Essays in Econometrics

by

Marta Boczoń

BA in Quantitative Methods in Economics & Information Systems
Warsaw School of Economics, 2012

MS in Statistics
Humboldt University of Berlin, 2015

MA in Economics
University of Pittsburgh, 2020

Submitted to the Graduate Faculty of

the Dietrich School of Arts and Sciences in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2021

UNIVERSITY OF PITTSBURGH

DIETRICH SCHOOL OF ARTS AND SCIENCES

This dissertation was presented

by

Marta Boczoń

It was defended on

April 2nd, 2021

and approved by

Thesis Advisor, Professor Jean-François Richard
Department of Economics, University of Pittsburgh

and

Professor Arie Beresteanu
Department of Economics, University of Pittsburgh

Professor Alistair J. Wilson
Department of Economics, University of Pittsburgh

Professor Roman Liesenfeld
Institute of Econometrics and Statistics, University of Cologne

Essays in Econometrics

Marta Boczoń, PhD

University of Pittsburgh, 2021

This dissertation consists of three chapters and is inspired by current economic issues affecting the majority (if not all) of economic agents, such as recession tracking, matching algorithms and, economic inequality.

In the first chapter, I measure the uncertainty affecting estimates of economic inequality in the US and investigate how accounting for properly estimated standard errors can affect the results of empirical and structural macroeconomic studies. While focusing on income and wealth shares within the top 10 percent, my results suggest that ignoring uncertainties in estimated shares can lead to statistically imprecise conclusions about the past and current levels of top income and wealth inequality, and therefore, lead to inaccurate predictions and potentially ineffective policy recommendations.

In the second chapter, Professor Jean-François Richard and I propose a hybrid version of Dynamic Stochastic General Equilibrium models with emphasis on parameter invariance and tracking performance at times of rapid changes (recessions). Our methodology is illustrated by an application to a pilot Real Business Cycle model for the US economy from 1948 to 2019, where we highlight the model's parameters invariance, tracking, and 1- to 3-step ahead forecasting performance, outperforming those of an unconstrained benchmark Vector AutoRegressive model.

In the third and final chapter, Professor Alistair J. Wilson and I analyze an existing matching procedure designed to solve a complex constrained assignment problem for one of the most successful pan-European ventures: the UEFA Champions League. Relying upon a combination of theory, structural estimation, and simulation, we outline a quantitative methodology aimed at assessing a highly transparent (but combinatorically complex) tournament's assignment procedure and provide evidence that the UEFA assignment rule is effectively a "constrained-best" in terms of pairwise independence.

Table of Contents

List of Tables

List of Figures

Preface

Foremost, I would like to express my sincere gratitude to my supervisor Professor Jean-François Richard for his unwavering support from the very beginning until the very last day of my PhD. studies. Particularly, I am grateful for his dedication, patience, and immense knowledge. He has always pushed me to be the best version of myself, not only as a researcher but also as a teacher, colleague, and friend. I consider Professor Richard to be my greatest mentor, who was there for me every step of the way, even when I was losing sight of the finish line, and to whom, in return, I would like to dedicate this dissertation.

Second, I am extremely grateful to be a student of Professor Alistair J. Wilson who has introduced me to new fields of study in economics and as such, has had a pivotal impact on my research agenda and this dissertation in particular. Professor Wilson is an excellent researcher and a wonderful colleague who I am more than fortunate to have met.

In addition to Professors Richard and Wilson, I would like to thank the rest of my thesis committee: Professors Arie Beresteanu and Roman Liesenfeld for their time and research support. In particular, I am grateful for their insightful comments and suggestions with respect to my job market paper, which constitutes the first chapter of this dissertation.

I would also like to extend my gratitude to the Federal Reserve Board of Governors Dissertation Fellowship program for research support. In particular, I would like to thank Jesse Bricker, Kevin Moore, and William Peterman for their valuable advice, practical suggestions, and unwavering guidance. Moreover, I want to thank John Czajka, Daniel Feenberg, and Allison Schertzer for their research support without which the completion of this dissertation would not have been possible.

Throughout the years as a PhD student I have interacted with various professors in the Department of Economics at the University of Pittsburgh, each time receiving valuable research advice and useful suggestions. In particular, I appreciate the discussions I have had with Professors Stefania Albanesi, Marla Ripoll, and Osea Giuntella.

Furthermore, I am indebted to everyone who made it possible for me to raise and train my service dog Watson during my PhD studies. Specifically, I am grateful to the Disability

Resources and Services at the University of Pittsburgh for providing me with assistance and guidance and to the Department's faculty, staff, and graduate students (including my office mates Prottoy Aman Akbar, Neeraja Gupta, and Yuriy Podvysotskiy) for their continuous support and understanding. Furthermore, I want to thank Humane Animal Rescue and Lilian Atkin for her invaluable help with Watson's training.

Though unforgettable and highly rewarding, I consider my time as a PhD student challenging. Therefore, my preface would not be complete without thanking Doctors Marnie Greenwald, Mary Koch Ruiz, and Frank Fetterolf for not giving up on me and constantly teaching me how to keep moving forward. I can never thank you enough for all your services.

Finally, I want to thank my parents Małgorzata and Andrzej Boczoń, my sister Magdalena Boczoń, and my grandparents for their unconditional support and endless love. In particular, I am eternally grateful and humbled by their selflessness at supporting my dreams and putting my career aspirations first.

1.0 Quantifying Uncertainties in Estimates of Income and Wealth Inequality

1.1 Introduction

As of 2018, more than 40 percent of income in the US was earned by the top 10 percent, and more than 30 percent by the top 5 percent. In relation to wealth, as of 2019, more than 75 percent was owned by the top 10 percent, and more than 65 percent by the top 5 percent. According to Saez (2017), the last time we observed comparably high levels of top income and wealth inequality was in the years leading to the 1929–1933 Great Depression.

Inequality endangers the economy in a number of ways: it threatens the integrity of economic systems and the impartiality of political institutions and, eventually, can lead to a rise of extremism or even oligarchies. So far, rising income and wealth inequality in the US has been linked to falling trust and fading civic participation (Uslaner and Brown, 2005), lower levels of agreeableness (de Vries et al., 2011), declining life expectancy (Clarkwest, 2008), and a rise in obesity (Pickett and Wilkinson, 2012), mental illness (Ribeiro et al., 2017), homicide (Daly, 2016), teenage pregnancy (Kearney and Levine, 2012), and drug overdose (Wilkinson and Pickett, 2007). For example, Chetty et al. (2016) find that the richest Americans in the top 1 percent of income distribution live almost 15 years longer than the poorest ones from the bottom 1 percent. According to the World Bank data, this represents a gap in life expectancy comparable to that between the US and Liberia, a country with GDP per capita of less than 1.2 percent of the US level. Furthermore, rising income and wealth concentration in the US endangers equal distribution of economic resources around the world. In particular, between 1970 and 1992 an increase in the number of "globally rich" in the US, defined as those with more than twenty times the mean world income, accounted for half of the worldwide increase, making "a perceptible difference to the world distribution" (Atkinson et al., 2011).

While numerous studies, including Piketty and Saez (2003), Atkinson et al. (2011), Bricker et al. (2016), Bricker et al. (2018), Smith et al. (2019), and Saez and Zucman (2016, 2020a,b) estimate income and wealth inequality, the novelty of my study lies in investigat-

ing the extent to which statistical uncertainty affects empirical and structural research on inequality. Regardless of whether we rely on self-reported surveys or administrative records, all data are subject to errors. Importantly, if unaccounted for such errors can result in statistically imprecise conclusions regarding the past and current levels of top income and wealth inequality, and therefore, lead to inaccurate predictions and potentially ineffective policy recommendations.

During the past two years, economists and politicians have discussed various measures aimed at combating inequality: instituting a wealth tax on millionaires, raising the top income tax rate, reducing exemptions and increasing tax rates on large estates. However, whether such policies would prove effective at closing the gap between rich and poor depends primarily on our ability to produce statistically accurate estimates of income and wealth inequality. Otherwise, the government might collect either too little in tax revenue, unable to provide the poor with adequate government-funded child care and paid-leave, or too much, causing a sudden and sharp decline in economic growth.

Moreover, since income and wealth inequality has been at the center of attention during the 2020 Democratic Party presidential primaries, it is imperative to provide the general public with an idea of the accuracy of these estimates. Otherwise, the public could easily be misled to either under- or overestimate the severity of the ongoing crisis linked to rising inequality. This, in turn, may result in voters misconstruing the effectiveness of current policies and lead them to express support for more conservative proposals.

In this paper, I estimate the uncertainties in estimates of the six income and wealth shares of the top 10, 5, 1, 0.5, 0.1, and 0.01 percent (the American upper middle and upper classes) and assess their impact on both empirical and structural macroeconomic studies. To do this, I first investigate which data set proves most reliable for studying income and wealth concentration, accounting for sampling, nonsampling, and modeling errors. Second, I determine what can and cannot be concluded about levels and trends in economic inequality once we account, at least partially, for data-driven errors in the constructed estimates. Finally, I examine to what extent uncertainties in calibration targets affect outcomes of structural macroeconomic modeling, in the context of a random growth model of income.

The present paper contributes to the existing literature in five main ways. First, it adds

to our understanding of the types of errors that are most prevalent in estimates of income and wealth concentration. Second, it provides a comparative analysis of studying economic inequality using survey data (with a wide range of both financial and nonfinancial variables but a small number of observations) versus administrative tax records (with a large number of observations but no information on taxpayers' wealth or socio-economic characteristics). Third, it introduces an approach to estimating sampling error using administrative tax data, which can be used to assess sampling accuracy of any estimator of any population parameter of interest and, in the present paper, is illustrated in the context of income and wealth concentration. Fourth, it is the first research project to estimate the long-term dynamics in economic inequality while accounting for uncertainties in the constructed estimates. Finally, the paper discusses the consequences of utilizing error-prone data for structural analysis, including data tracking and projection.

In this paper, I define income as gross income comprising all income items except for capital gains and wealth as all assets less all liabilities. For both the empirical and structural analysis, I use two data sets: the Survey of Consumer Finances (SCF)—a triennial survey of US household financial condition—and the Individual Tax Model Public Use File (PUF)—an annual sample of US individual income tax returns. The SCF survey data range from 1989 to 2019, with wealth measured in a current year (hence, running from 1989 to 2019) and income measured in a year before (hence, running from 1988 to 2018). The PUF tax data range from 1991 to 2012, with both income and wealth measured in a current year. Consequently, my analysis focuses on the over twenty-year long period that follows the Tax Reform Act of 1986, which lowered federal income tax rates and, in particular, reduced the top tax rate from 50 to 28 percent. Note that the 2019 SCF and the 2012 PUF are the latest available data sets at the time of writing.

In order to determine which data source is more reliable for studying top income and wealth inequality, I compare the SCF survey data and the PUF administrative data with respect to a number of criteria, such as the number of observations available for the estimation and the size of relative standard errors. For studying long-term dynamics in income and wealth inequality, I use weighted least squares with weights defined as reciprocals of squared standard errors. In addition to long-term trends, I focus on short-term dynamics and, in

particular, examine how income and wealth concentration changed between the onset and the aftermath of the 2007–2009 Great Recessions.

In addition to accounting for data-driven errors in empirical analysis of top-decile income and wealth shares, I investigate how data deficiencies affect outcomes of structural macroeconomic models. I consider the augmented random growth model of income proposed by Gabaix et al. (2016) and use Monte-Carlo simulation techniques to analyze how errors in inputs to this model impact the precision of the model's outcomes. Specifically, for each of the two data sets under consideration, I calibrate Gabaix et al. (2016)'s model multiple times, each time using a different value randomly drawn from a confidence interval constructed around the point estimate of the model's calibration target. Then, by averaging over the range of generated model outcomes, I determine the extent to which the model's outputs are affected by the uncertainty in the model's inputs.

My empirical analysis of top income inequality suggests that estimates constructed using the administrative data are considerably better than those constructed using the survey data. Some of the advantages of using the PUF are higher data frequency and much larger number of observations, which results in considerably smaller standard errors. Specifically, I find that sampling error in the estimated income shares constructed using the PUF is, on average, *six* times smaller than sampling error in the analogues estimates constructed using the SCF. While these data features are not critical when examining long-term dynamics of income inequality, they become a decisive factor when choosing between the SCF and the PUF in a study that analyzes short-time horizons and survey-to-survey/year-to-year changes. Moreover, the small number of observations above the 99.9 and 99.99 income fractiles in the SCF makes the SCF estimates of the two most granular income shares of the top 0.1 and 0.01 percent extremely volatile and, thus, uninformative.

In relation to wealth inequality, my empirical results indicate that neither the survey data nor the administrative data can be used without caution. For the less granular wealth shares of the top 10 to the top 0.5 percent, I find the SCF more reliable than the PUF. In the SCF, respondents are *asked* about their asset and liability holdings, whereas in the PUF, the wealth of every individual in the sample is *estimated* from their reported income using a capitalization model. Since such models are heavily dependent on numerous (and often

arbitrary) assumptions imposed on assets' rates of return, so are the resulting estimates of top wealth shares. Therefore, even though the SCF estimates have larger sampling errors than those constructed using the PUF, they are free from non-trivial and yet-to-be-fully-determined modeling errors arising in the process of inferring wealth from income. Lastly, regarding the wealth shares of the top 0.1 and 0.01 percent, I find that both the SCF and the PUF present difficult-to-overcome challenges (an insufficient number of observations and modeling errors, respectively) that result in highly unreliable estimates of the far right tail of wealth distribution of the top 0.1 and 0.01 percent.

In addition to identifying strengths and weaknesses of survey and administrative data in studying income and wealth concentration, this paper adds new insight to the ongoing debate regarding the magnitudes and trends in top income and wealth inequality. For income, my results confirm those from the related literature, indicating a statistically significant increase in income concentration between the early 1990s and the early 2010s, and suggest comparable levels and trends in the estimated income shares within the top 10 percent.

For wealth, since this paper finds the SCF a more reliable data source than the PUF and, moreover, accounts for data-driven errors in the constructed estimates, it portrays a different picture of top wealth inequality than the most widely-cited studies on wealth concentration estimated using either survey or administrative tax-level data. Specifically, while my study does suggest a statistically significant increase in the wealth shares of the top 10 and 5 percent between the early 1990s and the early 2010s, it implies a statistically insignificant increase in the wealth shares of the top 1 and 0.5 percent. Second, this paper neither supports not contradicts Saez and Zucman (2016)'s widely-cited conclusion regarding a 100 percent increase in the wealth shares of the top 0.1 percent between 1991 and 2012, a finding that has drawn substantial media coverage and major interest from politicians and policy makers. Lastly, the weighted linear regression analysis of the SCF point estimates does not suggest a larger increase in the wealth shares of the top 1 percent than in the wealth shares of the top 10 percent, casting doubts on a popular presumption that the observed rise in wealth inequality is driven particularly by the richest of the rich, leading to a surge in wealth disparity within the top 10 percent of wealth distribution.

My third set of results pertains to the consequences of modeling inequality using error-

prone data. In relation to the Gabaix et al. (2016)'s random growth model of income, I find that having precise estimates of calibration targets is critical for producing precise outcomes of structural analysis. This is the case since errors in calibration targets are carried over through the model and come to affect all outcomes of interest. Specifically, I find that the model calibrated to administrative tax data projects income shares of the top 1 percent in 2050 to be equal to 22.5 percent, with only negligible levels of uncertainty attributable to the data. On the other hand, the model calibrated to survey data is much less precise regarding the 2050 projection, with a 95 confidence interval ranging from 19 to 29 percent. Therefore, by relying upon administrative tax data for the model's calibration, as opposed to survey data, one can reduce the uncertainty in the model's outcome of interest by ten percentage points.

My paper is organized as follows. In Section 1.2, I discuss related literature. In Section 1.3, I provide a brief description of sources of error in survey and administrative data. In Section 1.4, I characterize the main features of the SCF and describe the estimation procedure of the SCF standard error. In Section 1.5, which follows the same format as Section 1.4, I first provide a brief description of the PUF and next, characterize the estimation of the PUF standard error. In Section 1.6, I define the concepts of income and wealth and discuss the estimation procedure of top-decile income and wealth shares. In Sections 1.7 and 1.8, I analyze the main empirical results centered around income and wealth inequality, respectively. In Section 1.9, I discuss the key outcomes of the structural analysis. Section 1.10 concludes. Online supplementary material with additional results and detailed discussions supporting my conclusions is available on https://martaboczon.com.

1.2 Literature Review

This paper contributes to four main strands of economic literature: economic inequality, survey statistics, SCF survey design, and PUF sample design.

First, since one of my objectives is to quantify uncertainties in estimates of income and wealth inequality within the top 10 percent, this paper contributes to the growing body of

literature on income and wealth inequality in the US. Specifically, it is closely-related to the work of Piketty and Saez (2003), where the authors rely upon tax returns statistics and micro-level data on individual-income tax returns in order to construct homogeneous series of top-decile income shares between 1913 and 1998. Another important reference the research is related to is Atkinson et al. (2011), which utilizes individual-income tax return statistics for numerous income brackets in order to provide a comprehensive overview and comparative analysis of historical and current trends in top income shares for multiple countries around the globe.

In addition, the current paper adds valuable insights to the literature on wealth inequality. In particular, it builds on Bricker et al. (2018), Smith et al. (2019), and Saez and Zucman (2016, 2020a,b), in which the authors rely upon micro-level data and aggregate statistics published in the Financial Accounts of the United States in order to estimate the distribution of wealth using a capitalization model. Moreover, it is indirectly related to Kopczuk and Saez (2004), who estimate wealth concentration using estate tax return data, in which individual estates are weighted by the inverse probability of death.

Since this paper analyzes various data-driven errors in surveys and other sample data, it constitutes a direct application of an important statistical concept related to the Total Survey Error (TSE) paradigm. Even though, TSE has been thoroughly discussed in the literature on survey statistics (see, e.g. Biemer and Lyberg, 2003; Groves et al., 2009), it remains largely ignored in economics. Therefore, this paper constitutes an example of how established and widely applied statistical concepts can benefit economic research.

In addition to adding to economic inequality and survey statistics literature, the present paper contributes to the literature on the SCF survey design (see, e.g. Kennickell, 1997, 1998, 2000, 2008). Specifically, except for research conducted by the Board of Governors of the Federal Reserve System (hereafter the Board), it is the first academic paper to consider data-driven errors in any macroeconomic estimates constructed using the SCF survey data.

Lastly, this paper contributes to the literature on the PUF sampling design (see Czajka et al., 2014; Bryant et al., 2014). Specifically, it proposes a bootstrapping technique that allows data users to estimate the PUF sampling error for any quantity of interest. As such, it provides an illustrative example of how the information regarding a complex sample selection

process can be incorporated into an economic analysis.

1.3 Total Survey Error

Since one of my primary objectives is to identify and quantify sources of data-driven errors in the estimates of income and wealth inequality, this paper is centered around the TSE paradigm—an umbrella term for a variety of error sources in survey data. Even though TSE pertains primarily to errors in surveys, it is also applicable to data consisting of administrative records. This is the case since administrative data are affected by the same sources of error as survey data. Specifically, as with any other sample, administrative data are subject to sampling error caused by drawing a sample rather than conducting a complete census. Moreover, the data are prone to nonsampling errors, which comprise all other sources of error arising in the process of designing, collecting, processing, and analyzing of sample data.

TSE consists of two main components: sampling error and nonsampling error, which can be further divided into specification, frame, nonresponse, measurement, and processing errors, all of which I briefly characterize in the remainder of the present section.[1] One of the advantages of decomposing TSE is that it allows me to differentiate between sources of error, and consequently, address them individually. Specifically, in the present paper, I estimate sampling and nonresponse errors in the SCF and sampling error in the PUF. As such, I do not account for all parts of TSE (which is beyond the scope of the present paper). Instead, I provide qualitative evidence on which components of TSE can be considered marginal for this particular analysis and which are likely to be non-negligible and therefore, to be accounted for in a follow-up research project.

Specification error occurs "when the concept implied by the survey question and the concept that should be measured in the survey differ" (Biemer and Lyberg, 2003). This often results from misunderstandings between the different parties involved in the survey process such as researchers, data analysts, survey sponsors, questionnaire designers, and

[1] In this paper, without loss of generality I rely upon a TSE decomposition from Biemer and Lyberg (2003). Alternative decomposition can be found, for example, in Groves et al. (2009).

others.

The other sources of nonsampling error are frame and processing errors. The former relates to the process of constructing, maintaining, and using the sampling frame for selecting the sample, whereas the latter occurs in data editing, coding, entry of survey responses, assignment of survey weights, tabulation and other data arrangements.[2]

Nonresponse encompasses unit nonresponse, item nonresponse, and incomplete response, and is considered "a fairly general source of error" (Biemer and Lyberg, 2003). A unit nonresponse occurs when a sampling unit does not participate in the survey; an item nonreponse when a participating unit leaves a blank answer to a specific survey question; and an incomplete response when the answer provided to a typically open-ended question is either incomplete or inadequate.

The fifth and final source of nonsampling error, measurement error, is considered "the most damaging source of error" (Biemer and Lyberg, 2003). It includes errors arising from respondents and interviewers, in addition to other factors such as the design of the questionnaire, mode of data collection, information system, and interview setting.

In summary, any estimator constructed using survey or sample data is subject to a variety of sampling and nonsampling errors. Therefore, an important question is that of the extent to which these errors can be reliably accounted for when estimating unknown population parameters (such as top income and wealth shares) using self-reported survey data and administrative records.

[2] A classic frame error occurred in a public opinion poll designed to predict the result of the 1936 presidential election between Alfred Landon and Franklin D. Roosevelt. Since the sample frame was heavily over-represented by individuals who identified as Democrats (phone owners, magazine subscribers, members of professional associations), the difference between the poll's prediction and the election's result was equal to 19 percentage points and as such, constitutes the largest error ever recorded in a major public opinion poll.

1.4 Uncertainties in Survey Data on Consumer Finances

The SCF is a triennial survey of household finances sponsored by the Board in cooperation with the Statistics of Income (SOI) Division of the Internal Revenue Service (IRS).[3] The objective of the survey is to characterize the financial situations of a set of households referred to as the Primary Economic Units (PEUs), where "the PEU consists of an economically dominant single individual or couple (married or living as partners) in a household and all other individuals in the household who are financially interdependent with that individual or couple."[4]

The SCF was initiated in 1983 and over the years has become one of the primary data sources in studying consumer finances. It provides exhaustive categorization and detailed information on a variety of household financial products.[5] While the most comprehensive data are collected on household portfolios, the survey also provides supplementary information on a wide range of demographic and socio-economic characteristics such as sex, age, race, ethnicity, family size, homeownership status, and employment history.

Since the survey oversamples the upper tail of wealth distribution, it is also one of the primary data sources used in studying economic inequality. However, as emphasized by the Board, "even under ideal operational conditions, the measurements of the survey are limited in a fundamental way by the fact that it is based on a sample of respondents rather than the entire population."[6]

In this paper, I use the SCF data between 1989 and 2019, where the 2019 SCF is the latest available data set at the time of writing. Since wealth is measured in current years and income is measured in previous years, wealth runs from 1989 to 2019, and income from 1988

[3]Until 1989 the survey data were collected by the Survey Research Center at the University of Michigan. Since 1992 the data collection process has been administered by the National Opinion Research Center at the University of Chicago.

[4]See the SCF codebook at https://www.federalreserve.gov/econres/files/codebk2016.txt (accessed on April 15, 2019).

[5]The SCF collects information on checking, brokerage, savings, and money market accounts; certificates of deposit; savings bonds and other types of bonds; mutual funds; publicly-traded stocks; annuities, trusts, and managed investment accounts; IRAs and Keogh accounts; life insurance policies; and other types of financial and non-financial assets; credit card debt; vehicle loans and other types of consumer loans; mortgages, lines of credit, and other loans.

[6]See the SCF codebook at https://www.federalreserve.gov/econres/files/codebk2016.txt (accessed on April 15, 2019).

to 2018. The data sets from 1983 and 1986 are not included as they do not provide enough information to reliably estimate the main sources of variation in the SCF point estimates. Lastly, note that my analysis is based on a public version of the SCF. While the non-public internal version of the survey would presumably have better top coverage (and less error in TSE framework), its access is highly restricted and granted only to economists from the statistical unit of the Board.

1.4.1 Estimation of the TSE

With respect to Section 1.3, SCF data users with access to publicly available data files and supplementary materials can estimate two types of error, sampling and nonresponse. Quantifying other types of error such as specification, processing, frame, and measurement would constitute a nontrivial task that, in most instances, would require access to undisclosed information regarding specifics of the data editing process or construction of the sample frame, and as such is beyond the scope of the present paper.

1.4.1.1 Sampling Error

In order to protect respondents' confidentiality, specifics regarding the SCF sampling design are not disclosed to the general public. This disclosure avoidance procedure has the objective of minimizing the risk of a third party revealing the identity of a survey respondent based on the sampling-specific information such as selection probability, sampling strata, and primary and secondary sampling units.

An important implication of this disclosure avoidance procedure is that sampling error cannot be estimated using standard statistical techniques or built-in functions in software such as STATA, SAS, SUDAN, or AM Statistical Software. Instead, I estimate the SCF sampling error using bootstrapped sample replicates, generated by the Board for all survey years between 1989 and 2019. The replicates are constructed based on the actual SCF sampling design (see Section A.1 in Appendix A of the Online Supplementary Material) and are provided to the general public in the form of materials supplementary to the main data set. The main reason for providing these replicates is to facilitate the estimation of sampling

error by all SCF data users, who lack access to the undisclosed and highly confidential information regarding the specifics of the SCF sampling design.

Let θ denote an unknown population parameter, and let $\hat{\theta}$ denote the estimate of θ computed in the main data set. Moreover, let $\hat{\theta}_l$ denote the estimate of θ obtained in the lth bootstrapped sample replicate (as opposed to the main data set), where $l : 1 \to L$, and $L = 999$.[7] It follows that a sampling error of $\hat{\theta}$ is given by a sample standard deviation of $\left\{\hat{\theta}_l\right\}_{l=1}^{L}$,

$$\hat{\sigma}_{1,\hat{\theta}} = \sqrt{\frac{1}{L-1}\sum_{l=1}^{L}\left(\hat{\theta}_l - \frac{1}{L}\sum_{l'=1}^{L}\hat{\theta}_{l'}\right)^2}. \tag{1}$$

For illustration, consider a problem of estimating the sampling error of the estimate of the top 10 percent income share. The estimation consists of two steps. In the first step, I estimate θ in each bootstrapped sample replicate, which yields a total of L replicate-dependent estimates $\hat{\theta}_l$. In the second step, I estimate the sampling error of $\hat{\theta}$ by computing the sample standard deviation of $\left\{\hat{\theta}_l\right\}_{l=1}^{L}$.

1.4.1.2 Nonresponse Error

In addition to sampling error, SCF data users can estimate two types of nonresponse: item nonresponse and incomplete response.[8] Across all survey years between 1989 and 2019, the SCF contains $M = 5$ multiple imputations for virtually all variables (dichotomous and continuous) initially coded as either partially or completely missing.[9,10,11] This data feature

[7] The reason for choosing 999 bootstrapped sample replicates instead of 1,000 is that, as indicated in Hall (1986), with 1,000 repetitions (unlike 999), coverage probability of 90 percent bootstrap confidence interval is biased by approximately 0.001.

[8] Unit nonresponse is compensated for with nonresponse-adjusted sampling weights computed by the Board based on undisclosed selection probabilities, specifics regarding the sample frame, and population totals estimated from the Current Population Survey (for more details see Kennickell, 1997).

[9] Note that since the M imputations are stored as five successive records for each survey respondent, the number of observations in each data set is mechanically inflated by a factor of five.

[10] For a general discussion on multiple imputation for nonresponse in surveys see Rubin (1987). For more information on multiple imputation in the SCF see Kennickell (1998).

[11] See Section A.2 in Appendix A of the Online Supplementary Material for more details on partially missing values in the SCF.

allows users to account for the uncertainty associated with completely and partially missing data by estimating the variability between the multiply imputed data sets as

$$\hat{\sigma}_{2,\hat{\theta}} = \sqrt{\frac{1}{M-1}\sum_{m=1}^{M}\left(\hat{\theta}_m - \hat{\theta}\right)^2}, \tag{2}$$

where $\hat{\theta}_m$, $m : 1 \to M$ denotes the estimate of θ in the mth imputation.

1.4.1.3 Standard Error

After estimating sampling and imputation errors, I estimate the standard error of $\hat{\theta}$ using Rubin's estimator, given by:

$$\hat{\sigma}_{\hat{\theta}} = \sqrt{\hat{\sigma}^2_{1,\hat{\theta}} + \hat{\sigma}^2_{2,\hat{\theta}}\,(1 + M^{-1})}. \tag{3}$$

More details regarding Rubin's variance estimator can be found in Section A.3 in Appendix A of the Online Supplementary Material.

Since the above estimator of the standard error of $\hat{\theta}$ (see equation 3) accounts for only two types of error, sampling and nonresponse, it can be thought of as a lower bound on the unknown standard error of $\hat{\theta}$, say $\sigma_{\hat{\theta}}$. Importantly, the degree to which $\hat{\sigma}_{\hat{\theta}}$ underestimates $\sigma_{\hat{\theta}}$ depends on the magnitudes of the four nonsampling errors that my estimation procedure does not account for. Given the high level of expertise of the Board in designing and supervising the survey, I assume three of these errors—specification, processing, and frame—to be fairly marginal. The only other type of error that may cause $\hat{\sigma}_{\hat{\theta}}$ to severely underestimate $\sigma_{\hat{\theta}}$ is measurement error, and more specifically, respondent-related measurement error, which I thoroughly discuss in Section A.4 in Appendix A of the Online Supplementary Material.

1.5 Uncertainties in Administrative Data on Tax Returns

Starting from the early 1960s, the SOI Division of the IRS began to draw an annual sample of the Individual and Sole Proprietorship (INSOLE) tax returns. The INSOLE sample contains detailed information on taxpayers' incomes, deductions, exemptions, taxes, and credits, and thereby constitutes a micro-level database for tax policy purposes. In its present form, the INSOLE sample contains highly sensitive information that, if made publicly available, could risk the exposure of taxpayers' identities. Therefore, in order to ensure full confidentiality of the entire sample, access to the INSOLE is highly restricted and only granted to a handful of agencies, such as the Treasury Department or Congress.

Since access to the INSOLE is strictly limited, the SOI Division annually creates another sample of tax returns commonly referred to as the PUF. The PUF is annually sub-sampled from the INSOLE and subjected to a number of disclosure avoidance procedures such as blurring, rounding, deleting, and modifying. These techniques have the objective of ensuring that no taxpayer can be identified from the PUF upon its release to the general public.[12] Hence, the PUF is accessible to much broader audiences, including academic and non-academic researchers, and constitutes one of the main data sets used in studies of inequality.

In the present paper, I rely upon the PUF data between 1991 and 2012, where the 2012 PUF is the latest available data set at the time of writing.[13] As such, I analyze a twenty-two-year period that follows the last formal redesign of the INSOLE from the late 1980s and includes four years of the PUF after its latest revision, which took place in 2009.[14]

1.5.1 Estimation of the TSE

Contrary to popular belief, the problem of data deficiencies pertains not only to self-reported survey data (such as the SCF) but also to data that comprise administrative records (such as the PUF). In fact, administrative data (which until very recently had been con-

[12] See Sections B.1 and B.2 in Appendix B of the Online Supplementary Material for a detailed description of the PUF and INSOLE sampling design.

[13] Due to COVID-19, the release of the 2013 PUF is not known at the moment.

[14] "The revised design modifies which returns in the INSOLE sample are excluded from the PUF, changes the way the INSOLE sample is subsampled for the PUF, and aggregates all returns with a 'large' value for any specified amount variable into a single record" (Bryant et al., 2014).

sidered free of any source of error) are subject to the same types of error as survey data (whose accuracy is known to be negatively affected by sampling and nonsampling errors). Therefore, as Groen (2012) indicates, "analyses of the quality of administrative data and reasons for differences between administrative data and survey data are greatly needed."[15]

In the present paper, I focus on the estimation of the PUF sampling error. As such, my analysis does not account for processing, nonresponse, measurement, specification, and frame errors. Whereas there is reason to assume that the first four types of error are marginal (see Section B.4 in Appendix B of the Online Supplementary Material for qualitative evidence supporting this claim), frame error may not be inconsequential.

In order to illustrate why the PUF frame error may matter, consider work by Piketty and Saez (2003), where the authors impute income of non-filers as a fixed fraction of filers' average income for all years between 1946 and 1998. Even though this particular imputation procedure has the objective of matching the ratio (of 75–80 percent) of gross income reported on tax returns and total personal income estimated in national accounts, it remains highly arbitrary. As such, this and other imputation and/or estimation procedures aimed at "filling the frame" with information on non-filers introduce additional and non-negligible sources of error into the analysis.

1.5.1.1 Sampling Error

Since the IRS does not provide data users with bootstrapped sample replicates, in order to estimate the PUF sampling error I first generate $L = 999$ bootstrapped sample replicates based on the publicly available information on taxpayers' strata and stratum-specific probability of selection. For brevity, let $\mathcal{S}$ denote the PUF sample of taxpayers, and assume that $\mathcal{S}$ comprises J mutually exclusive and collectively exhaustive strata such that

$$\mathcal{S} = \cup_{j=1}^{J} \mathcal{S}_j, \tag{4}$$

where $\mathcal{S}_j \cap \mathcal{S}_{j'} = \emptyset$ for all $j \neq j'$.

[15] See Section 1.3 for a general discussion on sampling and nonsampling errors and Section B.3 in Appendix A of the Online Supplementary Material for a detailed description of different sources of error in the PUF.

Moreover, let n denote the total sample size, and let n_j be the number of taxpayers selected for the sample from stratum j. Since the strata are mutually exclusive and collectively exhaustive, it follows from equation (4) that

$$n = \sum_{j=1}^{J} n_j. \tag{5}$$

Since across all tax years under consideration there exist strata with as low as 10 observations or fewer (see Table C.1 in Appendix C of the Online Supplementary Material), bootstrapping methods cannot be applied directly to $\{\mathcal{S}_j\}_{j=1}^{J}$. Instead, I first classify the J strata into $J^\star \ll J$ clusters using the Partitioning Around Medoids (PAM) clustering procedure (see Reynolds et al., 1992), where I determine the number of clusters in each tax year based on a silhouette analysis. The clustering procedure uses as an input three stratification variables originally designated for the INSOLE sample: gross income, presence or absence of special forms and schedules, and the return's potential usefulness for tax policy modeling. Since the income variable is ordinal (successive income brackets) whereas the latter two are nominal, I use the Gower distance measure, which is applicable to a mix of ordinal and nominal variables.

The clustering procedure results in a PUF sample of taxpayers $\mathcal{S}$ that comprises J^* mutually exclusive and collectively exhaustive clusters such that

$$\mathcal{S} = \cup_{j=1}^{J^\star} \mathcal{S}_j^\star, \tag{6}$$

where $\mathcal{S}_j^\star \cap \mathcal{S}_{j'}^\star = \emptyset$ for all $j \neq j'$.

Moreover, with $n_j^\star$ denoting the number of taxpayers in cluster j, it follows from equation (6) that

$$n = \sum_{j=1}^{J^\star} n_j^\star. \tag{7}$$

For example, in tax year 2008, I classify the 95 strata (with the minimum number of observations per stratum equal to 11) into 23 clusters (with the minimum number of observations per stratum equal to 184). Summary statistics for clustered strata in the remaining

tax years (1991 through 2012) can be found in Table C.1 in Appendix C of the Online Supplementary Material.

After classifying taxpayers into $J^\star$ clusters, I draw $L = 999$ independent bootstrapped sample replicates. Specifically, for each sample replicate $l : 1 \to L$, I draw with replacement $n_j^\star$ sample observations from each cluster j, such that the total number of observations in each sample replicate is equal to n.

Finally, in order to estimate the PUF sampling error, I follow the estimation procedure of the SCF sampling error outlined in Section 1.4.1.1. Let $\hat{\theta}$ denote the estimate of θ computed in the main data set, and let $\hat{\theta}_l$ denote the estimate of θ obtained in the lth bootstrapped sample replicate (as opposed to the main data set). I estimate the sampling error of $\hat{\theta}$ by a sample standard deviation of $\left\{\hat{\theta}_l\right\}_{l=1}^{L}$ as

$$\hat{\sigma}_{1,\hat{\theta}} = \sqrt{\frac{1}{L-1}\sum_{l=1}^{L}\left(\hat{\theta}_l - \frac{1}{L}\sum_{l'=1}^{L}\hat{\theta}_{l'}\right)^2}. \tag{8}$$

1.6 Income Inequality Measured in Survey and Administrative Data

In the following, I first define the concept of income and wealth used throughout the paper and next, describe the procedure for estimating top income and wealth shares in the SCF and the PUF. Note that for the PUF, the derivations that follow apply to every tax year between 1991 and 2012, and for the SCF, to all survey years between 1989 to 2019. The time subscript t is omitted for ease of notation.

1.6.1 Income

In this paper, I define income as gross income comprising all income items, except for capital gains, prior to deductions. The reason for excluding capital gains is that "realized capital gains are not an annual flow of income (in general, capital gains are realized by individuals in a lumpy way only once in a while) and form a very volatile component of

income with large aggregate variations from year to year depending on stock price variations" (Piketty and Saez, 2003).

The aforementioned income measure (as well as numerous alternative measures that could be applied without loss of generality, such as gross income *including* capital gains) can be constructed using each of the two data sets under consideration, without the need to rely upon supplementary data and/or econometric modeling. Specifically, the SCF collects information on households' income during the SCF interview process (*In total, what was your annual income from dividends, before deductions for taxes and anything else?*), whereas the PUF compiles income data from a sample of filed tax returns (Form 1040).[16,17] The resulting operational definitions of income are virtually identical for the two data sets, differing only with respect to two income components: the SCF reports both taxable and nontaxable IRA distributions and all other sources of income, the PUF reports only taxable amounts from line 15a on Form 1040 and does not provide information on all other sources of income from line 21.[18]

1.6.2 Wealth

Throughout my analysis, I define wealth as total assets less total debt. Since the SCF survey participants are asked detailed questions about their asset and liability holdings, measuring wealth in the SCF is straightforward and boils down to a simple accounting exercise.[19] In contrast, the PUF—a sample of individual income tax returns—provides limited informa-

[16]For the SCF, I compute gross income less capital gains as a sum of salaries and wages before deductions for taxes (variable X5702); income from a sole proprietorship or a farm (X5704); income from other businesses or investments; net rent, trusts, or royalties (X5714); income from non-taxable investments (X5706); income from other interest (X5708); income from dividends (X5710); income from Social Security or other pensions; annuities, or other disability or retirement programs (X5722); income from unemployment or worker's compensation (X5716); income from child support or alimony (X5718); and income from other sources (X5724).

