.3%	35.5%	45%

Daughters

Child brackets \ Parents brackets	0–0.2	0.2–0.4	0.4–0.6	0.6–0.8	0.8–1
0–0.2	41.4%	40%	37.8%	34.6%	28.1%
0.2–0.4	30.1%	29.2%	27.4%	27.3%	25.9%
0.4–0.6	14.1%	15.7%	18.6%	16.8%	18.5%
0.6–0.8	11.1%	9.4%	11.4%	13.2%	14.9%
0.8–1	3.3%	5.7%	4.7%	8.1%	12.6%

Notes: For both transition matrices, the ranks are given by the entire sample. Sons, in general, obtain higher ranks compared to daughters. Hence, transition probabilities for sons are higher in the upper quintiles. The opposite is true for daughters.

Figure 2.7: Decomposition Results



Notes: For all transition matrices, the ranks are given by the entire sample.

2.5 Conclusion

Intergenerational income mobility is a critical concept in addressing questions of inequality. We suggest a flexible semi-parametric approach to estimate the joint income distribution of parents and children using distribution regression and a copula with a copula parameter that is local in the value of the outcome. The new estimator allows us to control for different explanatory variables and gain deeper insights into drivers of intergenerational income mobility. The approach achieves this without imposing strong parametric assumptions and reduces the risk of dimensionality issues when compared to nonparametric approaches. Furthermore, decomposing the total difference into structural and compositional differences provides a more nuanced understanding of the drivers of intergenerational mobility.

The method is applied for two different countries, Switzerland, using the WiSiER data, and

the U.S., using the PSID data. The advantages of the new method are illustrated by focusing on differences in mobility outcomes between sons and daughters and the role of the income share contributed by fathers. We observe that sons experience higher upward mobility when the father contributes a larger share of the total parental income, whereas no such relation is detected for daughters. In family settings, where fathers are the main contributor to parents' income and mothers are working part-time, both the available time resources, as well as family culture and role models, are potential drivers of the observed differences in upward mobility. Furthermore, our findings indicate that the slope between parents' and children's ranks is similar, but daughters generally exhibit lower transition probabilities into higher ranks of the children's income distribution. The decomposition gives insights into the factors that drive these differences. We find that differences in observed characteristics, such as hours worked per week, explain some of the observed disparities between sons and daughters. However, a large part remains unexplained by observed traits. In summary, the suggested approach is well-suited to address key questions related to intergenerational mobility, allowing us to analyze mobility across various characteristics of individuals or groups.

2.A Appendix I - Supplementary Data Description

2.A.1 Additional Descriptives of the WiSiER Data

Descriptive Statistics for All Observations

Table 2.6: Descriptive Statistics

	Mean	SD	Median	Q5	Q95	N
Child's inocme	42,066.40	23,323.11	42,046	5,578.06	81999.06	767,222
Parents' income	75,936.36	31,986.47	69,975	36,575.75	134538.30	767,222
Income share	0.76	0.17	1	0.48	0.98	767,221
Son	0.51	0.50	1	0.00	1.00	767,222
Child's age	33.76	2.66	34	30.00	38.21	767,191
Parent's age	42.41	2.38	42	39.57	46.75	767,191
Income measurement age	13.83	3.68	14	7.50	19.50	767,222
Parents no. of children	2.56	0.89	2	1.00	4.00	755,778
Child's education	8.06	2.63	7	5.00	12.00	174,928
Parents education	6.33	2.67	6	2.00	12.00	323,077
Parents' wealth	629.70	798.93	389	0.00	2249.78	320,469
Parents' ISEI	45.83	17.34	44	18.00	71.00	135,939
Child's work-hours	36.38	12.11	41	10.00	50.00	182,402
Years worked	14.21	3.56	14	8.00	20.00	754,596
Child year	2,012.14	2.70	2,013	2,007.50	2016.00	767,222
Father year	1,990.84	4.09	1,990	1,985.00	1999.00	767,222
Mother year	1,993.83	5.31	1,993	1,986.00	2003.00	767,222
Child count	8.11	5.02	7	1.00	17.00	767,222
Father count	16.02	4.86	17	6.00	21.00	767,222
Mother count	13.51	6.07	14	2.00	21.00	767,222

Notes: The table displays the same statistics as table 2.2, but keeps all observations disregarding missing values.

Descriptives for Place of Birth

Table 2.7: Place of Birth

	All			Observed		
	Mother	Father	Child	Mother	Father	Child
ZH	90919	90177	126217	717	614	611
	(11.96)	(11.85)	(16.51)	(5.67)	(4.86)	(4.83)
BE	111358	112386	118208	2708	2790	3061
	(14.64)	(14.77)	(15.46)	(21.42)	(22.07)	(24.21)
LU	40922	40769	39260	1806	1869	2084
	(5.38)	(5.36)	(5.13)	(14.29)	(14.78)	(16.48)
UR	5621	5435	4033	35	34	6
	(0.74)	(0.71)	(0.53)	(0.28)	(0.27)	(0.05)
SZ	12742	12652	9883	97	68	27
	(1.68)	(1.66)	(1.29)	(0.77)	(0.54)	(0.21)
OW	4488	4622	3403	59	44	9
	(0.59)	(0.61)	(0.45)	(0.47)	(0.35)	(0.07)
NW	5146	4988	4394	85	75	33
	(0.68)	(0.66)	(0.57)	(0.67)	(0.59)	(0.26)
GL	5760	5733	4187	39	32	13
	(0.76)	(0.75)	(0.55)	(0.31)	(0.25)	(0.10)
ZG	7619	7473	13329	75	87	111
	(1.00)	(0.98)	(1.74)	(0.59)	(0.69)	(0.88)
FR	26541	27104	23294	117	114	41
	(3.49)	(3.56)	(3.05)	(0.93)	(0.90)	(0.32)
SO	26300	25985	23613	369	287	213
	(3.46)	(3.41)	(3.09)	(2.92)	(2.27)	(1.68)
BS	25162	25188	29056	412	413	436
	(3.31)	(3.31)	(3.80)	(3.26)	(3.27)	(3.45)
BL	8899	8971	15093	173	178	293

	(1.17)	(1.18)	(1.97)	(1.37)	(1.41)	(2.32)
SH	8250	8397	8670	52	60	15
	(1.08)	(1.10)	(1.13)	(0.41)	(0.47)	(0.12)
AR	7319	7417	4650	79	52	36
	(0.96)	(0.97)	(0.61)	(0.62)	(0.41)	(0.28)
AI	2612	2612	1525	19	15	1
	(0.34)	(0.34)	(0.20)	(0.15)	(0.12)	(0.01)
SG	55209	54262	53864	867	872	945
	(7.26)	(7.13)	(7.04)	(6.86)	(6.90)	(7.48)
GR	23722	24811	23501	145	114	74
	(3.12)	(3.26)	(3.07)	(1.15)	(0.90)	(0.59)
AG	49597	49087	51619	1851	1998	2318
	(6.52)	(6.45)	(6.75)	(14.64)	(15.80)	(18.34)
TG	21051	21431	19017	192	129	62
	(2.77)	(2.82)	(2.49)	(1.52)	(1.02)	(0.49)
TI	14434	15333	22156	35	32	25
	(1.90)	(2.01)	(2.90)	(0.28)	(0.25)	(0.20)
VD	34798	37070	49605	177	183	170
	(4.58)	(4.87)	(6.49)	(1.40)	(1.45)	(1.34)
VS	30638	31778	30348	589	597	568
	(4.03)	(4.18)	(3.97)	(4.66)	(4.72)	(4.49)
NE	13125	13748	16856	123	133	103
	(1.73)	(1.81)	(2.20)	(0.97)	(1.05)	(0.81)
GE	11836	12549	25164	369	395	864
	(1.56)	(1.65)	(3.29)	(2.92)	(3.12)	(6.83)
JU	9688	9946	7684	68	81	32
	(1.27)	(1.31)	(1.00)	(0.54)	(0.64)	(0.25)
SouthEastern EU	19008	18947	9986	226	217	115
	(2.50)	(2.49)	(1.31)	(1.79)	(1.72)	(0.91)
Africa	4969	5167	2836	83	95	39

	(0.65)	(0.68)	(0.37)	(0.66)	(0.75)	(0.31)
America	4940	2630	4120	71	60	42
	(0.65)	(0.35)	(0.54)	(0.56)	(0.47)	(0.33)
Canada/USA	1851	1066	2285	30	17	30
	(0.24)	(0.14)	(0.30)	(0.24)	(0.13)	(0.24)
Asia/Australia	7825	5780	6204	93	63	75
	(1.03)	(0.76)	(0.81)	(0.74)	(0.50)	(0.59)
NortWestern EU	68143	67438	10627	880	924	190
	(8.96)	(8.86)	(1.39)	(6.96)	(7.31)	(1.50)
Total	760492	760952	764687	12641	12642	12642
	(100.00)	(100.00)	(100.00)	(100.00)	(100.00)	(100.00)

Notes: The table displays the total number of observations with frequencies in brackets. "All" includes all observations regardless of missing values in some covariates used in the analysis. "Observed" includes only observations for which all covariates are non-missing and hence are included in the analysis. Since not all cantons participated in the WiSiER, frequencies in place of birth do not represent the Swiss population.

Descriptives for Father's Income Share

The findings shown in figure 2.8 reveal a traditional pattern in family structures, where the majority of fathers provide a higher share of the total parental income. Only for a few observations, the father's income share falls below 50%, and only about 2% exhibit a share below 40%. Additionally, a large number of mothers work part-time. (A typical full-time employment in Switzerland amounts to around 40-45 working hours per week.) Figure 2.9 illustrates that, on average, the father's income share is higher when mothers work fewer hours.

It is worth mentioning that observations used to create these figures only partly cover those included in the main analysis above. This is because average hours worked are sparsely observed, thus only included for children. Nevertheless, it can be expected that the findings remain the same for the observations included in the main analysis. There is no reason why the hours worked by parents would not be systematically (un-)observed.

Figure 2.8: Histograms

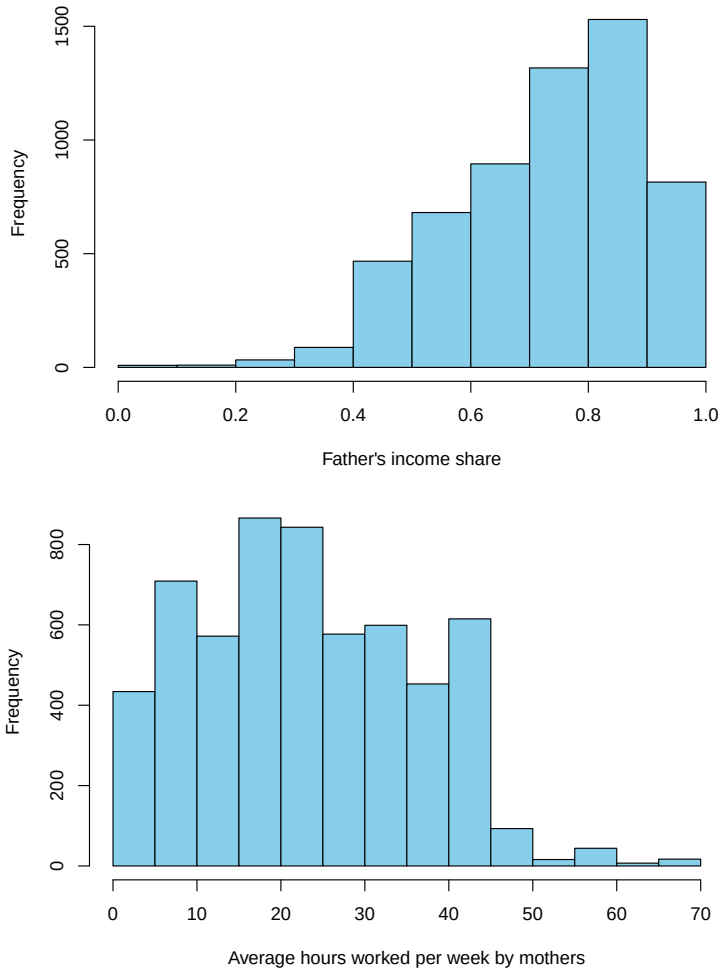
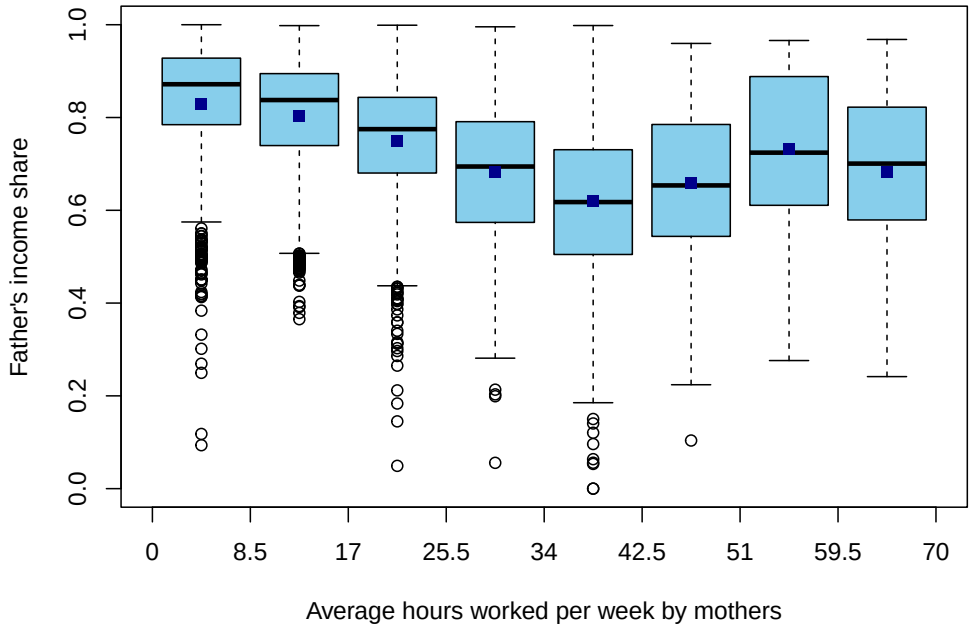


Figure 2.9: Boxplot



Notes: Average hours worked per week for mothers is categorized into eight categories. The x-axis displays the borders of these categories. The dark blue squares indicate the average father's income share for each category.

2.A.2 Detailed Description of the PSID Data

Construction of the PSID Data

Summary of Available Variables

Table 2.8 summarizes most variables available in the PSID, which are used to construct the PSID data. Further, it lists important properties of the variables and specific cleaning steps. Variables used for the identification of individuals or the assignment of roles in the family are not included in the table.

Table 2.8: Variable Descriptions, Availability, and Notes

Type	Variable Description	Availability	Notes & Cleaning steps
Wage income	Head's wage income (no bonuses, overtime, etc.)	1968 to 2019	
Labor income	Head's Labor Income (Wages plus other income, e.g., farming, business income, bonuses, overtime, tips)	1968 to 2019	The variable was capped, and 0 means no income.
	Spouse's Labor Income (Labor income from work)	1968 to 1993	A value of 0 means either no Spouse/Wife or no income from Spouse/Wife.
	Spouse's labor income excluding farm income and the labor portion of business income	1993 to 2019	A value of 0 means either no Spouse/Wife or no income from Spouse/Wife.
Family Income	Total family money income (Taxable income plus other transfer income)	1968 to 2019	The variable was capped at the bottom at 1 until 1993. Values of 0 are coded as missing.
Other income	Spouse's business income (Labor portion of business income)	1993 to 2019	A value of 0 means either no Spouse/Wife, no or negative business income, or the business was incorporated.
	Head's and Spouse's Income from Farming (Labor and asset portions)	1993 to 2019	Values are sometimes capped. A value of 0 means a broke even or not a farmer.
	Head's labor part of farm income	1970 to 1993	In the first years, reported as bracketed values.

Education	Indicator for college completion (head and spouse)	1975 to 2019	A value of 0 means no college or educated outside the US. From 2011 on, a value of 0 was also used if "NA, RF where head received education (ER51912=9)" and "NA, DK, RF whether attended college (ER51924=8 or 9)". Missings were replaced as follows: If an individual had no college degree in a certain year, then for all years prior missings are replaced as no college degree until a non-missing is met. If an individual had a college degree in a certain year, then for all years after, missings are replaced as a college degree until a non-missing is met. Values of 0 are kept as no college degree even though for later years, they started to mix other things in it, e.g., educated only outside of the US.
	Highest degree obtained (head and spouse)	1985 to 2019	A value of 0 means no college or educated outside the US. 2011, also "NA, RF where Head received education (ER57668=9)" and "DK, NA, or RF whether attended college (ER57680=8 or 9)". Missings were replaced as follows: If an individual had no college degree in a certain year, then for all years prior missings are replaced as no college degree until a non-missing is met. If an individual had a college degree or higher in a certain year, then all years after missings are replaced with that particular value until a non-missing is met. Values of 0 are kept as no college degree even though for later years, they started to mix other things in it, e.g., educated only outside of the US.

	Father's education (head)	1968 to 2019	Different coding schemes over time. 968; 1 is 0-5 grades, or DK, and could not read or write (Sidenote: Only a small fraction is coded as 0 in 1970. From 1970 - 2019; 0 is "Could not read or write, DK grade and could not read or write", 1 is 0-5 grades; illiterate. Furthermore, 0 sometimes denotes "Inap.: could not read or write; NA, DK grade and could not read or write". However, 0 is taken as "could not read or write" in those cases since there is always a category for NA (e.g. 9 or 99). The error of changing values across time was addressed by replacing values with each individual's mode. In the case of ties, the average between the smallest and largest mode was taken if they were no further apart than 2 categories. Otherwise, missings are generated.
	Mother's education (head)	1974 to 2019	Same coding scheme as father's education. Same cleaning as for father's education.
Food expenditures	Food at home, eating out, delivered food, food stamps (yearly)	Various years	All variables regarding food expenditures are constructed following the work by Blundell et al. (2008).
Age	The actual age at interview	1968 to 2019	Missing values were addressed by creating a new age variable as the difference between the interview year and the cleaned year of birth.
	The actual age of the reference person	1968 to 2019	
	Year of birth	1983 to 2019	The error of changing values across time was addressed by replacing values by the mode of each individual. In the case of ties, the average between the smallest and largest mode was taken if they were no further apart than 2 years. Otherwise, missings are generated.
	Month of birth	1983 to 2019	
Other	Year of first child's birth	All years	Missings are addressed by using the mode.

	Marital status (head)	1968 to 2019	No distinction between legally married and cohabiting.
	Race (head and spouse)	1968 to 2019 (head), 1985 to 2019 (spouse)	The error of changing values across time was addressed by replacing values by the mode of each individual. In the case of ties, the value is set to missing.
	Sex	All years	Missings are addressed by using the mode.
	Employment status (individual)	1979 to 2019	
	Father's occupation (head)	1970 to 2019	Variable changes frequently in 4-digit census occupation codes. It has been very complicated to compare over the years, and interpretations might remain unclear.
	Working hours as total annual working hours on all jobs for money (head and spouse)	1968 to 2019	A value of 0 means either no work or no spouse.
	Living state at the time of interview	1968 to 2019	
	Grewup location as urbanization categories of the place where the individual grew up	1968 to 2019 (head); 1976, 1985, 2009-2019 (spouse)	The error of changing values across time was addressed by replacing values by the mode of each individual. In the case of ties, the value is set to missing.
	Household size (actual number of persons in the family unit)	1968 to 2019	
	Number of children (individuals below 18 within the family unit, not necessarily actual children)	1968 to 2019	

Detailed Steps

In the first step, an indicator variable is created using the fims-data provided by the PSID. The variable links children to their (biological or adoptive) parents such that for each child identification number, a biological or adoptive father and/or mother identification number is created. The fims-data was chosen to be balanced, meaning that Individuals for which no such linkage was possible, i.e., for which no parents are in the data, are dropped. The number of observations is 71,241. These indicators are stored in a separate file. In a second step, all variables described in table 2.8 are downloaded from the PSID and loaded using the psidtools by Kohler (2023).

With all information available, the food expenditure variables are constructed following Blundell et al. (2008). Then, the data is restructured using the child and parents linkages indicator variables, such that one row comprises the information of one child and its parents within a specific wave. From the approximately 70,000 children for which a parent linkage indicator was created, about 19,000 could not be matched to any observation with any of the variables chosen, leaving us with about 51,000 children to start with. This is likely a result of the specific variables used. The PSID provides further variables and supplement data for which both children and a parent would be observed.

Some further data cleaning steps are similar across multiple variables and hence not listed in table 2.8. As suggested by the PSID, the sequence number is used to assign the correct values from the variables to the individuals. Except for the year 1968, the sequence number is always 1 if the individual is the head of the family unit and 2 if the individual is the spouse. For the year 1968, the variable "relation to the head" was used. For variables that record information about the head or spouse, the sequence number is used to assign values to an individual whenever the individual takes on the corresponding position. Other examples are adjustments to the relatively frequent changes in the PSID values. Categories or capped thresholds are frequently adjusted over the years, altering the same variable across years. Whenever possible, adjustments were made to allow for consistency across years. Details of these changes are openly accessible on the PSID website.

Ultimately, income variables need to be constructed for the analysis. The easiest accessible income variable is total family income, which is available across all years. Two additional variables were created to scale total family income according to the household size. The two equivalence scaling approaches implemented are the Oxford equivalence scaling and the Root scaling. However, total family income does not allow for the analysis of income for mothers and fathers separately. In order to conduct such an analysis, the focus was set on labor income, which in general is highly available too. For the head of a family unit, wage income and labor income, comprised of wage income plus other things like farming, business, bonuses, overtime, etc., are available for all years. Unfortunately, a similar labor income variable for the spouse

was only recorded until 1993. After 1993, the spouse's labor income is measured excluding farm income and the labor portion of business income. While business incomes continue to be easily available, the information on farming income is not. Thus, a comparable total labor income variable for the spouse over all years has to be constructed first. We proceed in two steps. First, a new variable for the spouse is created using total labor income until 1993 and the sum between labor income without farming and business income plus business income after 1993. In the second step, values from this new variable and the head's total labor income are set to missing whenever the family potentially generates farming income. Using the two variables on farm income, a new indicator variable is constructed, which is one for the whole family and all years whenever farming income within the family is detected for at least one year. That is, families for which farming income variables of the head or spouse for any year were nonzero are excluded, which amounts to about 760 individuals from the sample. Finally, it is important to note that for most income variables, income is only measured for those years in which the individual is the head or spouse in a family unit. However, total family income is also recorded for individuals who are still children living with their parents. Hence, the family income information of an individual of a particular year has to be dropped if the individual's sequence number does not indicate whether the individual is a head or spouse. For the other income variables, the property is held by the construction of the PSID. Lastly, all income variables are deflated using the CPI.

Consequently, we end up with a second data file that contains information for individuals across some years. If all years for which no sequence number is available are dropped. The number of individuals is 82,573, which is slightly less than if we used the unbalanced fims-data 85,542. Since each individual is potentially observed over multiple years, the total number of observations then is 904,796. However, the number of observations drastically vanishes once we focus on income information for parents and children. Of all individuals, 52,138 are at least once either the head or the spouse, and of those, 29,906 are children of parents available in the identifiers data from the PSID. However, for 10,308 children for which a link to parents is available in the identifiers data, no information is available regarding the chosen values, i.e., the parents are not recorded in the downloaded PSID data. Furthermore, for 7,373 children, only one parent is recorded, leaving us with 12,225 children for which we have at least two parents' information available.