[17]For the PUF, I compute gross income less capital gains as a sum of wages and salaries (line 7 on Form 1040); taxable interest (line 8a); tax-exempt interest (line 8b); ordinary dividends (line 9a); taxable refunds, credits, and offsets of state and local income taxes (line 10); alimony received (line 11); business and farm income (lines 12 and 18); IRA distributions, pensions and annuities, unemployment compensation, and social security benefits (lines 15b, 16a, 19, and 20a); and rental real estate, royalties, partnerships, S corporations, and trusts (line 17).

[18]Note that for comparability with the PUF, I do not include welfare assistance in SCF measure of wealth.

[19]See Figure C.1 in Appendix C of the Online Supplementary Material for a detailed description of the construction of the SCF wealth measure.

tion on taxpayers' wealth, which results from the fact that many asset and liability holdings are not reported on a tax form. For example, since for many taxpayers standard deductions are more effective at reducing financial burden than are itemized deductions, only a small fraction of homeowners deduct mortgage interests on their tax returns.

In order to construct comprehensive wealth measures using the PUF (as well as the INSOLE or any other tax-level data) it is necessary to rely upon auxiliary data sources and numerous modeling assumptions. In this paper, I utilize a capitalization model from Saez and Zucman (2016) and later re-visit in Bricker et al. (2018), where taxpayers' wealth is estimated by "capitalizing" asset income with an asset-specific rate of return. Specifically, as summarized in Saez and Zucman (2016), for each asset class I estimate a capitalization factor that maps the total flow of tax income to the amount of wealth from the household balance sheet of the Financial Accounts of the United States. Then, I estimate wealth of each tax payer by multiplying their reported incomes by the corresponding capitalization factors.

Let the wealth of taxpayer i be defined as

$$\widehat{wealth}_i = \widehat{nonfin}_i + \sum_{a=1}^{A} \frac{income_{i,a}}{\hat{r}_a}, \tag{9}$$

where $income_{i,a}$ denotes the income of taxpayer i generated by asset $a = 1, \cdots, A$, $\hat{r}_a$ denotes the estimated rate of return on asset a, and $\widehat{nonfin}_i$ is the estimate of taxpayer $i^{'}$s nonfinancial wealth.

The rate of return on asset a is estimated by computing a ratio of the household stock of asset a reported in the Financial Accounts, say FA_a, to the a's realized capital income measured in the PUF,

$$\hat{r}_a = \frac{\sum_{i=1}^{n} income_{i,a}}{FA_a}. \tag{10}$$

Following Saez and Zucman (2016), I organize assets from the Financial Accounts into seven categories: (1) taxable interest-bearing assets, (2) non-taxable interest-bearing assets, (3) dividend-generating assets, (4) assets generating profits of S corporations, (5) assets generating royalty income and profits of partnerships and C corporations, (6) tenant-occupied

real estate assets less mortgages, and (7) privately held and employer-sponsored pension assets.[20,21]

For illustration, consider the following problem of determining assets of taxpayer $i^\star$ with reported income from dividends of \$6,710. First, I estimate rate of return on dividend-generating assets, say $\hat{r}_{div}$, using equation (10), with the denominator FA_{div} computed as a sum of directly held equities (FL153064105), equities indirectly held through mutual funds (FL653064155), and the share of equities in money market funds (estimated based on FL153034005, FL634090005, and FL633062000), less equities held by nonprofit organizations and mutual funds held in IRAs. Then, I estimate dividend-generating assets of taxpayer $i^\star$ by dividing her/his dividend income of \$6,710 by the estimated rate of return. In 2012, the estimated rate of return was equal to 0.03411, implying a total of \$196,717 in dividend-generating assets for taxpayer $i^\star$.[22]

In this paper, I consider three sets of PUF estimates, one generated under a homogeneity assumption imposed on all rates of return under consideration, as in equation (10), and two sets of estimates constructed under a heterogeneity assumption, where I assume homogeneous rates of return on all income-generating assets except for those that generate taxable interests. As emphasized by Bricker et al. (2018) "implied rate of return on taxable interest-bearing assets in Saez and Zucman (2016) is much lower than market rates from the 10-year Treasury yield or Moody's Aaa corporate bond—the type of taxable interest-bearing assets that are held by wealthy families (Bricker et al., 2016; Kopczuk, 2015)." Therefore, following Bricker et al. (2018), I consider a scenario, where I assign a higher rate of return, say $\hat{r}_{a,\mathcal{A}}$, to the top 1 percent of the wealth distribution, and a lower rate, say $\hat{r}_{a,\mathcal{B}}$, to the bottom 99 percent, such that

$$FA_a = \frac{\sum_{i\in\mathcal{A}} income_{i,a}}{\hat{r}_{a,\mathcal{A}}} + \frac{\sum_{i\in\mathcal{B}} income_{i,a}}{\hat{r}_{a,\mathcal{B}}}, \tag{11}$$

where $\mathcal{A}$ denotes a set comprised of the top 1 percent of the wealth distribution and $\mathcal{B}$ a set comprised of the bottom 99 percent.

[20] In order to construct seven aggregate asset categories, it is often necessary to combine numerous lines from multiple tables reported in the Financial Accounts.

[21] Details regarding the construction of the remaining asset classes can be found on my personal website as well as in the Online Appendix to Saez and Zucman (2016).

[22] Since the Financial Accounts are revised on a quarterly basis, I use tables from the first quarter of 2020, which contain the latest data available at the time of writing.

Note that the resulting operational definitions of wealth in the SCF and the PUF differ with respect to three asset categories: defined benefit pension plans and term life insurance policies, which are included in the PUF but not in the SCF, and durable goods (e.g., vehicles), which are included in the SCF but not in the PUF.[23]

1.6.3 Estimation

In this section, I describe the estimation procedure for top income and wealth shares using the two data sets under consideration, the SCF survey data and the PUF sample of administrative tax records.

Let g_i denote either income or wealth of observation $i : 1 \to n$ (in either the SCF or the PUF), and let w_i denote sampling weight of i such that

$$\sum_{i=1}^{n} w_i = N, \tag{12}$$

where N denotes the number of units in the underlying population of interest. For the SCF, N is equal to the total number of PEUs. For the PUF, N is equal to the total number of taxpayers.

Let $g_{(j)}$ be the j^{th}-order statistic of g_i, with w_j denoting the sampling weight associated with $g_{(j)}$.[24] Moreover, let m_k denote the number of observations in the bottom $100k$ percent defined as

$$m_k = kN, \tag{13}$$

where $k \in \mathcal{K}$ and $\mathcal{K} = \{0.9, 0.95, 0.99, 0.995, 0.999, 0.9999\}$.

For example, consider tax year 2008, with the number of taxpayers equal to N = 142,580,866. It follows from equation (13) that $m_{0.9} = 128,322,779$ and $m_{0.9999} = 142,566,608$.

[23] Sabelhaus and Henriques Volz (2019) estimate defined benefit plans for the SCF survey participants using an actuarial present-value model. The model utilizes information on benefit amounts for those SCF respondents' currently collecting a pension, the expected timing and amount of future pension benefits from a past job for those who are entitled to a benefit but are not yet collecting it, and the current wage, age, sector (private or public), and the number of years in the plan of those who have a plan tied to their current job but are not yet receiving benefits.

[24] Order statistics of g_i are ranked in ascending order of magnitude (see, e.g., David and Nagaraja, 2003) with $g_{(1)} = \min_j \{g_i\}$ and $g_{(n)} = \max_j \{g_i\}$.

Next, let r_k denote the unknown (income or wealth) share of the *bottom* $100k$ percent, and let p_k be the unknown share of the *top* $100\,(1-k)$ percent defined as

$$p_k = 1 - r_k. \tag{14}$$

The estimation procedure consists of two main steps. In the first step, I determine an index value $j_k^\star \in \{1, \cdots, n-1\}$ such that

$$\sum_{j=1}^{j_k^\star} w_j \leq m_k \leq \sum_{j=1}^{j_k^\star+1} w_j, \tag{15}$$

which allows me to estimate the lower bound on r_k as

$$\underline{r}_k = \frac{\sum_{j=1}^{j_k^\star} w_j g_{(j)}}{\sum_{j=1}^{n} w_j g_{(j)}}, \tag{16}$$

and the upper bound on r_k as

$$\overline{r}_k = \frac{\sum_{j=1}^{j_k^\star+1} w_j g_{(j)}}{\sum_{j=1}^{n} w_j g_{(j)}}. \tag{17}$$

In the second step, I estimate r_k using linear interpolation between $\underline{r}_k$ and $\overline{r}_k$:

$$\hat{r}_k = \underline{r}_k + \omega_k \left(\overline{r}_k - \underline{r}_k\right), \tag{18}$$

where

$$\omega_k = \frac{m_k - \sum_{j=1}^{j_k^\star} w_j}{w_{j_k^\star+1}}. \tag{19}$$

It follows from equation (14) that the estimator of p_k is given by

$$\hat{p}_k = 1 - \hat{r}_k. \tag{20}$$

In relation to Section 1.4.1.2, it is important to note that since the SCF data are multiply imputed for missing values, the aforementioned estimation procedure needs to be repeated separately for each of the M imputed SCF data sets. This results in M estimates of p_k, which I denote as $\hat{p}_{k,m}$, $m : 1 \to M$. The grand estimate of p_k is then obtained by averaging over $\{\hat{p}_{k,m}\}_{m=1}^{M}$ such that:

$$\hat{p}_k = \frac{1}{M} \sum_{m=1}^{M} \hat{p}_{k,m}. \tag{21}$$

1.7 Empirical Results on Income Shares

In the following, I discuss my empirical results regarding the estimation of the income shares within the top 10 percent. In Section 1.7.1, I compare the number of observations above top income fractiles estimated using the the SCF and the PUF. In Sections 1.7.2–1.7.7, I focus on point estimates, standard errors, and the long- and short-term dynamics in income inequality. In Section 1.7.8, I establish a statistical link between the SCF and the PUF point estimates for top income shares. Finally, in Section 1.7.9, I summarize my key findings and conclude. I use triennial data on income between 1988 and 2018 from the 1989–2019 SCF and annual data on wealth between 1991 and 2012 from the 1991–2012 PUF.

1.7.1 Number of Observations

My first set of results pertains to the number of observations in the SCF and the PUF.[25] Across all of the years under study, I find large differences in the number of observations between the SCF and the PUF. For example, as indicated in Table 1.1, the number of observations available for the estimation of 2012 income in the SCF accounts for just 3.5 percent of the total number of observations available in the PUF. Moreover, since the number of observations increases proportionately across the two data sets, the ratio of the number of the SCF sample observations to the number of the PUF sample observations is fairly stable across years with a minimum of 3 and a maximum of 4.5 percent.[26]

In addition, I observe large differences not only in the total number of observations but also in the number of observations in the far right tail of income distribution. For example, the number of observations above the 90 income fractile in the 2013 SCF accounts for only 2.1 percent of the number of observations above the 90 income fractile in the 2012 PUF. More generally, between 1991 and 2012, this ratio varied from a minimum of 1.6 percent to a maximum of 3 percent.

The number of observations in the SCF is small not only in relative terms when compared

[25] Note that weight totals in the SCF and the PUF are not comparable because the two data sets rely upon different units of observation, and the number of tax units is always greater than the number of households.

[26] See Table C.2 in Appendix C of the Online Supplementary Material for the number of observations in the SCF and the PUF.

Table 1.1: Ratio of the SCF number of observations to the PUF number of observations available for the estimation of income

Year	Total	Above the k income fractile					
		90	**95**	**99**	**99.5**	**99.9**	**99.99**
1991	3.4	2.0	1.8	1.6	1.4	1.3	1.2
1994	4.5	2.8	2.5	2.0	1.8	1.4	1.3
1997	3.9	2.3	2.0	1.4	1.3	1.0	1.6
2000	3.0	1.6	1.4	1.0	0.9	0.8	1.5
2003	3.4	2.0	1.8	1.5	1.5	1.4	2.3
2006	3.0	1.9	1.7	1.6	1.6	1.9	7.2
2009	4.2	2.7	2.4	1.7	1.6	1.8	7.2
2012	3.5	2.1	1.7	1.3	1.3	2.1	5.8

Ratios are expressed in percent

to the abundance of data in the PUF but also in absolute terms. In most of the years under study, I observe the number of observations in the SCF above the 99.9 and 99.99 income fractiles to be less than three hundred and one hundred, respectively, and hence, to be insufficient for a reliable estimation of the income shares of the top 0.1 and 0.01 percent.[27] This result shows that the problem of a small number of observations in the SCF escalates when the focus of the analysis shifts from the mean or median to top-decile fractiles.

1.7.2 Point Estimates

Having compared the number of observations in the SCF and the PUF, in the following, I focus on point estimates. In particular, I compute ratios of the PUF to the SCF point estimates for the six income shares under consideration percent between 1991 and 2012.

[27] See Table C.3 in Appendix C of the Online Supplementary Material for the number of observations in the far right tail of the income distribution.

My analysis suggests that the SCF and the PUF point estimates concur with regard to less granular income shares (such as the top 10, 5, 1, and 0.5 percent), with the SCF point estimates only marginally above those obtained using the PUF. However, with respect to the more granular income shares of the top 0.1 and 0.01 percent, the two data sets greatly disagree. Specifically, ratios of point estimates vary from a minimum of 70 to a maximum of 210 percent, with the SCF point estimates being systematically below those obtained using the PUF. Whether this is driven by the small number of observations in the SCF (above the 99.9 and 99.99 income fractiles) or is related to a potential under-reporting of incomes by the SCF respondents from the top 0.1 and 0.01 percent of income distribution is beyond the scope of the current study. I will address this issue in a follow-up research project centered on measurement error arising from under- and over-reporting in financial surveys.

1.7.3 Relative Magnitudes of Sampling Error Across Data Sets

In the present section, I analyze the ratios of the Coefficients of Variation (CVs) in the PUF to those in the SCF for the six income shares within the top 10 percent between 1991 and 2012.[28] My results, illustrated in Figure 1.2, indicate that the SCF sampling errors are much larger than the PUF sampling errors for all years and income shares under consideration. In particular, the ratio of the PUF CV to the SCF CV is, on average, equal to 17 percent. In other words, my analysis suggests that the SCF CV is, on average, *six* times that of the PUF. Wheres I suspect that the observed unprecedented one-time increase in the PUF sampling error in 2009 may have resulted from the change to the PUF sub-sampling design that occurred in 2009 or be related to the impact of the 2007–2009 Great Recession on households' financial situations, more research is needed to provide a definite answer.

Another interesting question is about the amount of the reduction in the PUF sampling error that is driven purely by the larger sample size as opposed to a better sampling design. Since statistics typically follow a root–N asymptotic distribution, I conduct a simple "back-of-the-envelope" exercise, where I compute a square root of the ratio between the SCF and PUF sample sizes. By averaging across all years under consideration, I find the average

[28] CV is defined as sampling error standardized for point estimate.

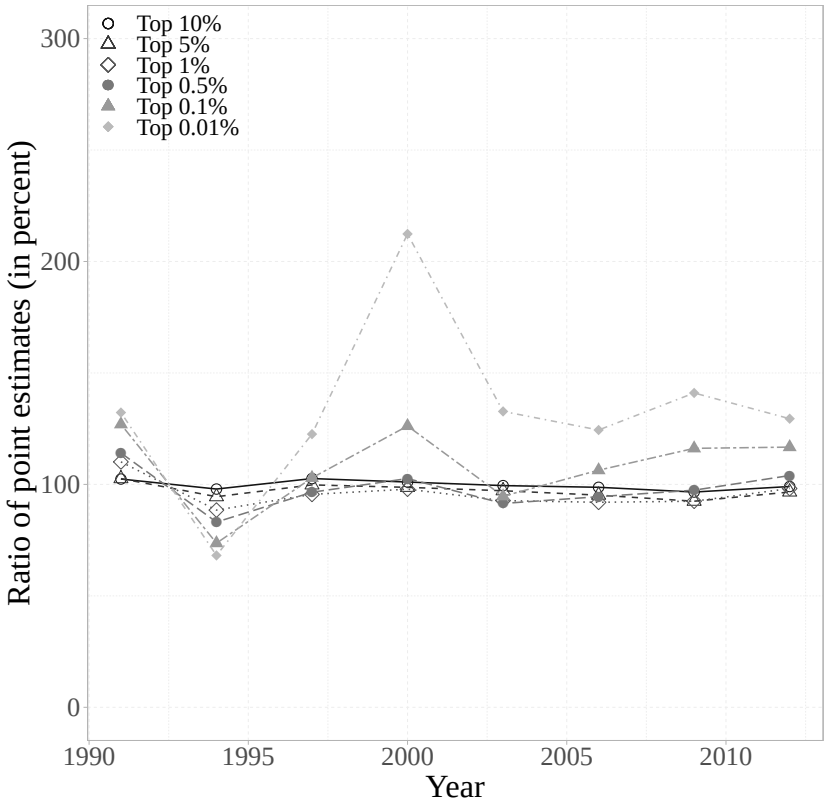

Figure 1.1: Ratio of the PUF point estimate to the SCF point estimate for the income shares within the top 10 percent

square root of the ratio in question to be equal to 19 percent. Since this value is reasonably close to the average ratio of the PUF CV to the SCF CV found above (17 percent), I conclude that it is sensible to assume that the SCF and the PUF sampling errors would be largely comparable if the sample sizes happened to be the same.

1.7.4 Relative Magnitudes of Sampling Error Across Income Shares

In the following, to complement Section 1.7.3, I compare the magnitudes of sampling errors *within* a data set but *across* estimated income shares. Specifically, for each of the two data sets under consideration, I compare the CVs for the income shares of the top 5, 1, 0.5, 0.1, and 0.01 percent to the CVs for the income share of the top 10 percent. As illustrated in

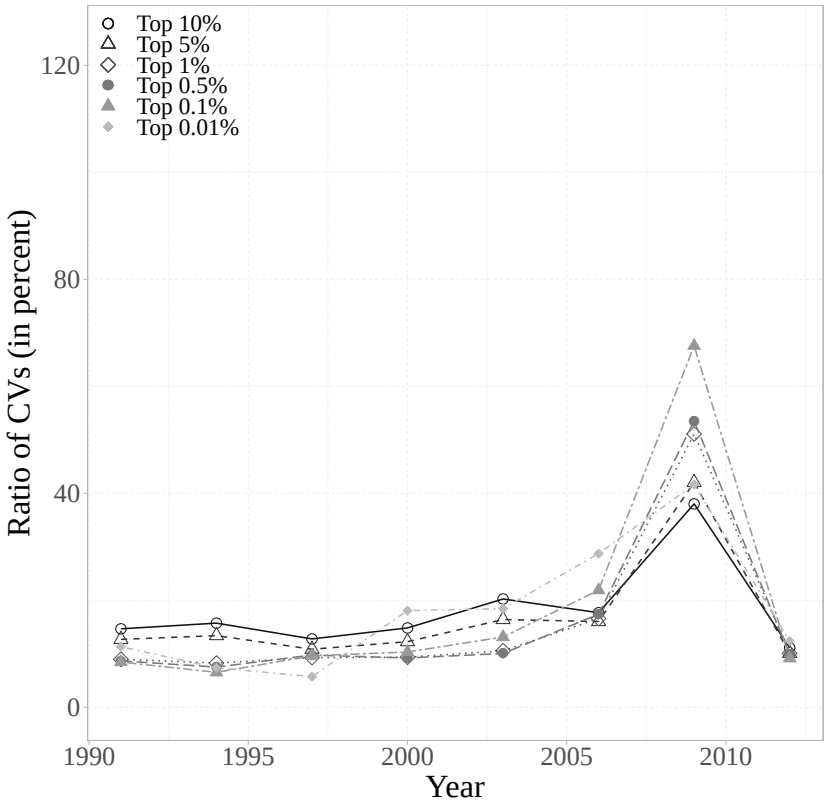

Figure 1.2: Ratio of the PUF coefficient of variation to the SCF coefficient of variation for the income shares within the top 10 percent

Figure 1.3, an increase in CVs as the estimated income shares become more granular is not only inevitable but also substantial. However, this increase is much less pronounced in the PUF than it is in the SCF. Therefore, the PUF estimates for more granular income shares when compared to those for less granular income shares are estimated with substantially smaller error than those estimated using the SCF.

1.7.5 Standard Error Decomposition

The present section breaks down SCF standard error for clearer analysis of leading sources of variation in the SCF estimates of top income inequality. In Table 1.2, I present ratios of sampling error to total standard error (that comprises both sampling and imputation errors)

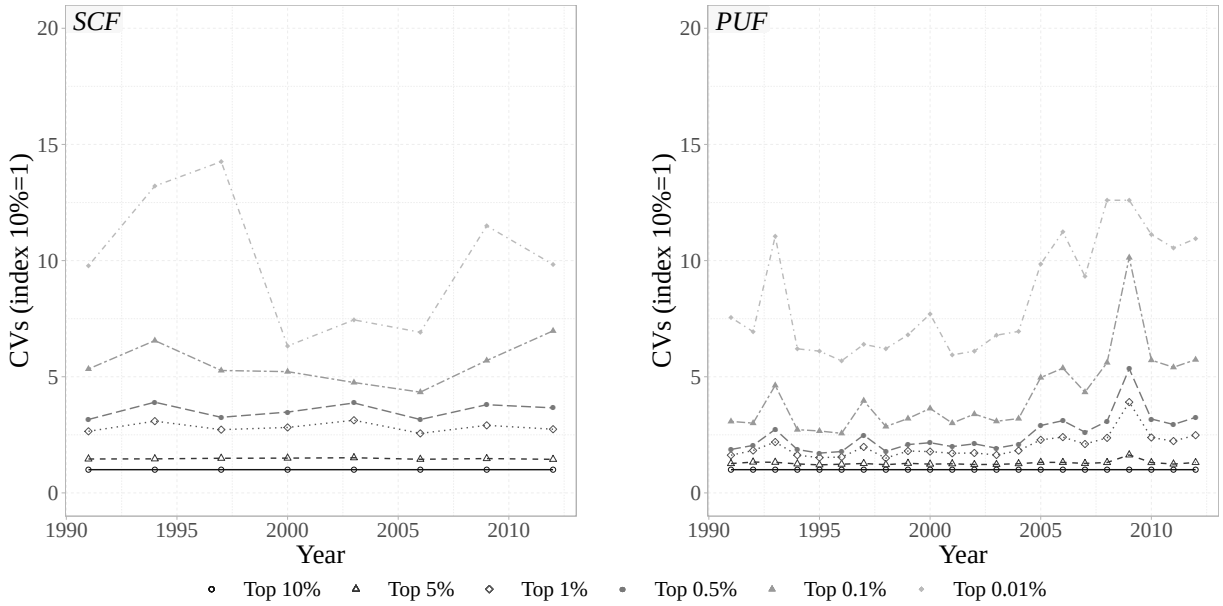

Figure 1.3: CVs in the SCF and the PUF for the income shares within the top 10 percent

for the six income shares within the top 10 percent between 1988 and 2018. My analysis suggests that even though sampling error is the main source of variation in the SCF estimates of top inequality, imputation error is not to be discarded. Specifically, until the early 2000s, imputation error accounted for at least 10 percent of total standard error, with the largest shares observed in the late 1980s and early 1990s.[29]

In the most recent years, the relative importance of imputation error has diminished, leaving sampling error as the sole source of variation in the SCF estimates of top income shares. Nevertheless, since imputation error was a significant contributor of the total variance in earlier years, the SCF standard error that accounts for both sampling and imputation errors is likely to be, on average, at least six times larger than that constructed using the PUF. Note that this is the case since, even though this paper does not estimate the PUF imputation/nonreponse error, qualitative evidence suggests this source of error to be incon-

[29]Since 1983, the SCF has gradually allowed for the possibility of reporting partial (range) information on dollar amounts in an effort to reduce item nonresponse. However, not until the 1995 SCF were the respondents allowed the possibility to provide a user-specified range of values or were they guided into a range response by a decision tree. This feature of the survey design may, at least partially, explain larger shares of imputation error in 1988 and 1991 than in 1994 and onward.

Table 1.2: Ratio of sampling error to total standard error in the SCF estimates of the income shares within the top 10 percent

Year	Income share of the top k%					
	10%	**5%**	**1%**	**0.5%**	**0.1%**	**0.01%**
1988	49.3	50.8	53.5	52.9	53.8	87.5
1991	82.8	77.0	88.4	91.8	89.3	92.6
1994	91.3	95.0	93.3	93.8	96.0	98.9
1997	97.1	95.4	89.7	81.8	77.0	91.1
2000	92.6	94.0	86.5	82.1	66.6	50.9
2003	78.9	81.7	88.8	92.4	92.7	97.2
2006	91.0	89.3	92.1	91.2	84.3	63.2
2009	95.7	93.5	97.7	98.3	96.3	97.3
2012	93.6	95.5	96.7	97.4	97.2	99.6
2015	97.2	98.4	99.2	99.6	98.5	99.5
2018	95.2	98.3	99.7	99.3	97.6	98.6

Ratios are expressed in percent

sequential for the PUF.

1.7.6 Long-Term Trends in Income Inequality

This section discusses long-term trends in income inequality. My two main objectives are to compare the estimated trend lines of the SCF and the PUF and to determine whether an observed increase in top income inequality between the early 1990s and early 2010s is statistically significant. In this exercise, I use data from 1991 through 2012 and regress estimated income shares (of the top 10 to the top 0.01 percent) on a constant and linear time trend using weighted least squares, with weights defined as reciprocals of squared standard

errors.[30]

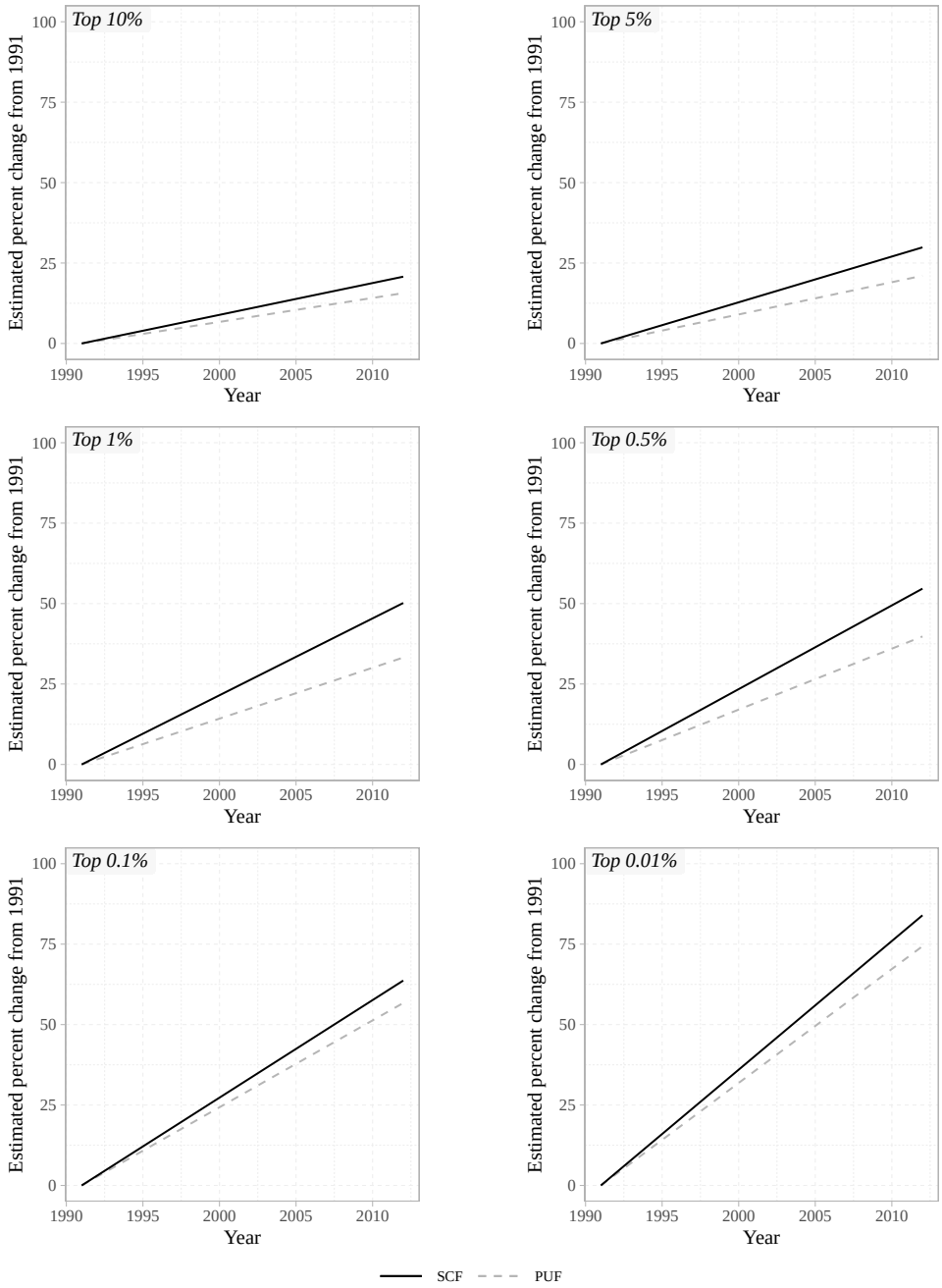

Figure 1.4: Estimated percentage change from 1991 in the six top income shares as measured by a weighted linear regression model

[30]Note that while the PUF standard error consists only of sampling error, the SCF standard error comprises both sampling and imputation errors.

As indicated in Figure 1.4, I find that both the SCF and the PUF suggest a statistically significant increase in all six top-decile income shares under consideration (see Table C.5 in Appendix C of the Online Supplementary Material for the estimation details.). However, the two data sets do not fully agree with respect to the estimated increase in income shares, with the SCF trend lines being consistently steeper than those constructed using the PUF. Nevertheless, since the observed discrepancies are moderate and by no means extensive, the two data sets imply an increase of comparable magnitude in top income inequality between the early 1990s and early 2012.

1.7.7 Income Inequality and the Great Recession

After discussing the long-term dynamics in income inequality, I focus on short-term trends during periods of rapid economic changes such as the 2007–2009 Great Recession. Specifically, in Figure 1.5, I present point estimates and their 95 percent confidence intervals for income shares of the top 10 percent, constructed using the SCF and the PUF between 2006 and 2012. I observe that the PUF estimates suggest a sharp and statistically significant decrease in income shares of the top 10 percent from 2007 to 2009, followed by a three-year long recovery to pre-recession levels. As noted by Thompson et al. (2018) "the factors explaining the rise and fall in income concentration are not fully understood, but some of the most prominent explanations for rising top incomes highlight the role played by individuals—who may be 'superstars' (Rosen, 1981) or 'rent seekers' (Bivens and Mishel, 2013)—whose compensation is relatively volatile from one year to the next (Bebchuk and Fried, 2003; Kaplan and Rauh, 2013, among others)." By contrast, the SCF does not support the claim of a statistically significant change in income shares of the top 10 percent in relation to the impact of the Great Recession.

Therefore, while the SCF can be used to analyze long-term dynamics and detect changes in income shares over longer-time horizons, it lacks statistical power to determine changes in income inequality over shorter-time horizons, including recessions and economic expansions. Furthermore, note that the PUF point estimates do not always lie within the SCF 95 percent confidence intervals. Specifically, the SCF and PUF 95 confidence intervals overlap in 2006

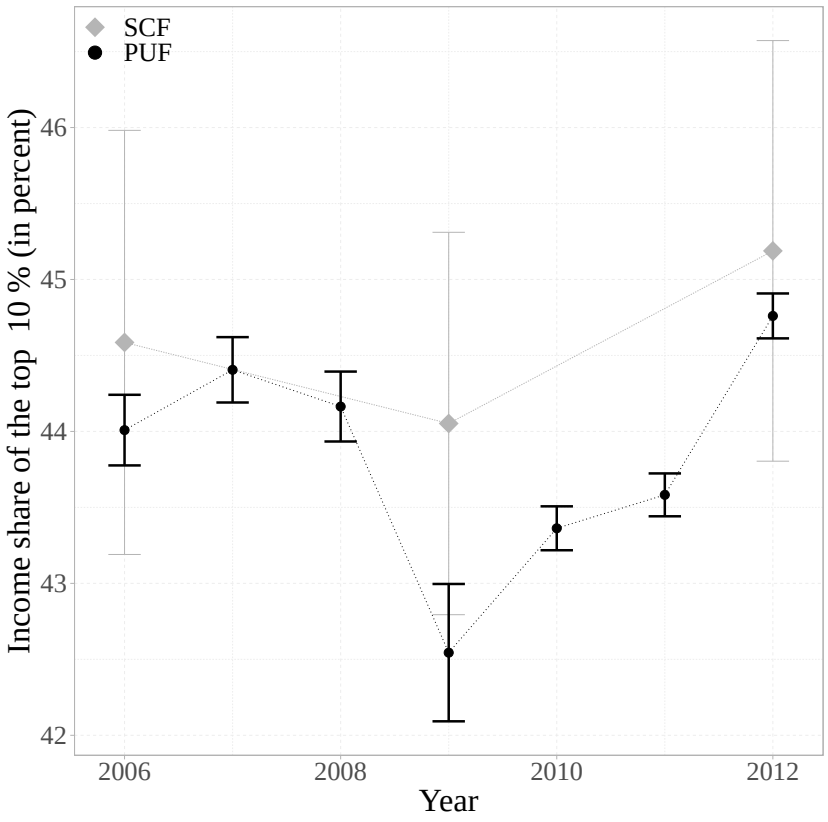

Figure 1.5: Income shares of the top 10 percent before and after the 2007–2009 Great Recession in the SCF and the PUF. Error bars indicate 95 percent confidence intervals around point estimates

and 2012, while the 2009 PUF estimate lies below the SCF confidence interval, implying a larger negative shock.

1.7.8 Regression Analysis

Lastly, I investigate whether there exists a statistical link between the SCF and the PUF. In this exercise, I regress the PUF point estimates on the SCF point estimates for each top-decile income share under consideration. As shown in Table 1.3, my analysis indicates strong correlation between the PUF and the SCF income shares of the top 10 to the top 0.5

percent.[31] This result is of great importance since it opens up the possibility of merging the two data sets into one, which would result in a superior data set with plenty of observations (see the PUF) and a rich set of demographic and socio-economic characteristics (see the SCF).

Table 1.3: Estimation results from regressing the PUF on the SCF for the income shares within the top 10 percent

Regressor	**PUF income share of the top k%**					
	10%	**5%**	**1%**	**0.5%**	**0.1%**	**0.01%**
const	0.075	0.061	0.030	0.021	0.024	0.017
	(1.793)	(1.990)	(1.486)	(0.892)	(1.126)	(2.086)
SCF income share	0.817***	0.772***	0.763***	0.789***	0.633	0.321
of the top k%	(8.204)	(7.799)	(5.968)	(3.837)	(1.673)	(0.727)
R^2	0.918	0.910	0.856	0.710	0.318	0.081

This table summarizes estimation results from six unweighted linear regressions of the PUF income shares on a constant and the SCF income shares. I report estimated coefficients and t-statistics in parentheses. "***" denotes statistical significance at the 99 percent significance level.

The two main data features that could be exploited while merging the SCF and the PUF are: (1) almost exact comparability of income categories between the two data sets and (2) similarities between the two data sets in reference to their sampling designs. As discussed in further detail in Section A.1 in Appendix A of the Online Supplementary Material, the SCF employs a dual-frame sample design consisting of a multi-stage national area-probability sample, designed to provide a good coverage of variables widely distributed in the population and a list sample, designed to provide sufficient coverage of variables largely concentrated in the upper tail of the wealth distribution. Since the selection of the SCF list sample is aided by the INSOLE, both the PUF and (partially) the SCF are based on the same frame of filed individual-income tax returns.

[31] For more granular income shares of the top 0.1 and 0.01 percent, the relationship in question is statistically insignificant.

1.7.9 Key Findings

A natural question is which data set, the SCF survey data or the PUF sample of tax records, proves more reliable in analyzing top income inequality? My study suggests that when interested primarily in estimates of top income shares, the PUF is better than the SCF. However, when interested in top income inequality more broadly defined, the answer depends on the research question at hand.

The main advantages of using the PUF lie in higher data frequency and larger number of observations, which, in turn, results in more precise estimates.[32] As discussed in detail in Section 1.7.3, the PUF sampling error is, on average, six times smaller than that of the SCF. Moreover, the SCF imputation error, briefly characterized in Section 1.7.5, introduces an additional and for the earlier years non-negligible layer of uncertainty to the SCF point estimates, whereas the PUF nonresponse error is likely to be inconsequential. Consequently, once all sources of error are accounted for, the SCF standard error is likely to be *at least* six times larger than that of the PUF. Note that though this study does not estimate all components of TSE, it provides qualitative evidence suggesting that, once accounted for, measurement error in the SCF is still likely to be much more substantial than in the PUF.

A natural question is whether more households could be surveyed for the SCF with the objective of producing more precise estimates of income concentration. Given that the cost of conducting the 2016 SCF was equal to $18 million, increasing the SCF sample size would be an expensive undertaking, and therefore, any such decision would require a detail cost-benefit analysis, which is beyond the scope of the current paper.

Since SCF estimates are considerably less precise than those constructed using the PUF, can they still be used to draw reliable conclusions regarding top income inequality? My study suggests that while the SCF can be used to answer most questions regarding the long-term dynamics in less granular income shares, the data are not well-suited to analyze short-term horizons. This is the case since using the SCF survey data for statistical testing is likely to result in a high probability of Type II error, leading to a failure to reject a null hypothesis

[32] Moreover, as emphasized by Atkinson et al. (2011), tax data are available for longer time horizons and many more countries than are survey data, which makes it possible to study structural shifts in income distributions spanning several decades and to conduct cross-country comparisons.

testing for a statistically significant change in income concentration from one survey year to another.

Regarding long-term dynamics, in Section 1.7.6, I find that both the SCF and the PUF suggest a statistically significant increase in all of the six top-decile income shares under consideration. However, as we move from less to more granular income shares, the accuracy of the SCF point estimates deteriorates (see Section 1.7.3). Moreover, as discussed in further detail in Section 1.7.4, relative increments in CVs in the SCF are much more pronounced than those in the PUF. Consequently, using the SCF for the estimation of more granular income shares leads to a greater loss in precision than if we were to use the PUF. Lastly, Section 1.7.1 shows that the number of observations in the SCF above the 99.9 and 99.99 income fractiles is insufficient in both absolute and relative terms, especially when compared to the large number of observations in the PUF. Therefore, even though the SCF may be useful for analyzing long-term dynamics in income shares of the top 10 or 5 percent, it should not be used in studies that focus on the top 0.1 or 0.01 percent.