Preprocessing the PSID Data

For the analysis, we focus on total labor income excluding farming as an outcome of interest, and in the following, total labor income excluding farming is simply referred to as income. Furthermore, we focus on biological parents.

Regarding the outcome of interest, we drop the highest and lowest 1% of yearly income to

avoid unrealistic values. Given the focus on biological parents and if we drop further missing values in the income variable, we end up with 10,258 child-parents observations. However, these include many children of young age with an income of 0. If we restrict the sample to only include income from individuals with age 30 to 50, the number of child-parents observations for which income is available drops to 6,041.

Regarding the covariates, some variables specific to the PSID interview, like sequence numbers or family relations, are dropped. Further, the information regarding the occupation of the father was dropped as well. Even though the variable is available across multiple years, the corresponding index changed four times, such that information is not comparable across years.

For time-invariant variables, the mode was applied to extend the information across all years to ensure that no information is lost. For most covariates, we are only interested in values resulting from the years for which income is measured as well. This is particularly the case for variables varying across time for example, working hours. Hence, we set all values to missing for those years in which no income is available.

Finally, the information of the biological mother and father is aggregated to a parent variable, if sensible. Variables for which creating a parent aggregate may not be particularly meaningful are sex, residence at the time of the interview, employment status, marital status, the location where the individual grew up, and race. For income as well as hours worked, the values are summed up. For all educational variables, the maximum was taken to focus on the highest education available within a family. For all other parent variables, the average values of the father and mother are used.

Descriptive Statistics for All Observations

Table 2.9: Descriptive Statistics for All Observations

	Mean	SD	Median	Q5	Q95	N
Child's Income	15,960.58	12,024.99	13,917	0.00	40173.50	6,041
Parents' Income	28,723.65	14,549.10	27,388	7,383.73	55415.29	6,041
Income share	0.75	0.21	1	0.38	1.00	6,031
Daughter	0.53	0.50	1	0.00	1.00	6,041
Race	1.44	0.88	1	1.00	2.00	5,912
Child Urbanization	2.36	0.65	2	1.00	3.00	5,814
Child's age	35.55	3.45	36	30.00	40.13	6,041
Parent's age	41.14	3.55	40	36.11	47.50	6,041
Income measurement age	15.82	4.61	16	8.00	23.00	6,041
Parents' no. of children	2.45	1.57	2	0.66	5.78	6,041
Parent's college degree	0.24	0.41	0	0.00	1.00	6,029
Mother's work-hours	1,027.89	708.63	1,041	0.00	2053.67	6,041
Father's work-hours	2,086.34	608.26	2,112	853.33	2955.86	6,041
Child's year	2,005.45	10.60	2,008	1,988.55	2019.00	6,041
Mother's year	1,984.94	10.49	1,985	1,971.00	2001.77	6,041
Father's year	1,982.86	10.36	1,982	1,969.00	2000.00	6,041
Child count	7.47	5.20	6	1.00	17.00	6,041
Mother count	14.22	5.26	15	4.00	21.00	6,041
Father count	12.67	6.06	13	2.00	21.00	6,041

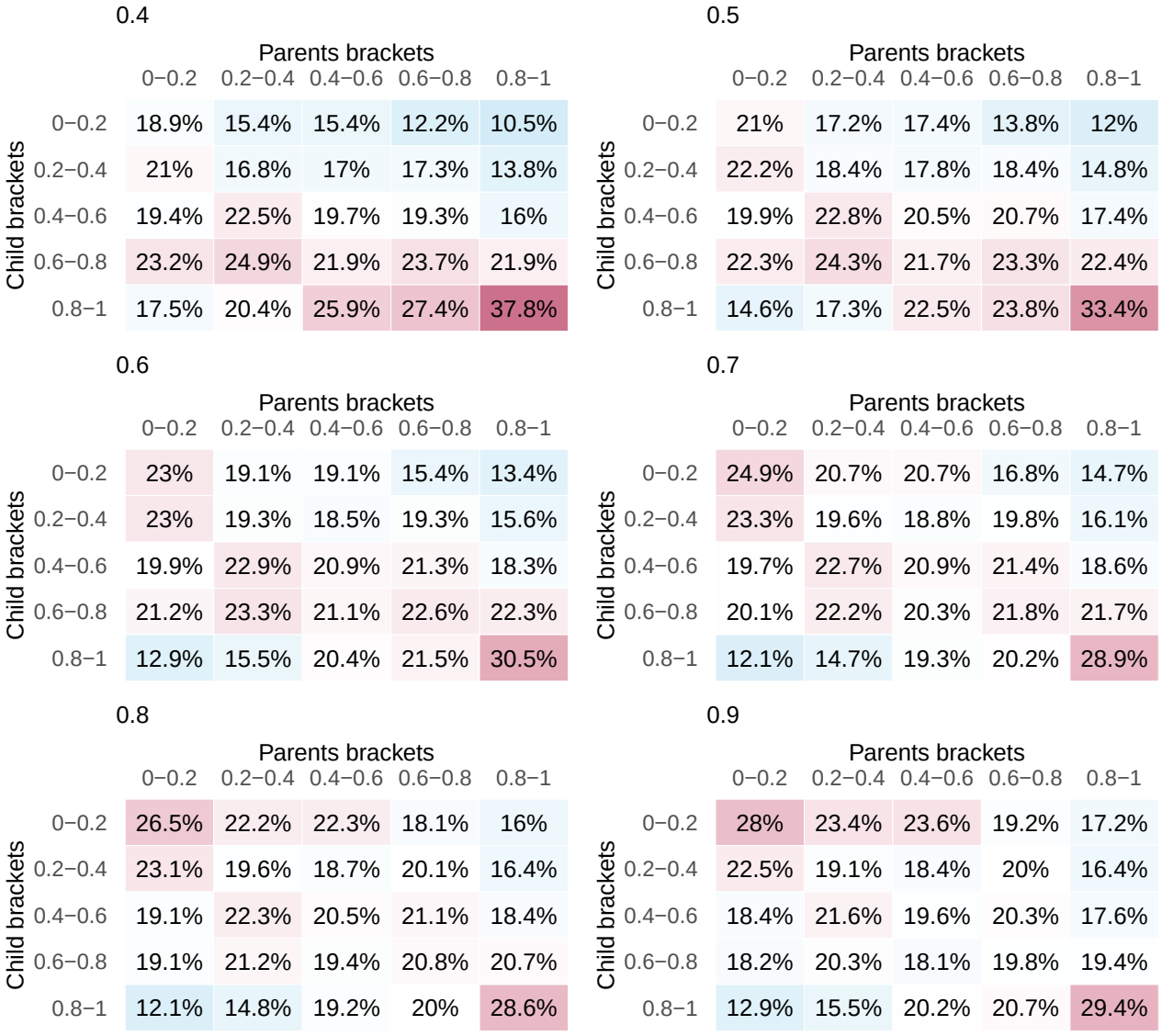
Notes: The table displays the same statistics as table 2.4, but keeps all observations disregarding missing values.

2.B Appendix II - Supplementary Mobility Results

2.B.1 Extended Values for Father's Income Share

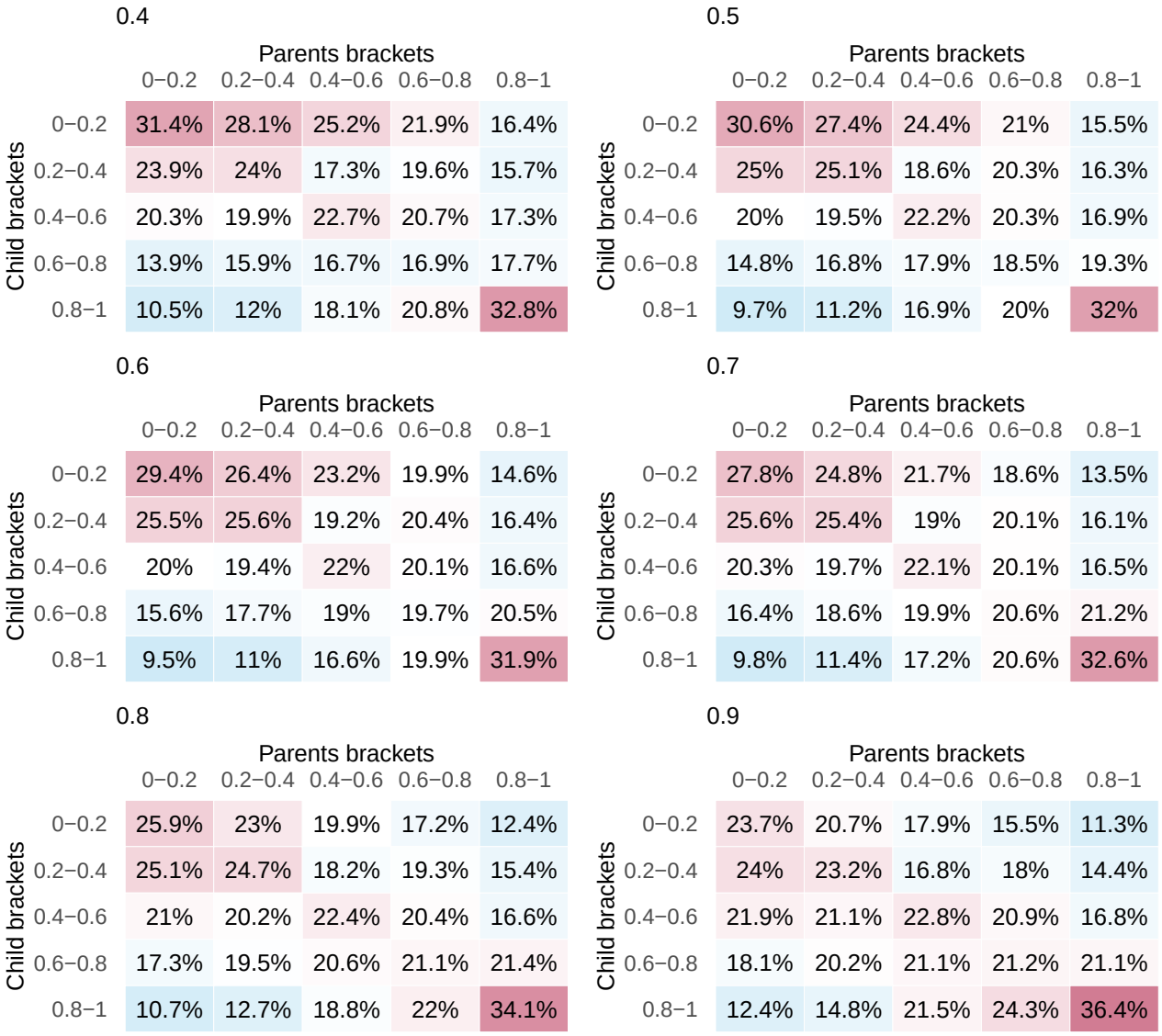
WiSiER

Figure 2.10: Daughters' Detailed Results of Father's Income Share *(WiSiER)*



Notes: The ranks are computed using the income distribution of daughters. Each matrix displays results for different father's income share. For example, the first matrix displays results if the father's income share is 40%.

Figure 2.11: Sons' Detailed Results of Father's Income Share *(WiSiER)*



Notes: The ranks are computed using the income distribution of sons. Each matrix displays results for different father's income share. For example, the first matrix displays results if the father's income share is 40%.

PSID

Figure 2.12: Daughters' Detailed Results of Father's Income Share *(PSID)*

Notes: The ranks are computed using the income distribution of daughters. Each matrix displays results for different father's income share. For example, the first matrix displays results if the father's income share is 40%.

Figure 2.13: Sons' Detailed Results of Father's Income Share *(PSID)*

Notes: The ranks are computed using the income distribution of sons. Each matrix displays results for different father's income share. For example, the first matrix displays results if the father's income share is 40%.

2.B.2 Exploring Covariate Variations and Transition Matrix Adjustments

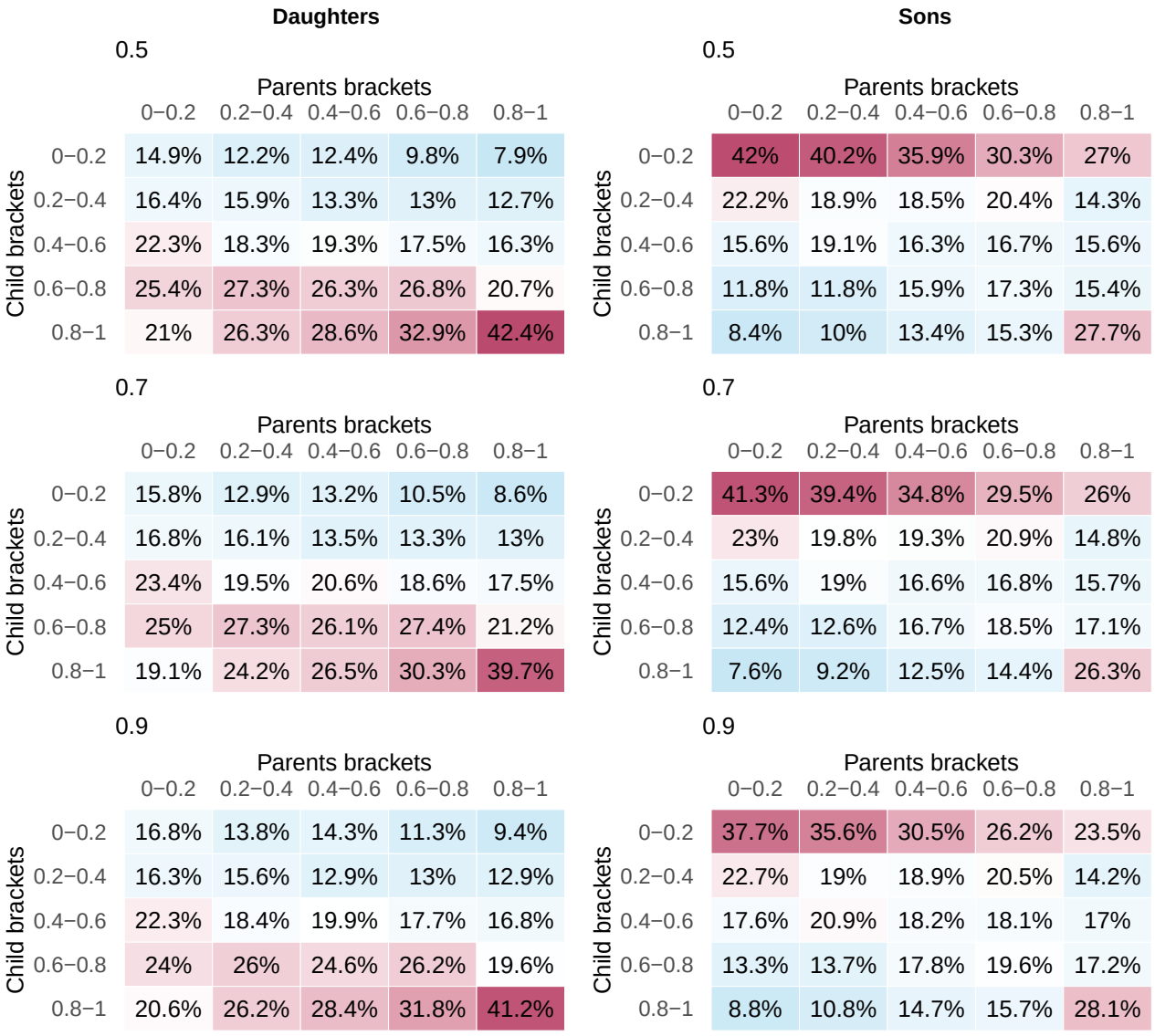
WiSiER

In this section, we analyze how the choice of covariates and definition of the brackets for the transition matrices alter the appearance of the results. Thereby, we follow the analysis regarding the father's income share using the WiSiER data from section 2.4.

Figure 2.14 displays the results, including all covariates. The transition matrices display transition probabilities for Sons and Daughters in their different income distribution. By controlling for variables like working hours, the estimated counterfactual distributions become more similar to the population's distribution, including all children. Hence, transition matrices based only daughters income distribution, exhibit higher probabilities at the top. The opposite is true for sons. Consequently, even though patterns from the main findings are still present, they are less visible. Figure 2.15 captures the results if, instead, transition matrices are based on the population's distribution. Compared to figure 2.14, the results are reverted but less strongly skewed. The reversal stems from the fact that, as seen in section 2.4, some part of the income gap between sons and daughters is not captured by the available covariates. The results from the main analysis are better visible now. However, for daughters (sons), higher probabilities are estimated for the bottom (top) of the children's income distribution regardless of the parents' income.

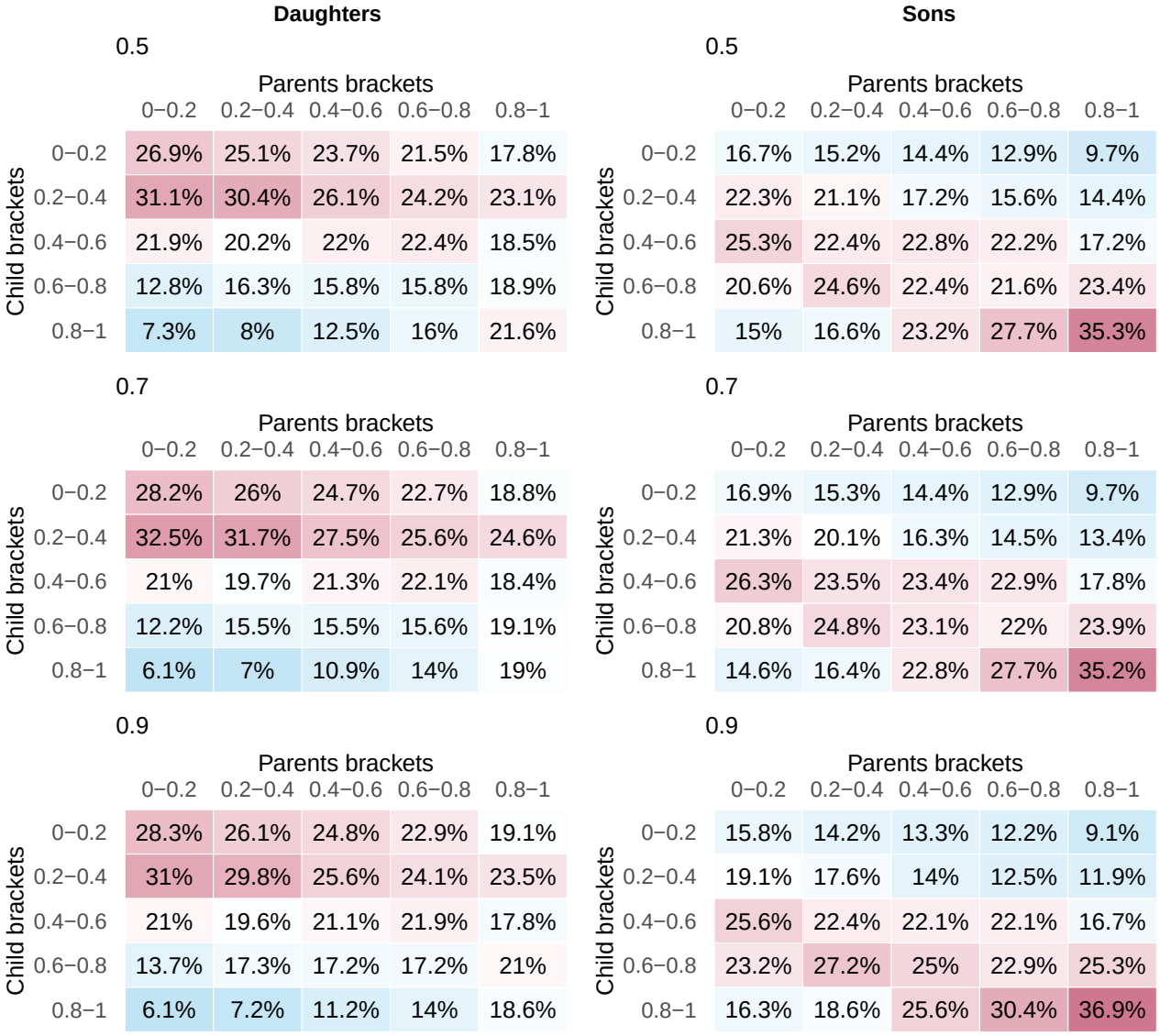
Figures 2.16 and 2.17 display transition matrices that are based on the estimated counterfactual distributions. For example, the first transition matrix represents mobility among daughters whose fathers contribute half to the total parents' income and not among daughters in general. In both cases, there is no noticeable difference in mobility between sons and daughters or between distinct values of the father's income share. This suggests that the father's income share does not affect mobility within subgroups. While different income shares may lead to advantages or disadvantages compared to children with other experiences, mobility remains unchanged when comparing children with identical characteristics.

Figure 2.14: Mobility Results including Endogenous Variables



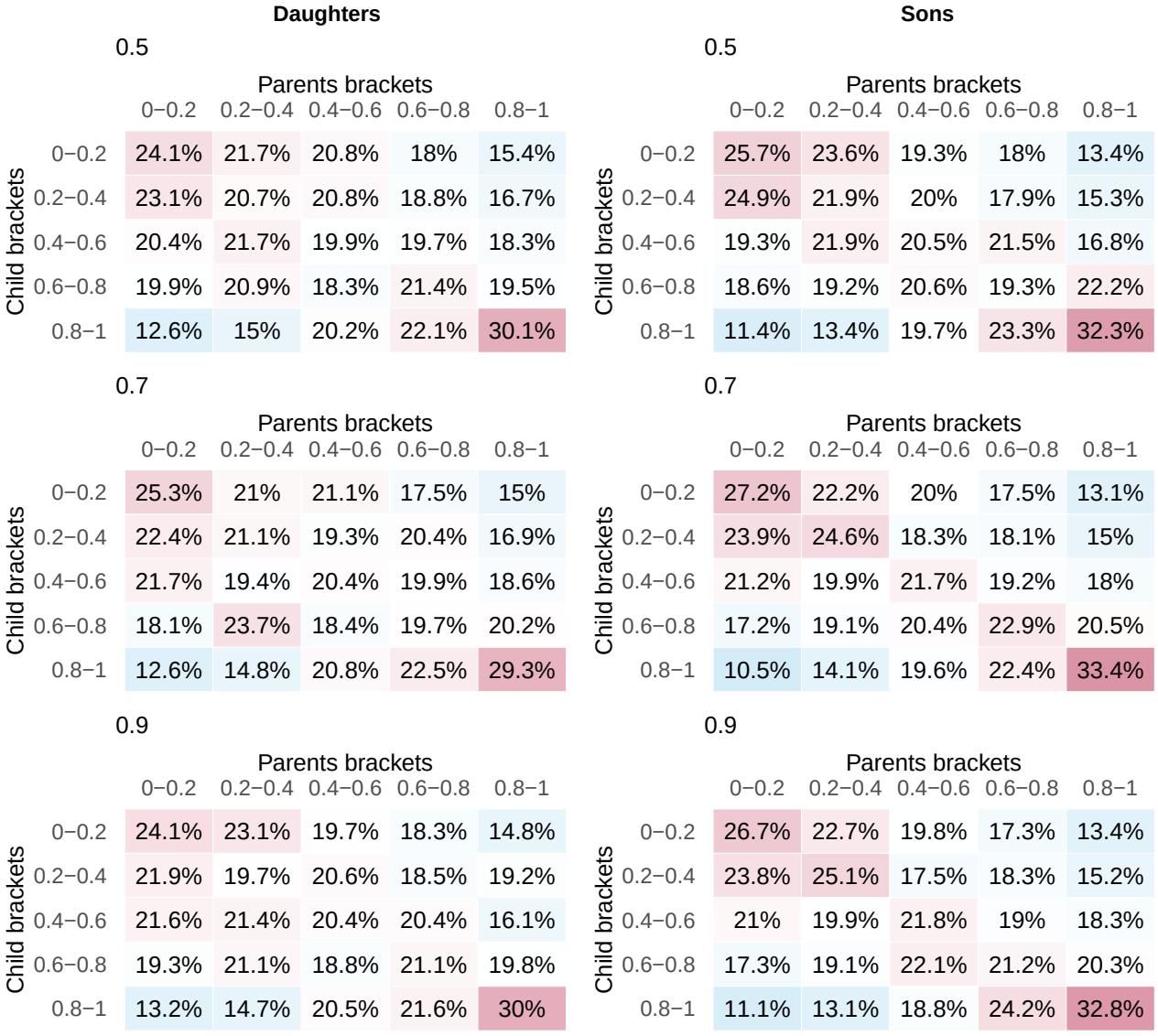
Notes: Compared to the main results in section 2.4, the endogenous variables for children's Education, Hours worked per week, and Years worked are included for estimation. Everything else remains the same. The left column displays results for daughters, and the right column displays results for sons. The ranks are computed by the respective subsample, i.e., the brackets from the transition matrices in the left column result from the daughter's income distribution. Each row displays results for different father's income share. For example, the first row displays results if the father's income share is 50%.

Figure 2.15: Mobility Results based on Population Distribution including Endogenous Variables



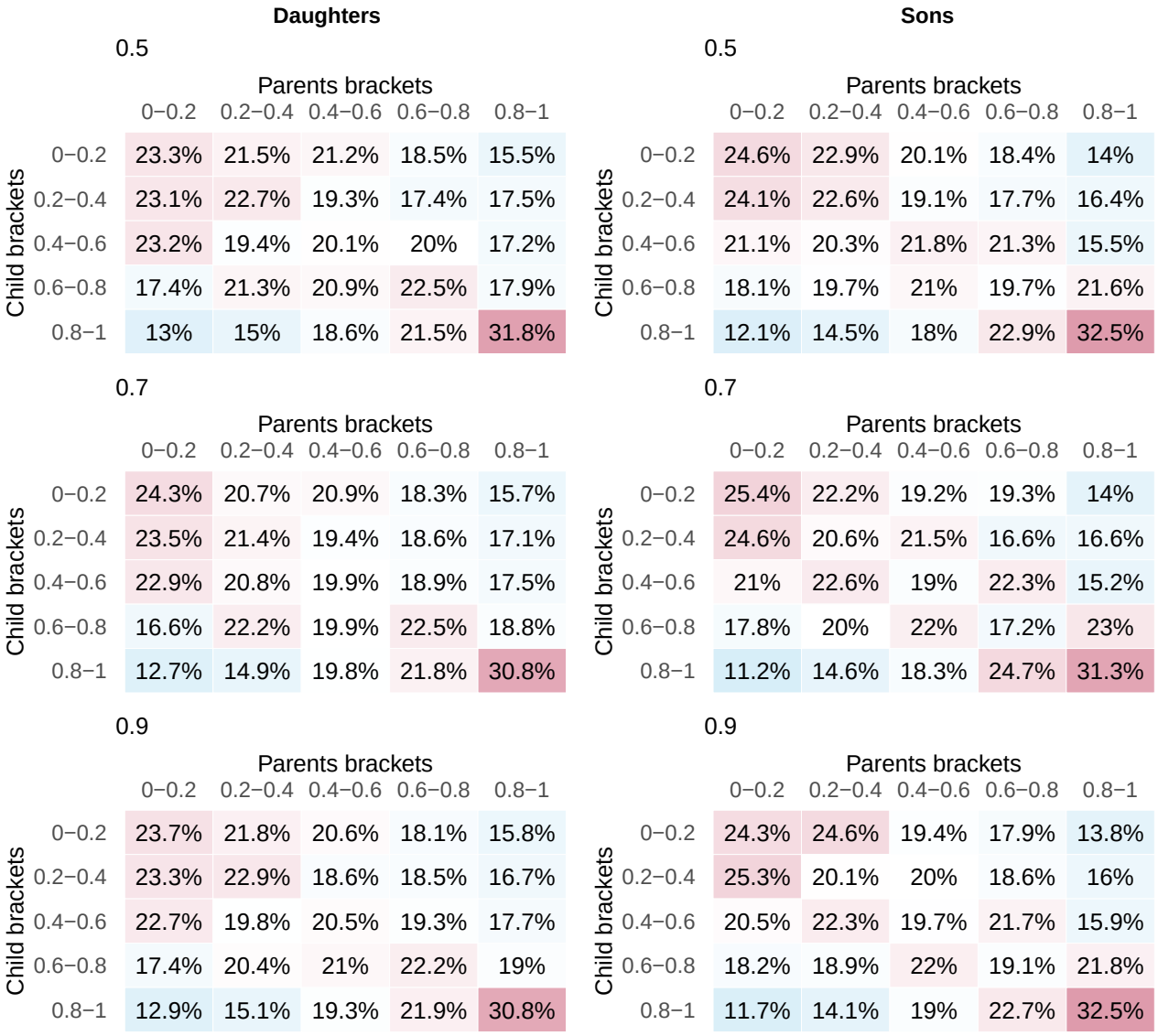
Notes: Compared to the main results in section 2.4, the endogenous variables for children's Education, Hours worked per week, and Years worked are included for estimation. Additionally, brackets in the transition matrices represent the overall income distribution of children (daughters and sons). Everything else remains the same. The left column displays results for daughters, and the right column displays results for sons. The ranks are computed by the respective subsample, i.e., the brackets from the transition matrices in the left column result from the daughter's income distribution. Each row displays results for different father's income share. For example, the first row displays results if the father's income share is 50%.

Figure 2.16: Mobility Results based on Counterfactual Distribution



Notes: Compared to the main results in section 2.4, the brackets in the transition matrices represent the counterfactual income distribution. For example, the first transition matrix represents mobility among daughters whose fathers contribute half to the total parents' income. Everything else remains the same. The left column displays results for daughters, and the right column displays results for sons. The ranks are computed by the respective subsample, i.e., the brackets from the transition matrices in the left column result from the daughter's income distribution. Each row displays results for different father's income share. For example, the first row displays results if the father's income share is 50%.

Figure 2.17: Mobility Results Based on Counterfactual Distribution including Endogenous Variables



Notes: Compared to the main results in section 2.4, the endogenous variables for children's Education, Hours worked per week, and Years worked are included for estimation. Additionally, the brackets in the transition matrices represent the counterfactual income distribution. For example, the first transition matrix represents mobility among daughters whose fathers contribute half to the total parents' income. Everything else remains the same. The left column displays results for daughters, and the right column displays results for sons. The ranks are computed by the respective subsample, i.e., the brackets from the transition matrices in the left column result from the daughter's income distribution. Each row displays results for different father's income share. For example, the first row displays results if the father's income share is 50%.

PSID

Figure 2.18 displays the same results as figure 2.5 presented in section 2.4, if all covariates described in section 2.3 are included in the analysis. Furthermore, the grid used to estimate the marginal and joint distributions is reduced to eleven points. Using a higher grid results in more volatile results, likely due to the smaller sample of the PSID.

The results in figure 2.18 yield the same interpretation and similar transition probabilities as presented above. One smaller difference is the transition probability for sons with parents in the second lowest bracket to reach the top of their income distribution. In contrast to the results presented above, the probability of reaching the cell decreases with a higher father's income share. However, the shift seems to result from a movement from the highest child bracket to the second highest child bracket. Summing up, the two cells still point to an increase in upward mobility with the father's income share.

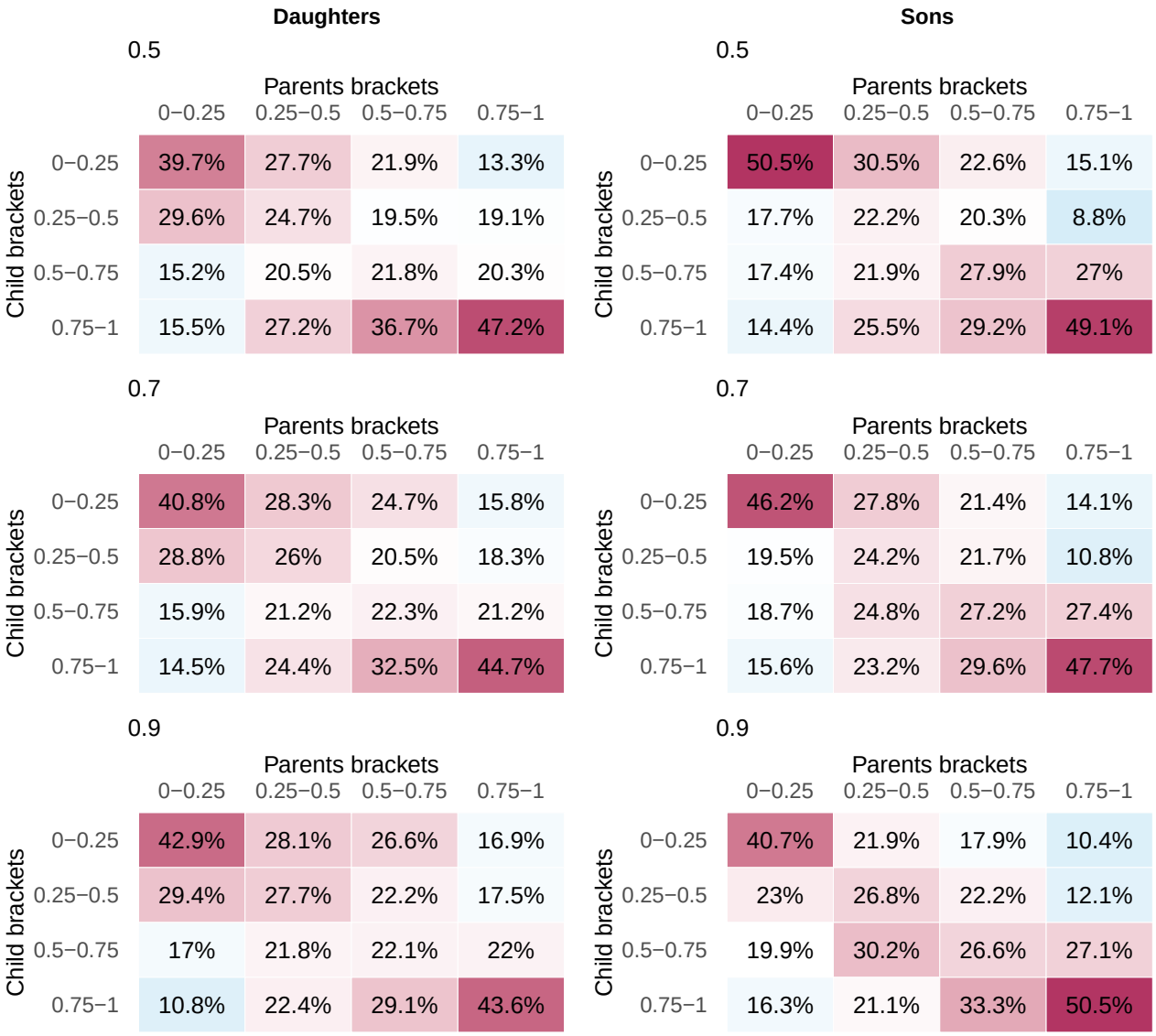
2.C Appendix III - Remarks on Intergenerational Mobility

The following section summarizes the theoretical framework developed to explain intergenerational mobility. The summary builds on work by Corak (2020), Björklund and Jäntti (2011), Corak (2013), Mogstad and Torsvik (2023), and Cholli and Durlauf (2022). As described in Corak (2020), the model by Becker and Tomes (1979) and Becker and Tomes (1986) has been the "workhorse model" providing the theoretical foundation of intergenerational mobility. It captures the broad idea that labor market outcomes of children, and hence mobility, are driven by human capital. Human capital, as defined in Mogstad and Torsvik (2023), is knowledge, skills, and attitudes that are acquired and that labor markets value. The Becker and Tomes model directly imposes a relationship between human capital and labor market outcomes, and both inherited endowments and investments in human capital. Endowments, encompassing cultural and genetic attributes passed from one generation to the next, are assumed to be immutable. However, recent research challenges this assumption, highlighting that while inherited genes impose a limit on the amount of human capital, other aspects of endowments are acquirable, such as cognitive and non-cognitive skills. Hence, the human capital a child brings to the labor market is less the result of a simple combination of endowments and investments and more a process where endowments and investments complement and determine each other.

The main driver of the human capital of a child is the investments in it. Investments in human capital are either made by parents or the public. Typically these investments have been thought of as education. However, the labor market also values other non-cognitive skills which are less related to educational achievements.

Investments by parents are constrained by the available resources, including both monetary

Figure 2.18: Mobility Results including All Variables



Notes: Compared to the main results in section 2.4, all covariates are included for the analysis. Everything else remains the same. The left column displays results for daughters, and the right column displays results for sons. The ranks are computed by the respective subsample, i.e., the brackets from the transition matrices in the left column result from the daughter's income distribution. Each row displays results for different father's income share. For example, the first row displays results if the father's income share is 50%.

and non-monetary resources. Typically, parents with high incomes have more resources and, hence, are more likely to invest. Non-monetary resources are typically linked to socioeconomic status, encompassing factors such as parental age or experience, the amount of time parents can invest in their children, educational attainment of parents, whether both biological parents live together or not, access to labor market information and networks, and more. According to the Becker-Tomes model, parents invest optimally in their children if they are not credit-constrained. However, if parents face a budget constraint, their investment level is below the optimal amount. This can result in nonlinearities, especially in the extremes of the income distribution. Poorer parents lack the income necessary to invest optimally, while richer parents have children with very high optimal investment levels that may not be feasible to attain.

Parents' decision on how much to invest in their children is not solely dependent on the availability of resources. A crucial determinant for parents making investment decisions is the returns on investments, as they want to invest an optimal amount. These returns can be thought of in two ways. Firstly, as the amount of additional human capital generated by each unit of investment, i.e., marginal return of investments. Secondly, as the amount of additional income generated by each unit of human capital, i.e., marginal return of human capital. For example, if a child possesses superior endowments, the marginal return on investments is higher. Assuming parents with a higher income have children with superior endowments leads to richer parents investing more, resulting in further persistence. Moreover, refined models, as proposed by Becker, Kominers, et al. (2018) and Solon (2004), suggest that higher labor market inequalities within a region incentivize parents to invest more, as the return on human capital is higher. High-income parents have more resources at their disposal to invest in their children's human capital, which in turn leads to greater persistence.

There are two aspects to consider. Firstly, the first example introduces a novel channel in which inherited endowments impact the marginal return of investments, creating a self-reinforcing cycle. Superior inherited endowments increase marginal returns of investments, leading to higher investments. However, this, in turn, enhances the part of endowments that can be altered, further increasing marginal returns. The literature on child development describes this channel in more detail. Socioeconomic status is recognized as an important factor that influences the labor market outcomes of children. The influence starts early in life, as it affects the endowments, for example, health or cognitive and non-cognitive skills, a child is born with. These endowments determine the returns on investments, which in turn affects investments. Available resources and the quality of the neighborhood and schools further shape early school outcomes, which then again determine later educational outcomes. (For an extensive review of the child development literature, see, for example, Knudsen et al., 2006 or Heckman and Mosso, 2014) Secondly, labor market inequalities are closely intertwined with the marginal return of human capital. This creates yet another circular relationship. Higher market inequalities

incentivize parents to invest more in their children, resulting in richer parents investing more, leading to a decline in mobility. This, again, leads to an increase in market inequality, which completes the circle.

Several other factors have been argued to impact intergenerational mobility. For example, public investments, social influences or timing of investments. Public Investments in human capital describe a pool of programs and opportunities to foster human capital available for everybody. How much the public is able to invest depends to some degree on the resources available to the parents living in the region. Choice-theoretical models, as proposed by Durlauf (1996) and Durlauf and Seshadri (2018) assign a critical role to social influences in shaping labor market outcomes, also describing 'poverty traps'. Labor market outcomes and the acquisition of human capital are directly affected by social influences, such as peer effects, role models, formation of personal identity, and more. Influential work by Chetty and Hendren (see, e.g., Chetty, Hendren, Kline, et al., 2014; Chetty and Hendren, 2018a; Chetty, Hendren, and Katz, 2016) has pushed differences in intergenerational mobility across regions within a country into focus. Possible contributing factors are differences in families' available resources and social influences, divergent public investments, and distinct labor market structures across regions. Recent literature highlights that not only the total amount invested but the timing of investments is crucial. Investments have varying effectiveness depending on the stage of childhood. For example, research by Carneiro et al. (2021) indicates that investments when children are between 6 and 11 years old are relatively less important compared to investments made during the early and later stages of childhood.

Embedded in this theoretical framework are the frequent discussions of Nature vs. Nurture and Family vs. Environment. The first discussion addresses the question of to what extent labor market outcomes are determined by inherited factors that can't be altered by investments. While results in this area are mixed and highly discussed, they suggest that the circumstances a child grows up in play an important role. The second discussion then debates how strongly labor market outcomes are determined by family circumstances or other circumstances. This interplay between family and environment is a crucial driver of intergenerational mobility, which has already been addressed above. The resources available, inherited endowments, and social influences are all characteristics associated with a family and affect the final labor market outcomes. For example, Lindbeck et al. (1999) discuss how social norms impact labor market attachment or Dahl et al. (2014) find that children reduce their labor market participation if their parents received disability benefits. To summarize, the family culture and the parent's own outcomes and goals have a notable impact on labor market outcomes. The environment, comprising public investments, available resources, returns to investments, and social influences as characteristics, also plays a vital role in intergenerational mobility. For instance, Havnes and Mogstad (2015) finds that the provision of high-quality childcare programs increases educational

and labor market attainment of children from low- and middle-income families in the long run. It is important to note that parents often have some control over the environment in which they raise their children, which can then result in neighborhood sorting.

Chapter 3

Using Natural Language Processing to Identify Monetary Policy Shocks

Abstract

Identifying the causal effects of monetary policy is challenging due to the endogeneity of policy decisions. In recent years, external instruments – particularly high-frequency monetary policy surprises – have been increasingly used for identification. However, market-based surprises around Federal Open Market Committee announcements often suffer from weak relevance and endogeneity. This paper improves upon these measures by incorporating policy-relevant speeches from Federal Reserve Board members and improving the relevance. Using state-of-the-art Natural Language Processing techniques, we predict changes in market expectations based on the text of Federal Open Market Committee statements and Federal Reserve speeches, isolating the component of surprises driven solely by central bank communication. Our results suggest that these language-driven monetary policy surprises mitigate endogeneity concerns and align more closely with economic theory than traditional market-based measures.

Acknowledgment: Joint work with Alexandra Piller and Larissa Schwaller. We are grateful to the University of Bern for their financial support. We thank Luca Benati, Martin Brown, Refet S. Gürkaynak, Peter Karadi, Daniel Kaufmann, Costanza Naguib, Galo Nuño, Oliver Pfäuti, Ulf Söderström, Dalibor Stevanovic, Xin Zhang and the Data Science Lab of the University of Bern for their comments. We also greatly benefited from comments from seminar participants at the JME-SNB-Gerzensee Conference on Informational Frictions in Macroeconomics 2024; the ESIF Economics and AI+ML Meeting 2024; the SSES Annual Congress 2024; the Norges Bank Research Workshop: Women in Central Banking 2023; the Brown Bag Seminar and the Macro PhD Seminar at the Department of Economics at the University of Bern.

3.1 Introduction

Identifying the causal effects of monetary policy is challenging due to the difficulty of isolating exogenous variations in policy indicators. Monetary policy decisions, such as interest rate adjustments, are endogenous — they both influence and are influenced by prevailing economic conditions. These decisions respond to economic expectations, such as anticipated slowdowns, economic activity accelerations, or projected inflation changes, to shape future economic outcomes. Because this bidirectional relationship complicates efforts to evaluate the causal impacts of monetary policy, economists increasingly rely on data-implied methods, particularly external instruments, to identify monetary policy shocks and disentangle these complex effects.

Over the last decade, much of the literature has thus moved away from using zero or sign restrictions to strategies employing external instruments.[1] The seminal work by Gertler and Karadi (2015) uses high-frequency price changes of futures contracts within narrow time windows around monetary policy announcements to construct so-called monetary policy surprises. These monetary policy surprises serve as external instruments to identify the effects of monetary policy decisions.[2] They are appealing proxies for monetary policy shocks because they focus on a narrow time window (typically 30 minutes) around Federal Open Market Committee (FOMC) announcements. The idea is that within these short windows, causality runs from monetary policy news to futures prices and not the other way around. Additionally, the short window helps to eliminate confounding factors, such as other news releases on the same day.

Despite their advantages, recent studies have highlighted that market-based monetary policy surprises still suffer from two issues: weakness of the instrument and endogeneity problems. Ramey (2016) points out that market-based monetary policy surprises are typically weak instruments, meaning they have limited relevance. Moreover, Cieslak and Schrimpf (2019), Miranda-Agrippino and Ricco (2021), Bauer and Swanson (2023a), and Bauer and Swanson (2023b) report substantial correlations between the surprises and publicly available macroeconomic or financial data preceding FOMC announcements.

To address the relevance issue, we propose enhancing the existing market-based monetary policy surprises observed around FOMC announcements. Building on Bauer and Swanson (2023a), we extend the dataset of surprises by including policy-relevant speeches by the Federal Reserve Board chair and vice chair. Compared to the existing work, we employ a different method to determine the relevance of these speeches for monetary policy. Specifically, we analyze the language used in each speech and retain only those that mention both the topics of inflation and labor, reflecting the dual mandate of the Federal Reserve. To identify words

[1] The important advantage of using external instrument strategies is that they impose less theoretically motivated restrictions.

[2] A positive surprise indicates that the monetary policy announcement shifted the expected path of short-term interest rates upwards, serving as a proxy for a contractionary monetary policy shock.

related to these categories, we utilize the dictionary published by Gardner et al. (2022). Our findings show that incorporating speeches we label as policy-relevant substantially enhances the strength of the surprises as instruments.