Yet Saez and Zucman (2016), Bricker et al. (2018), Saez and Zucman (2020b), and others continue to rely upon the SCF estimates of the most granular income and wealth shares of the top 0.1 and the top 0.01 percent. In particular, the estimates in question are used as reference points in assessing income and wealth concentration measures obtained using alternative data sets and/or estimation techniques. Most importantly, such comparisons are done without accounting for the SCF standard error, which affects not only the SCF point estimates, but foremost, the short- and long-term dynamics in income and wealth inequality. A notable exception is Kopczuk and Saez (2004) who acknowledge the small number of observations in the SCF and find the survey data unreliable for the estimation of wealth shares for groups smaller than the top 0.5 percent.

Regarding short-term dynamics, I find that large confidence intervals around the SCF point estimates may falsely suggest lack of a statistically significant increase in income shares from one SCF survey year to another. For instance, whereas the SCF does not suggest a statistically significant change in top income shares before, during, and after the 2007–2009 Great Recession, the PUF clearly illustrates an initial drop (2006–2009) followed by a statistically significant increase (2009–2012), to at or even above the pre-recession levels.

On the other hand, when it is necessary to use more-up-to-date data, control for demographic and socio-economic characteristics, or examine the composition of top earners by age, sex, or marital status, the PUF cannot be used, and instead, one must rely upon the SCF.[33,34] In order to arrive at credible results using the SCF, it is important to account for the survey's small number of observations, and, as such, refrain from estimating very granular income shares or analyzing small population subgroups. For instance, while the number of observations in the SCF is sufficient to estimate income shares of the top 10, 5, 1, and 0.5 percent of married households, it remains insufficient for conducting an analogous exercise for the smaller population of married household with minor dependents (see Table 1.4 below).

Table 1.4: Number of observations in the upper tail of income distribution by marital status and household composition

Population	**10%**	**5%**	**1%**	**0.5%**	**0.1%**	**0.01%**
All	1,455	1,105	644	490	263	84
Married	1,095	856	519	396	201	65
Married with dependents	452	343	196	145	79	24

Based on the 2019 SCF. Without loss of generality, I report the number of observations when using the first implicate. Note that the number of observations varies across the five imputations as well as across the 999 bootstrapped sample replicates.

An ideal data set for studying top income inequality would comprise a large number of observations and a rich set of demographic and socio-economic characteristics. Could such a data set be constructed from a merge of the data from the SCF and the PUF? While the answer to this question is beyond the scope of the current paper, as shown in Section 1.7.8, there exists an explicit statistical link between between the PUF and the SCF, which could be used in a follow-up research project to analyze how economic factors could impact PUF

[33]The PUF is released with a substantial lag when compared to the SCF. At the time of writing, the latest available PUF data set is from 2012 and the latest available SCF data set is from 2019.

[34]Even in the highly-restricted INSOLE sample data, information on taxpayers' demographics and socio-economic characteristics is limited to age and sex, which are merged into the INSOLE from Social Security records.

through its link with SCF.[35]

Specifically, combining information from the SCF and the PUF would be particularly advantageous for studying racial and ethnic income inequality. According to Bhutta et al. (2020), "the typical White family has eight times the wealth of the typical Black family and five times the wealth of the typical Hispanic family." Moreover, since COVID-19 is having disproportionate impact on people of color (blacks have the highest death toll per 100,000, in comparison to whites, Asians, Latinos, and indigenous Americans)[36] we can expect these longstanding racial and ethnic disparities in income to grow. So far, the Tax Policy Center and the Joint Committee of Taxation used administrative tax records in combination with the SCF survey data to examine the impact of the 2017 federal corporate tax cut (Trump's tax) on racial and ethnic income inequality.[37] Given the widening gap in income between black and white Americans, similar analyses are called for in the context of COVID-19, the 2007–2009 Great Recession, and any other major shock to the post-WWII US economy.

Consequently, this study does not portray the SCF and the PUF as supplements but rather as complementary data sources that can and should be used interchangeably in order to best answer a specific research question. However, this is often not possible due to a strictly limited access to individual-income tax returns, which is only granted to a handful of researchers. For this reason, this paper advocates for a broader access to various sources of administrative micro-level data (such as the PUF), supporting the calls by Card et al. (2010) and others.

1.7.9.1 Conclusion

Regarding top income inequality, I find a statistically significant increase in all six top-decile income shares under consideration in the twenty-two-year period between 1991 and 2012. Specifically, the weighted least square regression analysis of the PUF estimates suggests an increase in income shares of the top 10, 5, 1, 0.5, 0.1, and 0.01 percent by 16, 21,

[35]Moreover, since the PUF is released with a substantial delay when compared to the SCF (the latest available data set is from 2012), it is even more critical to statistically link the PUF and the SCF. While naive way would be through regressions, a more sophisticated approach belongs to future research.

[36]See https://www.apmresearchlab.org/covid/deaths-by-race (accessed on October 7, 2020).

[37]See https://www.nytimes.com/2018/10/11/business/trump-tax-cuts-white-americans.html (accessed on October 7, 2020).

33, 40, 56, and 75 percent, respectively. These results are in line with those published in the related literature, supporting the claim of the long-term trend toward growing income concentration.

1.8 Empirical Results on Wealth Shares

My analysis of the wealth shares within the top 10 percent is based on four sets of estimates: one constructed using the SCF and three constructed using the PUF. The first set of PUF estimates corresponds to wealth shares estimated under a homogeneity assumption imposed on all rates of return of the underlying capitalization model, whereas the other two sets allow for heterogeneous rates of return on taxable interest-bearing assets (see Section 1.6.2).

In Section 1.8.1, I examine the number of observations available for the estimation of top wealth inequality in the SCF and the PUF. In Sections 1.8.2–1.8.6, I discuss point estimates, standard errors, and the long- and short-term dynamics in top wealth inequality. Finally, Section 1.8.7 summarizes my main finding and concludes. I use triennial data on wealth between 1989 and 2019 from the 1989–2019 SCF and annual data on wealth between 1991 and 2012 from the 1991–2012 PUF.

1.8.1 Number of Observations

I start my analysis with a brief description of the number of observations available for the estimation of the six top-decile wealth shares under consideration in the SCF and the PUF between 1992 and 2010. Since neither of the two data sets contain any missing values, the number of observations available to study wealth is equal to the total number of observations in each of the two data sets. Consequently, as indicated in Table 1.5, there are, on average, 30 times more observations available in the PUF than there are in the SCF.

Large discrepancies also exist between the SCF and the PUF with respect to the number of observations in the far right tail of wealth distribution. In particular, even though the

Table 1.5: Ratio of the SCF number of observations to the PUF number of observations available for the estimation of wealth

Year	Total	Above the k wealth fractile					
		90	**95**	**99**	**99.5**	**99.9**	**99.99**
1992	4.2	2.8	2.7	2.2	2.1	2.3	3.0
1995	4.2	2.7	2.5	2.0	1.8	1.7	3.3
1998	3.5	2.0	1.9	1.4	1.3	1.4	2.0
2001	3.1	1.8	1.6	1.3	1.2	1.5	2.2
2004	3.0	1.6	1.5	1.3	1.2	1.5	2.9
2007	3.1	2.0	1.8	1.6	1.7	2.9	7.1
2010	4.1	2.4	2.1	1.8	1.8	2.8	7.5

Ratios are expressed in percent

SCF oversamples wealthy households, the number of observations above all six top-decile fractiles under consideration is much smaller in the SCF than in the PUF. Consequently, as discussed in more detail in Section 1.8.3, the SCF confidence intervals are substantially wider than those constructed using the PUF.

1.8.2 Point Estimates

My second set of results pertains to point estimates of the six wealth shares of the top 10 to the top 0.01 percent. The SCF and the PUF point estimates are compared over the seven-year period between 1992 and 2010; the homogeneous and heterogeneous PUF estimates are compared over the twenty-two year period between 1991–2012. As indicated in Figure 1.6, I find that the SCF and the PUF often disagree with respect to levels of the estimated wealth shares, with the SCF implying larger shares of the top 10, 5, 1, and 0.5 percent, and smaller shares of the top 0.1 and 0.01 percent. Since at the moment, I do not estimate defined benefit pensions using the SCF, it is beyond the scope of the present paper to determine whether

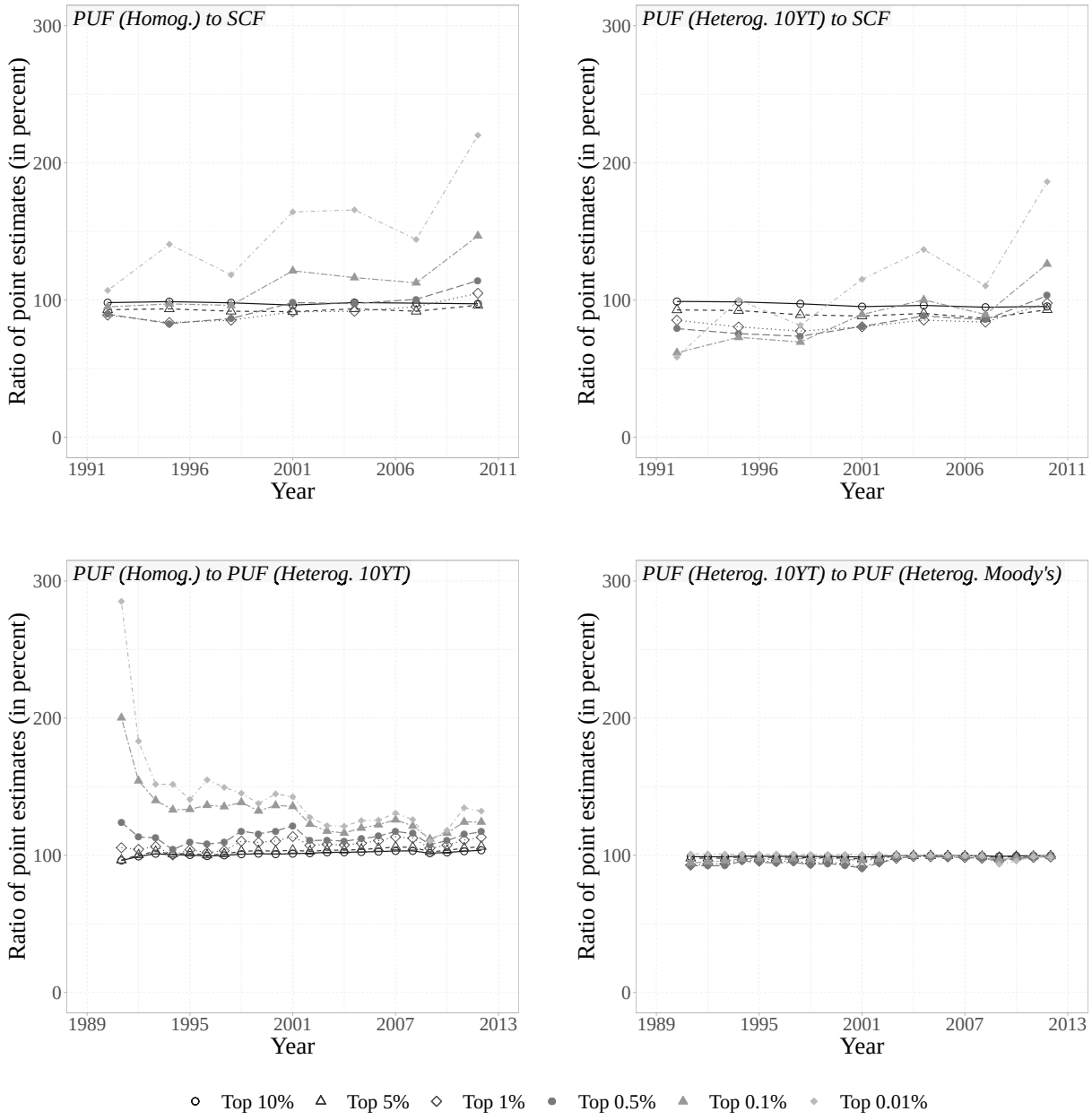

Figure 1.6: Ratio of point estimates for the wealth shares within the top 10 percent

some of the observed discrepancies in point estimates could be explained by differences in the operational definitions of wealth between the SCF and the PUF.

In addition to comparing the SCF and the PUF, I analyze ratios between the three

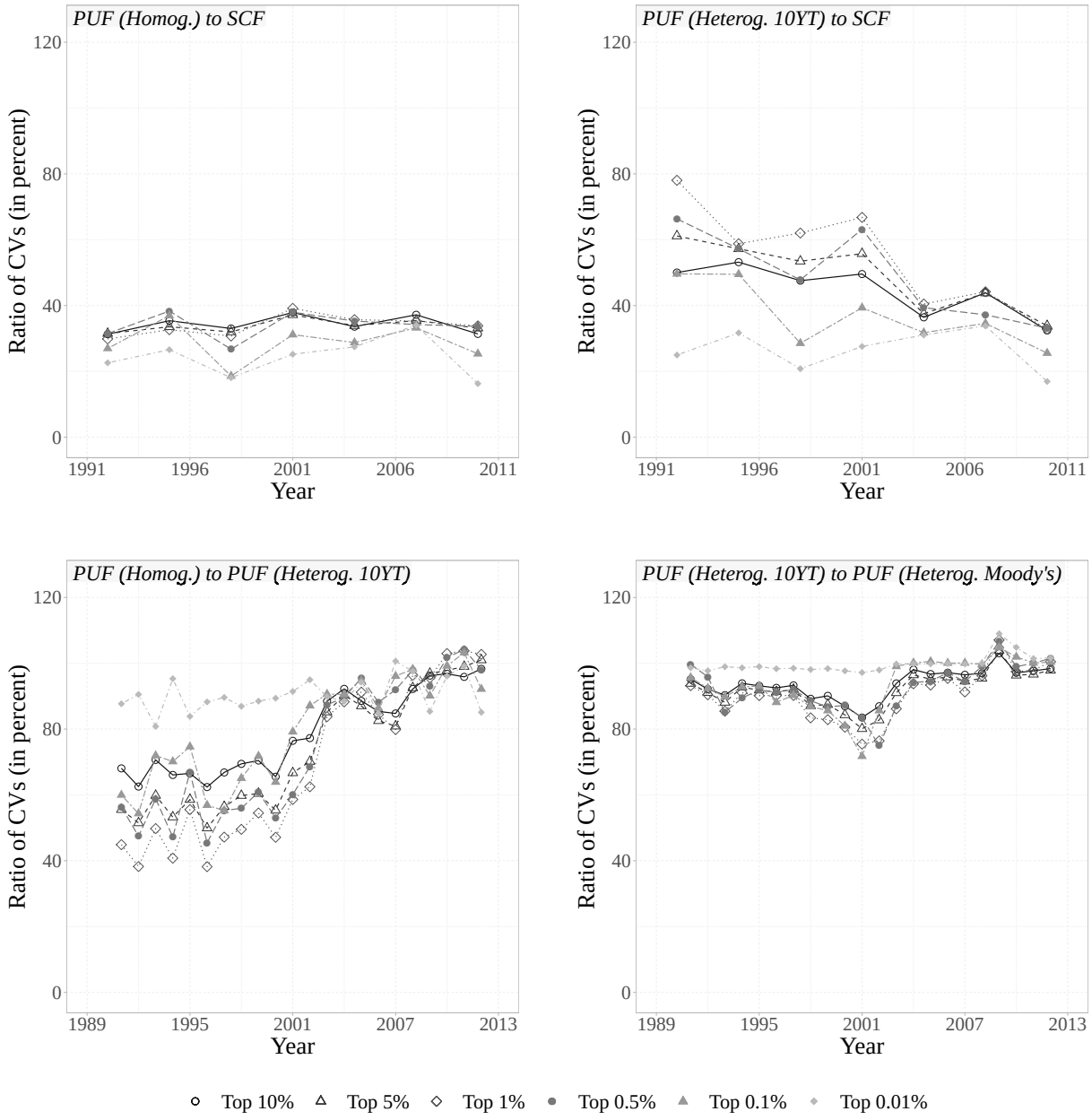

Figure 1.7: Ratio of CVs for the wealth shares within the top 10 percent

different sets of PUF estimates. Whereas I do not find significant differences in the estimated wealth shares between the two heterogeneous sets of PUF estimates, I observe non-trivial discrepancies between those estimated using homogeneous and heterogeneous models. For

less granular income shares, the observed discrepancies are small to moderate. However, for more granular wealth shares of the top 0.1 and 0.01 percent, the differences become large, especially over a ten-year period between the early 1990s and the early 2000s. For example, whereas the homogeneous model suggests an increase in wealth shares of the top 0.01 percent from 3.2 to 10.2 percent between 1991 and 2012, the heterogeneous model indicates an increase from 1.25 to 7.7. Overall, my analysis suggests that the differences between the homogeneous and heterogeneous PUF estimates are similar in magnitude to those observed between the SCF and the PUF. Consequently, except for wealth shares of the top 10 percent, there is little agreement between the different set of estimates, which makes the analysis of wealth inequality considerably more challenging than that of income.

1.8.3 Relative Magnitudes of Sampling Error Across Data Sets

In the present section, I focus on sampling errors. In particular, I analyze ratios of CVs computed for the six wealth shares of the top 10 to the top 0.01 percent. The SCF and the PUF point estimates are compared over the seven-year period between 1992 and 2010; the homogeneous and heterogeneous PUF estimates are compared over the twenty-two year period between 1991–2012. Like for top income inequality, I find sampling error in the SCF to be considerably bigger than that in the PUF—a result of the great discrepancy in the number of observations between the two data sets. As indicated in Figure 1.7, CVs for the SCF are 20–80 percent larger than those for the PUF. Therefore, if we were only concerned with the precision of the estimates, the PUF estimates would be preferred. However, in order to determine the most suitable set of estimates it is necessary to consider a wider range of factors, including estimates' accuracy and their possible dependence on wrong modeling assumptions.

1.8.4 SCF Standard Error Decomposition

In Table 1.6, I present ratios of sampling error to total standard error for the six wealth shares within the top 10 percent between 1989 and 2019. My analysis suggests that, unlike for income, sampling error in the SCF estimates of top wealth inequality does not constitute

Table 1.6: Ratio of sampling error to total standard error in the SCF estimates of the wealth shares within the top 10 percent

Year	Wealth share of the top k%					
	10%	**5%**	**1%**	**0.5%**	**0.1%**	**0.01%**
1989	68.6	78.4	56.3	59.0	54.9	73.8
1992	40.3	42.8	58.1	39.7	56.9	92.8
1995	88.1	88.5	85.2	70.5	72.1	90.1
1998	42.0	52.2	36.3	36.4	56.3	82.1
2001	55.7	66.9	71.8	73.7	77.3	89.2
2004	37.0	46.4	81.6	86.6	85.8	93.7
2007	67.6	70.8	66.1	72.8	96.2	96.2
2010	55.2	78.1	61.5	47.0	70.7	96.8
2013	91.2	87.3	95.8	97.3	94.5	95.1
2016	68.7	79.1	91.6	82.2	91.9	85.0
2019	85.3	65.3	76.5	83.1	93.0	95.7

Ratios are expressed in percent

the only considerable source of variation. In particular, I find that between 1989 and 2019, imputation error often accounted for between 30 and 50 percent of total standard error in the estimated wealth shares. From the perspective of policy-makers, this result suggests that revising the SCF imputation procedure and/or introducing new interview techniques aimed at mitigating item nonresponse could prove effective in reducing the uncertainty in the SCF estimates of top-wealth inequality. This is an important finding, since reducing the SCF sampling error by significantly increasing the number of households selected for the survey is nearly impossible due to the large costs and organizational complexity associated with conducting the survey. Furthermore, note that the observed imputation share, defined as the ratio of imputation error to the total standard error, generally declines as we move up

the wealth distribution, suggesting that revising the SCF imputation procedure would prove most effective in reducing the error in the less granular wealth shares of the top 10, 5, 1, and 0.5 percent.

1.8.5 Long-Term Dynamics

In order to analyze long-term dynamics in wealth inequality between 1992 and 2010, I estimate weighted linear regressions of top-decile wealth shares on a constant and linear time trend. As illustrated in Figure 1.8, the SCF regression results support the claim of a statistically significant increase in two out of the six wealth shares under consideration (see Table C.6 in Appendix C of the Online Supplementary Material for the estimation details). Specifically, I find a statistically significant increase in the wealth shares of the top 10 and the top 5 percent at the 1 percent significance level.

The PUF estimates, on the other hand, suggest a statistically significant increase in all of the six top-decile wealth shares under consideration, irrespective of the assumption imposed on the rate of return of taxable interest-bearing assets. However, there exist considerable differences in the estimated trend lines between the homogeneous and heterogeneous sets of estimates. For the wealth shares of the top 10 and 5 percent, the homogeneous model suggests a greater increase in inequality between 1992 and 2010 than the heterogeneous model. Yet for the wealth shares of the top 0.1 and 0.01 percent the opposite is true: the heterogeneous model indicates a much more pronounced increase than the homogeneous model. In particular, the estimated percent change in the wealth shares of the top 0.01 percent from 1992 to 2010 exceeds 150 percent for the heterogeneous model while being equal to over 75 percent for the homogeneous model. Therefore, without taking a strong stance on what model better describes the state of the economy, the PUF estimates remain inconclusive with respect to the magnitude by which the top wealth inequality has increased.

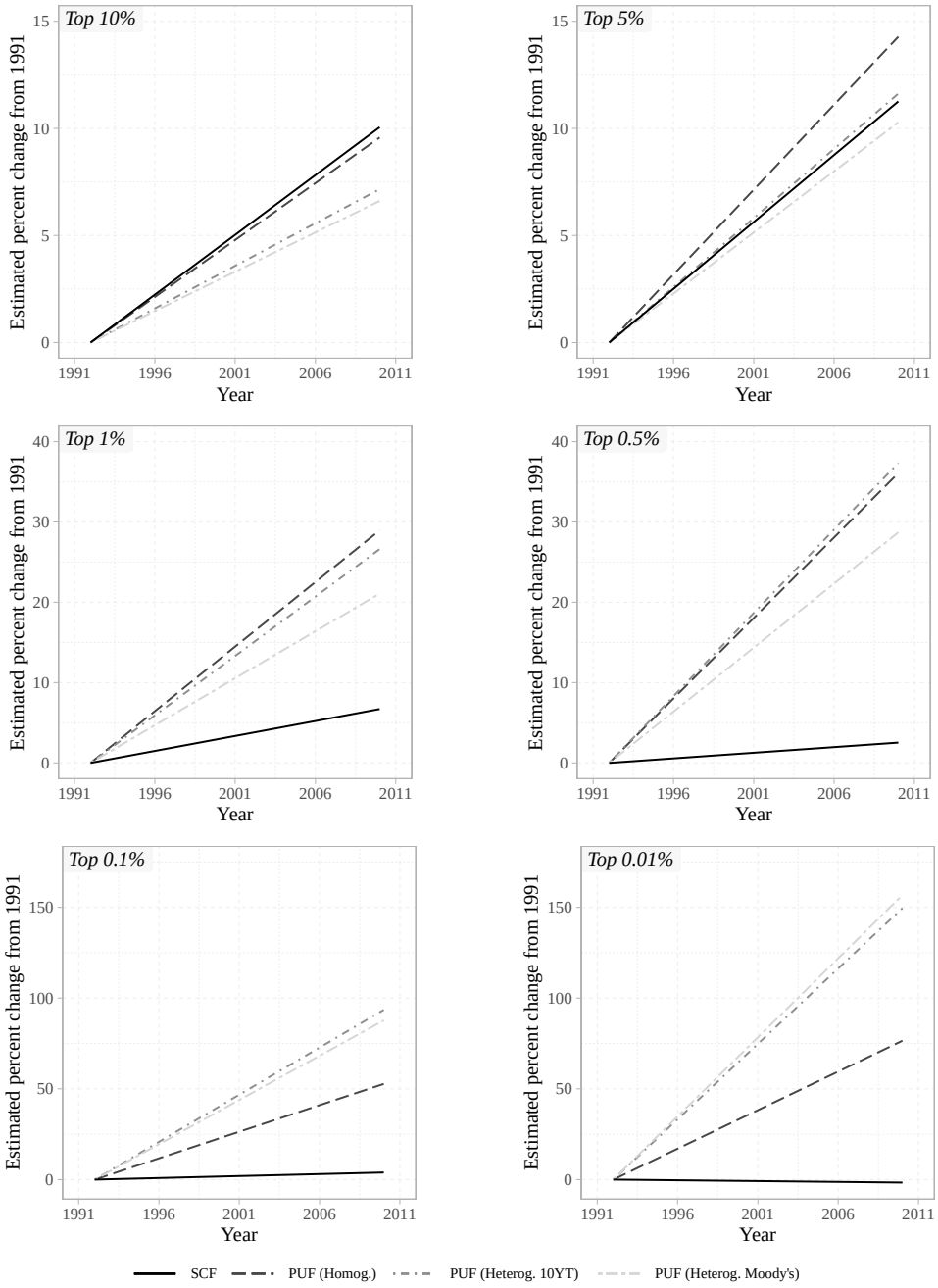

Figure 1.8: Estimated percentage change from 1991 in the six top wealth shares as measured by a weighted linear regression model

1.8.6 Wealth Inequality and the Great Recession

In the present section, I discuss the impact of the 2007–2009 Great Recession on wealth inequality, with the main focus on wealth shares of the top 10 percent between 2004 and 2013. As illustrated in Figure 1.9, I find considerable differences between the homogeneous and heterogeneous PUF estimates. Specifically, whereas homogeneous estimates had been steadily increasing in the aftermath of the Great Recession, heterogeneous estimates came to near standstill once the recession ended. In other words, the two sets of estimates lead to drastically different conclusions regarding the short-term dynamics in top wealth inequality. One set of estimates suggests a statistically significant increase of 2 percentage points between 2009 and 2012, whereas the other set implies no statistically significant change since the recession came to an end in 2009. Therefore, with reference to Section 1.8.5, I find the PUF estimates inconclusive not only in relation to the long-term dynamics in top wealth inequality, but also when examining shorter-time horizons and times of rapid economic changes, such as the 2007–2009 Great Recession.

1.8.7 Key Findings

My analysis suggests that the PUF estimates of top wealth inequality are largely inconclusive—a result of substantial differences in the estimates resulting from the set of assumptions imposed on the underlying capitalization model.

As indicated in Section 1.8.2, I find non-trivial differences in the PUF point estimates obtained under the homogeneity assumption and those computed using heterogeneous rates of return on taxable interest-bearing assets. In addition to disparities in point estimates, the two sets of estimates differ with respect to short- and long-run trends in top wealth inequality. First, regarding the long-term dynamics, the estimated trend lines lead to different conclusions regarding the severity of the ongoing crisis linked to rising inequality. Second, regarding the short-term dynamics, I find that whereas the homogeneous estimates support the claim of an increase in wealth inequality following the 2007–2009 Great Recession, the heterogeneous estimates do not indicate a statistically significant change.

Assuming different rates of return leads to considerably different dynamics in top wealth

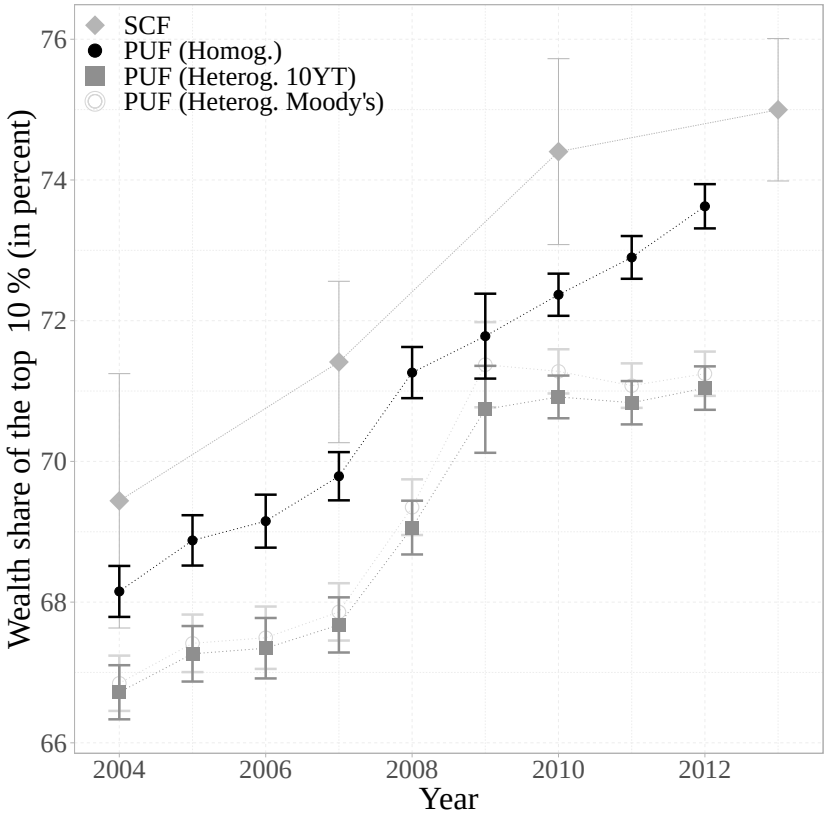

Figure 1.9: Wealth shares of the top 10 percent before and after the 2007–2009 Great Recession measured in the SCF and the PUF. Error bars indicate 95 percent confidence intervals around point estimates

inequality. Then, since the assumptions matter, a natural question is what rates of return should we be considering for each of the asset classes under consideration? Should they be mostly homo- or heterogeneous? Should we allow them to vary by income or wealth percentile? Should they depend on portfolio composition or regional macroeconomic conditions? In this paper, following Bricker et al. (2018), I consider a heterogeneous rate of return on only one class of assets, where I impose a hard, and rather unrealistic, cut-off between high and low rates of return. Therefore, another important question is whether a different and more realistic set of assumptions would result in distinct estimates of top wealth inequality. Based on the extent to which the PUF estimates obtained under the homogeneity assumption and

those computed using heterogeneous rates of return on taxable interest-bearing assets differ, I presume the resulting estimates would be considerably different from those currently obtained. This presumption is line with Kopczuk and Saez (2004), who, among many others, expressed concerns about the estimation of wealth using tax-based income data, resulting from "substantial and unobservable heterogeneity in the returns of many assets, especially corporate stock."

Since the PUF estimates of top wealth inequality are functions of numerous assumptions imposed on the underlying capitalization model, is the SCF a better alternative? My analysis suggests that for the less granular wealth shares of the top 10 to the top 0.5 percent the SCF appears more reliable. This results mainly from the fact that the SCF *measures* wealth, whereas the PUF *infers* wealth from data on income. However, using the SCF for the estimation of top wealth shares has its own limitations resulting from a small number of observations and large confidence intervals. Therefore, whereas the SCF can be used effectively to analyze the less granular wealth shares of the top 10 to the top 0.5 percent, the data remain inadequate to get a realistic picture of the wealth shares of the top 0.1 and 0.01 percent.[38]

It is important to note that while resolving the problem of the large standard errors in the SCF would require making changes to the survey design and/or the SCF imputation procedure, better estimates of wealth concentration could be obtained by merging the SCF and the PUF as discussed in the context of income inequality in Section 1.7.8. This approach would allow to reduce the SCF sampling error by constructing estimates out of a much larger data set combining both survey and tax data. At the same time, wealth measures would continue to be estimated using values of asset and liability holdings reported by the SCF interviewees, which would result in estimates of top-decile wealth shares free from modeling errors currently prevalent in the PUF estimates derived from an assumed capitalization model. Alternatively, future research could focus on refining the Saez and Zucman (2016)'s capitalization model, with the objective of producing more reliable estimates of wealth inequality using the PUF. A significant research effort has been already undertaken by Smith

[38]The wealth shares of the top 10 to the top 0.5 percent constructed using the SCF can be further refined by adding the wealth of the Forbes 400 wealthiest Americans and the value of defined benefit pensions as in Bricker et al. (2016).

et al. (2019) and Saez and Zucman (2020a,b), adding to our understanding of the numerous advantages and disadvantages embedded in studying wealth inequality using capitalization methods.

Since both the SCF survey data and the PUF individual-income tax returns pose major challenges for the estimation of top wealth inequality in the US, this paper supports Saez and Zucman (2020b) in calling for "more and improved statistics on inequality." The authors further emphasize that "we could and should do better to measure US wealth inequality than rely on a triennial survey of 6,200 families (the Survey of Consumer Finances) or indirectly infer asset ownership based on income flows (the capitalization method)."

1.8.7.1 Conclusions

Regarding the nineteen year period between 1992 and 2010, my analysis suggests a statistically significant increase in two out of the six wealth shares under consideration. Specifically, the weighted least square regression analysis of the SCF estimates suggests a statistically significant increase in wealth shares of the top 10 and 5 percent by 6.7 and 6.2 percentage points, respectively, and an insignificant increase in wealth shares of the top 1 and 0.5 percent. Since, as indicated above, neither the SCF nor the PUF proves credible to analyze more granular wealth shares of the top 0.1 and 0.01 percent, this study does not draw any conclusions related to top wealth inequality above the 99.5 wealth fractile. In particular, it neither supports not contradicts Saez and Zucman (2016)'s widely-cited conclusion regarding a 100 percent increase in the wealth shares of the top 0.1 percent between 1991 and 2012, a finding that has drawn substantial media coverage and major interest from politicians and policy makers. Lastly, the weighted linear regression analysis of the SCF point estimates does not suggest a larger increase in the wealth shares of the top 1 percent than in the wealth shares of the top 10 percent. On the contrary, I find that the estimated change in the wealth shares of the top 10 percent between 1991 and 2012 exceeds the estimated change in the wealth shares of the top 1 percent by more than 3 percentage points. This result is of great importance as it casts doubts on a popular presumption that the observed rise in wealth inequality is driven particularly by the richest of the rich, leading

to a surge in wealth disparity within the top 10 percent of wealth distribution.

1.9 Structural Exercise

In the following, I investigate how the SCF and the PUF data-driven errors in estimates of top income inequality affect outcomes of structural macroeconomic models. In Section 1.9.1, I introduce the theoretical foundation of my study, which is the augmented random growth model with type-dependence proposed by Gabaix et al. (2016). Next, in Section 1.9.2, I discuss details of my calibration strategy that, in contrast to the default approach, involves multiple model calibrations, each time to a different value of the calibration target randomly drawn from the estimated 95 percent confidence interval. Finally, in Section 1.9.3, I present the results of my structural exercise, and conclude on how data-driven errors in calibration targets may affect outcomes of macroeconomic and policy-oriented studies more generally.

1.9.1 Model

Consider a continuum of workers, where each worker i is either high- or low-type, with high-type workers having higher mean growth rate of income than low-type workers. Workers enter the labor market as high-types with probability θ and as low-types with probability $1-\theta$. Whereas no worker of low-type can ever become a high-type, high-type workers do switch to a low-type with probability α. Since a low-type is an absorbing state, high-type workers can switch to a low-type at most once in their life-time. Moreover, workers retire at rate δ and are replaced by new labor entrants with wages drawn from a known distribution ψ.

Next, let x_{it} denote a natural logarithm of income of worker i of type j at time t, and let the dynamics of x_{it} be given by a type-dependent random growth model as in Gabaix et al. (2016):

$$dx_{it} = \mu_j dt + \sigma_j dZ_{it} + \text{Injection} - \text{Death}, \quad j = \{H, L\}, \tag{22}$$

where Z_{it} is a standard Brownian motion, μ_j and σ_j^2 are type-dependent mean and variance of growth rate of log income, and where H and L are shorthand notations for high- and low-types, respectively.

Moreover, assume that the stationary distribution of log income has a Pareto tail,

$$P\left(x_{it} > x\right) \sim Ce^{-\xi x}, \tag{23}$$

where C is a constant and $\xi > 0$ is a power law exponent given by

$$\xi = \frac{-\mu_H + \sqrt{\mu_H^2 + 2\sigma_H^2\left(\delta - \alpha\right)}}{\sigma_H^2}. \tag{24}$$

Finally, impose that the economy is in a Pareto steady state with $\sigma_H = 0.15$, $\alpha = 1/6$, and $\delta = 1/30$, and assume that μ_H is calibrated from a one-to-one-mapping between the inverse of the power law exponent, say η, and the empirical ratio of two, top-decile income shares:

$$\eta = \frac{1}{\xi} = 1 + \log_{10}\left(\frac{p_{(k/10)}}{p_k}\right). \tag{25}$$

1.9.2 Calibration

Unlike Gabaix et al. (2016), where the authors consider a single value of the inverse of the power law exponent, η, I calibrate the model to multiple values of η drawn from a 95 percent confidence interval around the η point estimate. In particular, I conduct $B = 100$ random draws, where for each drawn value of η, I solve the model for the transition dynamics to a new steady state.

Gabaix et al. (2016) compute a point estimate of η using data on top income shares from the World Income Database (WID), whereas this paper's focus is on the SCF and the PUF. Moreover, the authors calibrate the model to the 1973 WID, whereas the SCF and the PUF data considered in this analysis start in 1988 and 1991, respectively. An obvious solution to this problem would be to estimate η using the 1989 SCF and/or the 1991 PUF. However, this strategy would not allow me to directly compare my results to those in Gabaix et al.

(2016), since I would be effectively investigating transition dynamics over a different time horizon: 1988–2065 (or 1991–2068) versus 1973–2050. Moreover, it would likely require me to consider a different magnitude of the shock and result in changes to the model parameters describing the initial Pareto steady state. Instead, I directly build upon the exercise in Gabaix et al. (2016) at the cost of making two assumptions regarding hypothetical values of the SCF and the PUF estimates of η in 1973.

Assumption 1: Let $\hat{\eta}_{\text{SCF},1973}$ and $\hat{\eta}_{\text{PUF},1973}$ denote the SCF and the PUF point estimates of η in 1973, and assume that $\hat{\eta}_{\text{SCF},1973}$ and $\hat{\eta}_{\text{PUF},1973}$ are equal to the Gabaix et al. (2016)'s point estimate of η obtained using the WID, say $\hat{\eta}_{\text{WID},1973}$:

$$\hat{\eta}_{\text{SCF},1973} = \hat{\eta}_{\text{PUF},1973} = \hat{\eta}_{\text{WID},1973} = 0.39. \tag{26}$$

Assumption 2: Let $\text{CV}_{\text{SCF},1973}$ and $\text{CV}_{\text{SCF},1988}$ denote CVs for the SCF estimates of η in 1973 and 1988, and let $\text{CV}_{\text{PUF},1973}$ and $\text{CV}_{\text{PUF},1991}$ denote CVs for the PUF estimates of η in 1973 and 1991. In the following, I assume the same CVs for the SCF and the PUF point estimates of η between the early 1970s and the early 1990s,

$$\text{CV}_{\text{SCF},1973} = \text{CV}_{\text{SCF},1988} \quad \text{and} \quad \text{CV}_{\text{PUF},1973} = \text{CV}_{\text{PUF},1991}. \tag{27}$$

Whereas the first assumption guarantees a direct comparison with Gabaix et al. (2016), the second conjecture provides me with conservative estimates of the SCF and the PUF standard errors in $\hat{\eta}_{1973}$.[39]

In the first step, I compute the point estimates and standard errors of η using the 1989 SCF and the 1991 PUF. For the SCF, I find $\hat{\eta}_{\text{SCF},1988} = 0.56$ and $\text{SE}\left(\hat{\eta}_{\text{SCF},1988}\right) = 0.03$, which results in $\text{CV}_{\text{SCF},1988} = 0.054$; for the PUF, I find $\hat{\eta}_{\text{PUF},1991} = 0.51$ and $\text{SE}\left(\hat{\eta}_{\text{PUF},1991}\right) = 0.001$, which results in $\text{CV}_{\text{PUF},1991} = 0.003$.