Moreover, we address the exogeneity problem by isolating the portion of market-based monetary policy surprises that can be predicted from the text of FOMC statements or Federal Reserve speech transcripts. It is well documented that central bank communication has become an increasingly important aspect of monetary policy (see, e.g., Woodford, 2005, Blinder et al., 2008, Gardner et al., 2022, and Kerssenfischer and Schmeling, 2024), especially since the Great Financial Crisis. From 2008 to 2015, the policy rate was at the zero lower bound, restricting the scope for traditional policy measures. Consequently, the importance of policy communication has grown, with FOMC statements becoming more detailed and extensive and Federal Reserve Board chair and vice chair speeches more frequent. Given the Federal Reserve's meticulous choice of language in its statements and speeches, these communication events play a crucial role in shaping market expectations. As a result, markets react not only to specific policy actions but also to the language used in Federal Reserve communications. This suggests that part of the market-based monetary policy surprises is likely influenced by the language employed and policy action communicated by the Federal Reserve. Our approach leverages this relationship by mapping the text of policy communications to market-based monetary policy surprises using state-of-the-art Natural Language Processing (NLP) methods. Specifically, we train a neural network to predict the market-based surprises from FOMC statements and Federal Reserve speech transcripts. For the training, we use a random subsample of the data. Using our trained model, we predict changes in market expectations based on all FOMC statements and speech transcripts. By doing so, we extract the component of market-based surprises solely resulting from FOMC statements or speech transcripts. The surprises are cleansed of any effects not directly related to the specific statement or speech. Factors such as market momentum or trader sentiment are filtered out. Additionally, since the communication texts do not systematically include raw economic and financial data, our approach significantly mitigates the endogeneity issue. The predicted values for all statements and speeches form our new series of language-driven monetary policy surprises, i.e., our new instrument for the analysis.

For the text analysis part, we employ a language model based on the transformer architecture. In their groundbreaking paper, Vaswani et al. (2017) have introduced transformer models, which utilize a simple network architecture based solely on attention mechanisms. Since then, models for NLP tasks such as information extraction, document classification, text generation, and translation have significantly improved. One such model is XLNet. It builds on the transformer architecture and is pre-trained on a vast amount of textual data. This extensive pre-training provides the model with a robust understanding of different languages and the relationships between words. XLNet can be trained for specific applications, a process called

fine-tuning on a downstream task. In our study, we fine-tune the basic XLNet model to predict high-frequency asset price changes from policy communication text.

First, we demonstrate that the endogeneity problem predominantly affects FOMC announcements and is of limited significance for speeches by the Federal Reserve Board chair and vice chair. Nonetheless, we advocate for cleansing these speeches using text analysis methods to filter out other issues related to financial market phenomena, such as market momentum or trader attitude. Second, our findings indicate that the language-driven monetary policy surprises are considerably less correlated with economic and financial indicators available before the respective policy announcement event. Therefore, our NLP approach substantially mitigates the endogeneity problem. Finally, we observe that the monetary policy shocks identified with our language-driven surprises produce impulse responses that align more closely with conventional economic theories compared to those obtained using purely market-based surprises as instruments.

Related Literature.

Our paper relates to two strands of the literature. First, it contributes to the vast line of research on the identification of monetary policy shocks using high-frequency futures data (see e.g. Gürkaynak, Sack, and Swanson, 2005b, Nakamura and Steinsson, 2018, Cieslak and Schrimpf, 2019, Miranda-Agrippino and Ricco, 2021, Bauer and Swanson, 2023a, and Bauer and Swanson, 2023b). Second, our work contributes to the rapidly growing literature on text analysis in monetary policy.

Miranda-Agrippino and Ricco (2021) is one of the papers pointing out that market-based monetary policy surprises suffer from an endogeneity problem. They argue that the issue arises because of information asymmetries between the central bank and the public. When the Federal Reserve makes an announcement, it not only releases information about monetary policy actions but also private information about the current state of the economy. To clean this so-called "Fed information effect" from the market-based monetary policy surprises, they project the surprises on Greenbook forecasts and forecast revisions for real output growth, inflation, and the unemployment rate.

Bauer and Swanson (2023a) and Bauer and Swanson (2023b) present evidence challenging the Fed information effect and propose a "Fed response to news" channel as the alternative explanation. The effect related to this channel is also based on information frictions, but not regarding the state of the economy but rather the responsiveness of the Federal Reserve. Specifically, the public does not know the true response intensity of the U.S. central bank and updates its estimate of the response intensity with every policy communication. Bauer and Swanson (2023a) create a new, orthogonalized series of monetary policy surprises by removing components correlated with pre-announcement economic and financial data. This new series

improves upon previous measures. However, concerns remain that other factors, such as market momentum or trader attitudes, influence federal funds futures prices even within narrow windows (see, e.g., Lucca and Moench, 2015 and Neuhierl and Weber, 2018). Additionally, Bauer and Swanson (2023a) only correct for a limited set of economic and financial variables, leaving the possibility that other pre-announcement data might still be correlated with the monetary policy surprises.

We contribute to this part of the literature by proposing an alternative approach to refining market-based monetary policy surprises. Recognizing the significant role of language in Federal Reserve communications, we suggest extracting the component of existing surprises that is predictable from FOMC statement text or Federal Reserve speech transcripts. By leveraging the relationship between Federal Reserve communication texts and changes in interest rate expectations, we develop a new surprise series that substantially mitigates the endogeneity problem.

Bauer and Swanson (2023a) address the weak instrument problem by expanding the set of surprises from FOMC announcements to policy-relevant speeches by the Federal Reserve Board chair or vice chair. They demonstrate that speeches by the chair contain important policy information and that incorporating these additional policy-relevant communication events improves the relevance of monetary policy surprises as proxies for monetary policy shocks. In their selection process, they first include post-FOMC press conferences, the semiannual monetary policy report testimonies to Congress, and the speeches by the Federal Reserve Board chair at the Jackson Hole Symposium. Second, from the remaining speeches by the Federal Reserve Board chair, they label those as policy-relevant that led to a substantial (3 basis points or more) reaction in the two-quarter-ahead Eurodollar futures contract and that had moved markets according to their reading of the market commentary in the *The Wall Street Journal* or *New York Times* that afternoon or the following morning. We perceive this labeling as somewhat subjective. The choice of a 3 basis points threshold in the two-quarter-ahead Eurodollar futures rate, along with the judgment of the market commentary for policy relevance, introduces non-negligible elements of discretion in their classification method.

Other studies, such as Jayawickrema and Swanson (2023) and Kerssenfischer and Schmeling (2024), also emphasize the importance of speeches by the Federal Reserve Board chair. Jayawickrema and Swanson (2023) find that speeches by the chair are more important than FOMC announcements for Treasury yields, stock prices, and all but the very shortest-maturity interest rate futures. They conclude that including these speeches is key to capturing the primary source of variation in U.S. monetary policy. Kerssenfischer and Schmeling (2024) analyze which types of news mainly drive asset prices, finding that chair speeches rank among the most important scheduled releases. Similar to Bauer and Swanson (2023a), they also filter out policy-relevant speeches. However, they employ an automatic approach to identify relevant speeches and count

the number of news reports mentioning each speech. Depending on the sample period, they retain speeches mentioned at least twice before 2010, at least three times between 2010 and 2019, and at least four times since 2020.

Our paper builds on previous work by implementing a slightly different method for labeling policy-relevant speeches. Utilizing the dictionary from Gardner et al. (2022), we analyze the words used in speech transcripts by the Federal Reserve Board chair and vice chair. We classify a speech as policy-relevant if it mentions at least one word related to inflation and one word related to labor. Given the Federal Reserve's dual mandate of price stability and maximum employment, checking the content for these two categories effectively indicates policy relevance. This method of word counting introduces fewer discretionary elements than some of the existing methodologies.

Our paper also connects to a new strand of literature that employs text analysis to construct monetary policy surprise series. For example, Ochs (2021) generates different sentiment measures for the FOMC minutes, which are issued with a slight lag after each meeting. Inspired by the approach taken by Romer and Romer (2004), he constructs a type of monetary policy surprise from the residuals of the regression of the change in the federal funds rate on the sentiment measure. His sentiment analysis, however, is based on pre-specified word combinations to which a sentiment class is assigned. In a similar paper, Aruoba and Drechsel (2024) generate a sentiment measure for the FOMC statements. Their sentiment analysis is based on a dictionary by Loughran and McDonald (2011). In contrast to these two papers, we use transformers as a tool to capture relations and nuances in communication, which might be missed otherwise and which have become an important part of policy implementation.

In another example, Doh et al. (2022) construct a new measure of monetary policy surprises based on the Universal Sentence Encoder algorithm, designed to capture contextual nuances in FOMC statements. Specifically, they exploit cross-sectional variations across alternative FOMC statements to identify the statement's tone and compare current and previous FOMC statements to obtain the novelty. They then combine the statement's tone and novelty to measure the monetary policy stance and extract its unexpected component to construct a new monetary policy surprise series.

Handlan (2022) uses the XLNet model to predict intraday changes in FFF contracts from FOMC statement text. To address the issue of the Fed information effect, Handlan (2022) additionally cleans her text shocks using alternative statements.

Our paper is most closely related to the work of Handlan (2022). However, our approach differs by not correcting for the information effect. As discussed earlier, Bauer and Swanson (2023a) challenge the notion that the Fed information effect is responsible for the exogeneity issues of monetary policy surprises. They demonstrate that market-based monetary policy surprises can be forecasted using publicly available macroeconomic and financial data pre-dating

the policy announcement. Furthermore, unlike Handlan (2022), our analysis also incorporates speech transcripts by Federal Reserve Board members, broadening the scope of our study to include more comprehensive central bank communications.

The remainder of this paper is structured as follows. Section 3.2 describes the data. Section 3.3 extends the dataset of market-based monetary policy surprises by incorporating speeches from the Federal Reserve Board chair and vice chair, assessing their impact on instrument strength. Section 3.4 details the text analysis methodology, including the NLP model and training process, and evaluates the newly constructed language-based monetary policy surprises. Finally, Section 3.5 concludes.

3.2 Data

In our analysis, we utilize three types of data: high-frequency financial data, text data, and monthly macroeconomic data. With the high-frequency financial data, we construct a dataset of market-based monetary policy surprises around FOMC announcements and Federal Reserve Board chair and vice chair speeches. To apply our text analysis approach and derive our language-driven monetary policy surprise series, we match these surprises with the corresponding FOMC statements or speech transcripts. The different monetary policy surprise series are evaluated by using each series as an instrument to identify the monetary policy shock and then assess the shock's impact on a selection of key macroeconomic variables.

It is important to distinguish between the terms we use: "FOMC announcements" or "Federal Reserve Board speeches" refer to the policy communication events, while "FOMC statements" and "speech transcripts" refer to the corresponding text documents published at these events.

3.2.1 FOMC Announcements

We consider FOMC announcements from January 1996 to December 2019, encompassing eight regularly scheduled meetings per year, typically spaced six to eight weeks apart. Occasionally, the FOMC also holds unscheduled meetings if unexpected action is required before the next scheduled meeting. We include both types of meetings in our analysis, resulting in a sample of 200 announcements. For the text analysis part of our study, we have to exclude 22 of these announcements as the corresponding statements are not available. The Federal Reserve began consistently publishing a press statement after each meeting starting in May 1999. Prior to that, from 1996 to 1998, the Federal Reserve only issued an explicit statement when there was a change in the federal funds rate target. Thus, our final sample consists of 178 FOMC announcements with press statements. These statements not only communicate the interest rate decision but also provide information on the future economic outlook, forward guidance,

and other unconventional policy measures. Over the years, the length of these statements has significantly increased, ranging from approximately 75 to 780 words during our sample period. All statements, including the announcement dates, are from the website of the Federal Reserve Board.

3.2.2 Federal Reserve Board Chair and Vice Chair Speeches

Building on Bauer and Swanson (2023a), we expand the set of policy events beyond FOMC announcements to include speeches by the Federal Reserve Board chair and vice chair. Our sample period aligns with that of the FOMC announcements, spanning from 1996 to 2019. The dataset covers a range of events, such as remarks at the Jackson Hole Economic Symposium, testimonies to Congress, and other chair and vice chair speeches. The number of speeches held each month varies greatly over time, ranging from none to as many as nine. During the Great Financial Crisis, the frequency of speeches was particularly high.

At the annual Jackson Hole Economic Policy Symposium, the Federal Reserve Board chair typically delivers an opening speech to an audience including central bankers, economists, financial market participants, academics, U.S. government representatives, and the media. This speech provides a comprehensive overview of the Federal Reserve's perspectives on the current state of the U.S. and global economies, highlighting key trends and important policy directions. The chair's address often outlines future policy trajectories and the challenges associated with the conduct of monetary policy. During our sample period, the chair delivered 22 speeches at Jackson Hole. However, because precise time stamps are unavailable for eight of these speeches, our dataset includes only 15. These symposium speeches range in length from approximately 1,850 to 7,750 words, reflecting the depth and breadth of the topics covered.

The Federal Reserve Board chair also gives semiannual testimonies to Congress. During these testimonies, the chair provides an overview of the current economic conditions and the rationale behind recent monetary policy decisions. He or she discusses issues such as inflation, employment, and economic growth, and addresses concerns related to financial stability and regulation. The testimony includes an introductory statement followed by a question-and-answer session, allowing for further clarification and discussion. These testimonies aim to ensure accountability and transparency of the Federal Reserve's actions and policies. The testimonies are held twice a year. Each time, the chair presents the testimony once to the Senate and once to the House of Representatives within a few days. Since the introductory statement remains unchanged, we only include the earlier date in our dataset. We assume that the question-and-answer session does not significantly impact interest rate expectations. Moreover, including the question-and-answer part would widen the event window considerably, increasing the risk of capturing effects unrelated to Federal Reserve communications. Given the

sample considered, the Federal Reserve Board chair gave 48 testimonies to Congress between 1996 to 2019. However, not all of the transcripts are available, reducing the number to 39. The testimonies contain between 1,200 to 5,700 words.

Additionally, we consider 582 other speeches, out of which 465 were given by the Federal Reserve Board chair and 117 by the vice chair. The length of these speeches varies from around 150 to 20,900 words.

Some policy communication events, such as FOMC announcements or testimonies to Congress, occur in well-defined settings, making it easy to determine when their information reaches financial markets. However, for some speeches, pinpointing this moment is less straightforward. In these cases, we used the timestamps provided on the documents to establish when the speech became publicly available. If no such information was available, the speech was excluded. Additionally, speeches by the Federal Reserve Board chair and vice chair are delivered across various locations in the U.S. and internationally. To ensure consistency, we converted all speech times to U.S. Central Time, aligning with the time zone of the financial market where Eurodollar futures contracts are traded. Similarly, timestamps for FOMC announcements were converted from U.S. Eastern Time to U.S. Central Time. Speech dates and transcripts are from the websites of the Federal Reserve Board and the Federal Reserve Bank of St. Louis.

3.2.3 High-Frequency Monetary Policy Surprises

To measure shifts in market expectations caused by central bank communication, we extract the high-frequency changes in the price of futures contracts around each announcement or speech. These price changes are often referred to as monetary policy surprises. The rationale for using changes in futures prices is based on the forward-looking nature of financial markets. The federal funds futures (FFF) market allows participants to hedge against fluctuations in the federal funds rate. On any given day, the FFF market continuously reflects the market's expectations of the average federal funds rate over the remainder of the month. Thus, upward or downward revisions in FFF rates following an FOMC announcement or a Federal Reserve Board chair or vice chair speech indicate that market participants were surprised by the policy announcement and had to adjust their expectations.

As highlighted by Nakamura and Steinsson (2018), financial markets are forward-looking and react only to unexpected components of policy decisions, not to anticipated changes. The construction of monetary policy surprises builds on this idea, measuring intraday price changes in FFF contracts within a narrow time window around Federal Reserve communication events. This approach aims to eliminate reverse causality, ensuring that any observed changes in the FFF rate are attributable solely to the policy announcement rather than any other economic event.

We use Eurodollar futures contracts instead of FFF contracts due to data availability.[3] Nonetheless, Eurodollar futures rates are a reasonable choice. According to Gertler and Karadi (2015), they are the best predictors of future federal funds rate values at horizons beyond six months and are as good as FFF at horizons of less than six months.

We purchased historical intraday financial market data from Tick Data, LLC, covering Eurodollar futures contracts from December 1981 to June 2023. Eurodollar futures settle based on the spot 90-day Eurodollar deposit rate at expiration, and we focus on contracts that expire approximately one quarter ahead.[4] We convert the raw data, which reports individual trades, into minute-by-minute data, recording the high and low prices for each minute.[5]

For the FOMC announcements, we follow Gürkaynak, Sack, and Swanson (2005b) and measure the change in the Eurodollar futures rate using a 30-minute window, starting 10 minutes before the announcement and ending 20 minutes after. To account for multiple trades within one minute, we use the midpoint between the high and low prices for the minutes marking the beginning and the end of the window. To calculate the surprises, we take the difference between the average price at the end of the window and the average price at the beginning of the window, and then multiply this difference by minus one. This scaling is necessary because we want the surprises to reflect changes in interest rate expectations: a decrease in the futures price indicates an increase in interest rate expectations.

For speeches by the Federal Reserve Board chair and vice chair, we consider a time window of 50 minutes, starting 10 minutes before the speech and ending 40 minutes after. These communications tend to be more extensive than FOMC statements and contain broader information, which may require investors more time to process. Although some speeches or testimonies can last over an hour, we avoid extending the window too much to minimize the risk of capturing fluctuations in futures rates unrelated to the monetary policy communication. Additionally, the transcript is typically uploaded to the Federal Reserve's website at the start of the speech, providing market participants immediate access to the entire document without the need to listen to the speech in real-time. Thus, we believe the 50-minute window is a reasonable choice. As with the FOMC announcements, we calculate the midpoint between the high and low prices for the minutes marking the beginning and end of the window and scale the change within the 50-minute time window by minus one.

Some of the speeches partially occur when markets are closed.[6] To address closed markets, we consider three scenarios. First, if the entire speech window falls outside trading hours, we

[3]Federal funds futures are not available in Tick Data until 2010, while Eurodollar futures are.

[4]Eurodollar futures expire on the International Monetary Market dates: the third Wednesday of March, June, September, and December.

[5]If there is only one trade in a given minute, or if all trades occur at the same price, then the high and low prices for that minute will be identical.

[6]Starting from July 2003, Tick Data includes almost around-the-clock electronic trading data, meaning these instances mainly occur in earlier years.

exclude the speech from our dataset as we cannot measure the corresponding change in market expectations. Second, if the speech begins outside trading hours but markets open while the speech is ongoing, we retain the data point if 70 percent[7] of the speech window falls within trading hours. Similarly, if the speech starts during trading hours but ends after markets have closed, we apply the same 70 percent rule, retaining the speech if at least 70 percent of the speech window occurs within trading hours.

Table 3.1 presents summary statistics for the surprises associated with the different types of U.S. monetary policy announcements: FOMC announcements, chair speeches at the Jackson Hole Symposium, chair testimonies to Congress, other chair speeches, and vice chair speeches. The table includes data for surprises based on the one-quarter-ahead Eurodollar futures rate (ED2) – the primary focus of our analysis – as well as surprises constructed from current-quarter, two-quarter-ahead, and three-quarter-ahead Eurodollar futures rates (ED1, ED3, and ED4, respectively). First, we observe that the statistics are relatively similar across all four horizons of the futures. The biggest differences are seen for surprises associated with the current-quarter Eurodollar futures rate. For this very short-term horizon, changes in the futures rate predominantly reflect surprises related to the effective change in the policy rate. For the other horizons, surprises capture additional elements such as forward guidance. Given our interest in capturing not only the effect of policy rate changes but also other effects transmitted through language, we focus on a different horizon. We have chosen the one-quarter-ahead Eurodollar futures rate (ED2) for our analysis, as it balances the immediate impact of policy decisions with anticipatory elements. Second, the standard deviations and the range of changes (minimum and maximum) indicate that chair speeches and, to a slightly lesser extent, testimonies to Congress are as impactful as FOMC announcements. The other two announcement types, Jackson Hole speeches, and vice chair speeches are considerably less important. Lastly, the mean changes for all five announcement types are close to zero, as expected. FOMC announcements show a slight easing bias of about 1 basis point, but this is relatively small compared to the standard deviations of these changes.

3.2.4 Macroeconomic Data

When evaluating the effects of FOMC announcements or Federal Reserve Board speeches on macroeconomic variables, we use monthly data on industrial production, the consumer price index, the excess bond premium[8], and the two-year Treasury yield. Industrial production and

[7]The 70 percent threshold was selected as it provided a balance between ensuring that a large part of the time window fell within trading hours and keeping a majority of the speeches. There are 28 speeches where the time window only partially overlaps with trading hours. Employing the threshold, 12 of these speeches are dropped.

[8]Gilchrist and Zakrajšek (2012) construct a corporate bond credit spread index – the so-called GZ credit spread, which is based on a large micro-level dataset. They then decompose the GZ credit spread into two parts:

the consumer price index are taken from the FRED database. The two-year Treasury yield is from Bauer and Swanson (2023a), who took it from the Gürkaynak, Sack, and Wright (2007) database on the Federal Reserve Board's website. The excess bond premium from Gilchrist and Zakrajšek (2012) is available on the Federal Reserve's website. The sample goes from January 1973 to December 2019. The start is determined by the earliest availability of the excess bond premium, while the end is chosen such as to exclude the dramatic swings of the COVID-19 pandemic and its aftermath.

Table 3.1: Summary Statistics for U.S. Monetary Policy Surprises

	FOMC announcements	Jackson Hole speeches	Testimonies to Congress	Chair speeches	Vice chair speeches
Number	200	15	39	465	117
Standard dev. (bp)					
ED1	4.9	0.6	2.3	1.4	0.8
ED2	5.2	1.3	4.2	2.3	1.3
ED3	5.8	2.1	5.8	2.7	1.5
ED4	5.8	2.7	6.4	3.0	1.6
Min. change (bp)					
ED1	-32.5	-2.0	-7.0	-13.0	-3.5
ED2	-27.3	-2.3	-8.3	-17.0	-4.0
ED3	-29.0	-2.5	-9.5	-19.0	-4.0
ED4	-24.0	-3.0	-12.5	-20.5	-4.5
Max. change (bp)					
ED1	18.3	0.8	4.3	6.5	4.3
ED2	12.0	2.5	9.0	18.0	8.0
ED3	17.8	5.0	15.5	22.3	7.8
ED4	24.3	7.3	15.0	26.3	9.3
Mean change (bp)					
ED1	-0.8	-0.1	-0.1	0.0	0.0
ED2	-1.0	0.1	0.5	0.0	0.2
ED3	-1.0	0.2	0.7	0.0	0.2
ED4	-1.0	0.4	0.7	0.0	0.1

Notes: Changes for ED1 to ED4 are in basis points. Sample period is 1996 to 2019.

one part capturing the systematic movements in default risk of individual firms and a residual component – the excess bond premium. The excess bond premium can be interpreted as the variation in the pricing of default risk, meaning it is a measure of the tightness of financial conditions.

3.3 Market-Based Monetary Policy Surprises

Most existing series of monetary policy surprises focus exclusively on the reactions of market participants to FOMC announcements (Gürkaynak, Sack, and Swanson, 2005a, Gertler and Karadi, 2015, Miranda-Agrippino and Ricco, 2021, Handlan, 2022). However, several studies have demonstrated that speeches by the Federal Reserve Board chair and vice chair also contain significant policy information that influences interest rate expectations (Bauer and Swanson, 2023a, Bauer and Swanson, 2023b, and Kerssenfischer and Schmeling, 2024). Additionally, Bauer and Swanson (2023a) show that including monetary policy surprises around Federal Reserve Board chair or vice chair speeches enhances the relevance of these surprises as instruments for identifying monetary policy shocks. Following Bauer and Swanson (2023a), we expand the set of existing market-based surprises to include relevant speeches. To identify the relevant speeches, we utilize a dictionary approach to analyze the content of the speech transcripts and include only those that contain policy-relevant topics.

3.3.1 Identification of Policy-Relevant Speeches

The Federal Reserve Board chair and vice chair deliver speeches on a wide range of topics, many of which extend beyond monetary policy. These can include ceremonial addresses or discussions on subjects such as bank regulation, securities market regulation, fiscal policy, and various other economic and financial issues. As shown by Bertsch et al. (2024), Federal Reserve communication often addresses topics such as financial stability, establishing it as a prominent and recurring theme for these speeches. For our analysis, we focus exclusively on central bank communications that have potential implications for U.S. monetary policy. To identify the speeches relevant to our study, we employ a dictionary-based approach. Specifically, we utilize the dictionary developed by Gardner et al. (2022), which includes lists of words related to inflation, labor, output, and financial topics. For instance, words in the inflation category include "inflation", "price", and "cost", while words in the labor category include "employment", "job losses", and "hiring". We count the occurrences of words related to these topics in each speech.

Figure 3.1 illustrates the evolution of the average frequency with which words from each category are mentioned per speech transcript. We see that around the Great Financial Crisis, financial topics were discussed more frequently than in other years. Additionally, inflation topics peaked right before the Great Financial Crisis and in 2018, while labor topics appeared more often starting from around 2011.

To determine the relevance of a speech for our application, we focus on the inflation and labor categories. If a speech transcript contains at least one word related to inflation and one word related to labor, we classify it as policy-relevant. Otherwise, it is labeled as non-relevant.

Given the Federal Reserve's dual mandate to promote maximum employment and stable prices, we assume that these two topics are always addressed when the speech pertains to monetary policy. Using these classification criteria, we identify a subset of 441 policy-relevant speeches.

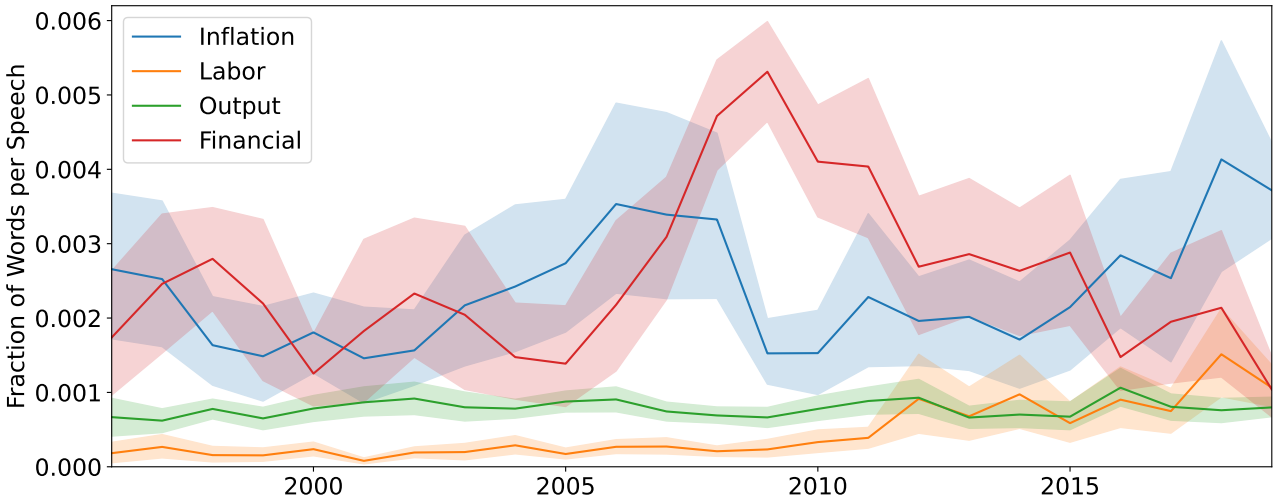
Figure 3.2 presents a histogram comparing the distribution of monetary policy surprises for all speeches against those based solely on policy-relevant speeches. By excluding non-policy-relevant speeches, we observe a significant decrease in the number of surprises centered around zero. This indicates that our classification method, which relies solely on input text, effectively filters out speeches lacking substantial monetary policy information, thereby refining our dataset to predominantly include those speeches that impact market expectations.

3.3.2 Monetary Policy Effects on Macroeconomic Variables

The primary objective of broadening the set of considered monetary policy communication events is to enhance the strength of the surprises as instruments for identifying monetary policy shocks. We evaluate this in a Proxy-Structural Vector Autoregression (Proxy-SVAR) framework, closely following Gertler and Karadi (2015). Additionally, we aim to assess how the identified monetary policy shock impacts key macroeconomic variables.

Following Bauer and Swanson (2023a), our VAR specification includes the log of industrial production, the log of the consumer price index, the Gilchrist and Zakrajšek (2012) excess bond premium, and the two-year Treasury yield. We include the excess bond premium because Caldara and Herbst (2019) find it to be necessary to identify monetary policy shocks correctly.

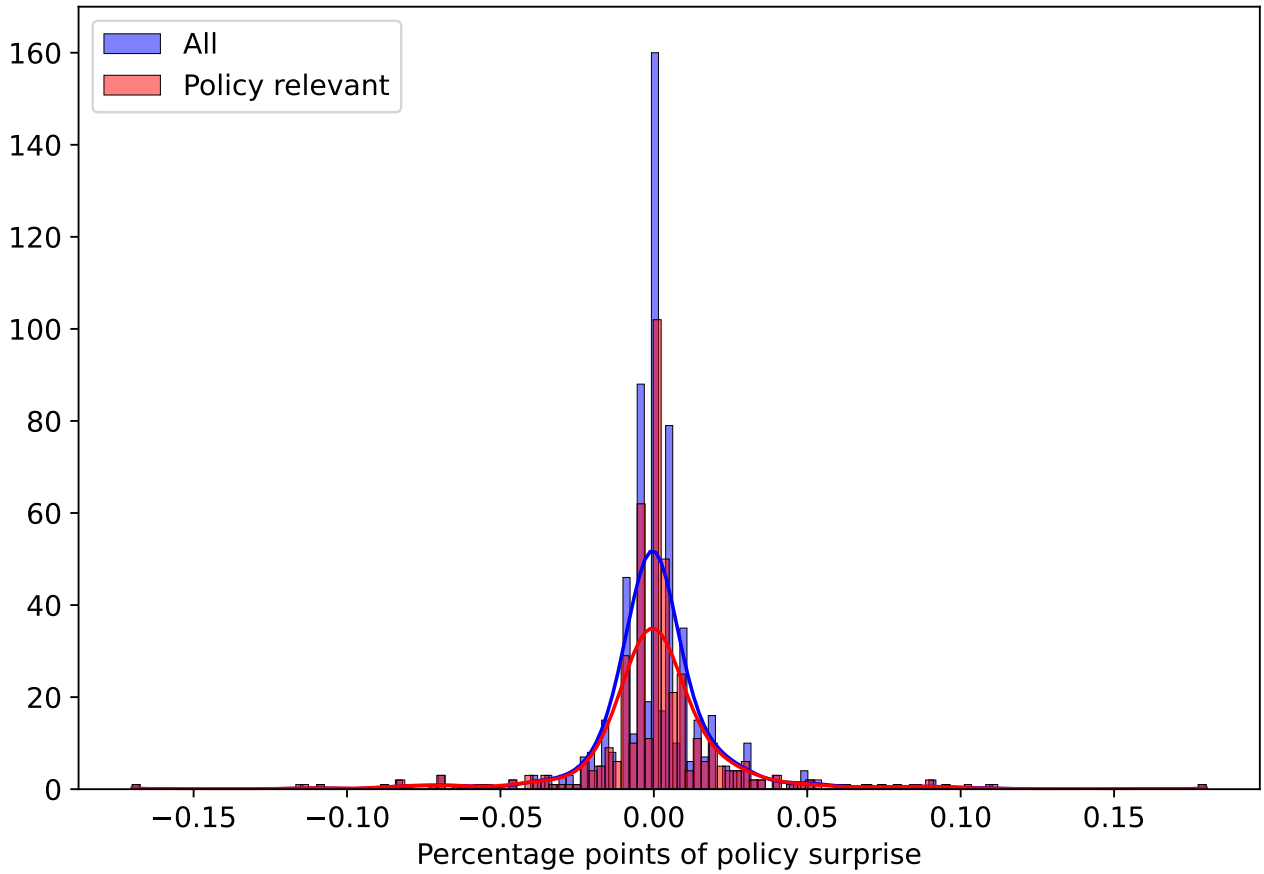
Figure 3.1: Chair and Vice Chair Speeches: Categories Sampled by Year



Notes: The categories are assigned using the dictionary by Gardner et al. (2022). Observations are sampled by year. The solid lines are the medians per year and the shaded areas show the entire distribution (from minium to maximum fractions).

Furthermore, as discussed in Gertler and Karadi (2015) and Bauer and Swanson (2023a), we use the two-year Treasury yield instead of the federal funds rate as the policy rate variable.[9] Unlike the federal funds rate, the two-year Treasury yield was largely unconstrained during the U.S. zero lower bound period from 2009 to 2015, making it a better measure of the stance of monetary policy. Moreover, an important advantage of using a government bond rate as the policy indicator is that its innovations do not only capture traditional monetary policy shocks, i.e., surprises related to the current federal funds rate, but also shocks to forward guidance. Swanson and Williams (2012) and Hanson and Stein (2015) argue that the Federal Reserve's forward guidance strategy operates with a roughly two-year horizon, which makes the two-year Treasury yield the preferred government bond rate.[10] Additionally, the speeches do not convey

Figure 3.2: Speech Distributions, All versus Policy Relevant



Notes: Distributions of market-based surprises from Federal Reserve Board chair and vice chair speeches considering all (blue) or only policy-relevant speeches (red).

[9] Although Gertler and Karadi (2015) advocate for the two-year Treasury yield, they use the one-year Treasury yield in their VAR due to an insufficiently large F-statistic for their first-stage instrumental variables regression with the two-year yield as the policy indicator.

[10] The Federal Reserve's forward guidance strategy, focusing on managing expectations of the path of the

information about changes in the current federal funds rate; rather, they influence market participants' expectations regarding future policy rate changes. Hence, employing the two-year Treasury yield in conjunction with our expanded market-based surprises as an instrument is a natural choice.

Proxy-SVAR Methodology

We start by considering the following structural VAR:

$$\mathbf{A}\mathbf{Y}_t = \sum_{j=1}^{p} \mathbf{C}_j \mathbf{Y}_{t-j} + \boldsymbol{\varepsilon}_t, \tag{3.3.1}$$

where $\mathbf{Y}_t$ is a vector of observables, $\mathbf{A}$ and $\mathbf{C}_j$ $\forall j \geq 1$ are conformable coefficient matrices, and $\boldsymbol{\varepsilon}_t$ is an $n \times 1$ vector of white noise structural shocks. When multiplying both sides with $\mathbf{A}^{-1}$, the reduced-form VAR representation follows:

$$\mathbf{Y}_t = \sum_{j=1}^{p} \mathbf{B}_j \mathbf{Y}_{t-j} + \mathbf{u}_t, \tag{3.3.2}$$

with $\mathbf{u}_t$ being the reduced-form VAR residuals, $\mathbf{B}_j = \mathbf{A}^{-1}\mathbf{C}_j$, and $\mathbb{E}[\mathbf{u}_t\mathbf{u}_t'] = \boldsymbol{\Sigma}$ for some positive definite matrix $\boldsymbol{\Sigma}$. The VAR residuals are modeled as linear combinations of the underlying structural shocks, namely

$$\mathbf{u}_t = \mathbf{S}\boldsymbol{\varepsilon}_t. \tag{3.3.3}$$

It follows that $\mathbf{S} = \mathbf{A}^{-1}$ and $\mathbb{E}[\mathbf{u}_t\mathbf{u}_t'] = \mathbb{E}[\mathbf{S}\mathbf{S}'] = \boldsymbol{\Sigma}$.

Let us then define $Y_t^p \in \mathbf{Y}_t$ to be the monetary policy indicator, i.e., the variable for which the exogenous variation is due to the monetary policy shock ε_t^p. To estimate the impulse responses to a monetary policy shock, we need to estimate the equation

$$\mathbf{Y}_t = \sum_{j=1}^{p} \mathbf{B}_j \mathbf{Y}_{t-j} + \mathbf{s}\varepsilon_t^p, \tag{3.3.4}$$

where $\mathbf{s}$ is the column of $\mathbf{S}$ associated with the effects of ε_t^p. Because we are only interested in the impulse responses to a monetary policy shock, it is sufficient to identify $\mathbf{s}$ and not the entire matrix $\mathbf{S}$. We use an external instruments strategy to obtain $\mathbf{s}$.

We define $\mathbf{m}_t$ as the $k \times 1$ vector of instruments. $\boldsymbol{\varepsilon}_t^q$ is a vector of structural shocks other than the monetary policy shock. For $\mathbf{m}_t$ to be a valid set of instruments, the exogeneity and

short rate two years into the future, supports the use of the two-year Treasury yield.

relevance conditions must be satisfied:

$$\begin{aligned} \mathbb{E}[\mathbf{m}_t \varepsilon_t^p] &\neq 0 \\ \mathbb{E}[\mathbf{m}_t (\boldsymbol{\varepsilon}_t^q)'] &= \mathbf{0}, \end{aligned} \tag{3.3.5}$$

meaning that the instruments are correlated with the monetary policy shock ε_t^p but orthogonal to any other structural shock $\boldsymbol{\varepsilon}_t^q$, where $q \neq p$. In the following application, we will always focus on a single instrument: some version of the monetary policy surprises. Notice that the market-based surprises or the language-based surprises in the next section are all intradaily changes in Eurodollar futures prices. To use them as an instrument in the Proxy-SVAR, we convert them to a monthly series by summing over all the high-frequency surprises within each month.[11]

The identification of $\mathbf{s}$ works as follows: First, we estimate the VAR using least squares estimation and get the reduced-form residuals $\mathbf{u}_t$. These residuals can then be split up into u_t^p, the residual associated with the equation of the policy indicator, and $\mathbf{u}_t^q$, the residuals of all other variables. Moreover, we define $s^p \in \mathbf{s}$ to be the response of u_t^p to a unit increase in ε_t^p. Similarly, $\mathbf{s}^q \in \mathbf{s}$ is the response of $\mathbf{u}_t^q$ to an increase of ε_t^p by one unit. Second, we perform a two-stage least squares regression. In the first stage, we regress u_t^p on the instrument $\mathbf{m}_t$. Consequently, the variation in the fitted value $\hat{u}_t^p$ is only due to the monetary policy shock ε_t^p. In the second stage, we regress $\mathbf{u}_t^q$ on $\hat{u}_t^p$:

$$\mathbf{u}_t^q = \frac{\mathbf{s}^q}{s^p} \hat{u}_t^p + \boldsymbol{\xi}_t. \tag{3.3.6}$$

This regression yields a consistent estimate of $\frac{\mathbf{s}^q}{s^p}$ because $\hat{u}_t^p$ is uncorrelated with the error term $\boldsymbol{\xi}_t$. An estimate for s^p can be obtained from the estimated variance-covariance matrix $\boldsymbol{\Sigma}$. In the next step, $\mathbf{s}^q$ can be computed. Based on the estimates of s^p, $\mathbf{s}^q$, and the VAR coefficients ($\mathbf{B}_j$s), we can calculate the impulse responses of all variables in $\mathbf{y}_t$ to a monetary policy shock ε_t^p.

We estimate the VAR using frequentist methods. To obtain confidence bands around the point estimates, we employ bootstrapping methods, with 10,000 bootstrap replications.[12] Moreover, we choose a lag order of $p = 12$. Based on the Ljung-Box Q-test, this lag order is the smallest for which the VAR residuals are no longer serially correlated. Additionally, this lag order is in line with Gertler and Karadi (2015), Ramey (2016) and Bauer and Swanson (2023a).[13]

[11] In months for which no surprise occurs, i.e., without FOMC announcements or Federal Reserve Board chair and vice chair speeches, the monthly surprises are equal to zero.

[12] We are using the wild bootstrap procedure of Mertens and Ravn (2013) and Gertler and Karadi (2015).

[13] The results remain qualitatively and quantitatively very similar if we choose a lag order of $p = 6$.

Results

We identify the monetary policy shock using three different surprise series as instruments. The first series includes only FOMC announcements. The second series expands to include surprises from both FOMC announcements and all speeches by the Federal Reserve Board chair and vice chair. The third series refines upon the second by incorporating only those speeches labeled as policy relevant.

Table 3.2 reports the F-statistics of the first stage regression for each of the three surprise series. When considering only FOMC announcements, the F-statistic is 5.78. Including all speeches raises the F-statistic to 13.70. Finally, when adding only policy-relevant speeches to the FOMC announcements, the F-statistic further increases to 16.50[14]. This indicates that broadening the set of market-based monetary policy surprises significantly improves their strength as instruments for identifying the monetary policy shock.

Figure 3.3 depicts the corresponding impulse responses, normalized to reflect a 25 basis point increase in the two-year Treasury yield. The responses are shown for the same three sets of surprises discussed above. The impulse responses obtained with the surprise series that include speeches differ markedly from those based solely on FOMC announcement surprises. The latter responses align with conventional wisdom: a monetary policy shock leads to a decline in the consumer price index and a contraction in industrial production. Additionally, the excess bond premium rises, indicating tighter financial conditions. In contrast, the impulse responses derived from the other two surprise series, which contain speeches, reveal notable deviations. The consumer price index responds positively, exhibiting the so-called "price puzzle". While industrial production still decreases, as expected, the excess bond premium shows little to no significant reaction, a finding that contradicts traditional economic theories.

To sum up, we find that adding Federal Reserve Board chair and vice chair speeches to the surprise dataset improves the F-statistic substantially. However, the dynamic responses to a monetary policy shock also change. When considering speeches in the instrument series, the identified monetary policy shock gives rise to a price puzzle. This finding is counterintuitive according to standard economic theory.

Table 3.2: F-statistics for Market-Based Surprise Series

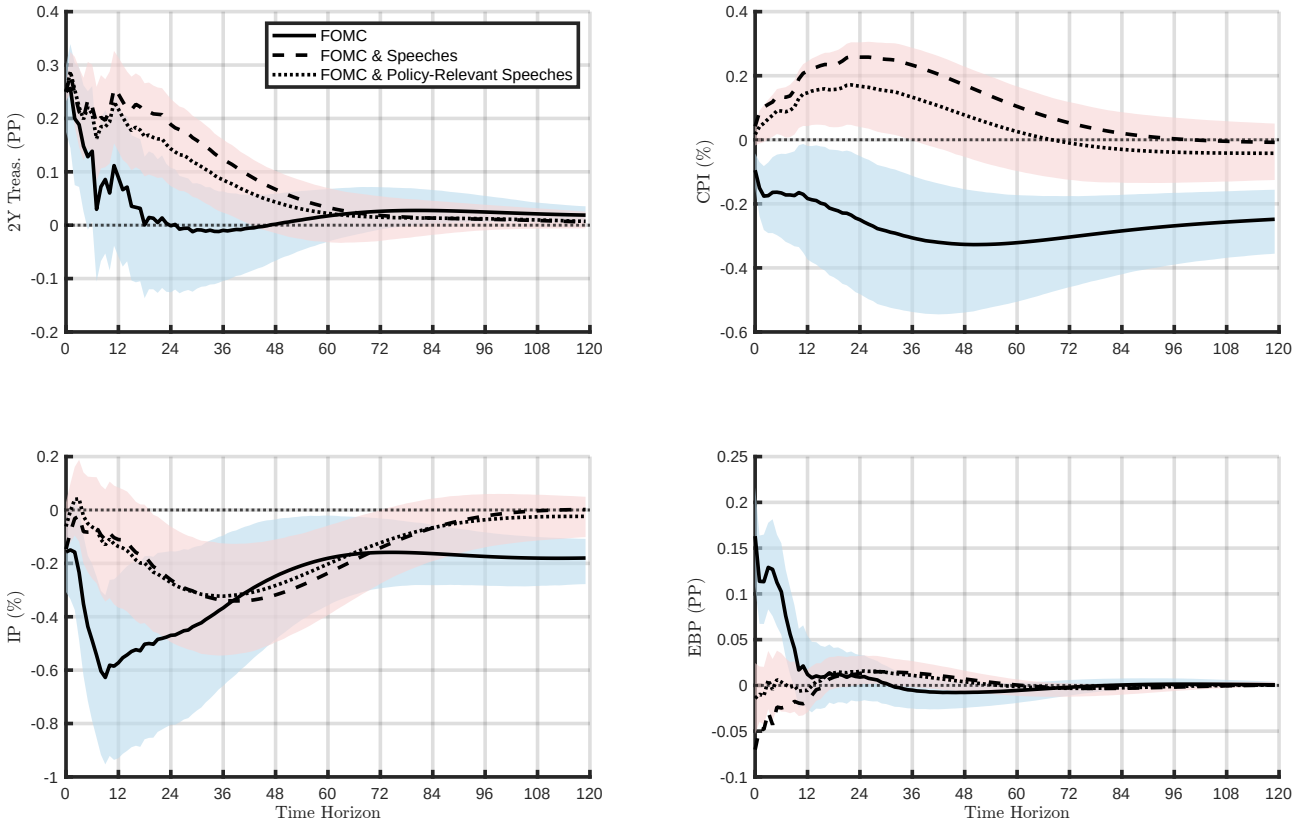
	F-statistic
FOMC announcements	5.78
FOMC announcements and *all* speeches	13.70
FOMC announcements and *policy-relevant* speeches	16.50

[14]Stock and Watson (2012) propose a rule of thumb according to which the instrument is weak if the first-stage F-statistics in the two-stage least squares regression is smaller than 10.

3.4 Language-Based Monetary Policy Surprises

In this section, we address the predictability of market-based monetary policy surprises using central bank communication texts. Advances in NLP enable us to directly link central bank statements and speech transcripts to financial market reactions. In particular, we map policy communication texts to market-based monetary policy surprises, isolating the portion of the surprises driven exclusively by FOMC statements and speech transcripts[15]. This approach allows us to construct language-based monetary policy surprises that capture solely the impact of central bank communication while abstracting from other influencing factors. We assess the exogeneity of our newly constructed language-based monetary policy surprises using past economic information.

Figure 3.3: Impulse Responses to a Monetary Policy Shock (Market-Based Surprises)



Notes: The blue shaded areas represent the 5-95 percentiles of the impulse responses identified with only FOMC announcements and the red shaded areas the 5-95 percentiles of the impulse responses identified with *policy-relevant* speeches. The impulse responses are normalized to a 25 basis point increase in the two-year Treasury yield. Horizontal axis: time horizon in months.

[15] Appendix 3.A.4 presents the results only using FOMC statements for training.

3.4.1 Natural Language Processing

To construct the language-based surprises, we follow a four-step process. First, we select a pre-trained neural network capable of understanding English. We opt for the XLNet-Base model. Second, we pre-process the central bank communication texts to create structured inputs for our NLP model. Third, we fine-tune the pre-trained model on our specific task. That is, we train the model to understand the relationship between central bank communication texts and market-based high-frequency policy surprises. Finally, using the trained model, we predict the changes in interest rate expectations associated with each FOMC statement or speech transcript. These predictions are what we call the language-driven monetary policy surprises.