In the second step, I use Assumptions 1 and 2 to compute the standard errors in the SCF and the PUF estimates of η in 1973:

$$\text{SE}\left(\hat{\eta}_{\text{SCF},1973}\right) = \text{CV}_{\text{SCF},1973} \times \hat{\eta}_{\text{SCF},1973} = 0.023 \tag{28}$$

[39]The estimates are conservative since it is likely that the standard errors in the SCF and the PUF estimates of η would have substantially decreased over the twenty-year period between the early 1970s and the early 1990s.

$$\text{SE}\left(\hat{\eta}_{\text{PUF},1973}\right) = \text{CV}_{\text{PUF},1973} \times \hat{\eta}_{\text{PUF},1973} = 0.001. \tag{29}$$

Having estimated $\text{SE}\left(\hat{\eta}_{\text{SCF},1973}\right)$ and $\text{SE}\left(\hat{\eta}_{\text{PUF},1973}\right)$, I construct 95 percent confidence intervals around $\hat{\eta}_{\text{SCF},1973}$ and $\hat{\eta}_{\text{PUF},1973}$. For the SCF, the 95 percent confidence interval is given by $[0.36, 0.45]$; for the PUF, the 95 percent confidence interval is given by $[0.40, 0.41]$.

In the third and final step of my analysis, I conduct $B = 100$ random draws from $[0.36, 0.45]$ and $[0.40, 0.41]$, where for each drawn value of η, I solve for the model's transition dynamics to a new steady state. For each of the two data sets, this exercise results in $B = 100$ different transition dynamics to a new steady state, where the observed variability in the model's outcomes is driven solely by the underlying uncertainty in the estimate of η.

In addition to the uncertainty in the estimate of η, the model is subject to errors arising from inaccurately assumed values of the remaining model parameters. In the following, I focus on σ_H, which Gabaix et al. (2016) set equal to 0.15, while pointing out that $\sigma_H = 0.15$ is a conservative estimate since "the growth rates of parts of the population may be much more volatile (think of startups)." While an ideal way to account for this additional layer of uncertainty would be to estimate σ_H from the data and, then calibrate the model to a 95 percent confidence interval around the point estimate of σ_H, I leave this approach for future research. In this paper, I consider a simpler exercise, in which I assume a set of possible values of σ_H given by $\{0.15, 0.175, 0.2\}$. Note that this set is constructed in accordance with the Gabaix et al. (2016)'s presumption that the growth rate of log income among high types is likely to exceed 0.15. Then, for each value of σ_H and for each of the two data sets under consideration, I calibrate the model $B = 100$ times, each time using a different value of η randomly drawn from a 95 percent confidence interval around the η's point estimate. This approach allows me to construct a "feasible region" of the model's transition dynamics to a new steady state while incorporating uncertainties arising from data-driven errors in two out of the six model's parameters.

1.9.3 Results

In the following, I discuss two sets of results. First, I analyze the transition dynamics accounting only for the variation in the calibration target η. Then, I introduce an additional

layer of uncertainty and repeat my analysis for different values of σ_H, a parameter that governs the volatility of the growth rate of log income among high types.

1.9.3.1 Varying η

My first set of results shows that data-driven errors in estimates of key macroeconomic aggregates impact outcomes of not only empirical but also structural analysis. Errors in calibration targets are carried over through the model and come to affect all outcomes of interest, including transition dynamics and the speed of convergence to a new steady state. Therefore, I find that having precise estimates of calibration targets is critical for producing precise outcomes of structural analysis. This becomes evident when comparing transition

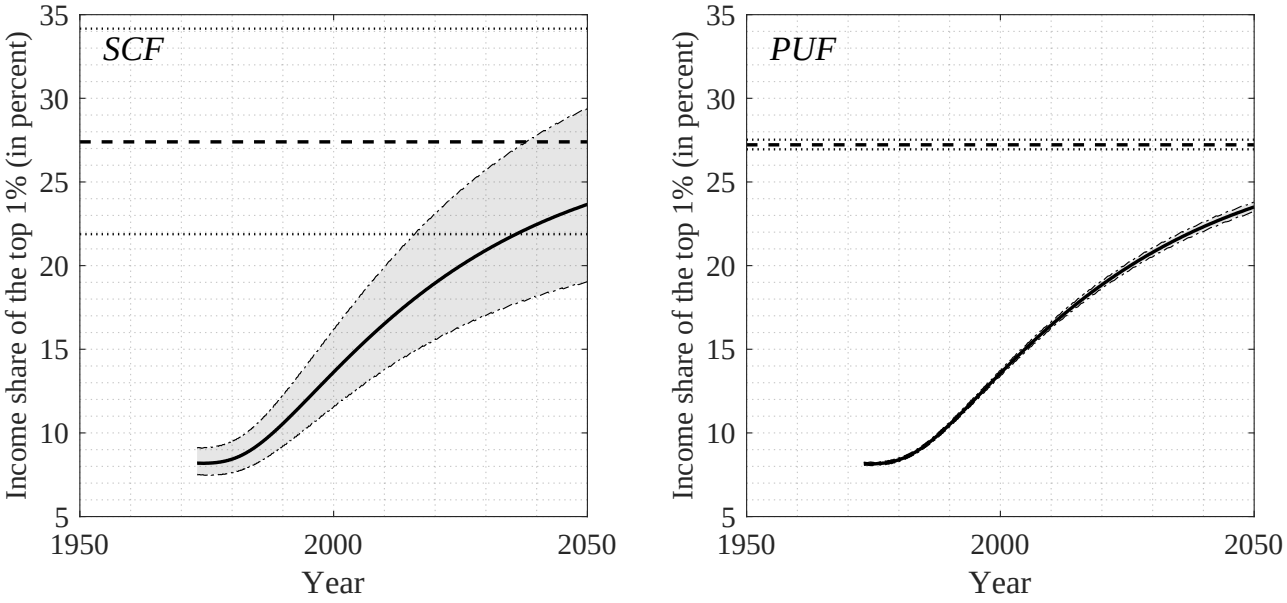

Figure 1.10: Transition dynamics, varying η

dynamics to a new steady state produced by the exact same model differing, only with respect to the level of uncertainty surrounding a single calibration target. As indicated in Figure 1.10, the PUF estimate of η has a negligible standard error, which results in a precise estimation of the model's dynamics to a new steady state. The SCF estimate of η, on the

other hand, has a sizable standard error, which results in a large 95 percent confidence envelop around the estimated transition dynamics and in a highly imprecise estimate of a new steady state.

1.9.3.2 Varying σ_H

Accounting for an additional layer of uncertainty creates an even clearer picture of the importance of conducting structural analysis using precisely estimated calibration targets. In Figure 1.11, for each of the three assumed values of σ_H, I present a 95 percent confidence envelope constructed around the estimated transition dynamics to a new steady state. By construction, varying the volatility of the growth rate of log income among high types has the same effect on the model's outputs, regardless of the data set under consideration. However, when combined with the data-set-specific uncertainty in the point estimate of η, the advantage of relying upon the PUF for the model's calibration becomes evident. Using the SCF, I arrive at projections of the income shares of the top 1 percent, ranging from 15 to 30 percent as of 2050. Since such a wide range of possible values is largely uninformative, it presents no real value to policy makers evaluating proposals targeted at combating rising inequality. On the other hand, the model's projections obtained using the PUF are much more precise. The 95 percent confidence envelope around the projected income shares varies from 20 to less than 25 percent, providing policy makers with an informative rage of values, while accounting for uncertainty in two out of the six model parameters.

1.9.3.3 COVID-19

The above analysis is conducted using data that precede COVID-19, and does not account for the impact of the ongoing pandemic on income inequality. Therefore, Figures 1.10 and 1.11 demonstrate largely outdated projections. This is the case since COVID-19 is likely to have far more long-lasting impacts on inequality than any other post-WWII recession, including the 2007–2009 Great Recession. Particularly, the forced shutdown of large parts of the US economy caused a dramatic spike in unemployment rates, especially among minorities and low-educated service workers in high-interaction jobs at restaurants, pubs, hotels,

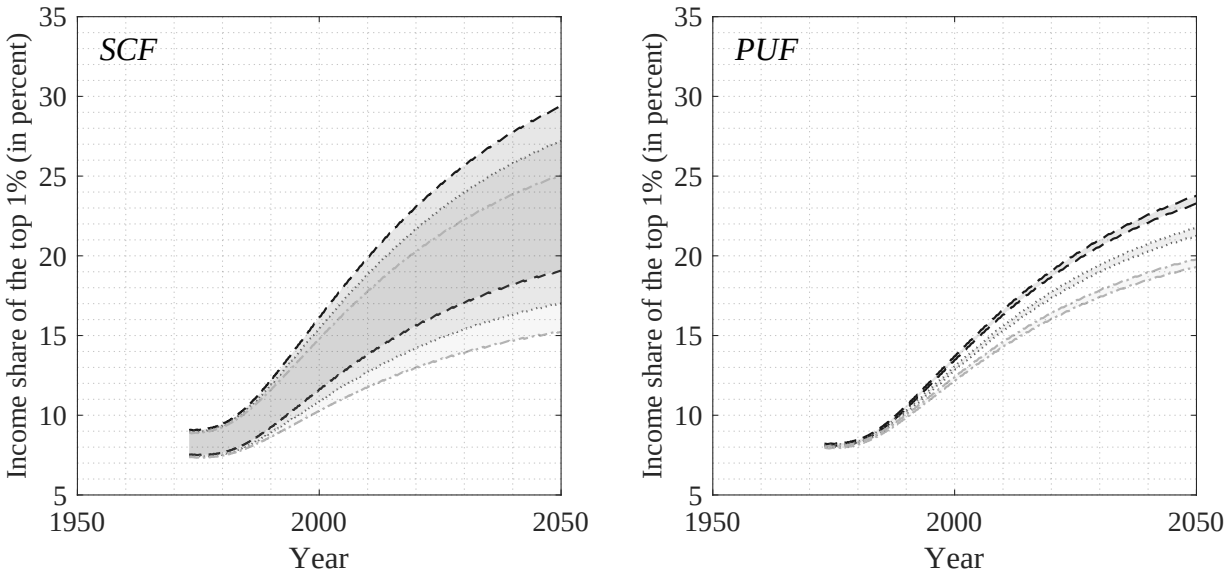

Figure 1.11: Transition dynamics, varying σ_H

and entertainment venues. Moreover, unlike other parts of the world, where governments covered employees' wages for the duration of the crises, US employees were laid off without a guarantee of being re-hired. Therefore, despite the fact that measuring impact of COVID-19 on inequality would require data that are not yet available, I find it an important topic for future research, which I briefly discuss in Section 1.10.[40]

1.10 Conclusions

This paper discusses various sources of uncertainty in studying economic inequality within and across data sets. With the focus on the six income and wealth shares within

[40]Furthermore, the current projections do not account for recent changes to the US tax policy, including the Trump's 2017 tax cut that lowered the corporate tax rate from 35 to 21 percent. Since such reductions in taxes are likely to widen the gap between rich and poor, extending the current analysis by accounting for not only the impact of COVID-19 but also the 2017 Trump's tax cut would deliver substantially more realistic projections that those currently obtained.

the top 10 percent, I investigate how sampling, nonsampling, and modeling errors affect outcomes of empirical analysis conducted using the SCF and the PUF. Regarding income inequality, I find that the PUF estimates are substantially better than the SCF estimates, and consequently, whenever possible, should be relied upon for both empirical and structural analysis. On the other hand, for top wealth inequality, neither of the two data sets can be used without caution. Regarding wealth shares of the top 10 to the top 0.5 percent, I find the SCF estimates more reliable—a result of unexpected and yet-to-be-accounted-for differences among the PUF estimates that arise from varying assumptions imposed on the underlying capitalization model. However, for the more granular wealth shares of the top 0.1 and 0.01 percent, neither of the two data sets proves credible. The PUF estimates lead to different conclusions depending on which capitalization model one applies, while the SCF estimates are unreliable as a result of the very sparse number of observations in the far right tail of wealth distribution.

In addition, using the random growth model of income from Gabaix et al. (2016), I illustrate how data-driven errors in calibration targets affect the outcomes of structural macroeconomic models. All in all, I find that in order for a structural analysis to be both conclusive and informative, it is necessary to rely upon precisely estimated calibration targets. As shown in Section 1.9, large uncertainties in the SCF estimates result in wide confidence intervals around the transition dynamics of the income shares of the top 1 percent, whereas small standard errors in the PUF estimates lead to estimates with negligible levels of uncertainty.

Regarding future research, one could extend my analysis by estimating the SCF measurement error, with the main focus on errors arising in the survey response process as described in Groves et al. (2009). Since other sources of error are either already accounted for in the present paper or shown to be fairly marginal, computing measurement error would allow estimation of an upper bound on the SCF total standard error.

Another potentially promising avenue for future research is centered around my structural exercise. Since my analysis does not account for the impact of COVID-19, the generated projections of income shares of the top 1 percent until 2050 are largely outdated. This is the case since data-driven errors in calibration targets are most likely to be of second order of importance relative to a COVID-19 shock. Since post COVID-19 data will not be

available until 2023 (the expected release date of the 2021 SCF), I plan on implementing a "hypothetical" COVID-19-recession shock in the model. Specifically, I intend to create a hypothetical scenario whereby the impact of COVID-19 is equivalent to or greater than that of the 2007–2009 Great Recession. Because of the severity of the ongoing pandemic and its uneven impact on US society, I expect the COVID-19 shock to have substantial and long-lasting repercussions on inequality, even when compared to those of the Great Recession.

Finally, since, in its present form, the PUF provides information on neither taxpayers' asset and liability holdings nor taxpayers' demographic and socio-economic characteristics, I plan to design an imputation procedure that would allow me to embed the SCF within the PUF. This modeling approach would result in a novel data set containing as many observations as in the PUF as well as a broad range of financial and nonfinanacial variables, which would be imputed from the SCF. Constructing such a data set would be advantageous for studying both income and wealth inequality. First, with such a rich data set containing numerous potential control variables, one could answer questions on income and wealth inequality that currently cannot be addressed using the SCF or the PUF individually. For illustration, consider a problem of examining income or wealth inequality among young individuals, the working-age population, and retirees. The lack of information on age in the PUF and a small number of observations in the SCF concerning various population subgroups would make conducting such an analysis virtually impossible. Second, merging the SCF and the PUF would allow to significantly improve the precision of the SCF estimates of top-decile wealth shares by increasing the number of observations while continuing to utilize information on households' asset and liability holdings reported during the SCF interview as opposed to relying upon wealth measures estimated from an assumed capitalization model.

2.0 Balanced Growth Approach to Tracking Recessions

Joint with Professor Jean-François Richard

2.1 Introduction

Dynamic Stochastic General Equilibrium (hereafter DSGE) models are generally justified on the grounds that they provide a structural foundation for policy analysis and are indeed widely used for that purpose. However, their tracking failures in times of rapid changes (such as the 2007–09 Great Recession) raise concerns relative to their relevance for policy recommendations in such times when they are most critically needed. Hence, a widely recognized need for greater diversification of the macroeconomics toolbox with models that focus on improved recession tracking performance, possibly at the cost of loosening the theoretical straitjacket of DSGE models.

In the present paper, we propose a generic procedure to transform DSGE models into hybrid versions thereof in a way that preserves their policy relevance while significantly improving their recession tracking performance. In particular, the approach we propose addresses the inherent "trade-off between theoretical and empirical coherence" (Pagan, 2003) and can be applied to a wide range of DSGE models, covering various sectors of the economy. For empirical coherence, we rely upon an Error Correction Mechanism (hereafter ECM), which has repeatedly proved highly successful in modeling agents' pursuit of moving targets represented by time-varying cointegration relationships. Simultaneously, in order to preserve theoretical coherence, we derive these targets as (moving) balanced growth solutions to the assumed model. This can be achieved without significantly weakening empirical coherence as theory models are designed to rationalize observed behavior and hence, there typically exists a close match between empirically derived cointegrating relationships and theory derived solutions.

In order to transform DSGE models into hybrid versions thereof, we implement four key modifications. First, we abandon assumptions of trend stationarity and rely instead on real (per capita) data that except for being seasonally adjusted are neither detrended nor Hodrick-

Prescott (hereafter HP) filtered. Second, instead of computing conventional DSGE solutions based upon model consistent expectations of future values, we compute balanced growth solutions based upon agents' perception of the growth scenario at any given point in time. Next, as we rely upon real data, we account for the fact that the balanced growth ratios vary significantly over time, to the extent that we are effectively treating these (theory-derived) moving targets as time-varying cointegration relationships. It follows that an appropriate subset of the model structural parameters can no longer be treated as time invariant. Instead, it is modeled as a set of state variables driven by a VAR process. Last but not least, we assume that agents rely upon an ECM process to track their moving targets.

In our approach, we draw a clear distinction between forecasting recessions and tracking them. As discussed further in our literature review, DSGE models that rely upon model consistent expectations are highly vulnerable to unexpected shocks. This is particularly critical for recessions since, as surveyed in Section 2.3, each postwar recession was triggered by a unique set of circumstances. This fundamentally prevents ex-ante econometric estimation of such potential triggers. Moreover, recession predictive failures can also extend to poor recession tracking, for the very same reason that model-based expectations are inherently slow to react to unexpected shocks.

This is where our proposed approach has its greatest potential in that balanced growth solutions can respond significantly faster to shocks impacting the agents' moving targets. In order to highlight this critical advantage we apply our hybrid methodology to a standard Real Business Cycle (hereafter RBC) model, selected for the ease of exposition since it allows for analytical derivations of the balanced growth solutions and, thereby, for a clearer presentation of the proposed methodology. By focusing on a fully ex-ante recursive analysis over a 35 year (137 quarter) validation period representing 47.9 percent of the full sample and, specifically, on narrow time windows around the last three recessions, we demonstrate that, while our model exhibits a delayed ex-ante forecasting performance similar to that of a benchmark unrestricted Vector AutoRegressive (hereafter VAR) model, it outperforms the latter in terms of recession tracking (based on commonly used metrics).

This is a remarkable result but there is more to that. The three dimensional state space we introduce for our RBC pilot model includes two structural parameters (in addition to

growth rate), that are known to have varied considerably during the postwar period. These play a central role in recession tracking in that they vary procyclically and are therefore key components to the model ability to respond quickly to unexpected shocks and, thereby, to improve its recession tracking performance. As we discuss further below, this opens numerous avenues for research on how to use these state variables as potential leading indicators as well as additional policy instruments.

Our paper is organized as follows: In Section 2.2, we provide a partial review of an extensive literature on DSGE models and related modeling issues in order to set the scene for our own proposal. In Section 2.3, we provide a brief description of the idiosyncratic causes of the 11 most recent US recessions in order to highlight the challenging environment one faces when trying to forecast economic downturns. In Section 2.4, we present a detailed generic description of the approach we propose. In Section 2.5, we provide an application to a pilot RBC model for the US postwar economy, where we detail the successive modeling steps, document an extensive recursive validation exercise, discuss modeling challenges when attempting to *ex-ante* predict the onset of the 2007–09 Great Recession, and conduct a policy experiment. Section 2.6 concludes. An Appendix to the online publication available at https://www.mdpi.com/2225-1146/8/2/14 presents a pseudo-code for the RBC application together with data description and auxiliary tracking and forecasting figures for the Great Recession. An Online Supplementary Material with additional results relative to parameter invariance and recession tracking/forecasting performance is available on https://www.martaboczon.com/.

2.2 Literature Review

DSGE models have become the workhorses of modern macroeconomics, providing a rigorous structural foundation for policy analysis. However, as recognized by a number of authors, even before the onset of the Great Recession, their high degree of theoretical coherence ("continuous and perfect optimization" Sims, 2007) produces dynamic structures that are typically too restrictive to capture the complexity of observed behavior, especially at

times of rapid changes. In order to obtain tractable solutions, DSGE models assume a stable long-run equilibrium trend path for the economy (Muellbauer, 2016), which is precisely why they often fail to encompass more densely parametrized and typically non-stationary VAR processes.[1] In this respect VAR reduced form models are more flexible and able to respond faster to large (unexpected) shocks. It is therefore, hardly surprising that there have been numerous attempts to link VAR and DSGE models, and our approach belongs to that important line of research.

Before the onset of the Great Recession, several authors had proposed innovative approaches linking VAR and DSGE models. For example, Hendry and Mizon (1993) implemented a modeling strategy starting from an unrestricted VAR and testing for cointegration relationships that would lead to a structural ECM.[2] Jusélius and Franchi (2007) translated assumptions underlying a DSGE model into testable assumptions on the long run structure of a cointegrated VAR model. Building upon an earlier contribution of Ingram and Whiteman (1994), Sims (2007) discussed the idea of combining a VAR model with a Bayesian prior distribution. Formal implementations of that concept can be found in Smets and Wouters (2005, 2007), or Del Negro and Schorfheide (2008).[3,4]

Smets and Wouters (2007) also incorporated several types of frictions and shocks into a small DSGE model of the US economy, and showed that their model is able to compete with Bayesian VAR in out-of-sample predictions. Along similar lines, Chari et al. (2007, 2009) proposed a method, labeled Business Cycle Accounting (hereafter BCA), that introduced frictions ("wedges") in a benchmark prototype model as a way of identifying classes of mechanisms through which "primitive" shocks lead to economic fluctuations. The use of wedges has since been criticized for lacking structural justification, flawed identification,

[1] A useful discussion of the inherent trade-off between theoretical and empirical coherence can be found in Pagan (2003).

[2] It follows that the ECM parsimounsly encompasses the initial VAR model. See Hendry and Richard (1982, 1989); Mizon (1984); Mizon and Richard (1986) for a discussion of the concept of encompassing and its relevance for econometric models.

[3] See also An and Schorfheide (2007) for a survey of Bayesian methods used to evaluate DSGE models and an extensive list of related references.

[4] In the present paper, we follow Pagan (2003) by using an unresricted VAR as a standard benchmark to assess the empirical relevance of our proposed model. Potential extensions to Bayesian VARs belong to future research (though imposing a DSGE-type prior density on VAR in order to improve its theoretical relevance could negatively impact its empirical performance).

and ignoring the fundamental shocks (e.g. financial) driving the wedge process (see e.g. Christiano and Davis, 2006; Romer, 2016). Nevertheless, BCA highlights a critical empirical issue—revisited in our approach—which is that structurally invariant trend stationary DSGE models are not flexible enough to accommodate rapid changes induced by unexpected shocks.

The debate about the future of DSGE models took a new urgency following their widespread tracking and forecasting failures on the occasion of the 2007–2009 Great Recession. The main emphasis has since been placed on the inherent inability of DSGE models to respond to unexpected shocks (see Caballero, 2010; Castle et al., 2010, 2016; Hendry and Mizon, 2014a,b; Hendry and Muellbauer, 2018; Stiglitz, 2018), on the recent advances and remaining challenges (see Christiano et al., 2018; Schorfheide, 2011) as well as on the need for DSGE models to share the scene with alternative approaches (see Wieland and Wolters, 2012; Blanchard, 2016; Korinek, 2017; Trichet, 2010).

Last but not least, the present paper is related to the literature on time-varying dynamic processes, and especially the emerging literature on time-varying (or locally stable) cointegrating relationships (see Bierens and Martins, 2010; Cardinali and Nason, 2010; Matteson et al., 2013). Another important reference is Canova and Pérez Forero (2015), where authors provide a generic procedure to estimate structural VAR processes with time-varying coefficients and successfully apply it to a study of the transmission of monetary policy shocks.

In conclusion of this brief literature survey, we do not intend to take side in the ongoing debate on the future of DSGE models. Instead, we propose a generic procedure to construct hybrid versions thereof with superior tracking performance in times of rapid changes (recessions and recoveries) by adopting a more flexible theoretical foundation based upon a concept of moving targets represented by time-varying cointegrating relationships. As such, we aim at offering an empirically performant complement, by no means a substitute, to DSGE models. As emphasized by Trichet (2010) "we need macroeconomic and financial models to discipline and structure our judgmental analysis. How should such models evolve? The key lesson I would draw from our experience [of the Great Recession] is the danger of relying on a single tool, methodology, or paradigm. Policymakers need to have input from various theoretical perspectives and from a range of empirical approaches. Open debate and a diversity of views must be cultivated - admittedly not always an easy task in an institution

such as a central bank. We do not need to throw out our DSGE and asset-pricing models: rather we need to develop complementary tools to improve the robustness of our overall framework."[5]

2.3 US Postwar Recessions

As discussed e.g. in Hendry and Mizon (2014a), a key issue with macroeconomic forecasting models is that of whether recessions constitute "unanticipated location shifts". More specifically, while one can generally identify indicators leading to a recession, the relevant econometric issue is that of whether or not such indicators can be incorporated ex-ante into the model and, foremost, whether or not their potential impact can be estimated prior to each recession onset, an issue we discuss further in Section 2.5.6 in the context of the Great Recession. As our initial attempt to address this fundamental issue, we provide a brief survey of the most likely causes for each of the US postwar recessions.

The 1945 recession was caused by the demobilization and the resulting transition from a wartime to a peacetime economy at the end of the Second World War. The separation of the Federal Reserve from the US Treasury is presumed to have caused the 1951 recession. The 1957 recession was likely triggered by an initial tightening of the monetary policy between 1955 and 1957, followed by its easing in 1957. Similar circumstances led to the 1960 recession. The 1969 recession was likely caused by initial attempts to close the budget deficits of the Vietnam War followed by another tightening of the monetary policy. The 1973 recession is commonly believed to originate from an unprecedented rise of 425 percent in oil prices, though many economists believe that the blame should be placed instead on the wage and

[5] A similar message was delivered by Powell (2018) in his swearing-in ceremony as the new Chair of the Federal Reserve: "The success of our institution is really the result of the way all of us carry out our responsibilities. We approach every issue through a rigorous evaluation of the facts, theory, empirical analysis and relevant research. We consider a range of external and internal views; our unique institutional structure, with a Board of Governors in Washington and 12 Reserve Banks around the country, ensures that we will have a diversity of perspectives at all times. We explain our actions to the public. We listen to feedback and give serious consideration to the possibility that we might be getting something wrong. There is great value in having thoughtful, well-informed critics." (see https://www.federalreserve.gov/newsevents/speech/powell20180213a.htm for the complete speech given during the ceremonial swearing-in on February 13, 2018.)

price control policies of 1971 that effectively prevented the economy from adjusting to market forces. The main reason for the double dip recession of the 1980s is believed to be an ill-timed Fed monetary policy aimed at reducing inflation. Large increases in federal funds rates achieved that objective but also led to a significant slowdown of the economic activity. There are several competing explanations for the 1990 recession. One was another rise of the federal funds rates to control inflation. The oil price shock following the Iraqi invasion of Kuwait and the uncertainties surrounding the crisis were likely contributing factors. Solvency problems in the savings and loan sector have also been blamed. The 2001 recession is believed to have been triggered by the collapse of the dot-com bubble. Last but not least, the Great Recession was caused by a global financial crisis in combination with the collapse of the housing bubble.

In summary, each postwar recession was triggered by idiosyncratic sets of circumstances, including but not limited to ill-timed monetary policies, oils shocks to aggregate demand and supply, and financial and housing crises. As we discuss further in Section 2.5.6 in the context of the Great Recession, such a variety of unique triggers makes its largely impossible to econometrically estimate their potential impact prior to the actual onset of each recession, a conclusion that supports the words of Trichet (2010), as quoted in Section 2.2.

A natural question is that of what will trigger the next US recession. In an interview given in May, 2019 Joseph Stiglitz emphasized political instability and economic stagnation in Europe, uneven growth in China, and President Trump's protectionism as the three main potential triggers. Alternatively, Robert Schiller, also interviewed in spring 2019, focused on growing polarization around President Trump's presidency and unforeseen consequences of the ongoing impeachment hearings.[6] He also emphasized that "recessions are hard to predict until they are upon you. Remember, we are trying to predict human behavior and humans thrive on surprising us, surprising each other." Similar concerns were recently expressed by Rogoff (2019): "To be sure, if the next crisis is exactly like the last one, any policymaker can simply follow the playbook created in 2008, and the response will be at least as effective. But what if the next crisis is completely different, resulting from say, a severe cyberattack, or an

[6]For more details see https://www.youtube.com/watch?v=lyzS7Vp5vaY (Stiglitz's interview) and https://www.youtube.com/watch?v=rUYk2DA8PH8 (Schiller's interview).

unexpectedly rapid rise in global real interest rates, which rocks fragile markets for high-risk debt?" Unfortunately, these concerns have turned out to be prescient with the dramatic and unexpected onset of COVID-19, which has triggered an unfolding deep worldwide recession that is creating unheard of challenges for policymakers. To conclude, the very fact that each recession is triggered by an idiosyncratic set of circumstances is the fundamental econometric reason why macroeconomic models will typically fail to ex-ante predict recession onset.

2.4 Hybrid Tracking Models

The transformation of a DSGE model into a hybrid version thereof relies upon four key modifications, which we first describe in generic terms before turning to specific implementation details in Section 2.4.2.

2.4.1 Key Features

Since our focus lies on tracking macroeconomic aggregates with emphasis on times of rapid changes (recessions and recoveries), we rely upon real seasonally adjusted per capita series that are neither detrended nor (HP) filtered.[7,8,9] Such data are non-stationary, which is precisely why they are frequently detrended and/or (HP) filtered in order to accommodate DSGE trend stationarity assumptions. In fact, the non-stationarity of the data allows us to anchor our methodology around the concept of cointegration, which has been shown in the literature to be a "powerful tool for robustifying inference" (Jusélius and Franchi, 2007).

[7]By doing so, we avoid producing "series with spurious dynamic relations that have no basis in the underlying data-generating process" (Hamilton, 2018) as well as "mistaken influences about the strength and dynamic patterns of relationships" (Wallis, 1974).

[8]There is no evidence that seasonality plays a determinant role in recessions and recoveries. Therefore, without loss of generality we rely upon seasonally adjusted data, instead of substantially increasing the number of model parameters by inserting quarterly dummies, potentially in every equation of the state VAR and/or ECM processes.

[9]NBER recession dating is based upon GDP growth, not per capita GDP growth. However, our objective is not that of dating recessions, for which there exists an extensive and expanding literature. Instead, our objective is that of tracking macroeconomic aggregates at times of rapid changes, and for that purpose per capita data can be used without loss of generality. Note that if needed per capita projections can be ex-post back-transformed into global projections.

Furthermore, instead of deriving DSGE intertemporal solutions based upon model consistent expectations of future values, we solve the model for (hypothetical) balanced growth ratios (hereafter great ratios) based upon the agents' current perception of a tentative growth scenario. We justify this modeling decision by noticing that such great ratios provide more obvious reference points for agents in an environment where statistics such as current and anticipated growth rates, saving ratios, interest rates and so on are widely accessible and easily comprehensible. Importantly, these cointegrating relationships are theory derived (thereby preserving theoretical coherence) rather than data derived as in some of the references cited in Section 2.2 (see Hendry and Mizon, 1993; Jusélius and Franchi, 2007).

Next, we introduce a vector of state variables to model the long term movements of the great ratios.[10] However, instead of introducing hard to identify frictions (wedges under the BCA) in the model equations, we allow for an appropriate subset of key structural parameters to vary over time and as such treat them as dynamic state variables (together with the benchmark growth rate). For example, with reference to our pilot RBC model, there is clear and well documented evidence that neither the capital share of output in the production function nor the consumers preference for consumption relative to leisure have remained constant between 1948 and 2019, a time period that witnessed extraordinary technological advances and major changes in lifestyle and consumption patterns. (See Section 2.5.1 for references.) Thus, our goal is that of selecting a subset of structural parameters to be treated as state variables and producing a state VAR process consistent with the long term trajectories of the great ratios.[11] Effectively, as mentioned above, this amounts to treating the great ratios as (theory derived) time-varying cointegrating relationships. It is also the key step toward improving the tracking performance of our hybrid models in times of rapid changes.

[10] Since we rely upon real data, it is apparent that the great ratios vary considerably over time. Most importantly, their long term dynamics appear to be largely synchronized with business cycles providing a solid basis for our main objective of tracking recessions.

[11] It is sometimes argued that in order to be interpreted as structural and/or to be instrumental for policy analysis, a parameter needs to be time invariant. We find such a narrow definition to be unnecessarily restrictive and often counterproductive. The very fact that some key structural parameters are found to vary over time in ways that are linked to the business cycles and can be inferred from a state VAR process paves the way for policy interventions on these variables, which might not be available under the more restricted interpretation of structural parameters. An example is provided in Section 2.5.7.

The final key feature of our hybrid approach consists of modeling how economic agents respond to the movements of their target great ratios. In state space terminology, this objective can be stated as that of producing a measurement process for the state variables. Specifically, we propose an ECM measurement process for the log differences of the relevant macroeconomic aggregates as a function of their lagged log differences, lagged differences of the state variables and, foremost, the lagged differences between the actual great ratios and their moving (balanced growth) target values.

2.4.2 Implementation Details

Next, we discuss the implementation details of our hybrid approach: description of the core model, specification of the VAR and ECM processes, estimation, calibration and validation.

2.4.2.1 Core Model

The core model specifies the components of a balanced growth optimization problem, which are essentially objective functions and accounting equations. While it can rely upon equations derived from a baseline DSGE model, it solves a different (and generally easier) optimization problem. Instead of computing trend stationary solutions under model consistent expectations of future values, it assumes that at time t, agents compute tentative balanced growth solutions based on their current perception of the growth scenario s_t they are facing. The vector s_t includes a tentative balanced growth rate g_t but also, as we discuss further below, additional state variables characterizing the target scenario at time t. Therefore, we are effectively assuming that agents are chasing a moving target.

Period t solutions to the agents' optimization problem produce two complementary sets of first order conditions. The first set consists of great ratios between the decision variables, subsequently re-interpreted as (theory derived) moving cointegrated targets. The second set provides laws of motion for the individual variables that would guarantee convergence towards a balanced growth equilibrium under a hypothetical scenario, whereby s_t would remain constant over time.

Using the superscript "*" to denote model solutions (as opposed to actual data), the two sets of first order conditions are denoted as:

$$\begin{pmatrix} r_t^* \\ \Delta x_t^* \end{pmatrix} = \begin{pmatrix} h_1(s_t;\lambda) \\ h_2(s_t, s_{t-1};\lambda) \end{pmatrix}, \tag{30}$$

where $r_t^* \in \mathbf{R}^p$ denotes great ratios, $\Delta x_t^* \in \mathbf{R}^{p+1}$ laws of motion for individual variables, λ a vector of time invariant parameters, and $s_t \in \mathbf{R}^q$ a state vector yet to be determined.[12]

Insofar as our approach assumes that agents aim at tracking the moving targets r_t^* through an ECM process, theory consistency implies that the long term movements of $(r_t, \Delta x_t)$ should track those of $(r_t^*, \Delta x_t^*)$ with time lags depending upon the implicit ECM adjustment costs. However, as we use data that are neither detrended nor (HP) filtered, it is apparent that $(r_t, \Delta x_t)$ vary considerably over time especially at critical junctures such as recessions and recoveries. It follows that we cannot meaningfully assume that the movements of $(r_t^*, \Delta x_t^*)$ are solely driven by variations in the tentative growth rate g_t. Therefore, we shall treat an appropriate subset of structural parameters as additional state variables and include them in s_t (together with g_t) instead of λ. The selection of such a subset is to be based on a combination of factors such as documented evidence, calibration of time invariant parameters, ex-post model validation and, foremost, recession tracking performance.

In Sections 2.4.2.2 and 2.4.2.3 below we describe the modelisation of the VAR and ECM processes for a given value of λ over an arbitrary time interval. Next, in Section 2.4.2.4 we introduce a recursive estimation procedure that will be used for model validation and calibration of λ.

2.4.2.2 The State VAR Process

The combination of a VAR process for s_t and an ECM process for Δx_t constitutes a dynamic state space model with non-linear Gaussian measurement equations. One could attempt to estimate such a model applying a Kalman filter to local period-by-period linearizations. In fact, we did so for the RBC model described in Section 2.5, but it exhibited

[12] Potential exogenous variables are omitted for the ease of notation.

inferior tracking performance relative to that of the benchmark VAR (to be introduced further below). Therefore, we decided to rely upon an alternative estimation approach whereby for any tentative value of the time-invariant structural parameters in λ, we first construct trajectories for s_t that provide the best fit for the first order conditions in equation (30).[13] Specifically, for any given value of λ (to be subsequently calibrated) we compute sequential point estimates for $\{s_t\}_{t=1}^{T}$ as follows

$$\hat{s}_t(\lambda) = \underset{s_t}{\operatorname{argmin}} ||\epsilon_t(s_t, \hat{s}_{t-1}(\lambda); \lambda)||_2, \quad t: 1 \to T \tag{31}$$

under an appropriate Euclidean L-2 norm and where $\epsilon_t(\cdot)$ denotes the differences $r_t - h_1(s_t; \lambda)$ and $\Delta x_t - h_2(s_t, s_{t-1}; \lambda)$ in equation (30).

Following estimation of s_t, we specify a state VAR(l) process for $\hat{s}_t(\lambda)$, say

$$\hat{s}_t(\lambda) = A_0 + \sum_{i=1}^{l} A_i \hat{s}_{t-i}(\lambda) + u_t, \tag{32}$$

where $u_t \sim \mathcal{IN}(0, \Sigma_A)$.

For example, for the RBC application described below, we ended selecting a VAR process of order $l = 2$.

2.4.2.3 The ECM Measurement Process

The ECM process is to be constructed in such a way that the economy would converge to the balanced growth equilibrium $h_1(s_*; \lambda)$ in a hypothetical scenario whereby $\hat{s}_t(\lambda)$ would remain equal to s_* over an extended period of time. In such a case Δx_t would naturally converge towards $h_2(s_*, s_*; \lambda)$ as the latter represents the law of motion that supports the s_* balanced growth equilibrium. In order to satisfy this theoretical convergence property, we specify the ECM as

$$\Delta x_t^o(\lambda) = D_0 + D_1 \Delta \hat{s}_t(\lambda) + D_2 \Delta x_{t-1}^o(\lambda) - D_3 r_{t-1}^o(\lambda) + v_t, \tag{33}$$

where

$$\Delta x_t^o(\lambda) = \Delta x_t - h_2(\hat{s}_t(\lambda), \hat{s}_{t-1}(\lambda); \lambda) \tag{34}$$

[13] It is also meant to be parsimonious in the sense that the number of state variables in s_t has to be less that the number of equations.

$$r^o_{t-1}(\lambda) = r_{t-1} - h_1(\hat{s}_{t-1}(\lambda);\lambda) \tag{35}$$

and $v_t \sim \mathcal{IN}(0, \Sigma_D)$.