Select Pre-Trained Neural Network for Text Processing

To capture the semantics of central bank communications, we use XLNet-Base, a state-of-the-art Natural Language Understanding algorithm developed by Yang et al. (2020).[16] XLNet combines two techniques – autoregressive language modeling and auto-encoding – to learn textual content. Both methods involve predicting missing words, but while autoregressive modeling predicts words at the end of a sequence, auto-encoding predicts missing words from anywhere within a sentence. Consequently, XLNet not only understands individual words but also captures sentence structure and longer textual contexts. Because XLNet learns representations through different approaches, it is highly versatile, making it suitable for tasks beyond simple word prediction. The model can be fine-tuned to map text to other text, categories, or continuous numbers. In our case, we want to teach the model to link central bank communication texts to changes in market expectations, which is a continuous financial market variable. We start from the base version's pre-trained network architecture and word representations, which have been trained on a vast corpus of text to acquire general English language skills.[17] Such off-the-shelf pre-trained models are widely used in NLP because training them from scratch requires substantial computational power and vast amounts of text data.

Pre-Processing Text Data

In the pre-processing step, we format the text of FOMC statements and speech transcripts in such a way that the language model can process it. For the FOMC statements, we make only minimal modifications to preserve the original wording. We replace long word combinations with abbreviations[18], standardize the formatting of numbers, and remove repetitive words,

[16] The base model consists of 12 attention heads/layers with 768 dimensions and two feed-forward layers with 768 and 3072 dimensions, resulting in approximately 117 million parameters.

[17] The model was trained on a text corpus from five sources: The Book Corpus and English Wikipedia (13GB), Giga5 text (16GB), Clue Web 2012-B (19GB), and a Common Crawl (110GB).

[18] We only use abbreviations that the Federal Reserve itself uses in at least one of the other statements.

dates, and committee member names to prevent the model from drawing incorrect conclusions during training. A full list of modifications is provided in Appendix 3.A.1.

The speech transcripts require additional adjustments. As described in Section 3.2, they tend to be significantly longer than FOMC statements, often spanning multiple pages. While neural networks can process long text inputs, their performance can deteriorate when exceeding a certain length – especially for smaller models like XLNet, which were trained on short sequences. To address this, we need to reduce the length of the speech transcripts. One possibility is to truncate each speech transcript at the model's recommended token[19] length, which is 512 tokens for the XLNet model. However, this approach risks omitting important information mentioned toward the end of the speech. Instead, we opt for summarization, ensuring that the most relevant content is preserved while keeping the input length manageable.

Summarization helps discipline the model by directing its focus toward the policy-relevant parts of a speech transcript. Unlike FOMC statements, which are short, well-structured, and carefully worded, speeches exhibit greater variability in both structure and phrasing. As a result, market participants are likely to focus on the key takeaways rather than the precise wording. By summarizing speech transcripts, we replicate this process, ensuring that our dataset better reflects how financial markets extract and interpret central bank communication. Furthermore, given the varying length and structure of speeches, summarization enhances consistency, reduces noise, and improves the model's ability to learn from these text documents.

We use the Mistral Large 2 model, a state-of-the-art language model with 123 billion parameters, for the summarization task. This model can process inputs of up to 128,000 tokens and outperformed alternative approaches,[20] making it the most suitable choice for our application. To ensure that the most relevant content is preserved, we instruct the model to: generate fluent, first-person summaries that maintain the speaker's voice rather than third-person bullet points; and (ii) focus on monetary policy topics, using a predefined dictionary from Gardner et al. (2022) along with additional key terms such as accommodative, contractionary, stance, and federal funds rate. The dictionary includes terms frequently appearing in FOMC statements from 2000 to 2020 related to labor markets, output, inflation, and financial conditions. While our classification criteria ensure that selected speeches are relevant to monetary policy, these topic-based instructions further refine the summaries to highlight the most pertinent content. Each summary is limited to a maximum of 15 sentences. In the end, we pre-process every summary in the same way as we pre-processed the FOMC statements.

[19]Tokens are the input feed to the language model. They capture the text and its words, where, as a rule of thumb, one token corresponds to 4 characters on average.

[20]We also tested Falcon-7B-Instruct, GPT-4, BART fine-tuned on CNN/Daily Mail, and a smaller variant of T5 fine-tuned for summarization.

Fine-Tune the Neural Network

As described earlier, every NLP model is associated with a specific vocabulary, a collection of distinct words, it was initially trained on. Through this initial training, the XLNet model has already gained an understanding of the words and their relationships among each other. In more technical terms, as part of the training, the words are converted into tokens so a computer can easily process them. The XLNet model has already mapped the tokens in an N-dimensional space where similar tokens lie closer together. For example, the tokens for apples and oranges lie closer together than the tokens for apples and inflation. To use the model by Yang et al. (2020), we convert the FOMC statements and the summaries of the speech transcripts into tokens used in the XLNet vocabulary.

Moreover, we want our model to learn the mapping from FOMC statements and speech transcripts to marked-based surprises. The objective is for the model to predict the surprises based on an unknown statement or speech. The model provided by Yang et al. (2020) can not yet execute this task. Thus, we add layers to their neural network structure that are suitable for obtaining continuous predictions. Specifically, we break down the text using convolutional layers so that the model can extract the relevant information and appropriately predict the surprises.[21] Appendix 3.A.2 presents our model architecture, and Appendix 3.A.3 explains the training algorithm in further detail.

As it is standard in the machine learning literature, we apply k-fold cross-validation. We split our dataset, consisting of the FOMC statements and speech transcripts, together with the respective marked-based surprise, into different parts. Notably, we always have a training set that the model adapts its parameters on and a test set that it runs the model on but does not adapt its parameters to. Such train and test splits are important in machine learning because the models are prone to overfitting, i.e., to learn too much from the training data, thereby being unable to predict data it has not yet seen. To train our model, we apply five-fold cross-validation. Thus, we always have 80 percent of our data in the training set and 20 percent in the test set. This procedure also implies that for the same hyperparameter described in Appendix 3.A.3, we have five different parameter values, depending on the data the model was trained on.

To find the optimal hyperparameters for our model, we experiment with different combinations of the number of epochs[22] and the learning rate.[23] During these experiments, we monitor the mean squared error (MSE) on the training (in-sample) and test (out-of-sample) data. The MSE is the mean squared difference between the prediction and the corresponding

[21] This procedure adds roughly 4 million parameters to the model. Thus, the final model counts around 221 million parameters. Training these additional parameters typically requires around three days, though the exact duration may vary depending on the specifications.

[22] The number of epochs defines how often the model sees the same data set to adapt its parameters.

[23] The learning rate defines by how much the neural network adapts its parameters after each iteration.

true marked-based surprise. Based on these test runs, we fix the learning rate of our model to 1e-5 and the number of epochs to 10. The MSE for this set of hyperparameters for the five-fold cross-validation is displayed in Table 3.3.

Table 3.3: MSE of 5-Fold Cross-Validation after 10 Epochs

Number of Split	In-Sample MSE	Out-of-Sample MSE
1	5.040045e-05	0.0023990888
2	0.00016885748	0.0015901675
3	0.00068839284	0.0011396597
4	0.0014133948	0.0010364184
5	0.00012260459	0.0022924726

Predict Surprises using Text Data

The predicted surprises for each FOMC statement or speech transcript are generated using the model parameters from the first split of the cross-validation after ten epochs, as described above.

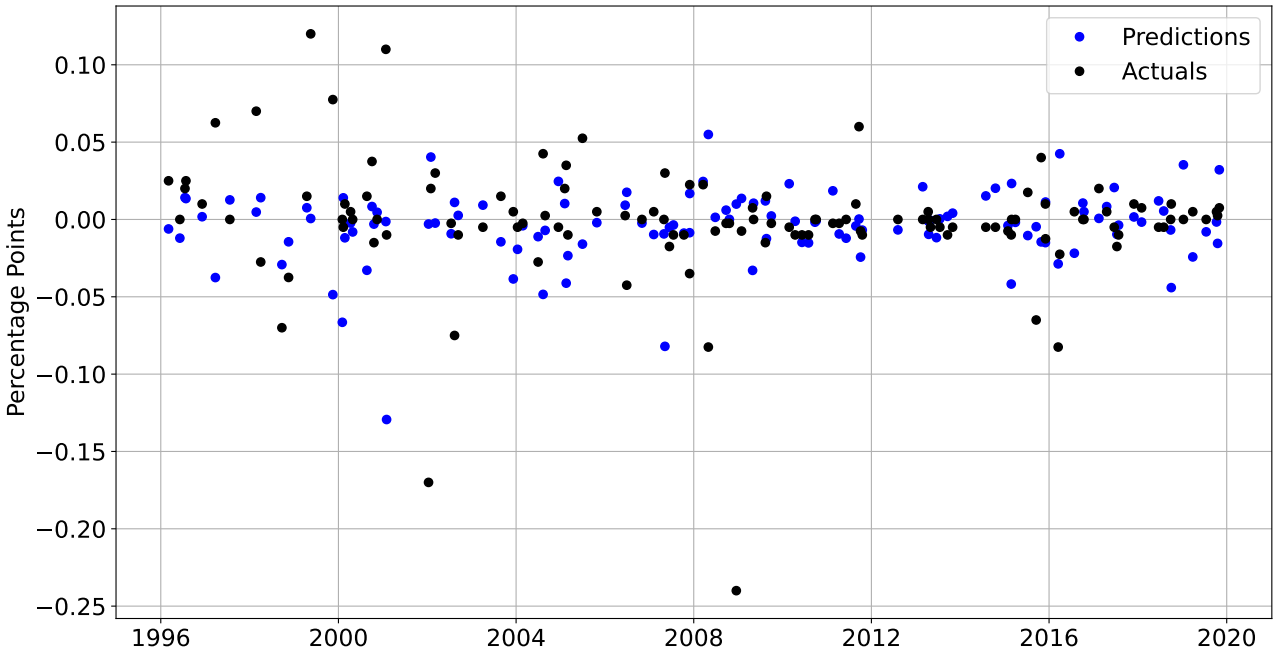
Figure 3.4 presents the out-of-sample MSE for this split, focusing on the test set of FOMC statements and speech transcripts. While the model captures changes in market-based surprises to some extent, it does not achieve a perfect fit. This result is unsurprising since the number of data points is small for a machine-learning task. However, considering the limited number of available texts, along with their length and complexity, the model performs remarkably well on the test set. Moreover, market-based surprises are expected to be influenced not only by the content of central bank communications but also by other factors such as market momentum and trader sentiment. As a result, some degree of deviation between predicted and actual surprises is both expected and even desirable. Overall, the results indicate that the model successfully learns patterns from the training data that allow it to predict market reactions.

3.4.2 Relation to Past Economic Information

The primary objective of our NLP task is to isolate the component of market-based surprises that stems solely from central bank communication. By doing so, we aim to remove the correlation between market-based surprises and past economic and financial data. To verify whether our language-based surprises are indeed uncorrelated to economic and financial information available before an announcement or speech, we conduct a regression analysis following Bauer and Swanson (2023a).

To capture past economic and financial conditions, we construct the following variables: (1) the most recent nonfarm payroll surprises (NFP_SURP), (2) the 12-month employment growth

Figure 3.4: Predictions and Marked-Based Surprises



Remark: Marked-based surprises and out-of-sample model predictions of the surprises are displayed in black and in blue, respectively.

in total nonfarm payrolls (NFP_12M), (3) the three-month growth in the S&P 500 stock market index (SP500_3M), (4) the three-month change in the slope of the yield curve (SLOPE_3M), (5) the three-month growth in the Bloomberg Commodity Spot Price index (BCOM_3M), and (6) the average skewness of the ten-year Treasury yield over the past month (TR_SKEW).[24] Except for nonfarm payroll surprises, constructing these variables is straightforward; further details can be found in Bauer and Swanson (2023a). For nonfarm payroll surprises, we take the difference between the actual nonfarm payroll release and the median forecast from a survey of financial market participants conducted before the release. Since we lack direct access to this survey (owned by Haver Analytics), we approximate the series as follows: first, we construct a time series of median expectations for months with FOMC meetings, where complete data is available. Then, for months without meetings, we estimate missing values using linear interpolation. This approach ensures that our series matches Bauer and Swanson (2023a) for FOMC statement dates while providing an approximation for speech dates.

Table 3.4 presents the regression results. The second column reports estimates for market-based surprises associated with FOMC announcements and Federal Reserve Board chair and

[24]The S&P 500 stock market index and the Bloomberg Commodity Spot Price index are from Datastream, provided by LSEG Data & Analytics and accessible via a University of Bern license. The implied skewness is based on the paper by Bauer and Chernov (2024) and obtained from the website of the Federal Reserve Bank of San Francisco. Total nonfarm payrolls and the yield curve slope are from the St. Louis FRED database.

vice chair speeches. Consistent with concerns raised in the literature, we find evidence that these surprises may capture factors beyond monetary policy shocks. Specifically, half of the coefficients on past economic or financial variables are statistically significant at the 5 percent level, suggesting that market-based surprises are correlation with economic and financial information available prior to the announcements or speeches.

In contrast, the first column displays results for our language-based surprise series. Here, with the exception of the ten-year Treasury yield skewness, we find no statistically significant relationships with past economic and financial information, which indicates an improvement over the market-based series. A similar pattern emerges in columns three and four, which repeat the analysis but restrict the sample to surprises associated with only FOMC statements. The last two columns, which focus solely on speech transcripts, show no big differences between market-based and language-driven surprises.

These results suggest that our language-driven approach provides a cleaner measure of monetary policy surprises by filtering out influences from prior economic information. Unlike

Table 3.4: Regression on MPS

	All		Statements		Speeches	
	LD	MB	LD	MB	LD	MB
NFP_SURP	0.0304	0.0280	0.0986	0.0875	0.0019	-0.0040
	(0.0914)	(0.1493)	(0.0393)	(0.0821)	(0.8957)	(0.7699)
NFP_12M	0.0010	0.0020	0.0024	0.0055	0.0007	0.0009
	(0.1686)	(0.0156)	(0.2893)	(0.0296)	(0.3065)	(0.1987)
SP500_3M	0.0309	0.0528	0.0334	0.1214	0.0206	0.0221
	(0.1570)	(0.0362)	(0.6211)	(0.0871)	(0.2198)	(0.2073)
SLOPE_3M	-0.0037	-0.0042	-0.0115	-0.0115	-0.0005	-0.0013
	(0.2847)	(0.2462)	(0.2475)	(0.2358)	(0.8195)	(0.5829)
BCOM_3M	-0.0139	0.0094	-0.0105	0.0623	-0.0147	-0.0082
	(0.4492)	(0.6633)	(0.8504)	(0.3445)	(0.3446)	(0.5867)
TR_SKEW	0.0118	0.0123	0.0302	0.0305	0.0033	0.0029
	(0.0162)	(0.0138)	(0.0171)	(0.0272)	(0.3987)	(0.3814)
N	619	619	178	178	441	441
R2	0.03	0.05	0.10	0.19	0.01	0.01

Notes: p-values in parenthesis. The abbreviations MB and LD stand for Market Based and Language Driven monetary policy surprise series, respectively. In the former case, the raw changes in market prices within the tight time window around the communication is used as surprise series. In the latter case, our predicted market reactions from the language model are uses as surprise series.

the market-based series, the language-driven surprises exhibit no systematic relationship with past economic conditions. Moreover, the few remaining significant coefficients likely stem from extreme events or outliers, to which OLS is particularly sensitive.

To control for such potential outliers, the same regression exercise is conducted using median regression.[25] The results again show an improvement for the language-driven surprise series. During crises, such as the dot-com bubble burst or the 2008 financial crisis, markets tend to underestimate the Federal Reserve's actions, leading to larger negative surprises in periods of heightened uncertainty and volatility. This increases the likelihood of influential outliers.[26] To address this concern, we repeat the regression analysis using median regression, which is less sensitive to outliers.[27] The findings confirm that the language-driven series exhibit a weaker correlation with past economic and financial data, enhancing their properties regarding exogeneity.

3.4.3 Monetary Policy Effects on Macroeconomic Variables

Analogous to the market-based monetary policy surprises, we assess the dynamics of the monetary policy shock identified when using language-driven surprises as instruments. First, we report the F-statistics of the first-stage regression for two instrument series: the language-driven surprises related to the FOMC statements and the language-driven surprises related to the FOMC statements and policy-relevant Federal Reserve Board chair and vice chair speech transcripts.

Table 3.5 reports the F-statistics for both surprise series. These values are only marginally lower than those obtained using market-based surprises, indicating that the application of our text analysis method has minimal impact on the strength of the instruments.

Table 3.5: F-statistics for Language-Driven Surprise Series

	F-statistic
FOMC statements	5.77
FOMC statements and *policy-relevant* speech transcripts	16.30

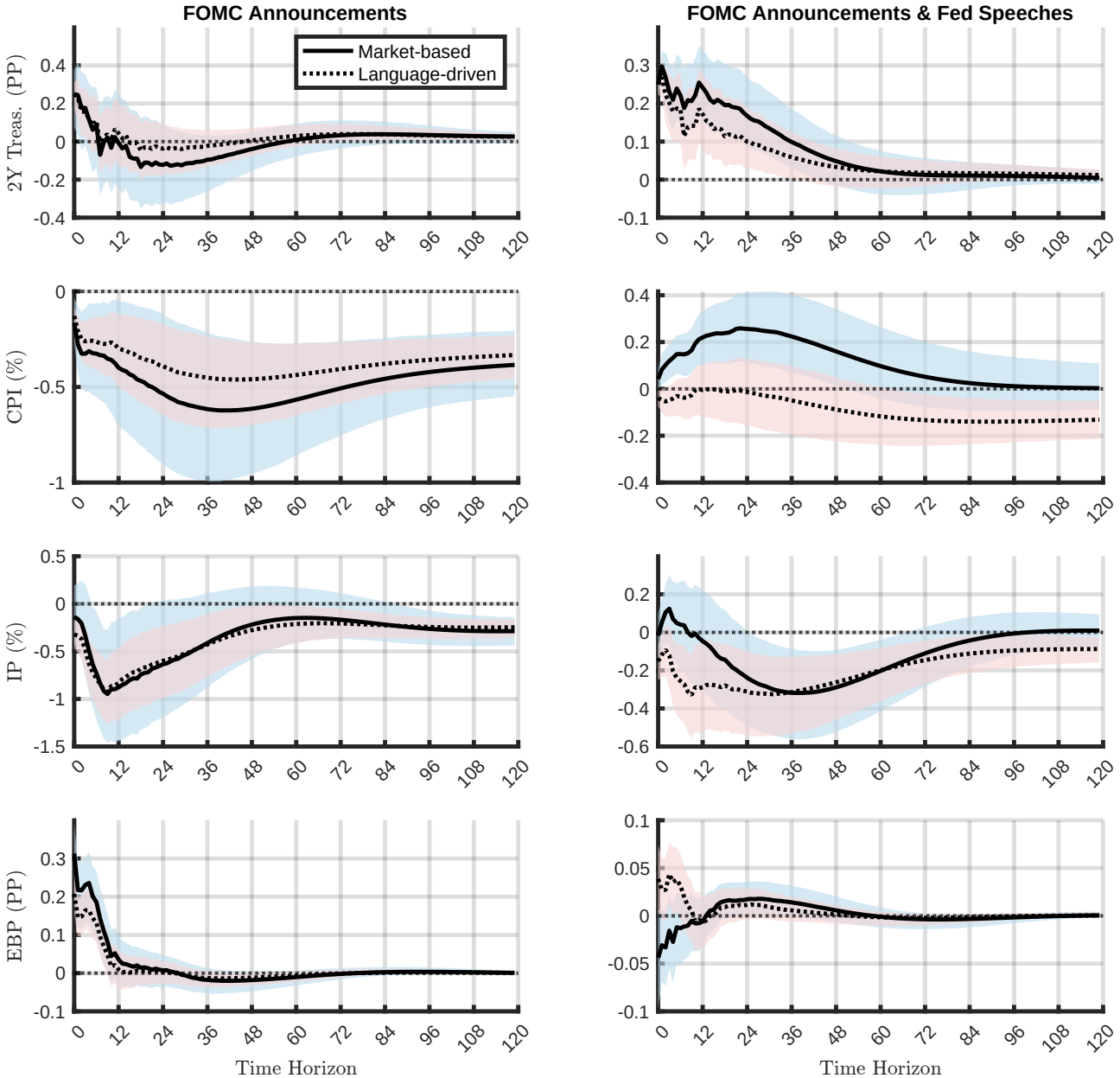
[25] The results are in Appendix 3.A.5

[26] Scatterplots for the two coefficients with the lowest p-values are shown in Appendix 3.A.5.

[27] Results are presented in Appendix 3.A.5.

Figure 3.5 displays the impulse responses to a monetary policy shock when using either market-based surprises or language-driven surprises as instruments for the identification. First, if we consider surprise series containing only FOMC announcements for identification, the dynamic responses of the macroeconomic variables differ only slightly.[28] Second, if we use the

Figure 3.5: Impulse Responses to a Monetary Policy Shock (Language-Driven Surprises)



Notes: The blue shaded areas represent the 5-95 percentiles of the impulse responses identified with market-based surprises and the red shaded areas the 5-95 percentiles of the impulse responses identified with language-driven surprises. The impulse responses are normalized to a 25 basis point increase in the two-year Treasury yield. Horizontal axis: time horizon in months.

[28]For the market-based surprises, we include only those dates for which a corresponding FOMC statement

surprises containing both FOMC announcements and Federal Reserve Board chair and vice chair speeches, the differences are bigger. With the language-based surprises, we no longer observe a price puzzle. The consumer price index does not react significantly on impact, but decreases below zero in the medium to long run. Moreover, industrial production decreases faster and stays negative for a prolonged period. Lastly, the excess bond premium increases (instead of decreasing), which is economically more intuitive.

3.5 Conclusion

This paper improves the identification of monetary policy shocks by combining NLP techniques with an expanded set of central bank communications. We extend the traditional market-based surprise measures—typically derived solely from FOMC announcements—to also include policy-relevant speeches by the Federal Reserve Board chair and vice chair. This expansion significantly improves the relevance of these surprises as instruments for identifying monetary policy shocks.

By leveraging a neural network trained on FOMC statements and speech transcripts, we construct a language-driven surprise series that isolates the component of market reactions driven purely by central bank communication. This approach mitigates endogeneity concerns inherent in traditional market-based surprises by filtering out confounding factors such as trader sentiment and market momentum. Our empirical findings confirm that language-based surprises produce impulse responses to monetary policy shocks that align more closely with economic theory.

Our results underscore the increasing importance of central bank communication as a monetary policy tool and demonstrate the potential of NLP in macroeconomic research. Future work could further refine our approach by incorporating additional forms of policy communication or testing alternative machine-learning architectures.

is available. This approach ensures that both surprise series are based on the same set of observations.

3.A Appendix I - Supplementary Material

3.A.1 Text Cleaning

We perform basic text cleaning by replacing repetitive and technical words with an abbreviation. Table 3.6 provides an overview of the abbreviations used. We use one word for the abbreviation except the *Federal Open Market Committee* is replaced with *Committee (FOMC)* because the FOMC statements usually refer to the *Committee* in their statements, whereas the Federal Reserve Board Chair and Vice Chair Speeches usually refer to the *FOMC*. However, both refer to the Federal Open Market Committee. Additionally, we restructure percentage numbers to match the following format: X.XX percent. Especially in the FOMC announcements, rate changes are sometimes marked in fractions, e.g. 1/4, making it hard to interpret for an NLP model. Thus, we similarly restructure all percentage numbers to facilitate comprehension. Finally, in the FOMC statements, we replaced the introductory sentence *Information received since the Committee (FOMC) met in January* with *Information received* for every month to prevent the model from learning from the timeline.

Table 3.6: Text Cleaning

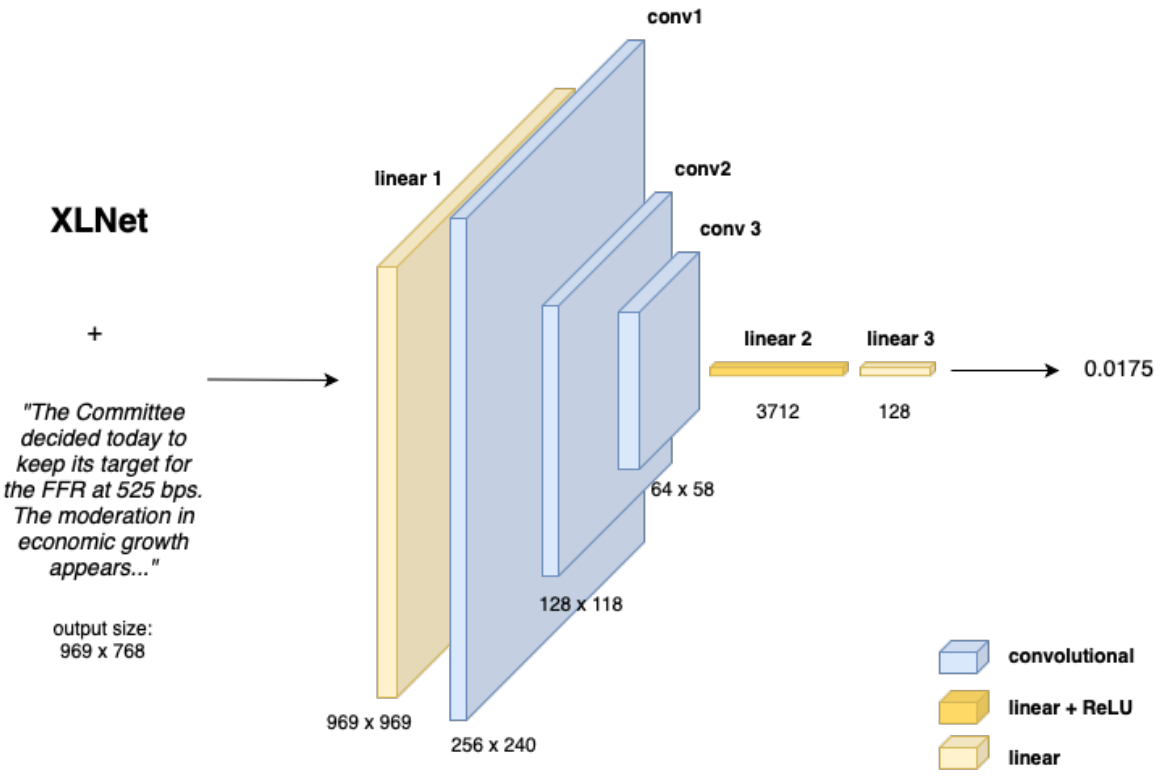
Words	Abbreviation
Federal Open Market Committee	Committe (FOMC)
federal funds rate	FFR
Board of Governors	BOG
Federal Reserve	FR
basis points	bps
basis point	bps
-basis-point	bps
mortgage-backed securities	mbs
Term Asset-Backed Securities	TABS

3.A.2 Neural Network Architecture

As mentioned, we use a pre-trained language model, XLNet-Base, developed and trained by Yang et al. (2020) and provided by the platform *Hugging Face*. The backbone neural network consists of 12 layers and 768 hidden states. On top of this, to train the model on our specific task, we add another six layers. A graphical representation of our additional structure is shown in figure 3.6. First, we increase the number of hidden states to match the length of our texts (number of tokens). Every token has a hidden state and associated weights in our first layer. Second, we add three convolutional layers, simultaneously breaking down the number of notes

and tokens. Convolutional neural networks (CNNs) were first proposed by Fukushima (1980) but gained greater attention in machine learning when Lecun et al. (1998) presented LeNet, an algorithm that detected handwritten numbers. CNNs learn via filter optimization and thus symmetrically reduce the number of notes and tokens. The notes remain fully connected using this procedure, making it prone to overfitting. Nonetheless, since we work with a small, heterogeneous text data set, we profit from the connectivity but must check that our model is not overfitting. Third, we add two linear layers, including activation functions, to decrease the number of notes to a single prediction. As an activation function, we use the Rectified Linear Unit (ReLU). In a neural network, the activation function transforms the summed weighted input from the node into the node's activation or output. ReLU is a piecewise linear function that will output the input directly if it is positive. Otherwise, it will output zero. It has become the default activation function for many types of neural networks because a model that uses it is easier to train and often performs better.

Figure 3.6: Neural Network Architecture



Remark: The graph presents the architecture of our neural network, taking as input the pre-trained XLNet and the FOMC statement. The input runs through different linear and convolutional layers, adapting its size to condense the information to a single number.

3.A.3 Overview of the Training Algorithm

1. Scale the continuous labels, i.e., the changes in the federal funds futures, removing the median and scaling the data according to the quantile range.
2. Define the hyperparameter:
    2.1. Learning rate: Defines how much the neuronal network should adapt its parameter after each iteration. We chose a rather low learning rate.
    2.2. Number of epochs: Defines how often the model sees the same data set to adapt its parameters.
    2.3. Loss function: Defines how the model should penalise its results compared to the true label. Since we work with linear prediction, we take the mean squared error.
    2.4. Batch size: Defines the number of statements we show to the model simultaneously.
3. Split the statements into training and test data. We use a 5-fold cross-validation, splitting the data set into five different subsets, always taking one of the splits as a test split and letting the model train with the other four splits.
4. Set the model to training mode to adapt its parameters.
5. Train the model by adapting its parameters such that the loss becomes decreases.
6. Stop updating the parameters
7. Evaluate the model using the test data.
8. Repeat from step four until the number of epochs (defined before) is reached. If the model is already overfitting, but the number of epochs is not yet reached, we should stop before.
9. Repeat from step three until all splits are tested.
10. Unscale the results.

3.A.4 Fine-Tune the Neural Network using FOMC statements

In this section, we map policy communication text to market-based monetary policy surprises, isolating the portion of the surprises driven exclusively by the FOMC statements. In contrast to Section 3.4, we abstain from using speech transcripts as part of the training set and use these only for predicting changes in market expectations. Truly, the statements are relatively short but comprise vocabulary that is highly policy-relevant.

Natural Language Processing

To obtain the prediction, we follow the same four steps explained in Section 3.4.1 with the only difference that we use a different dataset for fine-turning the neural network. We want our model to learn the mapping from FOMC statements to market-based surprises and apply this mapping to the speech transcripts. To train our model, we apply again a five-fold cross-validation using solely the FOMC statements. Thus, we have 80% of our FOMC statements in the training set and 20% in the test set. We use the same hyperparameters and algorithm as described in Appendix 3.A.2 and 3.A.3, respectively. Based on our test runs, we fixed the learning rate of our model to 1e-5 and the number of epochs to 8. The MSE of this set of hyperparameter for the five-fold cross-validation is displayed in Table 3.7.

Table 3.7: MSE of 5-Fold Cross-Validation after 8 Epochs

Number of Split	In-Sample MSE	Out-of-Sample MSE
1	0.0012598837	0.0012116632
2	0.0004515733	0.003678869
3	0.002779334	0.0034142113
4	0.0028541489	0.0032873761
5	0.0012911019	0.0020300713

After training our neural network, we obtain predictions from the FOMC statements and speech transcripts. In other words, we obtain a monetary policy surprise series that contains largely out-of-sample predictions.

Relation to Past Economic Information

After obtaining our predictions, we verify again whether our language-based surprises are uncorrelated to economic and financial information available before the announcement or speech. We conduct the same regression as in Section 3.4. The results are presented in Table 3.8 and are similar to the ones presented earlier in the paper. The only exception is the coefficient of nonfarm payroll surprises (NFP_SURP) that is here significant on the 5% level, for our combined results.

Similar to before, these results suggest that our language-driven approach exhibits no systematic relationship with past economic data. Hence, also using a limited training sample help to improve the surprise series to capture only the reaction to the central bank communication and abstract for other past information.

Table 3.8: Regression on MPS

	All		Statements		Speeches	
	LD	MB	LD	MB	LD	MB
NFP_SURP	0.0091	0.0280	0.0214	0.0875	0.0026	-0.0040
	(0.0259)	(0.1493)	(0.0434)	(0.0821)	(0.2913)	(0.7699)
NFP_12M	0.0002	0.0020	0.0010	0.0055	-0.0001	0.0009
	(0.3470)	(0.0156)	(0.1353)	(0.0296)	(0.4183)	(0.1987)
SP500_3M	0.0083	0.0528	0.0255	0.1214	0.0026	0.0221
	(0.1438)	(0.0362)	(0.1419)	(0.0871)	(0.4490)	(0.2073)
SLOPE_3M	-0.0003	-0.0042	-0.0036	-0.0115	0.0010	-0.0013
	(0.6716)	(0.2462)	(0.0707)	(0.2358)	(0.0248)	(0.5829)
BCOM_3M	0.0001	0.0094	0.0139	0.0623	-0.0048	-0.0082
	(0.9829)	(0.6633)	(0.3641)	(0.3445)	(0.1001)	(0.5867)
TR_SKEW	0.0023	0.0123	0.0052	0.0305	0.0007	0.0029
	(0.0099)	(0.0138)	(0.0327)	(0.0272)	(0.3539)	(0.3814)
N	619	619	178	178	441	441
R2	0.04	0.05	0.20	0.19	0.03	0.01

Notes: p-values in parenthesis. The abbreviations MB and LD stand for Market Based and Language Driven monetary policy surprise series, respectively. In the former case, the raw changes in market prices within the tight time window around the communication is used as surprise series. In the latter case, our predicted market reactions from the language model are uses as surprise series.

Monetary Policy Effects on Macroeconomic Variables

Analogous to the Section 3.3 and 3.4, we assess the dynamics of the monetary policy shock identified when using language-driven surprises as instruments. First, we report the F-statistics of the first-stage regression for two instrument series: the language-driven surprises related to the FOMC statements and the language-driven surprises related to the FOMC statements and policy-relevant Federal Reserve Board chair and vice chair speech transcripts.

Table 3.9 reports the F-statistics for both surprise series. These values, especially for the FOMC statements and policy-relevant speeches, are drastically lower compared to the results obtained in Section 3.3 and 3.4. This indicates that the model, trained only on FOMC statements, retrieves less information from the policy-relevant speeches.

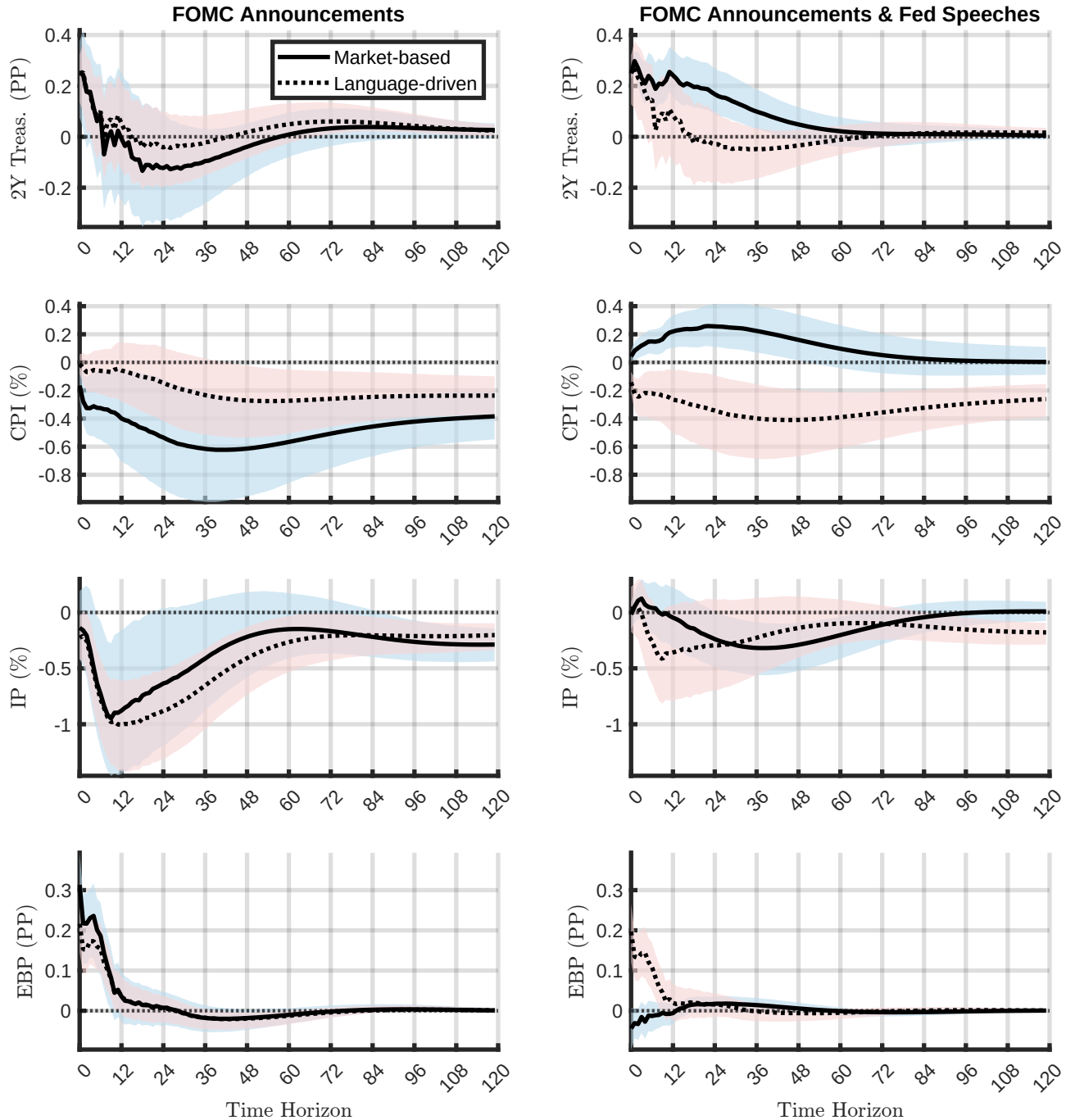
Figure 3.7 displays the impulse responses to a monetary policy shock when using either market-based surprises or language-driven surprises as instruments for identification. First, for the results using only FOMC announcements for identification, we observe marginally different results. With the language-based surprises, we obtain a negative impact on the CPI only after

Table 3.9: F-statistics for Language-Driven Surprise Series

	F-statistic
FOMC statements	3.31
FOMC statements and *policy-relevant* speech transcripts	2.97

some years. However, the reaction of industrial production is much stronger using our series. Second, for the results using both FOMC announcements and Federal Reserve Board chair and vice chair speeches, the deviation is even bigger. Using the language-driven surprises, the two-year treasury yield mean-reverts much quicker. Moreover, the CPI reacts negatively on impact and remains negative for all periods. In contrast, industrial production decreases some months after impact but reverts shortly after. Similar to the results obtained in Section 3.4, the excess bond premium increases in reaction to the monetary policy shock. In essence, the results are similar to the ones obtained in Section 3.4 with the big difference that our instrument remains weak even when adding the speech transcripts.

Figure 3.7: Impulse Responses to a Monetary Policy Shock (Language-Driven Surprises)



Notes: The blue shaded areas represent the 5-95 percentiles of the impulse responses identified with market-based surprises and the red shaded areas the 5-95 percentiles of the impulse responses identified with language-driven surprises. The impulse responses are normalized to a 25 basis point increase in the two-year Treasury yield. Horizontal axis: time horizon in months.

3.A.5 Robustness to Outliers

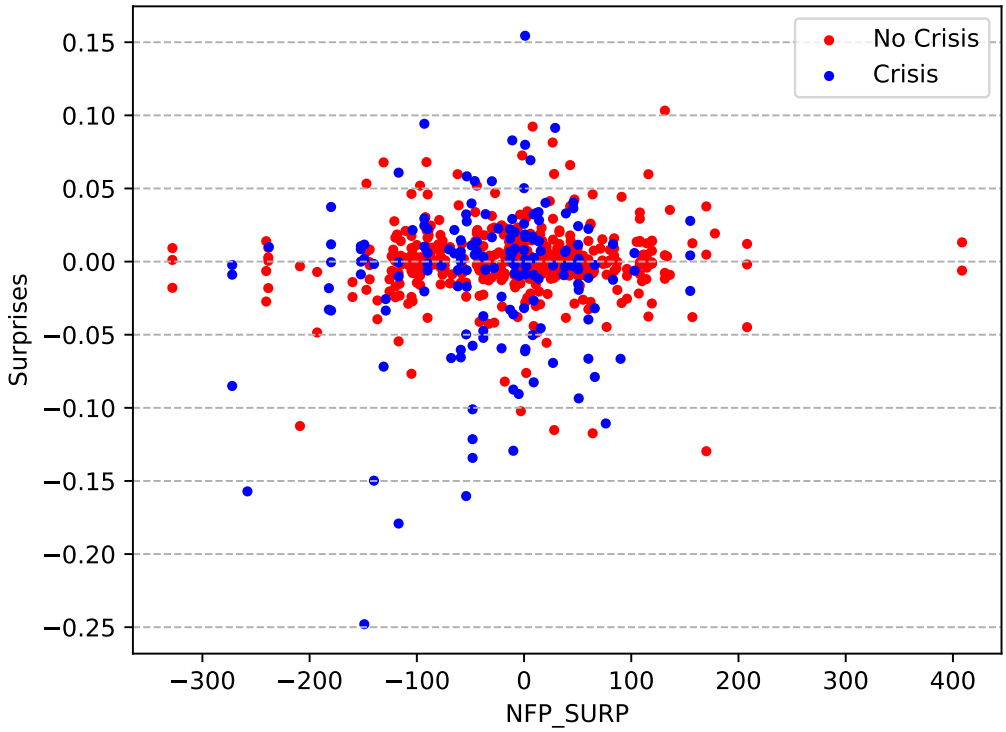
This section comprises the additional results described in section 3.4. Table 3.10 displays the results from the median regression, using the same specifications as for the OLS regression in section 3.4. Figures 3.8 and 3.9 plot the outcomes, the language-driven surprises, on the two covariates with the lowest p-values, the nonfarm payroll surprises and the average skewness of the ten-year treasury yields. Dates during times of crisis are plotted in blue, other dates are plotted in red. Times of crisis include the dot-com bubble burst between 1.1.2020 and 31.12.2002 and the financial crisis between 1.7.2007 and 1.1.2010.

Table 3.10: Median Regression on MPS

	All		Statements		Speeches	
	LD	MB	LD	MB	LD	MB
NFP_SURP	0.0049	-0.0013	0.0120	-0.0049	0.0010	-0.0000
	(0.5508)	(0.8398)	(0.7818)	(0.9002)	(0.9038)	(0.9998)
NFP_12M	0.0003	0.0004	0.0035	0.0035	0.0001	0.0000
	(0.5536)	(0.2395)	(0.1281)	(0.0913)	(0.7897)	(0.9994)
SP500_3M	0.0037	0.0016	-0.0140	0.0538	0.0071	0.0000
	(0.7184)	(0.8426)	(0.7820)	(0.2429)	(0.4955)	(0.9999)
SLOPE_3M	-0.0002	-0.0007	-0.0030	-0.0119	-0.0006	-0.0000
	(0.9058)	(0.5335)	(0.6497)	(0.0519)	(0.6782)	(0.9995)
BCOM_3M	-0.0069	0.0007	0.0042	0.0368	-0.0070	-0.0000
	(0.4230)	(0.9133)	(0.9174)	(0.3164)	(0.4239)	(1.0000)
TR_SKEW	0.0037	0.0005	0.0129	0.0148	0.0025	0.0000
	(0.1297)	(0.8011)	(0.2375)	(0.1370)	(0.3234)	(0.9999)
N	619	619	178	178	441	441

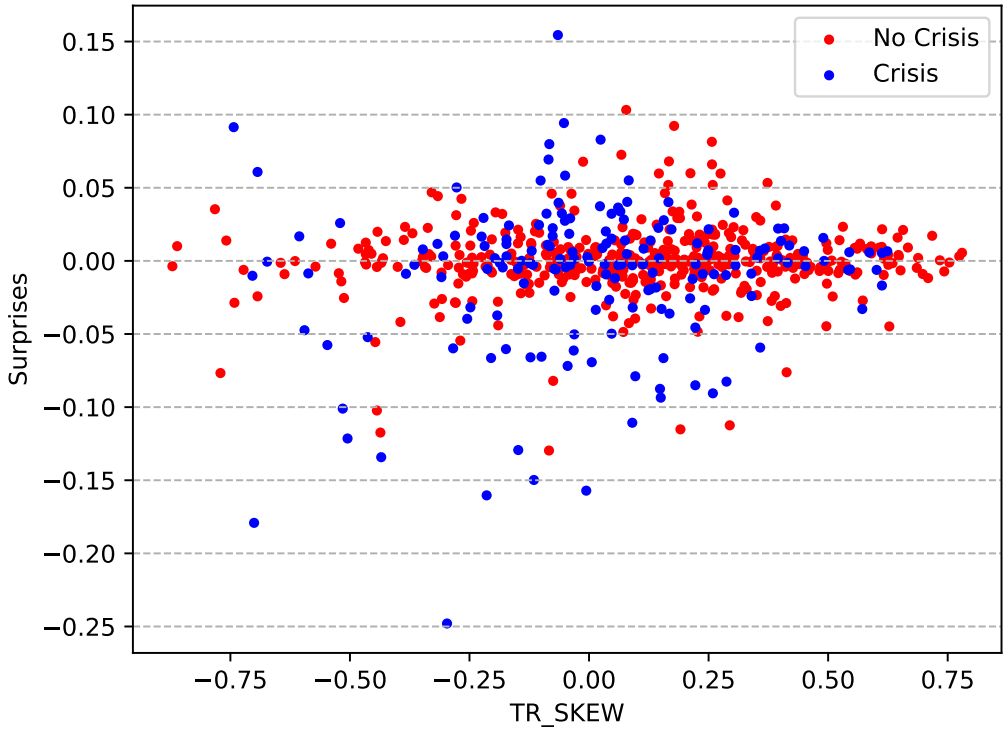
Notes: p-values in parenthesis. The abbreviations MB and LD stand for Market Based and Language Driven monetary policy surprise series, respectively. In the former case, the raw changes in market prices within the tight time window around the communication is used as surprise series. In the latter case, our predicted market reactions from the language model are uses as surprise series.

Figure 3.8: Language driven surprises on nonfarm payroll surprises



Notes: The graph presents the surprises used in the analysis on the y-axis and the most recent nonfarm payroll surprises on the x-axis.

Figure 3.9: Language driven surprises on average skewness of the treasury yield



Notes: The graph presents the surprises used in the analysis on the y-axis and the average skewness of the ten-year treasury yield in the last month on the x-axis.

Bibliography

Abadie, Alberto (2021). "Using Synthetic Controls: Feasibility, Data Requirements, and Methodological Aspects". In: *Journal of Economic Literature* 59.2, pp. 391–425.