Note that additional regressors could be added as needed as long as they would converge to zero in equilibrium. With $D_0 = 0$, the ECM specification in (33) fully preserves the theoretical consistency of the proposed model. Note that in practice, omitted variables, measurement errors, and other mis-specifications could produce non-zero estimates of D_0, as in our RBC pilot application, where we find marginally small though statistically significant estimates for D_0.

2.4.2.4 Recursive Estimation, Calibration, and Model Validation

The remaining critical step of our modeling approach is the calibration of λ, based on two criteria: parameter invariance and recession tracking performance. In order to achieve that twofold objective (to be compared to that of an unrestricted benchmark VAR process for x_t)[14] we rely upon a fully recursive implementation over an extended validation period (35 years for the RBC model). This recursive implementation, which we describe further below, is itself conditional on λ and is to be repeated as needed in order to produce an "optimal" value of λ, according to the aforementioned calibration criteria.

The recursive implementation proceeds as follows. Let T denote the actual sample size and define a validation period $[T_a, T]$, with $T_a \ll T$. First, for any $T_* \in [T_a, T]$, and conditionally on λ, we use only data from $t = 1$ to $t = T_*$ to compute the sequence $\{\hat{s}_t(T_*;\lambda)\}_{t=1}^{T_*}$. Then, using $\{\hat{s}_t(T_*;\lambda)\}_{t=1}^{T_*}$ we estimate the VAR and ECM processes, again using only data from $t = 1$ to $t = T_*$. Finally, based on these estimates, we compute (tracking) fitted values for $(\hat{s}_{T_*}(T_*;\lambda), \hat{x}_{T_*}(T_*;\lambda))$ as well as 1- to 3-step ahead *out-of-sample* forecasts for $\{\hat{s}_t(t;\lambda), \hat{x}_t(t;\lambda)\}_{t=T_*+1}^{T_*+3}$.

After storing the full sequence $(T_* : T_a \to T)$ of recursive estimates, fitted values and out-of-sample forecasts we repeat the *entire* recursive validation exercise for alternative values of λ in order to select an "optimal" value depending upon an appropriate mix of formal

[14] The benchmark VAR process for Δx_t is given by $\Delta x_t = Q_0 + Q_1 \Delta x_{t-1} + Q_2 x_{t-2} + w_t$

and informal calibration criteria. Specific criteria for the pilot RBC model are discussed in Section 2.5.4.

It is important to reiterate that while recursive step T_* (given λ) only relies on data up to T_*, the calibrated value of λ effectively depends on the very latest deseasonalized data set available at the time T (2019Q2 for our RBC application). Clearly, due to revisions and updates following T_*, these data are likely to be more accurate than those that were available at T_*. Moreover, pending further additions and revisions, they are the ones to be used to track the next recession. In other words, our calibrated value of λ might differ from the ones that would have been produced if our model had been used in the past to ex-ante track earlier recessions. But what matters is that a value of λ calibrated using the most recent data in order to assess past recursive performance is also the one which is most likely to provide optimal tracking performance on the occasion of the next recession.

2.5 Pilot Application to the RBC Model

2.5.1 Model Specification

In order to test both the feasibility and recession tracking performance of our approach, we reconsider a baseline RBC model taken from Rubio-Ramírez and Fernández-Villaverde (2005) and subsequently re-estimated by DeJong et al. (2013) as a conventional DSGE model, using HP filtered per capita data.

The model consists of a representative household that maximizes a discounted lifetime utility flow from consumption c_t and leisure l_t. The core balanced growth solution solves the following optimization problem

$$\max_{\{c_t,n_t,k_{t+1}\}_{t=0}^{\infty}} \sum_{t=0}^{\infty} \beta^t \frac{\left(c_t^{\varphi} l_t^{1-\varphi}\right)^{1-\phi} - 1}{1-\phi} \tag{36}$$

subject to a Cobb-Douglas production function

$$y_t = k_t^{\alpha} \left(n_t z_t\right)^{1-\alpha} = n_t z_t \left(\frac{k_t}{n_t z_t}\right)^{\alpha}, \quad \text{with} \;\; \Delta \ln z_t = g \tag{37}$$

and accounting identities

$$n_t = 1 - l_t, \quad k_{t+1} = (y_t - c_t) + \delta k_t, \tag{38}$$

where y_t, c_t, k_t denote real per capita (unfiltered) seasonally adjusted quarterly output, consumption, and capital, n_t per capita weekly hours as a fraction of discretionary time[15], and z_t latent stochastic productivity. α denotes the capital share of output, β the household discount rate, φ the relative importance of consumption versus leisure, ϕ the degree of relative risk aversion, and $1 - \delta$ the depreciation rate of capital.

For the subsequent ease of notation, we transform φ into

$$d = \ln\left(\frac{1-\varphi}{\varphi}\right), \quad \text{with} \quad \frac{\mathrm{d}d}{\mathrm{d}\varphi} = -\frac{1}{\varphi\,(1-\varphi)} < 0 \text{ on } [0,1]\,. \tag{39}$$

With reference to equation (30) the two sets of first order conditions are a two-dimensional vector of great ratios (or any linear combination thereof)

$$r_t^* = \left(\ln\left(\frac{y_t^*}{c_t^*}\right), \ln\left(\frac{1-n_t^*}{n_t^*}\right)\right)' \tag{40}$$

and a three-dimensional vector of laws of motion

$$\Delta x_t^* = \left(\Delta \ln\left(\frac{y_t^*}{n_t^*}\right), \Delta \ln\left(\frac{c_t^*}{n_t^*}\right), \Delta \ln\left(\frac{1-n_t^*}{n_t^*}\right)\right)'. \tag{41}$$

Note that under a hypothetical scenario, whereby s_t (to be defined below) would remain constant over time, all five components in $(r_t^*, \Delta x_t)$ would be constant with r_t^* being a function of $\xi = (g, d, \alpha, \beta, \delta, \phi)'$ and $\Delta x_t^* = (g, g, 0)'$.[16]

Next, we provide graphical illustrations of the sample values of Δx_t and r_t from 1948Q1 to 2019Q2 in Figures 2.1 and 2.2. It is immediately obvious that the components of r_t, and to a lesser extent, those of Δx_t are far from being constant over the postwar period. It follows that, since $(r_t, \Delta x_t)$ are assumed to be tracking their theoretical counterparts through an ECM process, $(r_t^*, \Delta x_t^*)$ must themselves have changed considerably over time. Moreover, the pattern of the observed variations suggests that $(r_t^*, \Delta x_t^*)$ cannot be a function of the sole state variable g_t. Therefore, we need to consider additional state variables (no more

[15] See Appendix A to the online publication for the full description of the data.
[16] See also DeJong and Dave (2011, Section 5.1.2).

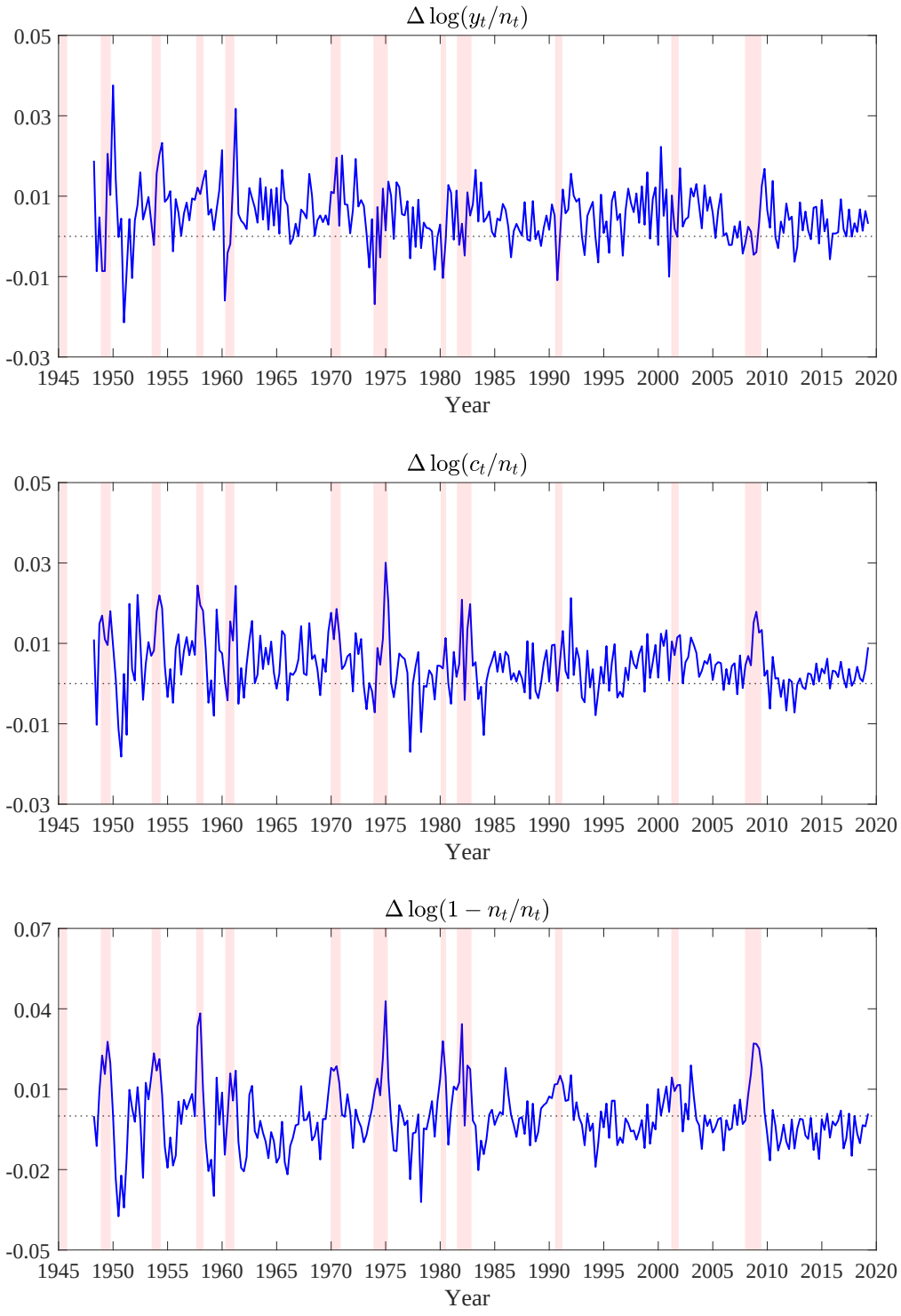

Figure 2.1: Laws of motion for individual variables. Shaded regions correspond to NBER recession dates

than three in order to avoid overfitting). The two natural candidates are α and d, as there exists evidence that neither of them has been constant over the postwar period.[17] Supporting evidence that the share of capital α has been steadily increasing (with cyclical variations) over the postwar period is highlighted in the following quote from Giandrea and Sprague (2017): "In the late 20th century—after many decades of relative stability—the labor share began to decline in the United States and many other economically advanced nations, and in the early 21st century it fell to unprecedented lows."

Similarly, our estimated state trajectory for φ (relative preference for consumption versus leisure), as illustrated in Figure 2.4 below, is broadly comparable with analyses of hours worked in the US. Specifically, Juster and Stafford (1991) document a reduction in hours per week between 1965 and 1981, whereas Jones and Wilson (2018) report a modest increase between 1979 and 2016, resulting from increased women labor participation.

Hence, we adopt the following partition

$$s_t = (g_t, d_t, \alpha_t)' \quad \text{and} \quad \lambda = (\beta, \delta, \phi)', \tag{42}$$

notwithstanding the fact that it will be ex-post fully validated by the model invariance and recession tracking performance.

According to the partition in (42) the great ratios are then given by[18]

$$r_t^* = \begin{pmatrix} \ln\left(\frac{y_t^*}{c_t^*}\right) \\ \\ \ln\left(\frac{y_t^*}{c_t^*} \cdot \frac{1-n_t^*}{n_t^*}\right) \end{pmatrix} = \begin{pmatrix} \ln\left(\frac{p(s_t;\lambda)}{q(s_y;\lambda)}\right) \\ \\ d_t - \ln\left(1-\alpha_t\right) \end{pmatrix} = h_1\left(s_t;\lambda\right) \tag{43}$$

with

$$\frac{y_t^*}{k_t^*} = p\left(s_t;\lambda\right) = \frac{1}{\alpha_t}\left[\frac{1}{\beta}\exp\left(g_t\left(1+\frac{\phi+e^{d_t}}{1+e^{d_t}}\right)\right)-\delta\right] \tag{44}$$

and

$$\frac{c_t^*}{k_t^*} = q\left(s_t;\lambda\right) = p\left(s_t;\lambda\right) - \left[e^{g_t}-\delta\right]. \tag{45}$$

[17] The risk aversion parameter ϕ could also be considered, except for the fact that it is loosely identified to the extent that letting ϕ vary over time serves no useful purpose, and worse, can negatively impact the subsequent recursive invariance of the model.

[18] For the ease of interpreation, the seond component of r_t^* is redefined as the sum of the original great ratios in equation (40).

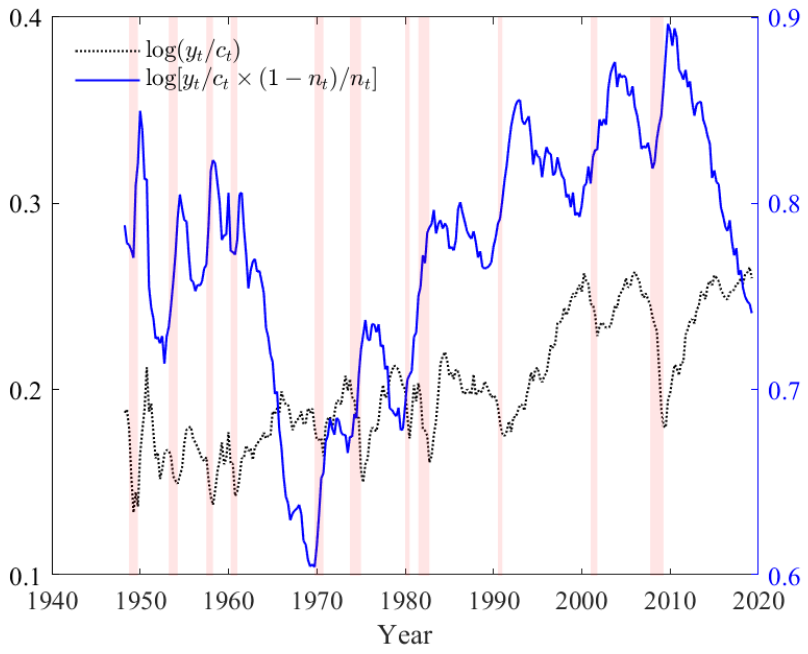

Figure 2.2: Balanced growth ratios. The dotted line's vertical axis is on the left and that of the solid line's on the right. Shaded regions correspond to NBER recession dates

As discussed earlier in Section 2.4.1 the great ratios in equation (43) initially represent theory derived time-varying cointegration relationships before being transformed into empirically relevant relationships with the introduction of the state vector s_t.[19]

As for the law of motions Δx_t^*, it follows from equations (37) and (44)-(45) that

$$\frac{y_t^*}{n_t^* z_t} = \left[p\left(s_t;\lambda\right)\right]^{\frac{\alpha_t}{\alpha_t - 1}} \quad \text{and} \quad \frac{c_t^*}{n_t^* z_t} = q\left(s_t;\lambda\right) \times \left[p\left(s_t;\lambda\right)\right]^{\frac{1}{\alpha_t - 1}}.$$

Next, taking log differences in order to eliminate z_t, we obtain

$$\Delta \ln\left(\frac{y_t^*}{n_t^*}\right) = g_t + \Delta\left(\frac{\alpha_t}{\alpha_t - 1}\ln\left(p\left(s_t;\lambda\right)\right)\right) \tag{46}$$

[19]It follows that standard cointegration rank tests are not applicable in this context. Bierens and Martins (2010) propose a vector ECM likelihood ratio test for time-invariant cointegration against time-varying cointegration. However, it is not applicable as such to our two stage model and, foremost, Figure 2.2 offers clear empirical evidence in favor of time-varying cointegration.

and

$$\Delta \ln \left(\frac{c_t^*}{n_t^*} \right) = g_t + \Delta \left(\frac{1}{\alpha_t - 1} \ln \left(p \left(s_t; \lambda \right) \right) + \ln \left(q \left(s_t; \lambda \right) \right) \right) \tag{47}$$

and, by differentiating the second great ratio in r_t^* in equation (43), we arrive at

$$\Delta \ln \left(\frac{1 - n_t^*}{n_t^*} \right) = \Delta \left(d_t - \ln \left(1 - \alpha_t \right) - \ln \left(\frac{p \left(s_t; \lambda \right)}{q \left(s_t; \lambda \right)} \right) \right), \tag{48}$$

which completes the derivation of the (moving) balanced growth solutions.

Before we proceed to the next section where we discuss recursive estimation of the model, note that the steps that follow are also conditional on a tentative value of λ, to be calibrated ex-post based upon the recursive validation exercise.

2.5.2 Recursive Estimation (Conditional on λ)

Recursive estimation over the validation period $[T_a, T]$, where T_a = 1985Q2 and T = 2019Q2 proceeds as described in Section 2.4.2.4 and consists of three steps, to be repeated for any tentative value of λ, and for all successive values of $T_* \in [T_a, T]$.

We start the recursive estimation exercise by computing recursive estimates for the state trajectories $\{\hat{s}_t \left(T_*; \lambda \right)\}_{t=1}^{T_*}$. First, we estimate $\{g_t\}_{t=1}^{T_*}$ as the (recursive) principal component of $\Delta \ln \left(\frac{y_t}{n_t} \right)$ and $\Delta \ln \left(\frac{c_t}{n_t} \right)$ for $t : 1 \rightarrow T_*$, where this particular choice guarantees consistency with the theory interpretation of g_t from equations (37), (46), and (47). Next, given estimates of $\{g_t\}_{t=1}^{T_*}$, we rely upon equation (31) in order to compute estimates of $\{(\alpha_t, d_t)\}_{t=1}^{T_*}$. Therefore, as shown below in equation (49), the resulting estimates of $\{g_t\}_{t=1}^{T_*}$ depend solely on T_*, whereas those of $\{(\alpha_t, d_t)\}_{t=1}^{T_*}$ depend on $T_*; \lambda$, and $\{g_t\}_{t=1}^{T_*}$, say

$$\{\hat{s}_t \left(T_*; \lambda \right)\}_{t=1}^{T_*} = \left\{ \hat{g}_t \left(T_* \right), \hat{d}_t \left(\hat{g}_t \left(T_* \right), T_*; \lambda \right), \hat{\alpha}_t \left(\hat{g}_t \left(T_* \right), T_*; \lambda \right) \right\}_{t=1}^{T_*}, \tag{49}$$

for all T_* in $[T_a, T]$. For illustration, we plot the trajectories of $\left\{ \hat{s}_t \left(T, \hat{\lambda} \right) \right\}_{t=1}^{T}$ for $T_* = T$ and $\lambda = \hat{\lambda}$ as dotted lines in Figures 2.3, 2.4, and 2.5, where $\hat{\lambda}$ denotes the calibrated value of λ, as given further below in equation (56). The trajectory of $\{\hat{g}_t \left(T \right)\}_{t=1}^{T}$ highlights a critical feature of the data, which is that $\hat{g}_t \left(T \right)$ typically increases during recessions, and especially during the Great Recession. This apparently surprising feature of the data follows

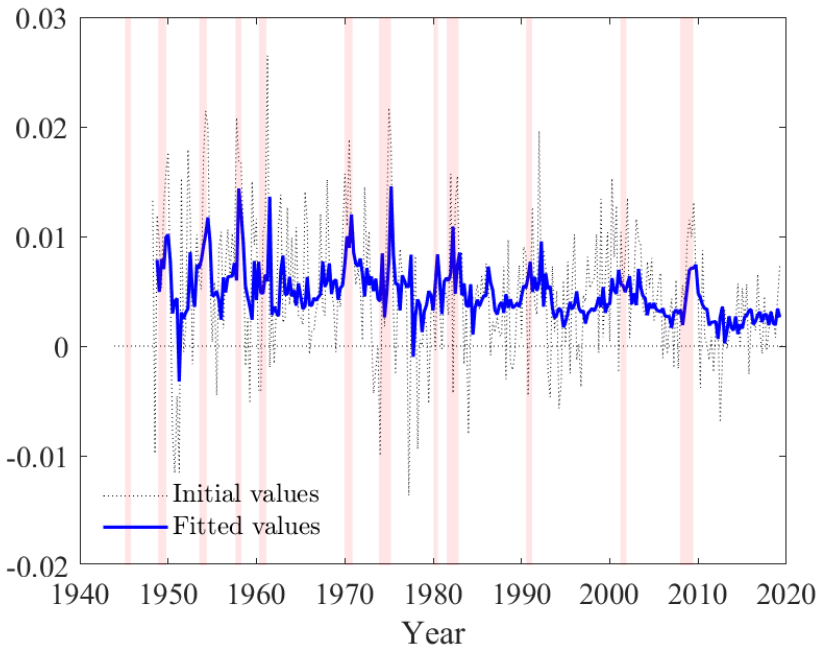

Figure 2.3: Estimated trajectory of state variable g_t. Fitted values result from an unrestricted SURE estimation of the state VAR model. Shaded regions correspond to NBER recession dates

from our theory consistent definition of $\hat{g}_T(T)$. Recall that in accordance with equations (46) and (47), $\hat{g}_T(T)$ is computed as the principal component of $\left(\Delta \ln\left(\frac{y_t}{n_t}\right), \Delta \ln\left(\frac{c_t}{n_t}\right)\right)$, rather than that of $(\Delta \ln y_t, \Delta \ln c_t)$. Hence, the increase of $\hat{g}_t(T)$ during recessions reflects a common feature of the data which is that n_t typically decreases faster than y_t and c_t during economic downturns. Importantly, this particular behavior proves to be a critical component of the parameter invariance and recession tracking performance of our model, which both outperform those produced when relying instead on the principal component of $(\Delta \ln y_t, \Delta \ln c_t)$, which is only theory consistent in equilibrium.

Once we have the trajectories of $\{\hat{s}_t(T_*;\lambda)\}_{t=1}^{T_*}$, our next step is the recursive estimation of the 48 VAR-ECM parameters $\theta = (\theta_{\text{VAR}}, \theta_{\text{ECM}})$, where $\theta_{\text{VAR}} = (A_0, A_1, A_2)$ and $\theta_{\text{ECM}} = (D_0, D_1, D_2, D_3)$. The outcome of this recursive exercise (conditional on λ) is a full set of

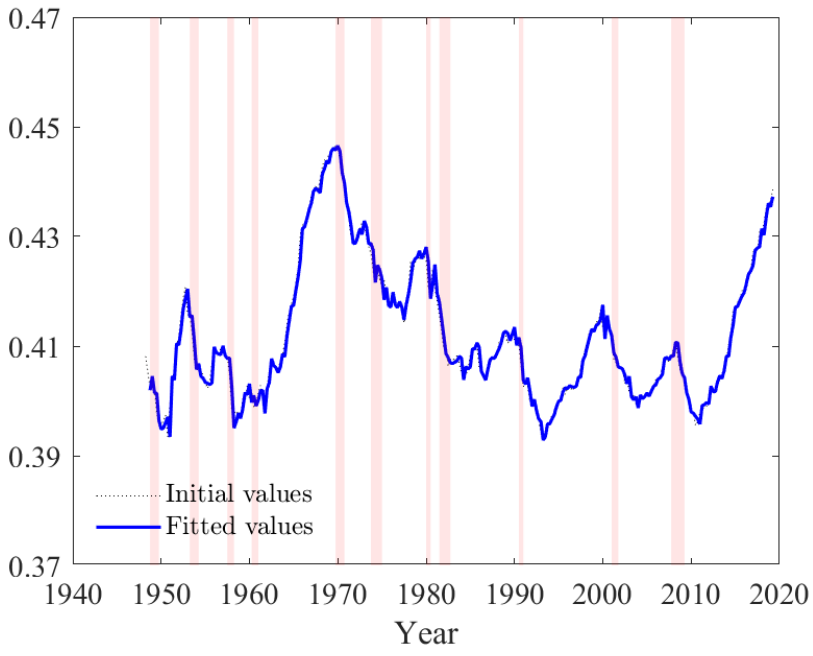

Figure 2.4: Estimated trajectory of state variable $\varphi_t = (\exp(d_t) + 1)^{-1}$. Fitted values result from an unrestricted SURE estimation of the state VAR model. Shaded regions correspond to NBER recession dates

recursive estimates given by

$$\left\{\hat{\theta}\left(T_*;\lambda\right)\right\}_{T_*=T_a}^{T} = \left\{\hat{\theta}_{\text{VAR}}\left(T_*;\lambda\right), \hat{\theta}_{\text{ECM}}\left(T_*;\lambda\right)\right\}_{T_*=T_a}^{T}. \tag{50}$$

Recursive estimates $\hat{\theta}_{\text{VAR}}\left(T_*;\lambda\right)$ for $T_* : T_a \to T$ are obtained using unrestricted OLS, where we find that conditionally on the subsequently calibrated value of $\hat{\lambda}$, the recursive estimates of θ_{VAR} are time invariant and statistically significant (though borderline for the intercept). The OLS estimates for $T_* = T$ and $\lambda = \hat{\lambda}$ are presented in Table 2.1 below, whereas the full set of recursive estimates $\left\{\hat{\theta}_{\text{VAR}}\left(T_*,\hat{\lambda}\right)\right\}_{T_*=T_a}^{T}$ is illustrated in Figure 2 of the Online Supplementary Material.

Recursive OLS estimation of the 27 ECM coefficients in $\theta_{\text{ECM}}\left(T_*;\hat{\lambda}\right)$ initially produced a majority (19 out of 27) of insignificant coefficients. However, eliminations of insignificant variables has to be assessed not only on the basis of standard test statistics but, also and

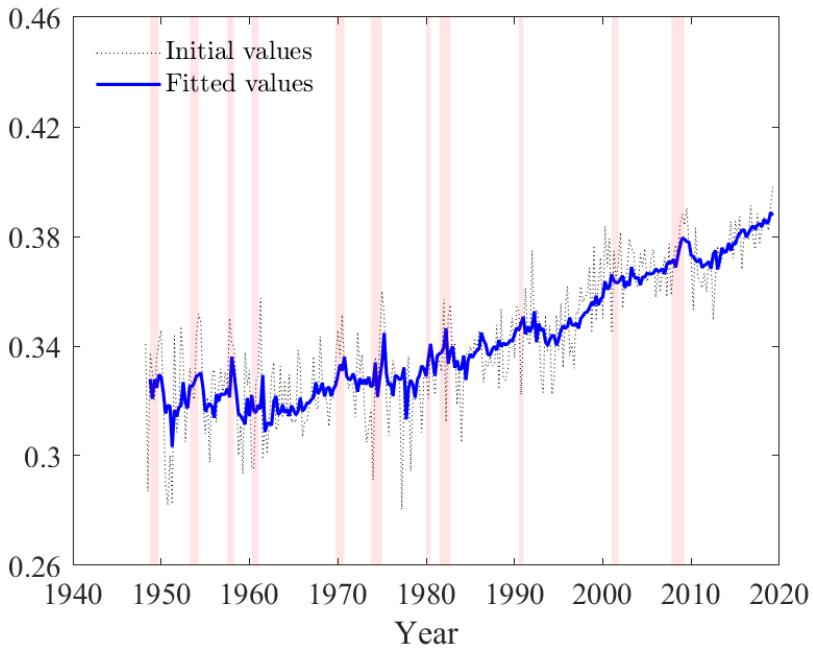

Figure 2.5: Estimated trajectory of state variable α_t. Fitted values result from an unrestricted SURE estimation of the state VAR model. Shaded regions correspond to NBER recession dates

foremost, on the basis of recursive parameter invariance and recession tracking performance. Hence, we decided to rely upon a sequential system elimination procedure, whereby we sequentially eliminate variables that are insignificant in all three equations, while continuously monitoring the recursive performance of the model.[20] This streamline procedure led first to the elimination of the ECM term associated with the great ratio $\ln\left(\frac{y_t}{c_t}\cdot\frac{1-n_t}{n_t}\right)$ followed by $\Delta\hat{d}_t$.[21] At this stage we were left with seven insignificant coefficients at the two-sided t-test (only three insignificant coefficients at the one-sided t-test). Since the remaining estimates

[20] Individual elimination would be undermined by the fact that the estimated residual covariance matrix $\hat{\Sigma}_D$ is ill-conditioned with condition numbers of the order of 2.4×10^5, which raises concerns about the validity of asymptotic critical values for system test statistics. One advantage of the sequential system elimination is that we can rely upon standard single equation t- and F-test statistics.

[21] Both eliminations appear to be meaningful. First, equilibrium adjustments in n_t are undoubtedly impeded by factors beyond agents control. Second, the elimination of $\Delta\hat{d}_t$ is likely driven by the fact that the quarterly variations of $\hat{d}_t$ are too small to have a significant impact on $\Delta x_t^o(\lambda)$.

Table 2.1: Estimation results for the VAR process

	Dependent variable					
Regressor	$\hat{\boldsymbol{s}}_{\boldsymbol{t,1}}$		$\hat{\boldsymbol{s}}_{\boldsymbol{t,2}}$		$\hat{\boldsymbol{s}}_{\boldsymbol{t,3}}$	
const	0.014	(2.66)	-0.002	(-0.36)	0.031	(2.60)
$\hat{s}_{t-1,1}$	2.422	(3.94)	2.715	(3.87)	3.768	(2.79)
$\hat{s}_{t-1,2}$	0.133	(3.10)	1.314	(26.81)	0.281	(2.97)
$\hat{s}_{t-1,3}$	-1.019	(-3.68)	-1.287	(-4.07)	-1.565	(-2.57)
$\hat{s}_{t-2,1}$	-2.162	(-3.41)	-2.388	(-3.30)	-5.470	(-3.92)
$\hat{s}_{t-2,2}$	-0.139	(-3.21)	-0.324	(-6.56)	-0.294	(-3.09)
$\hat{s}_{t-2,3}$	0.994	(3.58)	1.300	(4.10)	2.517	(4.11)

The estimation results are obtained for $T_* = T$ and $\lambda = \hat{\lambda}$ and consist of estimated coefficients of the VAR process together with corresponding t-statistics in parentheses.

were highly significant in at least one of the system equations, no other variable was removed from the system. The restricted OLS estimates are presented in Table 2.2, together with t- and F-test statistics. The full set of recursive estimates $\left\{\hat{\theta}_{\text{ECM}}\left(T_*, \hat{\lambda}\right)\right\}_{T_*=T_a}^{T}$ is illustrated in Figure 3 of the Online Supplementary Material.

At this stage we decided that additional (system or individual) coefficient eliminations were unwarranted. As discussed further below, the recursive performance evaluation of the VAR-ECM model already matches that of the benchmark VAR. Therefore, we have already achieved our main objective with this pilot application which was to demonstrate that under our proposed approach, there no longer is an inherent trade-off between theoretical and empirical coherence and that we can achieve both simultaneously.

2.5.3 Recursive Tracking/Forecasting (Conditional on λ)

In this section, we assess the recursive performance of the estimated model conditionally on tentative values of λ (final calibration of λ is discussed in Section 2.5.4 below).

Table 2.2: Estimation results for the ECM process

	Dependent variable					
Regressor	$\boldsymbol{\Delta x^o_{t,1}}$		$\boldsymbol{\Delta x^o_{t,2}}$		$\boldsymbol{\Delta x^o_{t,3}}$	
const	0.002	(4.41)	-0.002	(-3.98)	-0.002	(-4.18)
$r^o_{t-1,1}$	-0.078	(-5.78)	0.079	(4.68)	0.078	(5.16)
$\Delta x^o_{t-1,1}$	-0.258	(-1.92)	1.115	(6.70)	0.677	(4.53)
$\Delta x^o_{t-1,2}$	-0.224	(-1.83)	1.004	(6.62)	0.605	(4.44)
$\Delta x^o_{t-1,3}$	-0.053	(-1.66)	0.026	(0.67)	0.540	(15.14)
$\Delta \hat{s}^o_{t,1}$	0.169	(0.40)	3.865	(7.39)	1.941	(4.13)
$\Delta \hat{s}^o_{t,3}$	-0.991	(-5.15)	-0.461	(-1.93)	0.224	(1.05)
F-statistic	1.64		2.30		1.92	

The estimation results are obtained for $T_* = T$ and $\lambda = \hat{\lambda}$ and consist of estimated coefficients of the ECM process together with corresponding t-statistics in parentheses. The F-test statistics test the null hypothesis that the coefficients of $r^o_{t-1,2}$ and $\Delta \hat{s}_{t,2}$ are jointly zero. The corresponding critical value at the 5 percent significance level is $F_{(2,\infty)} = 2.99$. The F-statistic for excluding $r^o_{t-1,2}$ and $\Delta \hat{s}_{t,2}$ across all three equations is equal to 0.62. The corresponding critical value at the 5 percent significance level is $F_{(6,\infty)} = 2.10$.

First, for each $T_* \in [T_a, T]$, we compute fitted values $\hat{x}_{T_*}(T_*; \lambda)$ for $x = (y, c, n)'$ based upon $\hat{\theta}(T_*; \lambda)$, $\hat{s}_{T_*}(T_*; \lambda)$, $\hat{s}_{T_*-1}(T_*; \lambda)$, and $\hat{s}_{T_*-2}(T_*; \lambda)$. This produces a set of recursive fitted values, say

$$\{\hat{x}_{T_*}(T_*; \lambda)\}_{T_*=T_a}^{T}. \tag{51}$$

Similarly and relying upon $N = 1,000$ Monte Carlo (hereafter MC) simulations, we produce a full set of recursive i-step ahead *out-of-sample* point forecasts[22]

$$\left\{\hat{x}^n_{T_*+i}(T_*; \lambda)\right\}_{T_*=T_a}^{T}, \quad i = 1, 2, 3, \quad n : 1 \to N, \tag{52}$$

[22]We conduct the simulations using auxiliary draws from the error terms in equations (32) and (33) using recursive estimates for Σ_A and Σ_D.

from which we compute *mean forecast estimates* given by

$$\hat{x}_{T_*+i}\left(T_*;\lambda\right)=\frac{1}{N}\sum_{n=1}^{N}\hat{x}^{n}_{T_*+i}\left(T_*;\lambda\right),\quad i=1,2,3,\;\;T_*:T_a\rightarrow T. \tag{53}$$

Since our focus lies on the model's tracking and forecasting performance in times of rapid changes, we assess the accuracy of the estimates $\{\hat{x}_{T_*+i}\left(T_*;\lambda\right)\}_{T_*=T_a}^{T-3}$ for $i=0,1,2,3$ around the three recessions included in the validation period (1990–91, 2001, and the Great Recession of 2007–09). Specifically, for each recession j we construct a time window W_j consisting of two quarters before recession j, recession j, and six quarters following recession j (as dated by the NBER), for a total of N_j quarters:[23]

$$\begin{aligned}
W_1&=\left(\text{1990Q1 to 1992Q3}\right),\quad N_1=11,\\
W_2&=\left(\text{2000Q3 to 2003Q2}\right),\quad N_2=12,\\
W_3&=\left(\text{2007Q2 to 2010Q4}\right),\quad N_3=15.
\end{aligned}$$

Next, we assess tracking and forecasting accuracy over W_j using two commonly used metrics: the Mean Absolute Error (hereafter MAE) and the Root Mean Square Error (hereafter RMSE),

$$\text{MAE}\left(j,i;\lambda\right)=\frac{1}{N_j}\sum_{T_*\in W_j}\left|\hat{x}_{T_*+i}\left(T_*;\lambda\right)-x_{T_*+i}\right|,\;\;j=1,2,3,\;\;i=0,\ldots,3, \tag{54}$$

$$\text{RMSE}\left(j,i;\lambda\right)=\left[\frac{1}{N_j}\sum_{T_*\in W_j}\left(\hat{x}_{T_*+i}\left(T_*;\lambda\right)-x_{T_*+i}\right)^2\right]^{1/2},\;\;j=1,2,3,\;\;i=0,\ldots,3, \tag{55}$$

which, as discussed next, play a central role in the subsequent calibration of λ.[24,25]

[23] We investigated a number of alternative time windows and arrived at similar qualitative results.

[24] Depending upon an eventual decision context, alternative metrics could be used (see Elliott and Timmermann, 2016).

[25] It is important to note that the MAE and RMSE have inherent shortcomings because they measure a single variable's forecast properties at a single horizon (see Clements and Hendry, 1993). While measures do exist for assessing forecast accuracy for multiple series across multiple horizons, we believe that they would not impact our conclusions in view of the evidence provided further below (tables, figures, and hedgehog graphs).

2.5.4 Calibration of λ

The final step of our modeling approach consists of calibrating the time invariant parameters $\lambda = (\beta, \delta, \phi)'$ in accordance with the calibration procedure described above in Section 2.4.2.4.

Estimates of β, δ, and ϕ are widely available in the related literature, with β and δ generally tightly estimated in the $(0.95, 0.99)$ range, and ϕ often loosely identified on a significantly wider interval ranging from 0.1 to 3.0. Searching on those ranges, the calibration of λ is based upon a combination of informal and formal criteria thought to be critical for accurate recession tracking. The informal criteria consists of the time invariance of the recursive parameter estimates $\left\{\hat{\theta}\left(T_*; \lambda\right)\right\}_{T_*=T_a}^{T}$, with special attention paid to the three non-zero coefficients of the ECM correction term, D_3 in formula (33). The reason for emphasizing this invariance criterion is that tracking and forecasting in the presence of (suspected) structural breaks raises significant complications such as the selection of estimation windows (see for example Pesaran et al., 2006; Pesaran and Timmermann, 2007). The formal criteria are the signs of the three non-zero coefficients of the ECM correction term as well as the MAE and RMSE computed for the three recession windows W_j included in the validation period as described in Section 2.5.3.

The combination of these two sets of criteria led to the following choice of λ

$$\hat{\lambda} = \left(\hat{\beta}, \hat{\delta}, \hat{\phi}\right) = (0.97, 0.98, 1.3)\,. \tag{56}$$

We note that even though the calibrated value of β equal to 0.97 is relatively low for a quarterly model, it supports the argument raised by Carroll (2000) and Deaton (1991) that consumers appear to have shorter horizons than frequently thought.

2.5.5 Results

We now discuss the results obtained for our pilot RBC application conditional on the calibrated $\hat{\lambda}$, given in equation (56). The first set of results pertain to the invariance of the recursive estimates $\left\{\hat{\theta}\left(T_*; \hat{\lambda}\right)\right\}$ for T_* ranging from T_a = 1985Q2 to T = 2019Q2. In Figure 2.6, we illustrate the invariance of the three non-zero ECM equilibrium correction

coefficients in D_3, together with recursive 95 percent confidence intervals. All three coefficients are statistically significant, time invariant, and with the expected signs suggested by the economic theory: when the great ratio $\ln(y_{t-1}/c_{t-1})$ exceeds its target value as defined in equation (43), the equilibrium corrections are negative for $\Delta \ln(y_t/n_t)$ and positive for $\Delta \ln(c_t/n_t)$ and $\Delta \ln(1 - n_t/n_t)$—that is negative for $\Delta \ln n_t$.[26] Moreover, we find that the quarterly ECM adjustments toward equilibrium are of the order of 8 percent, suggesting a relatively rapid adjustment to the target movements. This is likely a key component in the model quick response to recessions and would guarantee quick convergence to a balanced growth equilibrium were s_t to remain constant for a few years.

Next, we discuss the tracking and forecasting performance of our hybrid RBC model. In the Appendix C to the online publication, we present the Fred data, together with recursive fitted values ($h = 0$) and 1- to 3-step ahead recursive out-of-sample MC mean forecasts over the W_3 time window for the Great Recession.[27,28] The key message we draw from these figures is that, while both the VAR-ECM and the benchmark VAR track closely the Great Recession and the subsequent economic recovery, they are unable to ex-ante predict its onset and to a lesser extent the subsequent recovery. On a more positive note, we find that the mean forecasts produced by the VAR-ECM outperform those obtained from the benchmark VAR.

For illustration we present summary statistics for the tracking accuracy of the fitted values ($h = 0$) and the forecasting accuracy of 1- to 3-step ahead mean forecasts ($h = 1, 2, 3$) for both the VAR-ECM and the VAR benchmark models over the three recession time windows W_j ($j = 1, 2, 3$) in Table 2.3. The first two measures under consideration are the MAE and RMSE introduced in equations (54) and (55), whereas the third metric is the Continuous Rank Probability Score (hereafter CRPS) commonly used by professional forecasters to

[26] Analogous figures for all other coefficients of the VAR-ECM model and of VAR benchmark are presented in Figures 2 to 4 of the Online Supplementary Material and confirm the overall recursive invariance of our estimates and those of the VAR benchmark.

[27] Analogous figures for the other two recessions are presented in Figures 17 to 22 of the Online Supplementary Material.

[28] Note that the 95 percent confidence intervals are those of the 1,000 *individual* MC draws. The mean forecasts are much more accurate with standard deviations divided by the square root of 1,000.

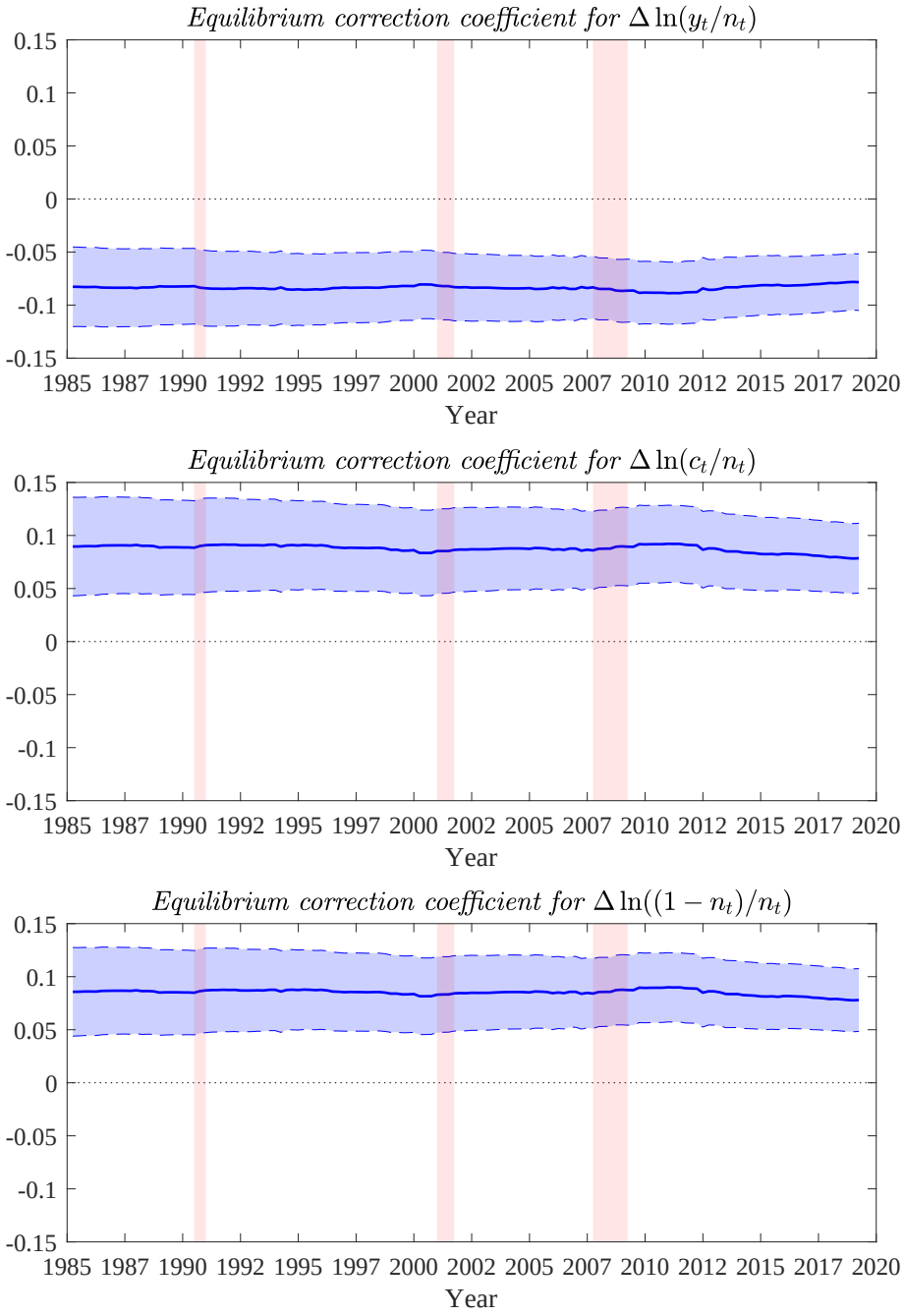

Figure 2.6: Recursive equilibrium correction coefficients in the hybrid RBC model. The solid lines represent the recursive parameter estimates and dashed lines the corresponding 95% confidence intervals. Vertical shaded regions correspond to NBER recession dates

Table 2.3: Tracking and forecasting accuracy of the baseline VAR-ECM and benchmark VAR

	Mean Absolute Error						Root Mean Square Error						Continuous Rank Probability Score					
	VAR-ECM			Benchmark VAR			VAR-ECM			Benchmark VAR			VAR-ECM			Benchmark VAR		
	y	c	n	y	c	n	y	c	n	y	c	n	y	c	n	y	c	n
Recession 1990–91																		
$h=0$	50	11	0.44	104	73	0.99	59	13	0.51	137	88	1.20	-	-	-	-	-	-
$h=1$	98	67	1.10	108	74	1.04	135	86	1.38	140	89	1.24	334	436	6.32	210	345	6.33
$h=2$	141	97	1.87	193	111	2.01	222	116	2.36	271	141	2.45	1155	719	4.51	1377	846	4.50
$h=3$	207	89	3.17	275	124	3.45	277	117	3.77	359	167	4.04	5731	3803	5.13	5857	3828	5.31
Recession 2001																		
$h=0$	62	16	0.43	138	38	1.52	79	22	0.56	167	49	1.96	-	-	-	-	-	-
$h=1$	106	45	1.44	144	39	1.60	134	57	1.74	173	51	2.05	4736	3140	3.79	4586	3063	3.72
$h=2$	164	60	2.45	268	75	3.05	237	90	2.71	319	98	3.53	1867	1400	5.33	1514	1306	6.75
$h=3$	211	68	3.07	376	107	4.46	292	112	3.33	455	138	5.36	2501	1768	3.35	2736	1849	3.30
Recession 2007–09																		
$h=0$	117	27	0.72	145	49	1.42	161	38	1.00	189	64	1.62	-	-	-	-	-	-
$h=1$	139	65	1.65	153	51	1.47	217	85	1.87	197	66	1.66	7103	5246	6.40	7074	5199	7.37
$h=2$	268	103	3.29	335	84	2.89	431	126	3.91	433	114	3.58	2442	2522	19.4	2116	2444	22.0
$h=3$	410	130	5.05	528	138	4.57	625	162	6.20	676	175	6.05	514	129	9.64	576	165	9.84

h denotes forecast horizon. n is expressed in 10^{-3}. All metrics are computed based on a time window covering 2 quarters before and 6 quarters after each of the three recessions. In black we indicate the smaller number and in light gray the larger number for each pairwise comparison between the VAR-ECM and benchmark VAR.

evaluate probabilistic predictions.[29,30] Based on the MAE and RMSE we find that the VAR-ECM model outperforms the benchmark VAR on virtually all counts (44 out of 48 pairwise comparisons) for the first two recessions, whereas the overall performances of the two models are comparable for the Great Recession (13 out of 24 pairwise comparisons).

The CRPS comparison, on the other hand, is more balanced, reflecting in part the fact that the VAR-ECM forecasts depend on two sources of error (u_t in the VAR process and v_t in the ECM process), which naturally translates into wider confidence intervals relative to that of the benchmark VAR (14 out of 27 pairwise comparisons).

As an alternative way of visualizing these comparisons, we provide in the Appendix C to the online publication take-off versions of hedgehog graphs for the VAR-ECM and benchmark VAR models, where "spine" T_* represents $\left\{\hat{x}_{T_*+i}\left(T_*;\hat{\lambda}\right)\right\}, T_*+i \in W_j, i=0,1,2,3$. See Castle et al. (2010) or Ericsson and Martinez (2019) for related images of such graphs and additional details.

Overall, the results prove that it is possible to preserve theoretical coherence and yet match the empirical performance of the unrestricted VAR model to the effect that, with reference to Figure 1 in Pagan (2003), there might be no inherent trade-off between the the approaches.

2.5.6 Great Recession and Financial Series

Our results indicate that while our RBC model tracks the Great Recession, it fails to ex-ante forecast its onset as the VAR-ECM forecasts respond with a time delay essentially equal to the forecast horizon h. Therefore, a natural question is that of whether we can improve ex-ante forecasting of (the onset of) the Great Recession by incorporating auxiliary macroeconomic aggregates to our baseline RBC model.

As discussed in Section 2.3, the Great Recession was triggered by the combination of a global financial crisis with the collapse of the housing bubble. This raises the possibility that

[29]The average CRPS is given by $\text{CRPS}\left(j,i;\hat{\lambda}\right)=\frac{1}{N_j}\sum_{T_*\in W_j}\int_R\left[\hat{F}_m\left(\hat{x}_{T_*+i}\right)-\mathbf{1}\left(\hat{x}_{T_*+i}\geq x_{T_*+i}\right)\right]^2 \mathrm{d}\hat{x}_{T_*+i}$, where $\hat{x}_{T_*+i}$ stands for $\hat{x}_{T_*+i}\left(T_*;\hat{\lambda}\right)$ and $\hat{F}_m$ denotes the predictive CDF. See Grimit et al. (2006, formula (3)) for the discrete version of the CRPS.

[30]The CRPS accounts for the full predictive CDF and as such was not used as one of the calibration criteria for $\hat{\lambda}$ since our objective is that of producing *mean* rather than *point* forecasts.

we might improve ex-ante forecasting by incorporating financial and/or housing variables into the baseline RBC model. However, from an econometric prospective, this approach suffers from three critical limitations.

First and foremost, there exists no precedent to the Great Recession during the postwar period, which inherently limits the possibility of ex-ante estimation of the potential impact of such auxiliary variables. Next, most relevant series have been collected over significantly shorter periods of time than the postwar period for y, c, and n, with start dates mostly from the early sixties to the mid seventies for housing series and from late seventies to mid eighties for financial series. In fact, some of the potentially most relevant series have only been collected from 2007 onward, after their potential relevance for the Great Recession became apparent (for example "Net Percentage of Domestic Banks Reporting Stronger Demand for Subprime Mortgage Loans"). Last but not least, even if it were possible to add financial variables into the model it is unclear whether they would improve the ex-ante forecasting performance since such series are themselves notoriously hard to forecast.

Nevertheless, we decided to analyze whether we might be able to improve the Great Recession ex-ante ($h = 1, 2, 3$) forecasting performance by incorporating additional variables into our baseline RBC model. First, we selected a total of 20 representative series (10 for the housing sector and 10 for the financial sector) based on their relevance as potential leading indicators to the Great Recession. Second, in order to avoid adding an additional layer of randomness into the VAR-ECM model, we incorporated our auxiliary variables lagged by 4 quarters one at a time as a single additional regressor in the state VAR process.[31] Finally, instead of shortening the estimation period as a way of addressing late starting dates of the majority of the auxiliary variables, we set the missing values of the added series equal to zero.[32] This approach allows us to provide meaningful comparisons with the

[31] A four quarter lag allows us to produce 4-step ahead forecasts, without ex-ante forecasting any of the auxiliary series added into the VAR-ECM baseline model. 4-step ahead forecasts are available upon request and were not included in the paper as they only confirm further the ex-ante forecasting delays already illustrated in the Appendix to the online publication.

[32] The history of earlier postwar recessions unambiguously suggest that even if such series were available for the entire postwar period, they would likely fail to explain earlier recessions and would, therefore, be irrelevant at that time. Hence, we believe that any potential bias resulting from the missing data would also be insignificant. This is confirmed further by the fact that the auxiliary series incorporated into the VAR component of the VAR-ECM model turn out to be largely insignificant for the Great Recession, even though they are directly related to its cause.

results in Table 2.3, notwithstanding the fact that dramatically shortening the estimation period would inevitably reduce the statistical accuracy, parameter invariance and, foremost, recession tracking performance of the model. The results of this exercise for each of the 20 selected series are presented in Table 2.4 for the ex-ante forecasting windows $h = 1, 2, 3$ in a format comparable to that used in Table 2.3 for the Great Recession.

Most additions result in deterioration of the forecast accuracy measured by the MAE and RMSE, as one might expect form the incorporation of insignificant variables. The four notable exceptions are the Chicago Fed National Financial Condition Index and three housing variables related to the issuance of building permits, housing starts, and the supply of houses (Housing Starts: New Privately Owned Housing Units Started; New Private Housing Units Authorized by Building Permits; and, to a lesser extent, Monthly Supply of Houses). However, the observed reductions in the corresponding MAE and RMSE do not translate into a mitigation of the delayed responses of ex-ante forecasts.

All together, these results appear to confirm that, as expected, there is a limited scope for structural models to ex-ante forecast recessions as each one has been triggered by unprecedented sets of circumstances. Nevertheless, it remains critical to closely track recessions, as these are precisely times when rapid policy interventions are most critically needed.

We understand that there is much ongoing research aimed at incorporating a financial sector into DSGE models. Twenty years after the Great Recession and accounting for the dramatic impact of the financial crisis, we have no doubts that such efforts will produce models that can better explain how the Great Recession unfolded and, thereby, provide additional policy instruments. However, such ex-post rationalization would not have been possible prior to the recession's onset. Nor it is likely to improve the ex-ante forecasting of the next recession as there are increasing evidence that it will be triggered by very different circumstances (see Section 2.3 for a brief discussion on possible triggers of the next US recession).

Table 2.4: Forecast accuracy of the augmented VAR-ECM

	Housing variables						Financial variables					
	Mean Absolute Error			Root Mean Square Error			Mean Absolute Error			Root Mean Square Error		
	y	c	n	y	c	n	y	c	n	y	c	n
	Housing Starts: Total: New Privately Owned Housing Units Started (1959Q1)						*Chicago Fed National Financial Conditions Index* (1971Q1)					
$h=1$	-3.8	-18.6	-5.5	-3.7	-19.8	0.3	-6.6	-12.4	-4.9	-5.2	-10.3	-3.1
$h=2$	-1.2	-28.5	3.0	-6.1	-26.4	-3.6	-4.2	-19.2	-4.2	-5.9	-15.9	-5.8
$h=3$	-5.5	-26.2	5.9	-5.9	-24.4	-4.8	-5.0	-21.5	-3.6	-5.5	-17.6	-5.7
	Median Number of Months on Sales Market for Newly Completed Homes (1975Q1)						*Delinquency Rate on Commercial and Industrial Loans at All Commercial Banks* (1987Q1)					
$h=1$	2.6	-17.2	0.2	-2.0	-10.5	0.6	3.3	1.3	0.3	2.3	2.8	0.1
$h=2$	11.4	-14.8	-0.1	-1.7	-10.2	-0.2	8.2	7.6	-0.4	2.5	8.2	1.5
$h=3$	7.6	-5.0	2.7	-1.1	-8.7	0.4	7.5	14.8	1.2	4.0	13.0	2.6
	Median Sales Price of Houses Sold (1963Q1)						*Delinquency Rate on Consumer Loans at All Commercial Banks* (1987Q1)					
$h=1$	7.9	9.8	-7.0	-2.9	17.7	-7.4	-1.0	-3.9	-0.4	-0.7	-2.0	0.1
$h=2$	17.5	13.4	-18.0	-2.4	24.3	-9.5	-1.0	-2.9	0.1	-1.6	-2.9	0.4
$h=3$	13.2	14.6	-14.2	-0.3	16.1	-7.9	-1.9	1.6	1.4	-1.6	-1.6	1.0
	Monthly Supply of Houses (1963Q1)						*Delinquency Rate on Loans Secured by Real Estate at All Commercial Banks* (1987Q1)					
$h=1$	-4.9	-9.2	0.9	-3.9	-5.4	1.3	4.2	-11.9	0.3	-1.5	-9.7	0.4
$h=2$	-1.9	-10.5	2.2	-3.7	-8.0	2.0	9.4	-15.8	0.3	-1.6	-10.8	-0.4
$h=3$	-4.0	-11.5	2.9	-4.0	-10.4	2.7	4.5	-12.8	3.1	-1.0	-10.5	0.2
	New One Family Homes for Sale (1963Q1)						*Household Financial Obligations as a Percent of Disposable Personal Income* (1980Q1)					
$h=1$	2.5	0.3	-1.7	2.2	-0.1	-0.4	4.0	3.3	0.5	2.9	3.8	0.8
$h=2$	0.1	1.6	-1.0	0.2	-0.5	-2.1	6.4	19.3	-0.5	3.2	15.1	3.7
$h=3$	0.9	8.9	1.9	1.3	5.3	-2.2	7.0	32.1	-1.6	4.8	24.3	4.6
	New One Family Houses Sold(1963Q1)						*Mortgage Debt Service Payments as a Percent of Disposable Personal Income* (1980Q1)					
$h=1$	8.0	-0.5	20.9	-14.4	-9.8	28.1	2.7	1.9	-0.3	2.0	2.1	0.4
$h=2$	11.0	-9.9	26.8	-18.1	-11.3	26.2	3.4	13.5	0.2	1.9	9.9	2.2
$h=3$	-0.8	-13.2	30.5	-22.2	-17.1	21.2	4.1	21.9	-1.0	3.1	16.7	2.6
	New Private Housing Units Authorized by Building Permits (1960Q1)						*Mortgage Real Estate Investment Trusts: Liability Level of Debt Securities* (1969Q2)					
$h=1$	-6.6	-21.7	-6.9	-8.4	-26.7	1.5	3.0	6.5	-0.9	0.7	7.8	-2.4
$h=2$	1.0	-38.4	3.5	-11.1	-37.6	-2.7	12.1	25.1	-6.0	1.8	23.3	2.5
$h=3$	-8.6	-39.0	8.3	-12.0	-40.4	-4.3	12.0	43.6	-4.8	3.5	31.7	5.7
	New Privately-Owned Housing Units Completed (1968Q1)						*Mortgage Real Estate Investment Trusts: Liability Level of Mortgage-Backed Bonds* (1984Q1)					
$h=1$	23.5	30.0	-0.5	15.2	23.9	-0.7	1.3	5.1	-1.8	0.3	6.0	-3.3
$h=2$	18.9	54.0	0.8	13.3	44.0	1.8	8.7	17.7	-8.7	0.7	16.7	-0.2
$h=3$	21.0	80.2	3.2	17.7	65.7	2.8	7.6	31.9	-6.4	1.8	22.0	2.5
	New Privately-Owned Housing Units Under Construction (1970Q1)						*10-Year Treasury Constant Maturity Minus 3-Month Treasury Constant Maturity* (1982Q1)					
$h=1$	14.7	17.6	-0.6	8.4	11.9	-0.4	-0.7	-1.9	0.4	0.3	-2.3	1.4
$h=2$	9.2	32.6	0.8	6.7	24.8	0.8	-0.5	0.3	6.0	-0.2	-0.8	3.3
$h=3$	10.3	46.5	2.5	9.3	38.6	1.3	-0.6	5.4	7.4	0.1	1.5	4.7
	S&P/Case-Shiller U.S. National Home Price Index (1987Q1)						*10-Year Treasury Constant Maturity Minus 2-Year Treasury Constant Maturity* (1976Q3)					
$h=1$	8.0	4.8	-0.3	3.2	5.5	-1.2	5.8	17.8	0.9	3.6	9.6	2.5
$h=2$	15.6	19.3	-3.3	3.8	17.5	0.1	5.4	21.0	8.5	3.3	16.0	5.8
$h=3$	16.7	37.6	-2.0	6.4	26.5	1.2	5.1	28.8	8.0	4.2	21.6	7.8

Each number is expressed in percentage and corresponds to the relative difference in the MAE and RMSE calculated under the augmented and baseline VAR-ECM models. Negative numbers (depicted in black) indicate a forecasting performance of the augmented VAR-ECM better than that of the baseline VAR-ECM. Positive numbers (depicted in light gray) indicate a forecasting performance of the augmented VAR-ECM worse than that of the baseline VAR-ECM. h denotes forecast horizon. n is expressed in 10^{-3}. Both metrics are computed based on a time window covering 2 quarters before and 6 quarters after the 2007–09 Great Recession. All auxiliary variables are introduced one at a time as a fourth lag into the state VAR equation. In parenthesis we indicate each series starting date.

2.5.7 Policy Experiment

As we mentioned in the introduction, treating some key structural parameters as time varying state variables allows for state-based policy interventions aimed at mitigating the impact of a recession. Clearly, with a model as simplistic as our pilot RBC model, the scope for realistic policy interventions is very limited. Nevertheless and for illustration purposes, we consider two sets of policy interventions. The first one consists of raising the capital share of output in order to stimulate production in accordance with equation (37). The other consists of raising the relative importance of consumption versus leisure φ in equation (36) or, equivalently lowering d is equation (39), in order to stimulate consumption.

Keeping in mind that the Cobb-Douglas production function in equation (37) is highly aggregated to the effect that α_t covers a wide range of industries with very different shares of capital, raising α_t would require shifting production from sectors with low α's to sectors with high α's (such as capital intensive infrastructure projects).

As for φ_t (or, equivalently, d_t) in equation (36) and in the absence of a labor market, low relative preference for consumption versus leisure at this aggregate level covers circumstances that are beyond agents' control, such as depressed income or involuntarily unemployment. Therefore, one should be able to raise φ_t by carefully drafted wages or other employment policies. The combination of the α_t and φ_t policies would therefore provide a highly stylized version of the New Deal enacted by F. D. Roosevelt between 1933 and 1939.

In the present paper, we implement these two artificial policies separately. Since our model fails to ex-ante forecast the onset of the Great Recession but tracks it closely, we consider two versions of each policy: one implemented at the onset of the Great Recession (2007Q4) and the other one delayed by four quarters (2008Q4). The corresponding policies are labeled 1a and 1b for α_t and 2a and 2b for d_t, which is a monotonic transformation of φ_t, as defined in equation (39). All the policies are progressive and last for either five or nine quarters (depending on the policy). This progressive design has the objective of smoothing the transition from the initial impact of a negative shock to the economy until the subsequent recovery. The specific implementation details are illustrated in Table 2.5 and the results are presented in Figure 2.7. With reference to Figures 2.4 and 2.5, we note that the cumulative

Table 2.5: Quarterly state-based policy interventions for the Great Recession

Year	Quarter	Policy 1a	Policy 1b	Policy 2a	Policy 2b
2007	Q4	0.0005	-	-0.0030	-
2008	Q1	0.0005	-	-0.0030	-
2008	Q2	0.0020	-	-0.0100	-
2008	Q3	0.0030	-	-0.0200	-
2008	Q4	0.0040	0.0060	-0.0200	-0.0200
2009	Q1	0.0040	0.0060	-0.0200	-0.0200
2009	Q2	0.0040	0.0060	-0.0200	-0.0200
2009	Q3	0.0030	0.0030	-0.0100	-0.0100
2009	Q4	0.0020	0.0030	-0.0100	-0.0100

Policies 1a and 1b pertain to interventions for $\hat{\alpha}$. Policies 2a and 2b pertain to interventions for $\hat{d}$.

sum of the two interventions represent around three times the size of the estimated changes in α_t and φ_t from 2007Q4 to 2009Q2. Therefore, and in relation to long-term variations in α_t and φ_t, the relative sizes of the two interventions are moderate.

We note that both policies significantly mitigate the impact of the recession on output and consumption. While such conclusion would require deeper analysis in the context of a more realistic model, including in particular a labor market, we find these results to be promising indications of the added policy dimensions resulting from interventions at the level of the additional state variables that would otherwise be treated as constant structural parameters within a conventional DSGE framework. Hence, the results in Figure 2.7 highlight the potential of (more realistic) implementations of both policies and the importance of their appropriate timing (hence the importance of tracking).

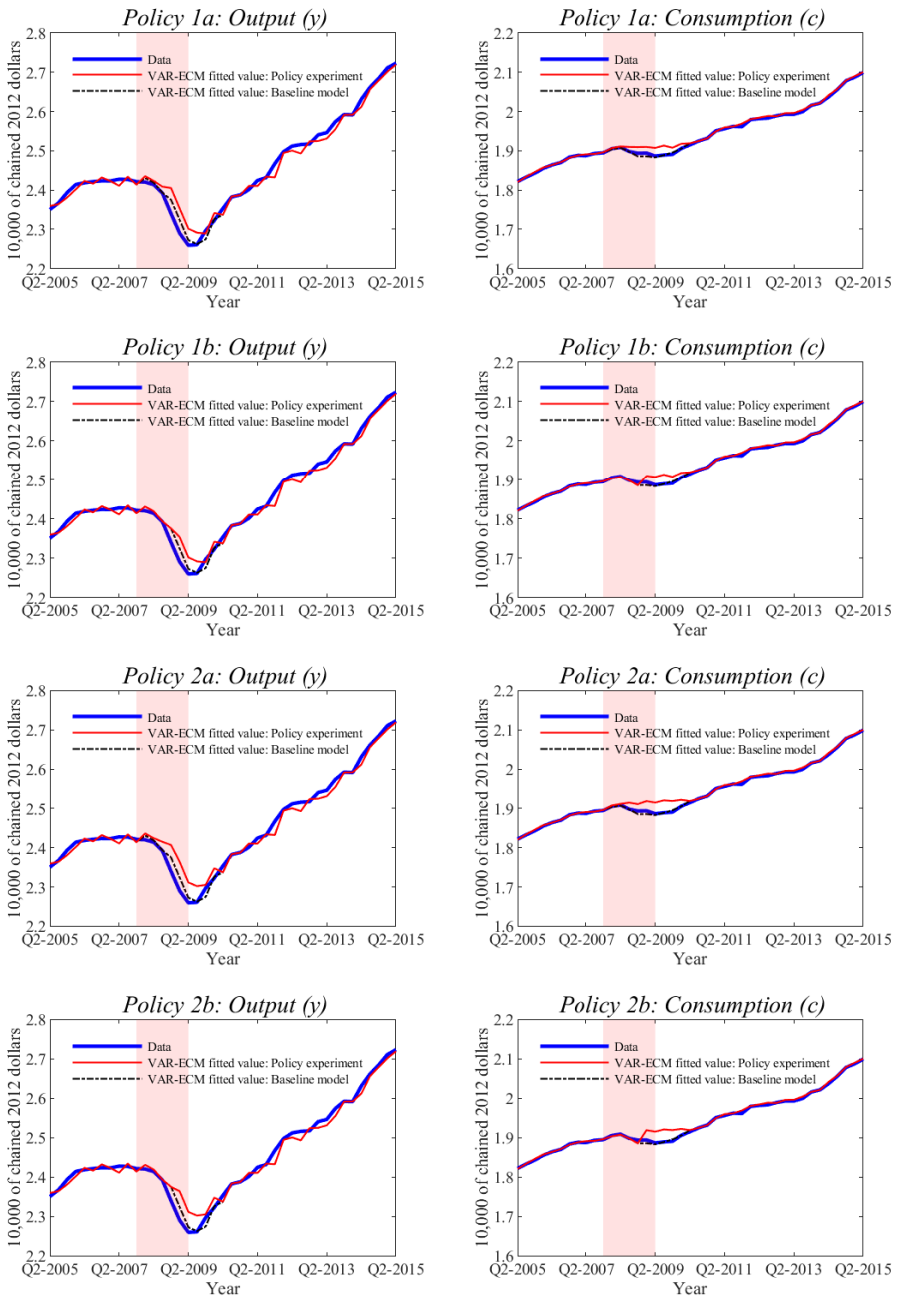

Figure 2.7: Effects of policy interventions for $\hat{\alpha}$ and $\hat{d}$ designed to mitigate the impact of the Great Recession on output and consumption. Policies 1a and 1b pertain to interventions for $\hat{\alpha}$. Policies 2a and 2b pertain to interventions for $\hat{d}$. Shaded regions correspond to NBER recession dates

2.6 Conclusions

We have proposed a generic approach for improving the empirical coherence of structural (DSGE) models with emphasis on parameter invariance and recession tracking performance while preserving the model's theoretical coherence.

The key components of our hybrid approach are the use of data that are neither filtered nor detrended, reliance upon (hypothetical) balanced growth solutions interpreted as agents' theory-derived time-varying cointegrating relationships (moving targets), the use of a state VAR process treating an appropriate subset of structural parameters as state variables, and finally, reliance upon an ECM process to model agents' responses to their moving targets.

Our application to a pilot RBC model demonstrates the potential of our approach in that it preserves the theoretical coherence of the model and yet matches or even outperforms the empirical performance of an unrestricted VAR benchmark model. Most importantly, our hybrid RBC model closely tracks (y, c, n) during the last three postwar recessions, including foremost the 2007–09 Great Recession, a performance largely unmatched by DSGE models and one that is critical for policy interventions at times when they are most needed. In other words, with reference to Pagan (2003) we do not find an inherent trade-off between theoretical and empirical coherence. Our hybrid RBC model achieves both simultaneously.

We also find that, as expected, ex-ante forecasting of recessions is likely to remain econometrically limited using structural models in view of the idiosyncratic nature of recession triggers preventing ex-ante estimation of their potential impact. Hence, the quote from Trichet (2010), as cited in Section 2.2, remains as relevant as ever. While structural models remain essential for policy analysis and, as we have shown, can match the recession tracking performance of the unrestricted VAR benchmark, they will likely remain inherently limited in their capacity to ex-ante forecast major unexpected shifts. Fortunately, there exists "complementary tools" such as leading indicators, that can bridge that gap.

Last but not least, a potentially promising avenue for future research is one inspired by DeJong et al. (2005), where the authors develop a (reduced form) non-linear model of GDP growth under which regime changes are triggered stochastically by a *tension index* constructed as a geometric sum of deviations of GDP growth from a *sustainable rate*. A

quick look at Figures 2.3 to 2.5 suggests that a similar index could be derived from the state variables, where the key issue would be that of incorporating such a trigger within the VAR component of our hybrid model.

3.0 Goals, Constraints, and Transparent Assignment: A Field Study of the UEFA Champions League

Joint with Professor Alistair J. Wilson

3.1 Introduction

Managers are typically the party tasked with deciding upon workable solutions to assignment problems. For example, consider a choice over a monthly scheduling procedure to allocate workers to a set of shifts, taking into account constraints on the required mix of worker skills, shifting availabilities, and previous allocations. While managers might be able to quantify their design objectives, implemented assignment procedures are often ad hoc or informal (turn-taking, drawing names from a hat, asking for volunteers) rather than the outcome of an explicit optimization over the process. One reason for this may be that assignment problems—particularly those with nontrivial constraints—are inherently challenging to optimize over. However, another important institutional factor for many assignment problems is the need for the employed procedure to appear unbiased to the various stakeholders (employees, the media, regulatory agencies, etc.). Even qualitative descriptions of an objective-optimizing algorithm can be complex, particularly to lay people. Moreover, the process of realizing randomized assignments through an algorithm can be opaque, as if produced by a black box. Consequently, managers might subordinate other design objectives to ensure that the resulting assignment process is transparent. Still, it is natural to wonder what the losses from such choices are; whether there might be alternative assignment procedures that better-achieve a manager's objectives, while maintaining transparency.

Our paper conducts an analysis of a constrained assignment with huge public scrutiny, hundreds of thousands following the matching procedure, and with millions of euros at stake from the realizations. The developed procedure provides a solution to a combinatorically complex constrained-assignment problem that prioritizes transparency of the process, conducting all randomization through simple urn draws. After characterizing theoretical properties of the chosen assignment rule, we outline how a recent market-design tool (Bud-

ish et al., 2013) allows us to quantify the loss from this transparent procedure, relative to an optimal one designed to provide all pairwise comparable tournament participants with identical distributions over prizes. In particular, focusing on such defined measure of fairness we show that to all practical extents the developed procedure is very close to a constrained best.

Our application is a public drawing of football-team pairs in the Union of European Football Association's (UEFA) Champions League (UCL). The UCL is one of the most successful pan-European ventures, and certainly the one with the most enthusiasm from the general public. The tournament brings together football clubs from across the continent (and beyond) that normally play within their own country-level associations. Selection into the competition is limited to the highest-performing clubs from each nation, where a series of initial qualifying rounds whittle the number of participating teams down to 32 group-stage participants. From there, half of the clubs advance to a knockout stage that begins with the Round of 16 (R16), followed by four quarter-finals (QF), two semi-finals (SF), and a final (F) that determines a European champion. Outside of the World Cup, the UCL final game is one of the most-watched global sporting events, eclipsing even the viewership of the Superbowl.

Because the UCL is under a magnifying glass—from the teams, sponsors, fans, and the media—UEFA has a clear interest in creating impartial and meritocratic assignment rules for the tournament, with credible randomizations. While a fair public drawing could be trivially designed in many situations (for instance, matching teams through urn draws without replacement) the tournament's design problem is complicated due to three constraints imposed on the R16 match pairs: (i) Each pairing must be between a group winner and a group runner-up (*the bipartite constraint*). (ii) Teams that played one another in the prior group stage cannot be matched (*the group constraint*). (iii) Teams from the same national association cannot be matched (*the association constraint*). Note that these three constraints provide some degree of meritocratic seeding (bipartite), promote variability in the realized match-ups (group), and maintain an international character to the tournament (association).

The chosen randomization assembles each R16 team pairing through a dynamic public draw of balls from an urn, where the draw composition is adapted dynamically by a computer

in order to respect the constraints. Our paper sets out to analyze the properties of this ad hoc assignment rule, using the tools of market design: theory, estimation, and simulation (Roth, 2002).

We first theoretically characterize the simple to follow (but combinatorically complex) UEFA draw procedure. Next, we quantitatively analyze the constraint effects, showing that the association constraint generates large distortions, of up to two million euro. Of particular analytical interest are the indirect effects of the matching constraints, where spillovers from imposed team-pair exclusions disproportionately affect the likelihood of other teams' matches. A natural question is whether an alternative procedure exists that respects the constraints but results in a randomization with less distorted likelihoods of unconstrained team-pair matches. Using an objective function focused on disparate treatment of pairwise comparable teams, we use the Budish et al. (2013) result to relax (without loss of generality) the complex problem of looking for randomizations over constrained assignments; instead we focus on the far-more-tractable problem of finding *expected* assignments that satisfy the constraints. The conclusion from our analysis of the UEFA assignment rule across the past sixteen years (directly preceding the outbreak of the COVID-19 pandemic) shows that even though marginal improvements *are* possible, the tournament's assignemt rule comes very close to achieving the same outcomes as optimal procedures specifically designed to reduce distortions in the likelihoods of unconstrained team-pair matches.

While our search for a superior constrained-assignment procedure suggests only minimal scope for improvement (and large potential costs from forgoing the transparent procedure UEFA developed), a related question is the extent to which better outcomes are possible from slacking the imposed constraints. As a constructive exercise, we conclude the paper by showing that a relatively small relaxation of the association constraint can substantially reduce the distortions. Importantly, this can be done with only minimal adjustments to the current procedure, retaining its transparent design.

As an assignment rule, our application's design stands out as one that focuses on fuzzier attributes: simplicity to lay people, transparency in the process, etc. The analytical tools we bring to bear are certainly capable of deriving an optimal procedure for any given objective. However, rather than as a constructive design tool, our analysis instead focuses on

quantifying the losses from a design tailored towards transparency.

In terms of the paper's organization, Section 3.2 provides a brief review of related literature. In Section 3.3, we describe the application and outline the UCL R16 draw procedure. In Section 3.4, we discuss the constraint effects on teams' expected prize money and the probability of reaching the semi-final stages of the competition (and beyond). In Section 3.5, we show near-optimality of the current UEFA procedure. In Section 3.6, we discuss the extent to which weakening the current set of matching constraints reduces distortions in the likelihoods of unconstrained team-pair matches. Finally, Section 3.7 concludes.[1]

3.2 Literature Review

Our paper contributes to three main strands of economic literature: OR, market design, and tournaments. While there is large theoretic literature on the incentive effects of tournaments (see Prendergast, 1999) our paper is more closely related to a growing body of applied work exploiting public and well-structured sports-tournament outcomes as naturally occurring experiments. In recent years, data from football to cricket to golf has been used to provide evidence supporting both standard theory (Walker and Wooders, 2001; Chiappori et al., 2002; Palacios-Huerta, 2003) and behavioral biases (Bhaskar, 2008; Apesteguia and Palacios-Huerta, 2010; Pope and Schweitzer, 2011; Foellmi et al., 2016).

Where the literature on sports tournaments has been centered around various positive aspects of individual' behavior, our paper instead emphasizes normative features of the tournament itself. In this sense, our work is more closely related to the market design literature, and a handful of applied papers examining well-structured environments. Key examples here are: Fréchette et al. (2007), demonstrating the problem of inefficient unraveling in a decentralized market through US college football bowls; Anbarci et al. (2015), designing a fairer mechanism for penalty shootouts in football tournaments; Baccara et al. (2012), investigating spillovers and inefficiency in a faculty office-assignment procedure; and Budish and

[1]Full data, programs, and the Online Supplementary Material are available at https://www.martaboczon.com.

Cantillon (2012), studying the superiority of a manipulable mechanism to the strategy-proof mechanism using data from a business school's course assignment procedure. In each, quantitative market-design methodologies are developed through a combination of theory with a structural analysis of the application. Similarly, our paper approaches the assignment design question through a combination of theory tools, estimation techniques, and simulation within the application (see Roth, 2002).