Abadie, Alberto, Alexis Diamond, and Jens Hainmueller (2010). "Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program". In: *Journal of the American statistical Association* 105.490, pp. 493–505.

— (2015). "Comparative Politics and the Synthetic Control Method". In: *American Journal of Political Science* 59.2, pp. 495–510.

Abadie, Alberto and Javier Gardeazabal (2003). "The economic costs of conflict: A case study of the Basque Country". In: *American economic review* 93.1, pp. 113–132.

Abadie, Alberto and Jérémy L'Hour (2, 2021). "A Penalized Synthetic Control Estimator for Disaggregated Data". In: *Journal of the American Statistical Association* 116.536, pp. 1817–1834.

Abramitzky, Ran, Leah Boustan, Elisa Jacome, and Santiago Perez (2021). "Intergenerational Mobility of Immigrants in the United States over Two Centuries". In: *American Economic Review* 111.2, pp. 580–608.

Adermon, Adrian, Mikael Lindahl, and Daniel Waldenström (2018). "Intergenerational Wealth Mobility and the Role of Inheritance: Evidence from Multiple Generations". In: *The Economic Journal* 128.612, F482–F513.

Arellano, Manuel, Richard Blundell, and Stéphane Bonhomme (2017). "Earnings and Consumption Dynamics: A Nonlinear Panel Data Framework". In: *Econometrica* 85.3, pp. 693–734.

Arkhangelsky, Dmitry, Susan Athey, David A. Hirshberg, Guido W. Imbens, and Stefan Wager (1, 2021). "Synthetic Difference-in-Differences". In: *American Economic Review* 111.12, pp. 4088–4118.

Aruoba, S. Boragan and Thomas Drechsel (2024). "Identifying Monetary Policy Shocks: A Natural Language Approach". In: *NBER Working Paper 32417*.

Athey, Susan, Mohsen Bayati, Nikolay Doudchenko, Guido Imbens, and Khashayar Khosravi (2, 2021). "Matrix Completion Methods for Causal Panel Data Models". In: *Journal of the American Statistical Association* 116.536, pp. 1716–1730.

Athey, Susan and Guido W. Imbens (2006). "Identification and Inference in Nonlinear Difference-in-Differences Models". In: *Econometrica* 74.2, pp. 431–497.

Bauer, Michael and Mikhail Chernov (2024). "Interest Rate Skewness and Biased Beliefs". In: *Journal of Finance* 79.1, pp. 173–217.

Bauer, Michael D and Eric T Swanson (2023a). "A Reassessment of Monetary Policy Surprises and High-Frequency Identification". In: *NBER Macroeconomics Annual* 37.

— (2023b). "An Alternative Explanation for the "FED Information Effect"". In: *American Economic Review* 113.3, pp. 664–700.

Bauer, Philipp and Regina T. Riphahn (2006). "Timing of school tracking as a determinant of intergenerational transmission of education". In: *Economics Letters* 91.1, pp. 90–97.

Becker, Gary S., Scott Duke Kominers, Kevin M. Murphy, and Jörg L. Spenkuch (2018). "A Theory of Intergenerational Mobility". In: *Journal of Political Economy* 126 (S1), S7–S25.

Becker, Gary S. and Nigel Tomes (1979). "An Equilibrium Theory of the Distribution of Income and Intergenerational Mobility". In: *Journal of Political Economy* 87.6, pp. 1153–1189.

— (1986). "Human Capital and the Rise and Fall of Families". In: *Journal of Labor Economics* 4.3, S1–S39.

Ben-Michael, Eli, Avi Feller, and Jesse Rothstein (2, 2021). "The Augmented Synthetic Control Method". In: *Journal of the American Statistical Association* 116.536, pp. 1789–1803.

Berger, Marius and Bruno Lanz (25, 2020). "Minimum wage regulation in Switzerland: survey evidence for restaurants in the canton of Neuchâtel". In: *Swiss Journal of Economics and Statistics* 156.1, p. 20.

Bertsch, Christoph, Isaiah Hull, Robin L Lumsdaine, Xin Zhang, Sveriges Riksbank, Francesco Bianchi, James Chapman, Ricardo Correa, Michael Ehrmann, Juri Marcucci, and Tho Pham (2024). *Central Bank Mandates and Monetary Policy Stances: through the Lens of Federal Reserve Speeches **. Tech. rep.

Bhattacharya, Debopam and Bhashkar Mazumder (2011). "A nonparametric analysis of black-white differences in intergenerational income mobility in the United States: Intergenerational income mobility". In: *Quantitative Economics* 2.3, pp. 335–379.

Biewen, Martin, Bernd Fitzenberger, and Marian Rümmele (15, 2022). *Using Distribution Regression Difference-in-Differences to Evaluate the Effects of a Minimum Wage Introduction on the Distribution of Hourly Wages and Hours Worked*. Rochester, NY.

Björklund, Anders and Laura Chadwick (1, 2003). "Intergenerational income mobility in permanent and separated families". In: *Economics Letters* 80.2, pp. 239–246.

Björklund, Anders and Markus Jäntti (3, 2011). "Intergenerational Income Mobility and the Role of Family Background". In: *The Oxford Handbook of Economic Inequality*. Ed. by Brian Nolan, Wiemer Salverda, and Timothy M. Smeeding. Oxford University Press, p. 0.

Björklund, Anders, Jesper Roine, and Daniel Waldenström (1, 2012). "Intergenerational top income mobility in Sweden: Capitalist dynasties in the land of equal opportunity?" In: *Journal of Public Economics* 96.5, pp. 474–484.

Blanden, Jo, Paul Gregg, and Lindsey Macmillan (1, 2007). "Accounting for Intergenerational Income Persistence: Noncognitive Skills, Ability and Education". In: *The Economic Journal* 117.519, pp. C43–C60.

Blinder, Alan S, Michael Ehrmann, Marcel Fratzscher, Jakob De Haan, and David-Jan Jansen (2008). "Central Bank Communication and Monetary Policy: A Survey of Theory and Evidence". In: *Journal of Economic Literature* 46.4, pp. 910–945.

Blundell, Richard, Luigi Pistaferri, and Ian Preston (2008). "Consumption Inequality and Partial Insurance". In: *American Economic Review* 98.5, pp. 1887–1921.

Boar, Corina and Danial Lashkari (2021). "Occupational Choice and the Intergenerational Mobility of Welfare". In: NBER Working Paper Series.

Bonhomme, Stéphane and Jean-Marc Robin (1, 2009). "Assessing the Equalizing Force of Mobility Using Short Panels: France, 1990–2000". In: *The Review of Economic Studies* 76.1, pp. 63–92.

Caldara, Dario and Edward Herbst (2019). "Monetary Policy, Real Activity, and Credit Spreads: Evidence from Bayesian Proxy SVARs". In: *American Economic Journal: Macroeconomics* 11.1, pp. 157–192.

Callaway, Brantly and Weige Huang (2018). "Local Intergenerational Elasticities". In: *SSRN Electronic Journal*.

Callaway, Brantly and Tong Li (2019). "Quantile treatment effects in difference in differences models with panel data". In: *Quantitative Economics* 10.4, pp. 1579–1618.

Carneiro, Pedro, Italo López García, Kjell G. Salvanes, and Emma Tominey (2021). "Intergenerational Mobility and the Timing of Parental Income". In: *Journal of Political Economy* 129.3, pp. 757–788.

Cattaneo, Matias D., Yingjie Feng, and Rocio Titiunik (2, 2021). "Prediction Intervals for Synthetic Control Methods". In: *Journal of the American Statistical Association* 116.536, pp. 1865–1880.

Chadwick, Laura and Gary Solon (1, 2002). "Intergenerational Income Mobility Among Daughters". In: *American Economic Review* 92.1, pp. 335–344.

Chen, Yi-Ting (2020). "A distributional synthetic control method for policy evaluation". In: *Journal of Applied Econometrics* 35.5, pp. 505–525.

Chernozhukov, Victor, Iván Fernández-Val, and Siyi Luo (18, 2023). *Distribution Regression with Sample Selection, with an Application to Wage Decompositions in the UK*. arXiv: 1811.11603[econ].

Chernozhukov, Victor, Iván Fernández-Val, Jonas Meier, Aico van Vuuren, and Francis Vella (8, 2024). *Conditional Rank-Rank Regression*. arXiv: 2407.06387[econ].

Chernozhukov, Victor, Iván Fernández-Val, and Blaise Melly (2013). "Inference on Counterfactual Distributions". In: *Econometrica* 81.6, pp. 2205–2268.

Chernozhukov, Victor, Kaspar Wüthrich, and Yinchu Zhu (2021). "An exact and robust conformal inference method for counterfactual and synthetic controls". In: *Journal of the American Statistical Association*, pp. 1–16.

Chetty, Raj and Nathaniel Hendren (1, 2018a). "The Impacts of Neighborhoods on Intergenerational Mobility I: Childhood Exposure Effects*". In: *The Quarterly Journal of Economics* 133.3, pp. 1107–1162.

— (1, 2018b). "The Impacts of Neighborhoods on Intergenerational Mobility II: County-Level Estimates*". In: *The Quarterly Journal of Economics* 133.3, pp. 1163–1228.

Chetty, Raj, Nathaniel Hendren, Maggie R Jones, and Sonya R Porter (1, 2020). "Race and Economic Opportunity in the United States: an Intergenerational Perspective*". In: *The Quarterly Journal of Economics* 135.2, pp. 711–783.

Chetty, Raj, Nathaniel Hendren, and Lawrence F. Katz (2016). "The Effects of Exposure to Better Neighborhoods on Children: New Evidence from the Moving to Opportunity Experiment". In: *American Economic Review* 106.4, pp. 855–902.

Chetty, Raj, Nathaniel Hendren, Patrick Kline, and Emmanuel Saez (1, 2014). "Where is the land of Opportunity? The Geography of Intergenerational Mobility in the United States *". In: *The Quarterly Journal of Economics* 129.4, pp. 1553–1623.

Cholli, Neil A. and Steven N. Durlauf (2022). "Intergenerational Mobility". In: NBER Working Paper Series.

Chuard, Patrick and Veronica Grassi (2020). "Switzer-Land of Opportunity: Intergenerational Income Mobility in the Land of Vocational Education". In: *SSRN Electronic Journal*.

Cieslak, Anna and Andreas Schrimpf (2019). "Non-Monetary News in Central Bank Communication". In: *Journal of International Economics* 118, pp. 293–315.

Corak, Miles (2013). "Income Inequality, Equality of Opportunity, and Intergenerational Mobility". In: *Journal of Economic Perspectives* 27.3, pp. 79–102.

— (16, 2020). "The Canadian Geography of Intergenerational Income Mobility". In: *The Economic Journal* 130.631, pp. 2134–2174.

Dahl, Gordon B., Andreas Ravndal Kostøl, and Magne Mogstad (1, 2014). "Family Welfare Cultures *". In: *The Quarterly Journal of Economics* 129.4, pp. 1711–1752.

Deutscher, Nathan and Bhashkar Mazumder (2023). "Measuring Intergenerational Income Mobility: A Synthesis of Approaches". In: *Journal of Economic Literature* 61.3, pp. 988–1036.

Dickens, Richard (2000). "Caught in a Trap? Wage Mobility in Great Britain: 1975–1994". In: *Economica* 67.268, pp. 477–497.

Doh, Taeyoung, Dongho Song, and Shu-Kuei X Yang (2022). "Deciphering Federal Reserve Communication via Text Analysis of Alternative FOMC Statements". In: *Federal Reserve Bank of Kansas City Working Paper Forthcoming*.

Doudchenko, Nikolay and Guido W. Imbens (2016). *Balancing, regression, difference-in-differences and synthetic control methods: A synthesis*. National Bureau of Economic Research.

Durlauf, Steven N. (1, 1996). "A theory of persistent income inequality". In: *Journal of Economic Growth* 1.1, pp. 75–93.

Durlauf, Steven N. and Ananth Seshadri (2018). "Understanding the Great Gatsby Curve". In: *NBER Macroeconomics Annual* 32, pp. 333–393.

Favre, Giacomin, Joël Floris, and Ulrich Woitek (1, 2018). *Intergenerational Mobility in the 19th Century: Micro-Level Evidence from the City of Zurich*. Rochester, NY.

Ferman, Bruno and Cristine Pinto (2021). "Synthetic controls with imperfect pretreatment fit". In: *Quantitative Economics* 12.4, pp. 1197–1221.

Fernández-Val, Iván, Jonas Meier, Aico van Vuuren, and Francis Vella (3, 2024). *Distribution Regression Difference-In-Differences*. arXiv: 2409.02311[econ].

Fields, Gary Sheldon and Efe Ok (1999). "The measurement of income mobility: an introduction to the literature". In: *The Measurement of Income Mobility: An Introduction to the Literature | SpringerLink*.

Firpo, Sergio and Vitor Possebom (25, 2018). "Synthetic Control Method: Inference, Sensitivity Analysis and Confidence Sets". In: *Journal of Causal Inference* 6.2.

Fukushima, Kunihiko (1980). "Neocognitron: A Self-Organizing Neural Network Model for a Mechanism of Pattern Recognition Unaffected by Shift in Position". In: *Biological Cybernetics* 36.4, pp. 193–202.

Gardner, Ben, Chiara Scotti, and Clara Vega (2022). "Words Speak as Loudly as Actions: Central Bank Communication and the Response of Equity Prices to Macroeconomic Announcements". In: *Journal of Econometrics* 231.2, pp. 387–409.

Gertler, Mark and Peter Karadi (2015). "Monetary Policy Surprises, Credit Costs, and Economic Activity". In: *American Economic Journal: Macroeconomics* 7.1, pp. 44–76.

Gilchrist, Simon and Egon Zakrajšek (2012). "Credit Spreads and Business Cycle Fluctuations". In: *American Economic Review* 102.4, pp. 1692–1720.

Gunsilius, F. F. (2023). "Distributional Synthetic Controls". In: *Econometrica* 91.3, pp. 1105–1117.

Gürkaynak, Refet S, Brian Sack, and Eric Swanson (2005a). *The Sensitivity of Long-Term Interest Rates to Economic News: Evidence and Implications for Macroeconomic Models*. Tech. rep.

Gürkaynak, Refet S., Brian Sack, and Eric T. Swanson (2005b). "Do Actions Speak Louder Than Words? The Response of Asset Prices to Monetary Policy Actions and Statements". In: *International Journal of Central Banking* 1.1.

Gürkaynak, Refet S., Brian Sack, and Jonathan H. Wright (2007). "The U.S. Treasury Yield Curve: 1961 to the Present". In: *Journal of Monetary Economics* 54.8, pp. 2291–2304.

Handlan, Amy (2022). *Text Shocks and Monetary Surprises: Text Analysis of FOMC Statements with Machine Learning*. Tech. rep.

Hanson, Samuel G. and Jeremy C. Stein (2015). "Monetary Policy and Long-Term Real Rates". In: *Journal of Financial Economics* 115.3, pp. 429–448.

Havnes, Tarjei and Magne Mogstad (1, 2015). "Is universal child care leveling the playing field?" In: *Journal of Public Economics*. The Nordic Model 127, pp. 100–114.

Heckman, James J and Stefano Mosso (2014). "The Economics of Human Development and Social Mobility". In: *Annual Review of Economics* 6, pp. 689–733.

Jann, Ben (30, 2019). "iscogen: Stata module to translate ISCO codes." In: *Statistical Software Components*.

Jann, Ben and Simon Seiler (2014). "A new methodological approach for studying intergenerational mobility with an application to Swiss data". In: University of Bern Social Sciences Working Paper.

Jardim, Ekaterina, Mark C. Long, Robert Plotnick, Emma van Inwegen, Jacob Vigdor, and Hilary Wething (2022). "Minimum-Wage Increases and Low-Wage Employment: Evidence from Seattle". In: *American Economic Journal: Economic Policy* 14.2, pp. 263–314.

Jayawickrema, Vishuddhi and Eric Swanson (2023). *Speeches by the Fed Chair Are More Important Than FOMC Announcements: An Improved High-Frequency Measure of U.S. Monetary Policy Shocks*. Tech. rep.

Kalambaden, Preetha and Isabel Z Martınez (2021). "Intergenerational Mobility Along Multiple Dimensions Evidence from Switzerland". In: *Working Paper*, p. 84.

Kato, Masahiro, Akari Ohda, Masaaki Imaizumi, and Kenichiro McAlinn (24, 2023). *Synthetic Control Methods by Density Matching under Implicit Endogeneity*. arXiv: 2307.11127[cs, econ,stat].

Kerssenfischer, Mark and Maik Schmeling (2024). "What moves markets?" In: *Journal of Monetary Economics* 145.

Knudsen, Eric I., James J. Heckman, Judy L. Cameron, and Jack P. Shonkoff (5, 2006). "Economic, neurobiological, and behavioral perspectives on building America's future workforce". In: *Proceedings of the National Academy of Sciences* 103.27, pp. 10155–10162.

Kohler, Ulrich (2, 2023). "PSIDTOOLS: Stata module to facilitate access to Panel Study of Income Dynamics (PSID)". In: *Statistical Software Components*.

Lecun, Y., L. Bottou, Y. Bengio, and P. Haffner (1998). "Gradient-Based Learning Applied to Document Recognition". In: *Proceedings of the IEEE* 86.11, pp. 2278–2324.

Li, Kathleen T. (11, 2020). "Statistical Inference for Average Treatment Effects Estimated by Synthetic Control Methods". In: *Journal of the American Statistical Association* 115.532, pp. 2068–2083.

Lindbeck, Assar, Sten Nyberg, and Jörgen W. Weibull (1, 1999). "Social Norms and Economic Incentives in the Welfare State*". In: *The Quarterly Journal of Economics* 114.1, pp. 1–35.

Loughran, Tim and Bill McDonald (2011). "When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks". In: *Journal of Finance* 66.1, pp. 35–65.

Lucca, David O. and Emanuel Moench (2015). "The Pre-FOMC Announcement Drift". In: *The Journal of Finance* LXX.1, pp. 329–371.

Mertens, Karel and Morten O. Ravn (2013). "The Dynamic Effects of Personal and Corporate Income Tax Changes in the United States". In: *American Economic Review* 103.4, pp. 1212–1247.

Miranda-Agrippino, Silvia and Giovanni Ricco (2021). "The Transmission of Monetary Policy Shocks". In: *American Economic Journal: Macroeconomics* 13.3, pp. 74–107.

Mogstad, Magne and Gaute Torsvik (4, 2023). "Family Background, Neighborhoods and Intergenerational Mobility". In: *Handbook of the Economics of the Family*. 1st Edition. Handbooks in Economics.

Nakamura, Emi and Jón Steinsson (2018). "High-Frequency Identification of Monetary Non-Neutrality: The Information Effect". In: *The Quarterly Journal of Economics* 133.3, pp. 1283–1330.

Nelder, J. A. and R. Mead (1, 1965). "A Simplex Method for Function Minimization". In: *The Computer Journal* 7.4, pp. 308–313.

Nelsen, Roger B. (2006). *An introduction to copulas*. 2nd ed. Springer series in statistics. New York: Springer. 269 pp.

Neuhierl, Andreas and Michael Weber (2018). "Monetary Momentum". In: *NBER Working Paper 24748*.

Nybom, Martin and Jan Stuhler (2017). "Biases in Standard Measures of Intergenerational Income Dependence". In: *Journal of Human Resources* 52.3, pp. 800–825.

Ochs, Adrian C R (2021). "A New Monetary Policy Shock with Text Analysis". In: *Cambridge Working Papers in Economics*. Cambridge Working Papers in Economics.

Peri, Giovanni and Vasil Yasenov (2019). "The labor market effects of a refugee wave synthetic control method meets the mariel boatlift". In: *Journal of Human Resources* 54.2, pp. 267–309.

Ramey, V. A. (2016). "Macroeconomic Shocks and Their Propagation". In: *Handbook of Macroeconomics*. Vol. 2. Elsevier B.V., pp. 71–162.

Richey, Jeremiah and Alicia Rosburg (2018). "Decomposing economic mobility transition matrices". In: *Journal of Applied Econometrics* 33.1, pp. 91–108.

Robbins, Michael W., Jessica Saunders, and Beau Kilmer (2, 2017). "A Framework for Synthetic Control Methods With High-Dimensional, Micro-Level Data: Evaluating a Neighborhood-Specific Crime Intervention". In: *Journal of the American Statistical Association* 112.517, pp. 109–126.

Romer, Christina D and David H Romer (2004). "A New Measure of Monetary Shocks: Derivation and Implications". In: *American Economic Review* 94.4, pp. 1055–1084.

Shi, Claudia, Dhanya Sridhar, Vishal Misra, and David Blei (3, 2022). "On the Assumptions of Synthetic Control Methods". In: *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*. International Conference on Artificial Intelligence and Statistics. PMLR, pp. 7163–7175.

Shorrocks, A. F. (1978). "The Measurement of Mobility". In: *Econometrica* 46.5, pp. 1013–1024.

Solon, Gary (2004). "A model of intergenerational mobility variation over time and place". In: *Generational Income Mobility in North America and Europe*. Ed. by Miles Corak. Cambridge: Cambridge University Press, pp. 38–47.

Stepanyan, Ara and Jorge Salas (13, 2020). "Distributional Implications of Labor Market Reforms: Learning from Spain's Experience1". In: *IMF Working Papers* 2020.29.

Stock, J. H. and M. W. Watson (2012). "Disentangling the Channels of the 2007-09 Recession". In: *Brookings Papers on Economic Activity* 1, pp. 81–135.

Swanson, Eric T and John C Williams (2012). "Measuring the Effect of the Zero Lower Bound on Medium- and Longer-Term Interest Rates". In: *FRB San Francisco WP*.

Vaswani, Ashish, Google Brain, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin (2017). "Attention Is All You Need". In: *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems*.

Waldkirch, Andreas, Serena Ng, and Donald Cox (31, 2004). "Intergenerational Linkages in Consumption Behavior". In: *Journal of Human Resources* XXXIX.2, pp. 355–381.

Woodford, Michael (2005). "Central Bank Communication and Policy Effectiveness". In: *NBER Working Paper 11898*.

Yang, Zhilin, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le (2020). "XLNet: Generalized Autoregressive Pretraining for Language Understanding". In: *33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*.

Zhang, Yuhang, Yue Liu, and Zhihua Zhang (28, 2023). *Constructing Synthetic Treatment Groups without the Mean Exchangeability Assumption*. arXiv: 2309.16409[cs,stat].

Statement of Authorship

I declare herewith that I wrote this thesis on my own, without the help of others. Wherever I have used permitted sources of information, I have made this explicitly clear within my text and I have listed the referenced sources. I understand that if I do not follow these rules that the Senate of the University of Bern is authorized to revoke the title awarded on the basis of this thesis according to Article 36, paragraph 1, litera r of the University Act of September 5th, 1996.

Selbständigkeitserklärung

Ich erkläre hiermit, dass ich diese Arbeit selbständig verfasst und keine anderen als die angegebenen Quellen benutzt habe. Alle Koautorenschaften sowie alle Stellen, die wörtlich oder singemäss aus Quellen entnommen wurden, habe ich als solche gekennzeichnet. Mir ist bekannt, dass andernfalls der Senat gemäss Artikel 36 Absatz 1 Buchstabe o des Gesetzes vom 5. September 1996 über die Universität zum Entzug des aufgrund dieser Arbeit verliehenen Titels berechtigt ist.

Bern, 11. Februar 2025

Marc Schranz