The main normative insights for our application are made possible through the core theorem in Budish et al. (2013)—which shows that a consideration of the expected assignment matrix is sufficient for an analysis of the assignment design problem, as long as the constraints satisfy a biheirarchy separability condition.[2] This result allows us to show near-optimality of the existing UEFA assignment rule. To our knowledge our paper is one of the first ones to apply this market-design tool in a normative assessment of a procedure in the field. While our tournament setting is of standalone interest,[3] the focus on constraints in matching is related to topics in school choice such as the implementation of affirmative-action constraints (for example, Dur et al., 2016).

Lastly, our paper introduces a novel design consideration to a literature that is primarily focused on fairness, efficiency, and strategy proofness (see Abdulkadiroğlu and Sönmez, 2003, and references thereof). In our setting instead of manipulation by participants, the main design consideration is that of minimizing the potential for deceptive behavior on the part of principals, removing any opaque or manipulable features in the randomization. This leads to an additional and nontrivial challenge: designing a procedure that incorporates substantial combinatoric complexity through the constraints, where the randomization procedure is transparent and credible to the general public. In this sense, the paper is related to Akbarpour and Li (forthcoming) who examine a designer's credibility problems in implementing auction rules.[4]

[2]See also Hylland and Zeckhauser (1979) and Bogomolnaia and Moulin (2001), each of which considers assignments with individual choice, whereas our paper focuses on alternative randomization procedures from the designer's point of view.

[3]While our paper primarily serves as an application for the Budish et al. (2013) result, it also contributes to a literature on optimal tournament design. For example, see Dagaev and Sonin (2018), Guyon (2018, 2015), Ribeiro (2013), Scarf and Yusof (2011), Scarf et al. (2009), and Vong (2017).

[4]See also Bó and Chen (2019) on the importance of simplicity and transparency in a historical random assignment for civil-servants in Imperial China.

3.3 Application Background

The UCL is the most prestigious worldwide club competition in football. Its importance within Europe is similar to that of the Superbowl in the United States, though with stronger global viewership figures (details below). The tournament is played annually between late June and the end of May by the top teams from 53 national associations across Europe, featuring most of the sports' star players.

In the 2017 season, 6.8 million spectators attended UCL matches, with more than 65,000 at the final game.[5] In addition to the in-stadium audience the tournament has vast media exposure through television and the Internet. The UCL final game is globally the most-watched annual sporting event, where the 2015 final had an estimated 400 million viewers across 200 countries, with a live audience of 180 million. For comparison, the 2015 Super Bowl, which is the most-watched Super Bowl match in history, was viewed by 114 million fans.[6]

Revenue for UEFA is primarily generated (98.5 percent) by selling broadcasting and commercial rights for its club and national team competitions (such as the UCL, the UEFA Europa League, and the UEFA Super Cup). In the 2017 season, 50 percent of the generated revenue (2.8 billion euro) was distributed to the UCL participants as either prize money for tournament progression (60 percent) or financial benefits distributed through a market pool (40 percent).[7]

Fixed payments of 12.7 million euro (as of the 2017 season) were received by each of the 32 clubs that qualified for the group stage, where an additional performance bonus of 1.5 (0.5) million euro was awarded for a group stage win (draw). Beyond the group stage, teams are given additional payments for reaching each of the subsequent tournament stages.

[5]Since each UCL season spans across two calendar years, for clarity and concision we refer to a particular season by the year of its final game; so 2019 would indicate the 2018–19 season.

[6]Furthermore, the UCL proves a massive success on social media. In the 2017 season, the UCL official Facebook page became the worlds' most followed for a sporting competition with 63 million fans, 300 million video views, and 98 million interactions over the 2017 UCL final.

[7]The first-order beneficiaries are clubs participating in the group stage and onward. The remaining UEFA revenue is distributed among second- and third-order beneficiaries, which are clubs participating in the qualifying rounds and non-participating clubs, respectively. The solidarity payments made to the latter teams are distributed via national associations and allocated for the most part to youth training programs.

Teams that advanced to the R16 (in the 2017 season) garnered an additional 6 million euro, quarter-finalists 6.5 million more, semi-finalists 7.5 million more, and finalists a further 11 million (with the winner receiving a 4.5 million bonus). While prize money is fixed, benefits based on the market-pool share are in proportion to the value of the TV market for each club's games. Among the biggest winners of the 2017 season were Juventus, the runner-up with tournament revenue of 110 million euro (a 46:54 split between prize money and market pool), Leicester City, a quarter-finalist with 82 million euro (a 40:60 split), and the 2017 European champion Real Madrid with 81 million euro (a 64:36 split).

Introduced in 1955 as a European Champion Club's Cup (and consisting only of the national champion from each association) the tournament has evolved over the years to admit multiple entrants from each national association (at most five). The last major change to the tournament's design took place in the 2004 season. As such, in our empirical analysis we focus on the 2004–19 seasons.[8]

Since the 2004 season, the UCL consists of a number of pre-tournament qualifying rounds followed by a group and then a knockout stage, similar in format to the World Cup, but played concurrently with the national associations' leagues. In the group stage, 32 teams are divided into eight groups of four.[9] Beginning in September each team plays the other three group members twice (once at home, once away). At the end of the group stage in December, the two lowest-performing teams in each group are eliminated, while the group winner and runner-up advance to the knockout stage. The knockout stage (except for the final game) follows a two-legged format, in which each team plays one leg at home, one away. Teams that score more goals over the two legs advance to the next round, where the remaining teams are eliminated.[10]

[8]The 2020 season was disrupted by the COVID-19 pandemic and, therefore, followed a different, mostly one-legged, format whereas the 2021 season has not yet concluded.

[9]Prior to the 2015 season, seeding in the group stage was entirely determined by the UEFA club coefficients calculated based on clubs' historical performance, with the titleholder being automatically placed in Pot 1. Starting from the 2016 season, the titleholder together with the champions of the top-seven associations based on UEFA country coefficients are placed in Pot 1. The remaining teams are seeded to Pots 2-4 based on club coefficients. The eight groups are then assembled by making sequential draws from the four pots with a restriction that teams from the same association cannot be placed in the same group, enforced in a similar way to the R16 match we detail in the paper.

[10]Technically, the scoring rule is lexicographic over total goals, and goals away from home. A draw on both results in extra time, followed by a penalty shootout if a winner is still not determined.

The focus of our paper is on the assignment problem of matching the 16 teams at the beginning of the knockout phase into eight mutually disjoint pairs.[11,12] If the problem consisted simply of matching two equal-sized sets of teams under the bipartite constraint, the assignment could be conducted with two urns (one for group winners, one for runners-up) by sequentially drawing team pairs without replacement. However, the presence of the group and association constraints prohibits such a simple procedure for two reasons. First, after drawing a team to be partnered, the urn containing eligible partner draws must not contain any directly excluded teams. Second, a match with a non-excluded partner must not force an excluded match at a later point in the draw. While the first concern is easy to address, the second one requires a more-complicated combinatoric inference.

For illustration, consider the following dynamic draw from two urns. The first urn contains teams A, B and C; the second teams d, e and f. Suppose that the match-ups Ad and Be are directly excluded by the constraints. An initial draw from the first pot selects team A; as such, the process directly excludes d as a potential match partner for A. Assume f is selected from the second urn (containing e and f) to create the Af pair. In the second round of the draw, suppose C is chosen from the first urn; since C has no *directly* excluded partners, a draw from the second urn could be over d and e. However if Cd were formed, B would have no valid partner in the third round, as Be is directly excluded. This implies that for the process to work, in the second round C must be *indirectly* excluded from matching to d.

Although this logic is easy to follow in a three-to-three matching, with eight teams on each side and many more constraints, the combinatorics become involved. While matchings could be formed via fully computerized draws, UEFA instead opts to make the randomization transparent and ex-post verifiable through urn draws. The dynamic draw procedure UEFA developed randomizes the R16 tournament matching as follows: (i) eight blue balls representing eight runners-up are placed in the first urn and one runner-up ball is drawn without replacement; (ii) a computer algorithm determines the maximal feasible set of group

[11] In principle, a similar analysis could be conducted for the group-stage assignment problem. However, for tractability and clarity purposes, we focus on the R16 draw in isolation.

[12] The quarter- and semi-final draws are free from both seeding and the association constraint, and as such are conducted in a standard fashion by drawing balls from an urn without replacement.

winners that could match with the drawn runner-up given the constraints and all previous draws; (iii) white balls representing the feasible group winner matches are placed into a second urn, where one is drawn; (iv) a pairing of the two drawn teams (one winner and one runner-up) is added to the aggregate R16 matching. This procedure repeats until all eight matches are formed. In what follows, we refer to the above algorithm as the constrained $\mathcal{R}$-to-$\mathcal{W}$ dynamic draw, where $\mathcal{W}$ and $\mathcal{R}$ indicate the sets of group winners and runners-up, respectively.

Though the above assignment rule is combinatorically involved, the UEFA draw procedure has three useful features. First, all draws are conducted using an urn, and thus each individual randomization is dawn from a uniform distribution. Second, the number of possible realizations in each draw is always less than eight, implying that each individual draw is easy to comprehend, with a clear urn composition. Finally, all nontrivial elements of the draw conducted opaquely by the computer (the calculation of the set of valid partners at each step) can be checked ex post, and thus the draw procedure is fully verifiable by more-sophisticated viewers. In particular, the 20-minute R16 draw ceremony is streamed live by UEFA over the Internet and broadcast by many national media companies.[13] Consequently, as long as the urn draws are conducted publicly and without foul play,[14] it is not possible for the designer to cherry pick realizations, inoculating the mechanism against corruption on the part of managers and principals.

In the absence of the association constraint, the tournament has 14,833 possible R16 matchings, where each same-nation exclusion significantly reduces the number of valid assignments.[15] Across the 16 seasons under consideration, the number of valid assignments ranged from 2,988 in the 2009 season to 6,304 in 2011 to 9,200 in 2006.[16] We graph the

[13] A full rerun of the 2020 UCL R16 draw ceremony with over 1.3 million views is available on UEFA's YouTube channel.

[14] Unlike many state lotteries which use mechanical randomization devices to draw urn outcomes, the UEFA draw is conducted by human third-parties (typically famous footballers). Pointing to football fans' distrust in the process, the human draw has led to plausible allegations of UEFA rigging draws with hot/cold balls (here made by a former FIFA president Sepp Blatter in an interview with Argentine newspaper *La Nacion* on June 13$^{\text{th}}$, 2016).

[15] A political constraint also excludes Russian teams from being drawn against Ukrainian teams. In what follows, we re-interpret this restriction as an extended association constraint.

[16] In order to calculate the number of valid assignments in each season we proceed in three steps. First, we create a set of all assignments satisfying the bipartite constraint, where this set has $8! = 40,320$ possible matchings. In the second step, we delete the 25,487 assignments that violate the group constraint. Finally,

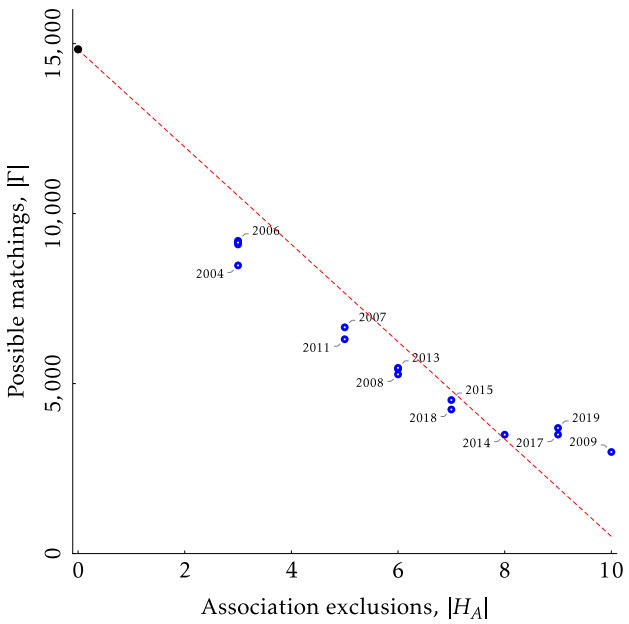

Figure 3.1: Number of possible matchings against the number of same-nation exclusions (2004–19). Red dashed line indicates a fitted linear relationship (intercept constrained to 14,833)

relationship between the number of possible matchings and the number of same-nation exclusions implied by the association constraint in Figure 3.1. While the number of valid assignments is not purely a function of the number of exclusions (it depends on their arrangement too) the relationship in question can be approximated by a linear function that decreases by 1,400 matchings for each same-nation exclusion.

3.4 Constraint Effects in the Current UEFA Procedure

In this section, we first describe a generalized version of the constrained dynamic assignment rule used by UEFA. Then, we quantify how the current procedure affects expected assignments in the UCL R16, and how the spillovers from imposed team-pair exclusions

from the remaining 14,833 assignments, we delete those that violate the association constraint.

impact the likelihoods of other team matches.

3.4.1 Theory for the Constrained Dynamic Draw

Let $\mathcal{W} = \{w_1, w_2, ..., w_K\}$ and $\mathcal{R} = \{r_1, r_2, ..., r_K\}$ denote the sets of group winners and runners-up, respectively. Let $\mathcal{V}$ be the set of all *possible* perfect (exhaustive one-to-one) matchings between $\mathcal{W}$ and $\mathcal{R}$. We examine a random assignment mechanism $\psi : 2^{\mathcal{V}} \to \Delta\mathcal{V}$ that takes as input $\Gamma \subseteq \mathcal{V}$ (a set of *admissible* matchings) and provides as output a probability distribution over the elements of Γ. A generalized version of the dynamic draw employed by UEFA proceeds as follows:

Algorithm (The constrained $\mathcal{R}$-to-$\mathcal{W}$ dynamic draw)**.** *Given an input set of admissible matchings $\Gamma \subseteq \mathcal{V}$, the algorithm selects a matching $\psi(\Gamma)$ in $K = |\mathcal{W}|$ steps.*

Initialization: *Set $\mathcal{R}_0 = \mathcal{R}$, and $\Gamma_0 = \Gamma$.*

***Step-**k: (for $k = 1$ to K)*

1. *Choose $R_k \in \mathcal{R}_{k-1}$ through a uniform draw over $\mathcal{R}_{k-1}$;*
2. *Choose $W_k \in \mathcal{W}_k := \{w \in \mathcal{W} \,|\exists V \in \Gamma_{k-1}$ s.t. $R_k w \in V\}$ (the set of admissible partners for R_k) through a uniform draw;*
3. *Define the currently unmatched runners-up $\mathcal{R}_k = \mathcal{R}_{k-1} \setminus \{R_k\}$, and valid assignments given the current draw $\Gamma_k = \{V \in \Gamma_{k-1} \,|R_k W_k \in V\}$.*

***Finalization**: After K steps the algorithm assembles a vector of K runner-up–winner pairs, $\mathbf{v} = (R_1W_1, \ldots, R_KW_K)$, where the realization of $\psi(\Gamma)$ is $\{R_1W_1, R_2W_2, \ldots, R_KW_K\} \in \Gamma$.*

This dynamic procedure as an interim output produces a sequence of K matches, $\mathbf{v}$. In order to characterize the probability of a specific matching V we define: (i) $\mathcal{P}(V)$, the set of possible sequence permutations for matching V; and (ii) $\mathcal{W}_k(\mathbf{v})$, the set of admissible match partners for runner-up R_k selected at Step-$k(i)$ in the permutation $\mathbf{v}$.[17]

Proposition 1. *Under the constrained $\mathcal{R}$-to-$\mathcal{W}$ dynamic draw the probability of any match-*

[17] That is, for the permutation $\mathbf{v} = (R_1W_1, \ldots, R_KW_K)$ the set of partners at step k is $\mathcal{W}_k(\mathbf{v}) := \left|\{w \in \mathcal{W} \,|\exists V \in \Gamma \text{ s.t. } R_k w \in V \text{ and } \wedge_{j=1}^{k-1} (R_j W_j \in V)\}\right|$.

ing $V \in \Gamma$ *is given by*

$$\Pr\{\psi(\Gamma) = V\} = \frac{1}{K!} \sum_{\mathbf{v} \in \mathcal{P}(V)} \prod_{k=1}^{K} \frac{1}{\mathcal{W}_k(\mathbf{v})}.$$

Proof. See Appendix to the Working Paper. □

Proposition 1 shows that the probability distribution over Γ requires $K! \times |\Gamma|$ calculations. Even though the cardinality of Γ can be substantially lower than $K!$, the exact computation of $\Pr\{V\}$ involves between $K!$ and $(K!)^2$ steps, and can be taxing even for our application with $K = 8$.

Given the characterization in Proposition 1, one remaining question is the extent to which the above calculation can be simplified. Defining two randomization procedures as *distinct* if they induce different probabilities over the matchings in $\mathcal{V}$, we can show that:

Proposition 2. *The constrained* $\mathcal{R}$*-to-*$\mathcal{W}$ *dynamic draw is distinct from:*

1. *A uniform draw over* Γ*;*
2. *constrained* $(\mathcal{W}, \mathcal{R})$ *dynamic draw;*[18]
3. *The constrained* $\mathcal{W}$*-to-*$\mathcal{R}$ *dynamic draw.*

Proof. See Appendix to the Working Paper. □

The first two parts of Proposition 2 are essentially negative results, indicating that our environment is not equivalent to algorithmically simpler fair draws over complete matchings or individual matches, where the third part shows that the procedure is asymmetric. While the main takeaway from Proposition 2 is negative, it does demonstrate three potentially constructive design channels, analogous to the reversal of the proposing sides in the National Resident Matching Program algorithm detailed in Roth and Peranson (1997).[19]

Above we characterize a generalized version of the dynamic draw employed by UEFA to assemble the R16 matching. Within this family of randomization procedures, the actual

[18] A constrained $(\mathcal{W}, \mathcal{R})$ dynamic draw in identical to the constrained $\mathcal{R}$-to-$\mathcal{W}$ dynamic draw except for the fact at each step k runner-up and winner are not selected sequentially but together as a pair.

[19] While distinct, we later show that in our particular setting the three draw procedures lead to only marginally different outcomes, where a uniform draw over Γ has a computational advantage for approximating assignment probabilities in fractions of second.

Table 3.1: Expected assignment matrix for the 2018 R16 draw

	Basel	*Bayern Munchen*	*Chelsea*	*Juventus*	*Sevilla*	*Shakhtar Donetsk*	*Porto*	*Real Madrid*
Manchester United	0 (H_G)	0.148	0 (H_A)	0.183	0.183	0.155	0.148	0.182
Paris Saint-Germain	0.109	0 (H_G)	0.294	0.128	0.128	0.108	0.105	0.128
Roma	0.159	0.151	0 (H_G)	0 (H_A)	0.189	0.160	0.152	0.189
Barcelona	0.149	0.144	0.413	0 (H_G)	0 (H_A)	0.150	0.144	0 (H_A)
Liverpool	0.159	0.151	0 (H_A)	0.189	0 (H_G)	0.160	0.152	0.189
Manchester City	0.156	0.148	0 (H_A)	0.183	0.184	0 (H_G)	0.148	0.183
Besiktas	0.109	0.105	0.293	0.128	0.128	0.108	0 (H_G)	0.129
Tottenham Hotspur	0.160	0.152	0 (H_A)	0.189	0.189	0.159	0.151	0 (H_G)

Probabilities derived from a simulation ($N = 10^6$) of the UEFA draw procedure. Exclusion constraints indicated by the group constraint (H_G) and the association constraint (H_A)

UEFA draw defines the admissible matchings Γ via a set of match exclusions $H \subset \mathcal{R} \times \mathcal{W}$, where the overall exclusion set $H = H_A \cup H_G$ is the union of the association-level exclusions H_A and group-level exclusions H_G. The precise set H varies across seasons depending on the group-level assignment and the composition of teams in the R16. The admissible matching set for the UEFA implementation of the constrained dynamic draw is defined by

$$\Gamma_H := \left\{ V \in \mathcal{V} \middle| V \cap H = \emptyset \right\},$$

where the draw induces the random matching $\psi(\Gamma_H)$.

3.4.2 Measures of Distortions

After discussing the theory for the current UEFA procedure, we start our empirical analysis with an illustrative example from the UCL R16 draw in 2018. The expected assignment matrix, $\hat{\mathbf{A}}$, under the current UEFA draw procedure is given in Table 3.1. Each row represents a group winner, and each column a runner-up, so the row-i–column-j cell indicates the probability that the (ij)-pair is selected within the realized R16 matching.[20]

[20]We calculate all probabilities with a Monte Carlo simulation of size $N = 10^6$. At this size, 95 percent confidence intervals for each probability are within ± 0.001 of the given coefficient (see Appendix to the Working Paper).

The constraints in the 2018 draw are as follows: First, along the diagonal, the probabilities of each match are zero, reflecting the eight group constraints.[21] Second, seven same-nation matches are excluded reflecting the 2018-specific association constraint. Finally, all rows and columns sum to exactly one, as each represents the marginal match distribution for the respective team through the bipartite constraint.[22]

Despite having random urn draws at each point in time, the likelihoods of two teams playing each other are not uniform due to asymmetry generated by the association constraint. For illustration, consider Paris Saint-Germain (PSG) in 2018, the second row of Table 3.1. As PSG is the only French team in the R16 in the 2018 season it has no same-nation exclusions and thus, seven feasible match partners. However, the likelihoods of each of the seven match-ups varies substantially—with the probability of PSG playing Chelsea almost three times larger than that of PSG playing either Basel, Shaktar, or Porto (columns 1, 6, and 7)

In the following, we quantify the *total* effect of the association constraint, consisting of both the *direct* effect on team i from the ij exclusion as well as the *indirect* spillover effect from others' constraints (i.e., how the jk exclusion affects the chances of the ik match-up, $i \neq j$). For illustration, consider the match probabilities for Real Madrid and Barcelona in the 2018 season (see Table 3.1, column 8 and row 4, respectively). While the Read-Madrid and Barcelona exclusion directly benefits the two Spanish teams (these are two of the three strongest teams that season according to our ability index), Barcelona additionally profits from substantial spillovers generated by six other same-nation exclusions. In particular, consider Barcelona and two other unconstrained group winners, PSG and Besiktas (each with seven potential match partners, rows 2 and 7, respectively). Even though none of the three aforementioned group winners are constrained from matching with the two lowest-performing teams (Basel and Shaktar Donetsk, columns 1 and 6) Barcelona is 1.35 times more likely to match to either of them than are either PSG or Besiktas, the less-constrained group winners.

[21] The bipartite and group constraints impose symmetric restrictions, leading to an equal probability of matching with every non-excluded partner. Consequently, without the association constraint, the expected assignment would have a one-in-seven chance for each off-diagonal entry.

[22] Note that the constraints are not mutually exclusive and consequently, even though the expected assignment is an 8×8 matrix, it has 34 degrees of freedom. The expected assignment matrices for the R16 draw in the other 15 seasons can be found in the Online Supplementary Material.

Holding constant the matching exclusions, distortions caused by the direct effects are an unavoidable consequence. However, the indirect effects have the potential to be ameliorated through better randomization. In order to quantify the size of the indirect effects of the association constraint we construct a fairness spillover measure designed to comparatively assess the matching chances of pairwise comparable team pairs.

We regard two teams i and j pairwise comparable with respect to third team k if both the ik and jk matches are not directly excluded. As such, by construction the designed spillover measure only compares team pairs that are either both group winners (members of $\mathcal{W}$) or both runners-up ($\mathcal{R}$). Given a set of match constraints H the set of admissible match comparisons is:

$$\Upsilon_H := \{(ik, jk) \,|i, j \in \mathcal{W}, k \in \mathcal{R}, ik, jk \notin H\} \bigcup \{(ki, kj) \,|k \in \mathcal{W}, i, j \in \mathcal{R}, ki, kj \notin H\}.$$

We define the ik and jk match-ups as being *fairer* the smaller the distance between the ik and jk likelihoods, where i and j are pairwise comparable with respect to k. For any expected assignment $\mathbf{A}$ our fairness objective measures the average absolute difference between all pairwise comparable teams as:

$$Q(\mathbf{A}; H) = \frac{1}{|\Upsilon_H|} \sum_{(ik,jk)\in\Upsilon_H} |a_{ik} - a_{jk}|.$$

In what follows, we quantify the total and indirect effects of distortions induced by the R16 matching constraints using a structural model of game outcomes estimated using historical (and out-of-sample) data from the 2004–19 UCL seasons. In particular, by simulating the R16 draw and all subsequent games within the tournament, we show that direct effects of tournament matching constraints have major effects on teams' expected earnings and progression probabilities within the tournament.

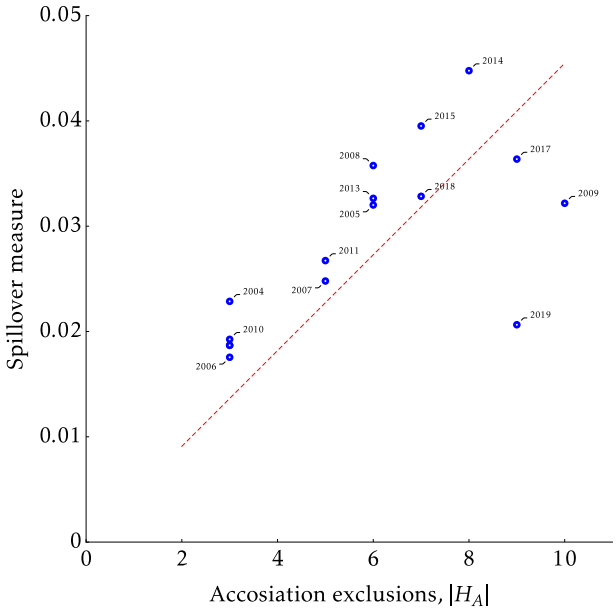

Figure 3.2: Spillover measure versus number of same-nation exclusions. Red dashed line indicates fitted linear relationship

3.4.3 Data and Estimation of Game-Outcome Model

In order to account for variation in teams' ability while examining potential effects driven by the tournament's constraints, we estimate a commonly used structural model for football-game outcomes: the bivariate Poisson (Maher, 1982; Dixon and Coles, 1997).

Model. Let S_i and S_j be the random variables indicating the number of goals scored by home-team i and guest-team j in a given game. In a bivariate Poisson model with parameters $(\lambda_1, \lambda_2, \lambda_3)$ the realized scoreline (s_i, s_j) has a joint probability distribution given by

$$\Pr_{(S_i,S_j)}(s_i, s_j) = \exp\{-(\lambda_1 + \lambda_2 + \lambda_3)\} \frac{\lambda_1^{s_i}}{s_i!} \frac{\lambda_2^{s_j}}{s_j!} \sum_{k=0}^{\min(s_i,s_j)} \binom{s_i}{k} \binom{s_j}{k} k! \left(\frac{\lambda_3}{\lambda_1\lambda_2}\right)^k,$$

where $\mathbf{E}[S_i] = \lambda_1 + \lambda_3$, $\mathbf{E}[S_j] = \lambda_2 + \lambda_3$ and $\mathbf{Cov}(S_i, S_j) = \lambda_3$.

In our specification we follow Karlis and Ntzoufras (2003) and assume that $\ln \lambda_1 = \mu^t + \eta^t + \alpha_i^t - \delta_j^t$, $\ln \lambda_2 = \mu^t + \alpha_i^t - \delta_j^t$, and $\lambda_3 = \rho^t$. In this specification α_k^t and δ_k^t measure the idiosyncratic offensive and defensive abilities for team k in season t (larger values indicating

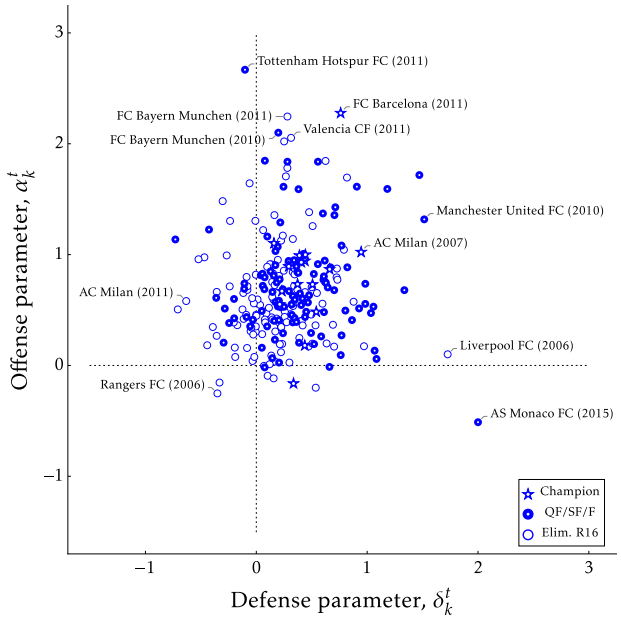

Figure 3.3: Estimated offense and defense parameters for the R16 teams (2004–19). Figure excludes estimated parameter pairs for teams that fail to reach the R16

greater ability), while μ^t denotes the season-specific constant and η^t the season-specific home-advantage parameter.

We estimate the above model via constrained maximum likelihood separately for each of the sixteen seasons t from 2004 to 2019. For scale identification we impose two sum-to-zero constraints in each season, forcing $\sum_k \alpha_k^t = \sum_k \delta_k^t = 0$. For estimation in season t we rely on game-level data from the group stage in season t and the group and knock-out stages (except for the final game which is played on a neutral soil) in seasons $t-1$ and $t-2$.[23]

In Figure 2 we graph the estimated parameters (defense on the horizontal axis, offense on the vertical) for the subset of teams reaching the R16 in the 2004–19 seasons, dropping those that fail to get past the group stage. The strongest teams have large positive values for both the offense and defense parameters; see for example Manchester United in 2010 and

[23]This results in a total of 408 game-level observations used in the estimation for the 2004 season, 376 observations for the 2005 season, and 348 observations for each season between 2006 and 2019. The differences in the number of observations across years result from a change to the tournament design in the 2004 season, where a second group stage feeding into the quarter-finals was replaced by the R16.

Table 3.2: Summary statistics for estimated parameters (R16 teams)

Stage	**Offense parameter**, α_k^t				**Defense parameter**, δ_k^t			
	Mean	Med.	Min	Max	Mean	Med.	Min	Max
Elim. in R16	0.62	0.51	-0.20	2.25	0.27	0.17	-0.58	10.99
Elim. in QF	0.69	0.64	-0.51	2.67	0.31	0.26	-0.73	2.00
Reach SF	0.78	0.74	-0.43	2.28	0.66	0.44	-0.43	11.52

FC Barcelona in 2011. Conversely, low-performing teams have either a negative value for the offense parameter (AS Monaco in 2015), the defense parameter (AC Milan in 2011), or both (Rangers in 2006). The large mass of teams are low-to-medium-strength with small but positive values over both the offense and defense parameters.[24]

Complementing the figure, Table 3.2 presents summary statistics for the estimated offense and defense parameters broken out by the realized stage reached within the tournament. We find that the eight teams that progress to the quarter-finals are stronger both offensively and defensively than the teams eliminated in the R16. Similarly, the four teams advancing to the semi-finals have better offensive and defensive performance relative to those knocked out in the quarter-finals. A two-sample Kolmogorov-Smirnov test confirms the pattern, indicating that the empirical distributions of the offense and defense parameters in the R16 and the semi-finals are statistically different ($p = 0.004$ for defense, $p < 0.001$ for offense).

3.4.4 Effects from the Constraints

In what follows, we separately quantify the *total* and *spillover* effects from the constraints on two main metrics for teams' outcomes: their expected prize money and the probability of

[24]See Appendix to the Working Paper for estimates for the constant term μ^t, the home-advantage parameter η^t, and the correlation coefficient between the number of goals scored by opposing teams ρ^t in the seasons 2004–19.

reaching the semi-final stages of the competition (and beyond).[25] In summary, we find that:

Result 1. *The association constraint in the R16 generates substantial effects by: (i) altering expected tournament prizes by millions of euro; (ii) significantly affecting the chances of reaching later stages of the tournament; and (iii) creating spillovers to the matching chances of pairwise comparable teams.*

Evidence: We combine our model of the exact draw procedure and the estimated bivariate Poisson model to calculate expected tournament outcomes for each R16 team i in season t. We conduct our draw simulations twice, once under the current set of UEFA constraints (valid assignments in $\Gamma_{H_G \cup H_A}$), and once under a counterfactual set of constraints that drops the association constraint entirely (valid assignments in Γ_{H_G}).[26,27]

For each team i in season t we calculate the *association-constraint effect* as the difference in expected prizes between the current draw mechanism and the counterfactual draw where we drop the association constraint entirely:

$$\Delta\pi_i^t := \hat{\mathbf{E}}\left[\pi_i^t \middle| \Gamma_{H_A \cup H_G}\right] - \hat{\mathbf{E}}\left[\pi_i^t \middle| \Gamma_{H_G}\right].$$

Teams with a positive value of $\Delta\pi_i^t$ are those benefiting from the association constraint, whereas those with a negative value are being disadvantaged. Across all 16 UCL seasons, the association-constraint effect has a standard deviation of 0.3 million euro (it is mean-zero by construction within each season) and a range of 2 million euro: a cost of 0.8 million euro to Arsenal in the 2014 season (eliminated in the R16) and a subsidy of 1.2 million euro to Real Madrid in the 2017 season (went on to become European champion). The effects from enforcing the association constraint are therefore substantial.

[25] As mentioned in Section 3.3, in addition to the direct prize money, teams' revenue stem from the market pool share and in-stadium game attendance. These other earnings scale with team's progression within the tournament, and their local media-markets. Therefore, our prize-money measure serves as a lower-bound on the true underlying effects, where UEFA financial reports indicate that total effects are approximately twice as large as the prize money effect (though this varies by local media markets).

[26] In detail, for both the actual and counterfactual calculations we proceed by first drawing $J = 1,000$ R16 matchings, $\{V_j\}_{j=1}^J$, using the relevant Γ_H-constrained $\mathcal{R}$-to-$\mathcal{W}$ dynamic draw. For each drawn R16 matching V_j we simulate the remaining tournament outcomes $S = 1,000$ times (the R16 home/away games, quarter- and semi-final home/away games, and the final game on neutral soil). Consequently, each season is simulated one million times.

[27] Given the double-loop for the simulation we also calculate accurate metrics for the expected prize money conditional on each R16 draw as $\hat{\mathbf{E}}\left[\pi_i^t \middle| V_j\right] := \frac{1}{S}\sum_{s=1}^S \pi_{i,js}^t$. The unconditional expected prize is then calculated as $\hat{\mathbf{E}}\left[\pi_i^t \middle| \Gamma_H\right] := \frac{1}{J}\sum_{j=1}^J \hat{\mathbf{E}}\left[\pi_i^t \middle| V_j\right]$.

In order to validate the measured association-constraint effect, we further demonstrate that it is predictive of realized outcomes even after controlling for teams' ability. For each season t we construct a zero-to-one ability index for the R16 teams using the estimated bivariate Poisson model.[28] By way of example, for the 2018 season, our ability index runs from Shaktar Donetsk at 0, FC Basel at 0.134 and Besiktas at 0.233, up to Real Madrid at 0.919, Liverpool at 0.999, and Barcelona at 1. Regressing the realized tournament prizes for each team-year observation on the ability index, we extract the fitted residuals as a measure of the prize-money outcome that is orthogonal to teams' ability.[29] Figure 3.4 illustrates the relationship between the association-constraint effect $\Delta\pi_i^t$ on the horizontal axis, and the fitted residuals from the regression of realized prizes on teams' ability on the vertical axis. Even though the measured association-constraint effect explains only 4.7 percent of the total variation in realized prizes after controlling for ability, our model-generated measure is statistically significant at any conventional significance level ($p < 0.001$).

A similar exercise conducted using a probit model, suggests that both the ability index and the association-constraint effect are significant predictors of teams' success in the final stages of the competition. While a unit shift in the ability index—moving from the worst to the best team—increases the likelihood of a semi-final appearance by 67.3 percent ($p < 0.000$), a measured one-million-euro subsidy from the association constraint increases the likelihood by 19.7 percent ($p = 0.004$).[30]

After characterizing the total effects from the constraints, in Figure 3.2, we graph the fairness objective for each season t under the current UEFA draw procedure, against the

[28] For each season t and team i, we calculate the average probability that team i wins a game against each of the other R16 teams that season using the bivariate Poisson model as:

$$\omega_i^t = \frac{1}{15}\sum_{j\neq i}^{15} \widehat{\Pr}\left(S_i > S_j\right).$$

Next, we re-scale the average probabilities to run from zero to one at the season level as $\tilde{\omega}_i^t = \frac{\omega_i^t - \underline{\omega}^t}{\overline{\omega}^t - \underline{\omega}^t}$ where $\overline{\omega}^t$ and $\underline{\omega}^t$ are the best and worst values for ω_i^t in season t.

[29] We find that a unit increase in the ability index (i.e. going from the worst to the best team in a given season) increases the expected prize money by 17.8 million euro.

[30] The association-constraint effects for reaching later stages of the tournament are diminishing in magnitude. Whereas a one-million-euro association-constraint subsidy increases the chances of reaching the quarter-finals by 31.2 percent ($p = 0.001$), the chances of reaching the final are raised by only 9 percent ($p = 0.086$).

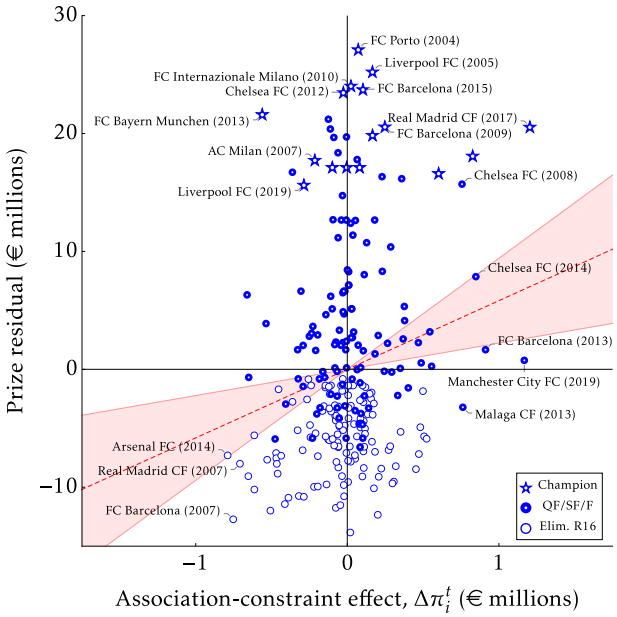

Figure 3.4: Prize-money residuals versus association-constraint effect. Red dashed line indicates a fitted linear regression; shaded band indicates 95 percent confidence interval on the estimated relationship

number of same-nation exclusions. Since the relationship in question is strongly positive, we conclude that the association constraint substantially distorts match chances of pairwise comparable teams. Specifically, the estimated slope coefficient from a regression of $Q\left(\hat{\mathbf{A}}^t\right)$ on the number of same-nation exclusions suggests an average wedge in match-likelihoods of approximately 5 percentage points for every ten exclusions.[31] This represents a relative swing of up to a third for unconstrained teams. Although this result points to quantitatively large spillovers even after accepting the constraints' direct effect, we next show that there is only limited scope to ameliorate the spillovers through better randomization procedures.

[31] In the absence of any same-nation exclusions our measure is zero by construction; we therefore estimate relationships without a constant.

3.5 Near-Optimality of the Current UEFA Procedure

A natural question raised by Result 1 is whether there exists a randomization procedure that generates less distortions than the one currently employed. In order to examine the extent to which fairer random assignments might exist we employ the core result in Budish et al. (2013) that guarantees the *existence* of an equivalent randomization over assignments in Γ_H for *every* feasible *expected* assignment. This allows us to relax the constrained assignment problem over discrete final matchings to one of finding *expected* assignments allowing for fractional (and continuous) assignment in the analysis.

3.5.1 Reduction of the Assignment Problem's Dimension

First, notice that any assignment V can be rewritten as a matrix $\mathbf{X}(V) \in \{0,1\}^{K\times K}$ with a generic entry $x_{ij}(V) = \mathbf{1}\{r_i w_j \in V\}$ indicating whether or not runner-up r_i is matched to winner w_j. Since V represents a perfect matching between $\mathcal{R}$ and $\mathcal{W}$, $\mathbf{X}(V)$ is a rook-matrix where each row and column have exactly one unit-valued entry with all other entries equal to zero. Second, for any random draw over Γ_H, the expected assignment matrix is defined as $\mathbf{A} := \mathbf{E}\mathbf{X}(V) = \sum_{V\in\Gamma_H} \Pr\{V\}\cdot\mathbf{X}(V)$, with the generic entry a_{ij} representing the probability of the $r_i w_j$ match.

An expected assignment matrix $\mathbf{A}$ in our setting *satisfies the matching constraints* if:

1. Each entry a_{ij} can be interpreted as the probability of $r_i w_j$ being part of V $(\forall i,j : 0 \leq a_{ij} \leq 1)$;
2. Excluded entries have zero probability $(\forall r_i w_j \in H : a_{ij} = 0)$;
3. Each row and column can be interpreted as the marginal probability distribution for the respective team $(\forall i,j : \sum_{k=1}^{K} a_{kj} = \sum_{k=1}^{K} a_{ik} = 1)$.

While satisfying these matching constraints is clearly a necessary condition for any expected assignment resulting from a randomization over the feasible assignment set Γ_H, the following indicates it is also sufficient.

Proposition 3 (Implementability)**.** *For any expected assignment matrix* $\mathbf{A}$ *satisfying the matching constraints there exists an equivalent randomization in* $\Delta\Gamma_H$*.*

Proof. See Appendix to the Working Paper. □

An implication of Proposition 3 is that for any maximization problem over Γ_H (a space with $O(K!)$ degrees of freedom), if we can express the objective in terms of expected assignments, then it is without loss of generality to consider maximization over the space of expected assignment matrices satisfying the constraints (a space with $O(K^2)$ degrees of freedom). Therefore, for our specific UEFA application with $K = 8$, the theorem reduces the degrees of freedom from 2,000–10,000 across the sixteen seasons we look at to 30–40.

3.5.2 Search for Optimal Assignment Rules

Proposition 3 in Section 3.5.1 states that a feasible assignment procedure exists for every feasible expected assignment matrix satisfying the constraints. This result substantially simplifies the search for a randomization procedure that generates less distortions than the one currently employed, by reducing the effective degrees of freedom by two orders of magnitude, allowing for a computationally tractable optimization over the expected assignment $\mathbf{A}_t^\star$.

We define the optimal expected assignment as one that solves the following optimization problem:

$$\mathbf{A}_t^\star := \underset{\mathbf{A}}{\operatorname{argmin}}\, Q(\mathbf{A}; H^t),$$

subject to the matching constraints:

(i) $\forall ij \in H^t : a_{ij} = 0$; (ii) $\forall ij : 0 \leq a_{ij} \leq 1$; (iii) $\forall i : \sum_k a_{ik} = \sum_k a_{ki} = 1$,

where H^t denotes the set of association and group match exclusions in season t and $Q(\cdot)$ is the fairness spillover measure.

By comparing the optimal expected assignments and expected assignments under the current procedure we arrive at the main result of this section which is:

Result 2. *While the UEFA assignment rule is not optimal given the constraints, it comes very close to the optimal procedure when considering the fairness spillover measure.*

In Figure 3.5(A) we graph the spillover measure for the optimal expected assignments $\mathbf{A}_t^\star$ against the spillover measure for the expected assignments $\hat{\mathbf{A}}_t$ under the current procedure, across the 2004–19 seasons. While some improvement is possible across the tournament years, the gains are marginal (on average a less than 10 percent relative reduction in the size of the spillovers).[32,33]

Against the small potential benefits, there are large prospective implementation costs associated with modifying the existing assignment rule, as procedures yielding the optimal expected assignments $\mathbf{A}_t^\star$ are potentially quite complex in comparison to the one currently employed.[34] Put against this transparency cost, a reduction in the match-up distortions from a 5 percent average difference under the current procedure to a 4.5 percent difference for an optimal procedure seems marginal.[35,36]

3.6 Weakening the Constraint Set of the Current UEFA Procedure

One response to Result 2 is to accept the current UEFA assignment rule and the distortions it generates. In this final section, however, we take a different approach, and examine the extent to which gains can be made by weakening the association constraint. Specifically, we investigate the extent to which the association constraint can be partially relaxed while still protecting the tournament from excessive R16 same-nation match-ups. There

[32]The size of the reduction is given by the estimated slope coefficient from a regression of the spillover measure under the optimal randomization against the current procedure for all years $t = 2004, ..., 2019$.

[33]One potential objection is that we are not using the right objective function $Q(\cdot)$. While we are amenable to suggestions for other objectives, we have additionally tried minimizing the square differences between match probabilities, the differences between the maximal and minimal positive-probability matches for each team, as well as a measure based on the Kullbeck-Leibler divergence. None of these showed economically meaningful gains from optimization, where the interpretation of the objective became harder than the "average difference" in match likelihoods interpretation for our chosen spillover measure.

[34]See Online Appendix B to Budish et al. (2013) for a construction.

[35]While there may exist a simple modification of the current assignment rule that would result in fairer match-ups, none of the distinct procedures detailed in Proposition 2 achieve such an end (see Appendix to the Working Paper). Therefore, simple modifications such as changing the order of the draw to $\mathcal{W}$-to-$\mathcal{R}$ or replacing sequential team draws with draws of team-pairs do not provide improvements over the current procedure.

[36]We should note that the inability to improve upon the expected assignments generated under the UEFA procedure is not driven by a limited scope in moving the expected assignments under the constraints. Taking the 2018 season as a (fairly representative) example, we can obtain any value for our spillover measure from a minimum of 0.03 to an upper bound of 0.32.

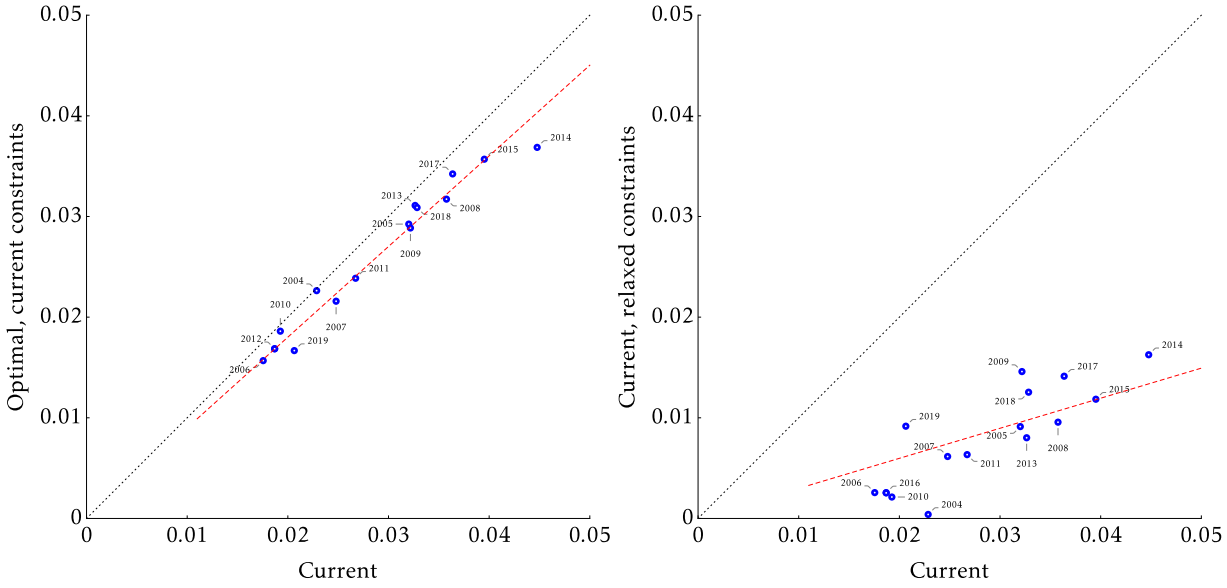

Figure 3.5: Spillover measures: Counterfactuals versus actual. Red dashed line indicates fitted linear relationship

are several practicable ways in which this could be accomplished, but in what follows, we focus on a procedure that relaxes the association constraint while only marginally modifying the current draw procedure. Specifically, we study an alternative to the current association constraint where we allow *at most one* same-nation match in the R16. Therefore, we can continue to use the constrained $\mathcal{R}$-to-$\mathcal{W}$ dynamic draw detailed in Section 3.4.1, with a sole modification to the (now expanded) admissible set given by

$$\tilde{\Gamma}_{H_A,H_G} := \Big\{ V \in \mathcal{V} \Big| V \cap H_G = \emptyset \text{ and } |V \cap H_A| \leq 1, \Big\}.$$

As such, the relaxation retains the desirable features of the current procedure: the randomization is transparent (random draws from a small-sample urn) and the more-opaque combinatoric check continues to be fully verifiable at all points during the draw.

Under this relaxation we find that:

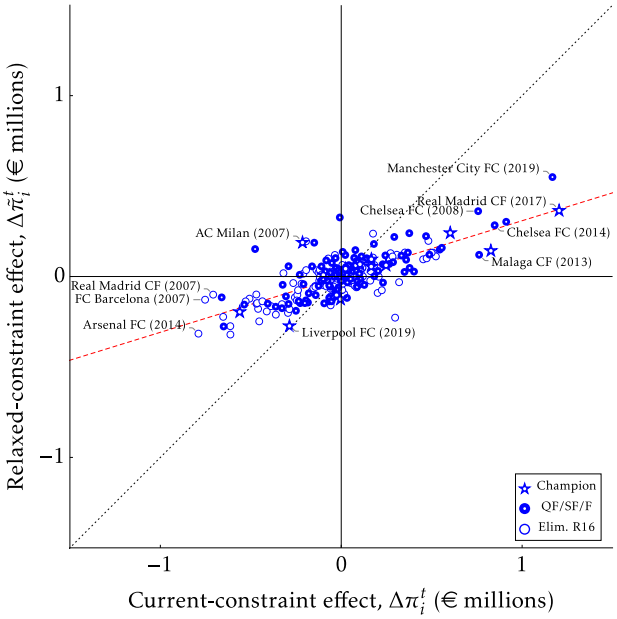

Figure 3.6: Relaxed-constraint effect versus current-constraint effect. Red dashed line indicates fitted linear relationship

Result 3. *Weakening the association constraint to allow for at most one same-association match in the R16 substantially reduces the distortions, while protecting associations from excessive same-nation match-ups. Moreover, as a secondary effect, weakening the association constraint (mechanically) reduces the number of same-nation games in the later stages of the tournament.*

Evidence: We start by constructing analog results to those presented in Section 3.4.4. In Figure 3.5(B), we illustrate the fairness spillover measure for the UEFA procedure with the at-most-one same-association constraint set on the vertical axis against that for the UEFA draw under the default constraint set. We find that allowing for a single same-nation match in the R16 decreases the total distortions by more than 70 percent. This is a sizable reduction, especially when compared to a 10 percent reduction obtained under the optimal assignment that maintains the current constraint structure (see Figure 3.5(A)).

Next, we define the *relaxed-constraint effect* as

$$\Delta\tilde{\pi}_i^t := \hat{\mathbf{E}}\left[\pi_i^t \middle| \tilde{\Gamma}_{H_A,H_G}\right] - \hat{\mathbf{E}}\left[\pi_i^t \middle| \Gamma_{H_G}\right],$$

where $\hat{\mathbf{E}}\left[\pi_i^t \middle| \tilde{\Gamma}_{H_A,H_G}\right]$ is the expected prize for team i in season t under the at-most-one same-association draw, and $\hat{\mathbf{E}}\left[\pi_i^t | \Gamma_{H_G}\right]$ is the expected prize in the absence of any association constraints as before.[37] In Figure 3.6, we illustrate the relationship between $\Delta\tilde{\pi}_i^t$ and the association constraint effect $\Delta\pi_i^t$ defined in Section 3.4.4. We conclude that allowing for a single same-association match in the R16 reduces the total prize distortions by 71 percent.[38]

The above demonstrates a substantial reduction in the matching distortions with only a slight relaxation of the association constraint. However, there are presumably nontrivial costs associated with allowing for same-nation R16 matches. Relaxing the association constraint as we have done leads to a single same-nation match-up in the R16 in approximately six out of every ten tournaments. In seasons with at least six same-nation exclusions, this ratio increases to seven-in-ten. Since the association constraint is imposed intentionally, UEFA likely has a clear underlying preference for the tournament to be primarily an international competition in its early stages.[39]

As a final point in favor of our relaxation approach, we show that, perversely, imposing same-nation exclusions at earlier stages of the tournament has the effect of increasing the likelihood of same-nation match-ups in the subsequent rounds. Using our estimated model of goal outcomes and the at-most-one same-association match mechanism, we assess the predicted change in the number of same-nation matches in later stages of the tournament. We find that for every same-association pairing generated in the R16, there is a 0.10 reduction in the same-nation games in later stages of the tournament.[40] In Figure 3.7 we illustrate these two compensating effects in all seasons between 2004 and 2019.

[37] We numerically calculate $\hat{\mathbf{E}}\left[\pi_i^t | \tilde{\Gamma}_{H_A,H_G}\right]$ via the same Monte Carlo simulation method used in Section 3.4.4 and described in footnotes 26 and 27.

[38] The estimated slope coefficient from a regression of the relaxed-constraint effect on the current-constraint effect, across all R16 teams in all seasons between 2004 and 2019, is equal to 0.29.

[39] This, however, cannot be determined, since "the identification of design constraints is usually more difficult because they are rarely communicated by the organizers" (Csató, 2018).

[40] The effect would be larger if we additionally accounted for the same-nation exclusions in the group stage that precedes the R16.

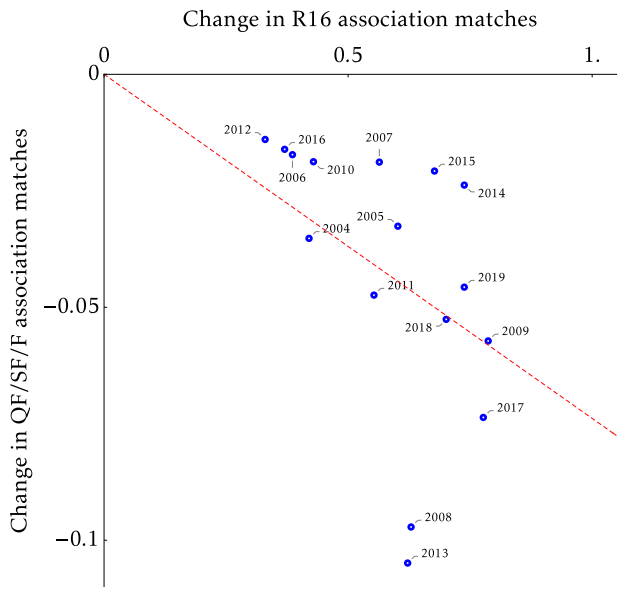

Figure 3.7: Change in same-nation match-ups in QF/SF/F verus the R16 after relaxing the association constraint. Red dashed line indicates fitted linear relationship (forced intercept)

3.7 Conclusion

We document a constrained-assignment problem—one with huge public interest, and with millions of euro at stake from the outcome—where the randomization is primarily focused on transparency and credibility, to both participants and to the general public. This is in contrast to many market-design solutions that are focused on more-quantifiable objectives: efficiency, fairness, and/or strategic compatibility.While many assignment procedures will choose rules that achieve or maximize these theoretical objectives, the resulting algorithms can seem opaque; to the extent that participants can misunderstand inherent features of the design (for example, not using dominant strategies in deferred acceptance, see Rees-Jones and Skowronek, 2018). In such situations, managers and principals may choose to focus instead on a transparent (but ad hoc) assignment rule; especially, when the designer lacks credibility with the procedure's participants.

In the present paper we ask the following questions: What are the losses from a procedure focused on transparency? How does a particular assignment compare to an optimal one? In our application, we illustrate how the tools of market design can be brought to bear on these questions. In our assessment of the UEFA draw mechanism we show that the enforced matching constraints significantly distort the tournament's outcomes. However, we also demonstrate that the chosen procedure is very close to a constrained best in terms of reducing distortions in the likelihoods of unconstrained team-pair matches.

Our methodology relies on a recently introduced market-design tool (Budish et al., 2013) that allows us to simplify the domain of analysis without loss of generality, switching to a consideration of expected assignments. This shift in domain not only allows us to more compactly specify a clear and easy-to-interpret objective, more importantly, it makes optimization possible by reducing the number of degrees of freedom by two orders of magnitude. Through this shift we are able to show that the UEFA mechanism is remarkable: it solves a complex combinatoric problem, but in a way that is both transparent and comprehensible to the general public and fully verifiable by more-sophisticated third parties. Not only that, the developed procedure comes very close to achieving a first-best outcome under the desired set of constraints.

Bibliography

A. Abdulkadiroğlu and T. Sönmez. School choice: A mechanism design approach. *American Economic Review*, 93(3):729–47, 2003.

M. Akbarpour and S. Li. Credible auctions: A trilemma. *Econometrica*, forthcoming.

S. An and F. Schorfheide. Bayesian analysis of DSGE models. *Econometric Review*, 26(2-4): 113–172, 2007.

N. Anbarci, C.-J. Sun, and M. U. Ünver. Designing fair tiebreak mechanisms: The case of fifa penalty shootouts. Boston College working paper, January 2015.

J. Apesteguia and I. Palacios-Huerta. Psychological pressure in competitive environments: Evidence from a randomized natural experiment. *American Economic Review*, 100(5): 2548–64, 2010.

A. B. Atkinson, T. Piketty, and E. Saez. Top incomes in the long run of history. *Journal of Economic Literature*, 49(1):3–71, 2011.

M. Baccara, A. İmrohoroğlu, A. J. Wilson, and L. Yariv. A field study on matching with network externalities. *American Economic Review*, 102(5):1773–804, 2012.

L. A. Bebchuk and J. M. Fried. Executive compensation as an agency problem. *Journal of Economic Perspective*, 17(3):71–92, 2003.

V. Bhaskar. Rational adversaries? evidence from randomised trials in one day cricket. *Economic Journal*, 119(534):1–23, 2008.

N. Bhutta, A. C. Chang, L. J. Dettling, and J. W. Hsu. Disparities in wealth by race and ethnicity in the 2019 survey of consumer finances. *FEDS Notes. Washington: Board of Governors of the Federal Reserve System*, 315(16), 2020.

P. Biemer and L. Lyberg. *Introduction to survey quality*. John Wiley & Sons, Hoboken, New Jersey, 2003.

H. J. Bierens and L. F. Martins. Time-varying cointegration. *Econometric Theory*, 26(5): 1453–1490, 2010.

J. Bivens and L. Mishel. The pay of corporate executives and financial professionals as evidence of rents in top 1 percent incomes. *Journal of Economic Perspective*, 27(3):57–78, 2013.

O. Blanchard. Do DSGE models have a future? *Peterson Institute for International Economics: Policy Brief 16-11*, 2016.

I. Bó and L. Chen. Designing heaven's will: Lessons in market design from the chinese imperial civil servants match. *Working Paper*, 2019.

A. Bogomolnaia and H. Moulin. A new solution to the random assignment problem. *Journal of Economic theory*, 100(2):295–328, 2001.

J. Bricker, A. Henriques, J. Krimmel, and J. Sabelhaus. Measuring income and wealth at the top using administrative and survey data. *Brookings Papers on Economic Activity*, Spring:261–331, 2016.

J. Bricker, A. Henriques, and P. Hansen. How much has wealth concentration grown in the united states? a re-examination of data from 2001-2013. *Finance and Economics Discussion Series*, 2018-024:1–52, 2018.

V. L. Bryant, J. L. Czajka, G. Ivsin, and J. Nunns. Design changes to the soi public use file (puf). *Prepared for the"New Resources for Microdata-Based Tax Analysis" Session of the 107th Annual Conference on Taxation, National Tax Association Santa Fe, New Mexico, November 15*, pages 1–19, 2014.

E. Budish and E. Cantillon. The multi-unit assignment problem: Theory and evidence from course allocation at harvard. *American Economic Review*, 102(5):2237–71, 2012.

E. Budish, Y.-K. Che, F. Kojima, and P. Milgrom. Designing random allocation mechanisms: Theory and applications. *American Economic Review*, 103(2):585–623, April 2013. doi: 10.1257/aer.103.2.585. URL http://www.aeaweb.org/articles?id=10.1257/aer.103.2.585.

R. J. Caballero. Macroeconomics after the crisis: Time to deal with the pretense-of-knowledge syndrome. *NBER Working Papers 16429*, 2010.

F. Canova and F. J. Pérez Forero. Estimating overidentified, nonrecursive, time-varying coefficients structural vector autoregressions. *Quantitative Economics*, 6(2):359–384, 2015.

D. Card, R. Chetty, M. Feldstein, and E. Saez. Expanding access to administrative data for research in the united states. NSF SBE 2020 White Paper, National Science Foundation Directorate of Social, Behavioral, and Economic Sciences, Arlington, VA, 2010.

A. Cardinali and G. P. Nason. Costationarity of locally stationary time series. *Journal of Time Series Econometrics*, 2(2):1–33, 2010.

C. D. Carroll. Saving and growth with habit formation. *American Economic Review*, 90(3): 341–355, 2000.

J. L. Castle, N. W. P. Fawcett, and D. F. Hendry. Forecasting with equilibrium-correction models during structural breaks. *Journal of Econometrics*, 158(1):25–36, 2010.

J. L. Castle, M. P. Clements, and D. F. Hendry. An overview of forecasting facing breaks. *Journal of Business Cycle Research*, 12(1):3–23, 2016.

V. V. Chari, P. J. Kehoe, and E. R. McGrattan. Business cycle accounting. *Econometrica*, 75(3):781–836, 2007.

V. V. Chari, P. J. Kehoe, and E. R. McGrattan. New Keynesian models: Not yet useful for policy analysis. *American Economic Journal: Macroeconomics*, 1(1):242–266, 2009.

R. Chetty, M. Stepner, and S. Abraham. The association between income and life expectancy in the united states, 2001-2014. *JAMA*, 315(16):1750–1766, 2016.

P.-A. Chiappori, S. Levitt, and T. Groseclose. Testing mixed-strategy equilibria when players are heterogeneous: The case of penalty kicks in soccer. *American Economic Review*, 92 (4):1138–51, 2002.

L. J. Christiano and J. M. Davis. Two flaws in business cycle accounting. *NBER Working Papers 12647*, 2006.

L. J. Christiano, M. S. Eichenbaum, and M. Trabandt. On dsge models. *Journal of Economic Perspectives*, 32(3):113–140, 2018.

A. Clarkwest. Neo-materialist theory and the temporal relationship between income inequality and longevity change. *Social Science & Medicine*, 66(9):1871–1881, 2008.

M. P. Clements and D. F. Hendry. On the limitations of comparing mean square forecast errors. *Journal of Forecasting*, 12(8):617–637, 1993.

L. Csató. A simulation comparison of tournament designs for world men's handball championships. arXiv:1803.10975v3, March 2018.

J. L. Czajka, B. Kirwan, and A. Sukasih. An assessment of the need for a redesign of the statistics of income individual tax sample. *Mathematica Policy Research*, pages 1–95, 2014.

D. Dagaev and K. Sonin. Winning by losing: Incentive incompatibility in multiple qualifiers. *Journal of Sports Economics*, 19(8):1122–46, 2018.

M. Daly. *Killing the competition: Economic inequality and homicide.* Routledge, London, 2016.

H. A. David and H. N. Nagaraja. *Order Statistics, Third Edition.* Hoboken, New York–Chichester–Brisbane–Toronto–Singapore, 2003.

R. de Vries, S. Gosling, and J. Potter. Income inequality and personality: Are less equal u.s. states less agreeable? *Social Science & Medicine*, 72(12):1978–1985, 2011.

A. Deaton. Saving and liquidity constraints. *Econometrica*, 59(5):1221–1248, 1991.

D. N. DeJong and C. Dave. *Structural macroeconomics: Second edition.* Princeton University Press, 2011.

D. N. DeJong, R. Liesenfeld, and J.-F. Richard. A nonlinear forecasting model of gdp growth. *The Review of Economics and Statistics*, 87(4):697–708, 2005.

D. N. DeJong, R. Liesenfeld, G. V. Moura, J.-F. Richard, and H. Dharmarajan. Efficient likelihood evaluation of state-space representations. *Review of Economic Studies*, 80(2): 538–567, 2013.

M. Del Negro and F. Schorfheide. Forming priors for DSGE models (and how it affects the assessment of nominal rigidities). *Journal of Monetary Economics*, 55(7):1191–1208, 2008.

M. J. Dixon and S. G. Coles. Modelling association football scores and inefficiencies in the football betting market. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 46(2):265–80, 1997.

U. Dur, P. A. Pathak, and T. Sönmez. Explicit vs. statistical preferential treatment in affirmative action: Theory and evidence from chicago's exam schools. Working Paper 22109, National Bureau of Economic Research, March 2016.

G. Elliott and A. Timmermann. Forecasting in economics and finance. *Annual Review of Economics*, 8:81–110, 2016.

N. R. Ericsson and A. B. Martinez. *Evaluating government budget forecasts, in: Williams D., Calabrese T. (ed.), The Palgrave Handbook of Government Budget Forecasting*. Palgrave Macmillan, Cham, Switzerland, 2019.

R. Foellmi, S. Legge, and L. Schmid. Do professionals get it right? limited attention and risk-taking behaviour. *Economic Journal*, 126(592):724–55, 2016.

G. R. Fréchette, A. E. Roth, and M. U. Ünver. Unraveling yields inefficient matchings: evidence from post-season college football bowls. *RAND Journal of Economics*, 38(4): 967–82, 2007.

X. Gabaix, J.-M. Lasry, P.-L. Lions, and B. Moll. The dynamics of inequality. *Econometrica*, 84(6):2071–2111, 2016.

M. D. Giandrea and S. Sprague. Estimating the u.s. labor share. *Monthly Labor Review*, 2017.

E. P. Grimit, T. Gneiting, V. J. Berrocal, and N. A. Johnson. The continuous ranked probability score for circular variables and its application to mesoscale forecast ensemble verification. *Quarterly Journal of the Royal Meteorological Society*, 132(621C):2925–2942, 2006.

J. A. Groen. Sources of error in survey and administrative sata: The importance of reporting procedures. *Journal of Official Statistics*, 28(2):173–198, 2012.

R. M. Groves, F. F. Fowler, M. P. Couper, J. M. Lepkowski, E. Singer, and R. Tourangeau. *Survey methodology*. John Wiley & Sons, Hoboken, New Jersey, 2009.

J. Guyon. Rethinking the fifa world cup final draw. *Journal of Quantitative Analysis in Sports*, 11(3):169–82, 2015.

J. Guyon. What a fairer 24 team uefa euro could look like. *Journal of Sports Analytics*, 2018. in press.

P. Hall. On the number of bootstrap simulations required to construct a confidence interval. *Journal of Economic Literature*, 14(4):1453–1462, 1986.

J. D. Hamilton. Why you should never use the hodrick-prescott filter. *Review of Economics and Statistics*, 100(5):831–843, 2018.

D. F. Hendry and G. E. Mizon. *Evaluating dynamic econometric models by encompassing the VAR. In: Phillips, P. C. B. (ed.), Models, Methods and Applications of Econometrics*. Blackwell, Cambridge, MA, 1993.

D. F. Hendry and G. E. Mizon. Unpredictability in economic analysis, econometric modeling and forecasting. *Journal of Econometrics*, 182(1):186–195, 2014a.

D. F. Hendry and G. E. Mizon. Why DSGEs crash during crises. *VOX CEPR's Policy Portal*, 2014b.

D. F. Hendry and J. N. J. Muellbauer. The future of macroeconomics: Macro theory and models at the Bank of England. *Oxford Review of Economic Policy*, 34(1-2):287–328, 2018.

D. F. Hendry and J.-F. Richard. On the formulation of empirical models in dynamic econometrics. *Journal of Econometrics*, 20(1):3–33, 1982.

D. F. Hendry and J.-F. Richard. *Recent developments in the theory of encompassing. In: Cornet, B., Tulkens H. (ed.), Contributions to Operation Research and Econometrics: The Twentieth Anniversary of CORE*. MIT press, Cambridge, MA, 1989.

A. Hylland and R. Zeckhauser. The efficient allocation of individuals to positions. *Journal of Political economy*, 87(2):293–314, 1979.

B. F. Ingram and C. H. Whiteman. Supplanting the 'Minnesota' prior: Forecasting macroeconomic time series using real business cycle model priors. *Journal of Monetary Economics*, 34(3):497–510, 1994.

J. Jones and V. Wilson. Working harder or finding it harder to work: Demographic trneds in annual work hours show an increasingly featured workforce. *Economic Policy Institute*, 2018.

K. Jusélius and M. Franchi. Taking a DSGE model to the data meaningfully. *Economics-ejournal*, 1(4):1–38, 2007. doi: http://www.economics-ejournal.org/economics/journalarticles/2007-4.

T. F. Juster and F. P. Stafford. The allocation of time: Empirical findings, behavioral models, and problems of measurement. *Journal of Economic Literature*, 29(2):471–522, 1991.

S. N. Kaplan and J. Rauh. It's the market: The broad-based rise in the return to top talent. *Journal of Economic Perspective*, 27(3):35–56, 2013.

D. Karlis and I. Ntzoufras. Analysis of sports data by using bivariate poisson models. *The Statistician*, 52(3):381–93, 2003.

M. S. Kearney and P. B. Levine. Why is the teen birth rate in the united states so high and why does it matter? *Journal of Economic Perspectives*, 26(2):141–163, 2012.

A. B. Kennickell. Consistent weight design for the 1989, 1992, and 1995 scfs, and the distribution of wealth. *Working Paper of the Board of Governors of the Federal Reseve System*, 1997.

A. B. Kennickell. Multiple imputation in the survey of consumer finances. *Working Paper of the Board of Governors of the Federal Reseve System*, pages 1–9, 1998.

A. B. Kennickell. Revisions of the variance estimation procedure for the scf. *Working Paper of the Board of Governors of the Federal Reseve System*, 2000.

A. B. Kennickell. The role of over-sampling of the wealthy in the survey of consumer finances. *Irving Fisher Committee Bulletin*, 28, 2008.

W. Kopczuk. What do we know about the evolution of top wealth shares in the united states? *Journal of Economic Perspectives*, 29(1):47–66, 2015.

W. Kopczuk and E. Saez. Top wealth shares in the united states, 1916–2000: Evidence from estate tax records. *National Tax Journal*, 57(2), 2004.

A. Korinek. Thoughts on DSGE macroeconomics: Matching the moment, but missing the point? *Paper prepared for the 2015 Festschrift Conference 'A Just Society' honoring Joseph Stiglitz's 50 years of teaching*, 2017.

M. J. Maher. Modelling association football scores. *Statistica Neerlandica*, 36(3):109–18, 1982.

D. S. Matteson, N. A. James, W. B. Nicholson, and L. C. Segalini. Locally stationary vector processes and adaptive multivariate modeling. *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 8722–8726, 2013.

G. E. Mizon. *The encompassing approach in econometrics. In: Hendry, D. F., Wallis, K. F. (ed.), Econometrics and Quantitative Economics*. Blackwell, Oxford, UK, 1984.

G. E. Mizon and J.-F. Richard. The encompassing principle and its application to testing non-nested hypotheses. *Econometrica*, 54(3):657–678, 1986.

J. N. J. Muellbauer. Macroeconomics and consumption: Why central bank models failed and how to repair them. *VOX CEPR's Policy Portal*, 2016.

A. Pagan. Report on modelling and forecasting at the Bank of England. *Bank of England Quarterly Bulletin*, 2003.

I. Palacios-Huerta. Professionals play minimax. *Review of Economic Studies*, 70(2):395–415, 2003.

M. H. Pesaran and A. Timmermann. Selection of estimation window in the presence of breaks. *Journal of Econometrics*, 137(1):134–161, 2007.

M. H. Pesaran, D. Pettenuzzo, and A. Timmermann. Forecasting time series subject to multiple structural breaks. *Review of Economic Studies*, 73(4):1057–1084, 2006.

K. E. Pickett and R. G. Wilkinson. *Income inequality and psychosocial pathways to obesity. In: Offer, A., Pechey, R., Ulijaszek, S. (ed.), Insecurity, Inequality, and Obesity in Affluent Societies*. British Academy, Oxford, 2012.

T. Piketty and E. Saez. Income inequality in the united states, 1913-1998. *Quarterly Journal of Economics*, 118(1):1–39, 2003.

D. G. Pope and M. E. Schweitzer. Is tiger woods loss averse? persistent bias in the face of experience, competition, and high stakes. *American Economic Review*, 101(1):129–57, 2011.

J. H. Powell. Remarks at the ceremonial swearing-in. 2018.

C. Prendergast. The provision of incentives in firms. *Journal of Economic Literature*, 37(1): 7–63, 1999.

A. Rees-Jones and S. Skowronek. An experimental investigation of preference misrepresentation in the residency match. *Proceedings of the National Academy of Sciences*, 115(45):11471–6, 2018. ISSN 0027-8424. doi: 10.1073/pnas.1803212115. URL https://www.pnas.org/content/115/45/11471.

A. P. Reynolds, G. Richards, B. de la Iglesia, and V. J. Rayward-Smith. Clustering rules: A comparison of partitioning and hierarchical clustering algorithms. *Journal of Mathematical Modelling and Algorithms*, 5:475–504, 1992.

C. C. Ribeiro. Sports scheduling: Problems and applications. *International Transactions in Operational Research*, 19(1-2):201–26, 2013.

W. S. Ribeiro, A. Bauer, and M. C. R. Andrande. Income inequality and mental illness-related morbidity and resilience: A systematic review and meta-analysis. *Lancet Psychiatry*, 4(7):554–562, 2017.

P. Romer. The trouble with macroeconomics. *Forthcoming in the American Economist*, 2016.

S. Rosen. The economics of superstars. *American Economic Review*, 71(5):845–858, 1981.

A. E. Roth. The economist as engineer: Game theory, experimentation, and computation as tools for design economics. *Econometrica*, 70(4):1341–78, 2002.

A. E. Roth and E. Peranson. The effects of the change in the nrmp matching algorithm. *Journal of the American Medical Association*, 278(9):729–32, 1997.

D. B. Rubin. *Multiple imputation for nonresponse in surveys*. John Wiley & Sons, New York–Chichester–Brisbane–Toronto–Singapore, 1987.

J. F. Rubio-Ramírez and J. Fernández-Villaverde. Estimating dynamic equilibrium economies: Linear versus nonlinear likelihood. *Journal of Applied Econometrics*, 20(7): 891–910, 2005.

J. Sabelhaus and A. Henriques Volz. Are disappearing employer pensions contributing to rising wealth inequality? *FEDS Notes. Washington: Board of Governors of the Federal Reserve System*, 2019.

E. Saez. Income and wealth inequality: Evidence and policy implications. *Contemporary Economic Policy*, 35(1), 2017.

E. Saez and G. Zucman. Wealth inequality in the united states since 1913: Evidence from capitalized tax income data. *Journal of Economic Literature*, 131(2):519–578, 2016.

E. Saez and G. Zucman. Comments on smith, zidar and zwick (2019). *Working Paper*, 2020a.

E. Saez and G. Zucman. Trends in us income and wealth inequality: Revisings after the revisionists. *Working Paper*, 2020b.

P. Scarf and M. M. Yusof. A numerical study of tournament structure and seeding policy for the soccer world cup finals. *Statistica Neerlandica*, 65(1):43–57, 2011.

P. Scarf, M. M. Yusof, and M. Bilbao. A numerical study of designs for sporting contests. *European Journal of Operational Research*, 198(1):190–8, 2009.

F. Schorfheide. Estimation and evaluation of DSGE models: Progress and challenges. *NBER Working Papers 16781*, 2011.

C. A. Sims. Monetary policy models. *CEPS Working Papers 155*, 2007.

F. Smets and R. Wouters. Comparing shocks and frictions in US and Euro area business cycles: A Bayesian DSGE approach. *Journal of Applied Econometrics*, 20(2):161–183, 2005.

F. Smets and R. Wouters. Shocks and frictions in US business cycles: A Bayesian DSGE approach. *American Economic Review*, 97(3):586–606, 2007.

M. Smith, O. Zidar, and E. Zwick. Top wealth in america: New estimates and implications for taxing the rich. *Working Paper*, 2019.

J. E. Stiglitz. Where modern macroeconomics went wrong. *Oxford Review of Economic Policy*, 34(1-2):70–106, 2018.

J. Thompson, M. Parisi, and J. Bricker. Top income concentration and volatility. *Finance and Economics Discussion Series 2018-010. Washington: Board of Governors of the Federal Reserve System*, 2018.

J.-C. Trichet. Reflections on the nature of monetary policy non-standard measures and finance theory. *Opening address at the ECB Central Banking Conference*, 2010.

E. Uslaner and M. M. Brown. Trust, inequality, and civic engagement. *American Politics Research*, 33(6):868–894, 2005.

A. Vong. Strategic manipulation in tournament games. *Games and Economic Behavior*, 102:562–7, 2017.

M. Walker and J. Wooders. Minimax play at wimbledon. *American Economic Review*, 91 (5):1521–538, 2001.

K. F. Wallis. Seasonal adjustment and relations between variables. *Journal of the American Statistical Association*, 69(345):18–31, 1974.

V. Wieland and M. Wolters. Macroeconomic model comparisons and forecast competitions. *VOX CEPR's Policy Portal*, 2012.

R. G. Wilkinson and K. E. Pickett. The problems of relative deprivation: Why some societies do better than others. *Social Science & Medicine*, 65(9):1965–1978, 2007.