École des Hautes Études en Sciences Sociales

École doctorale de l'EHESS

École doctorale n°465
Économie Panthéon Sorbonne

Discipline : Analyse et Politique Économique

Thomas Blanchet

Essais sur la distribution des revenus et des patrimoines:

méthodes, estimations et théorie

Thèse dirigée par : Thomas Piketty

Date de soutenance: le 21 janvier 2020

Rapporteurs : Frank Cowell, London School of Economics and Political Science
Emmanuel Saez, University of California, Berkeley

Jury : François Bourguignon, École des Hautes Études en Sciences Sociales
Xavier d'Haultfœuille, Centre de Recherche en Économie et Statistique
Muriel Roger, Université Panthéon–Sorbonne
Thomas Piketty, École des Hautes Études en Sciences Sociales

ÉCOLE DES HAUTES ÉTUDES DE SCIENCES SOCIALES

PHD THESIS

to obtain the title of

Doctor of Philosophy of the
École des Hautes Études en Sciences Sociales

Prepared and defended at the Paris School of Economics on 21st January 2020 by:

Thomas BLANCHET

Essays on the Distribution of Income and Wealth: Methods, Estimates and Theory

Thesis Supervisor: Thomas PIKETTY

Jury

Frank COWELL (London School of Economics and Political Science — Referee)
Emmanuel SAEZ (University of California, Berkeley — Referee)
François BOURGUIGNON (EHESS — Examiner)
Xavier D'HAULTFŒUILLE (CREST — Examiner)
Muriel ROGER (Université Panthéon–Sorbonne — Examiner)
Thomas PIKETTY (EHESS/PSE — Supervisor)

Remerciements

Cette thèse doit énormément à tous ceux qui m'ont accompagné ces dernières années. Je tiens, tout d'abord, à remercier Thomas Piketty, pour avoir été un directeur de thèse remarquable. Sa passion, sa vision d'ensemble, son approche de la recherche ont toujours été un exemple. Depuis que j'ai commencé à travailler avec lui en master, il a toujours été disponible pour me guider, m'encourager, et a suivi mes travaux avec enthousiasme. Je lui suis particulièrement reconnaissant d'avoir eu suffisamment confiance en mon travail pour m'avoir — à peine sorti du master — laissé prendre en main plusieurs aspects méthodologiques clés aujourd'hui en usage au sein du World Inequality Lab.

Je tiens ensuite à remercier Frank Cowell et Emmanuel Saez, pour m'avoir fait l'honneur d'accepter d'être les rapporteurs de cette thèse. Leurs travaux ont posé de nombreuses bases conceptuelles et méthodologiques sans lesquelles cette thèse n'aurait jamais pu exister.

Je souhaite aussi remercier François Bourguignon, Xavier d'Haultfœuille et Muriel Roger pour avoir accepté de participer au jury. François Bourguignon a longtemps su porter la question des inégalités mondiales, et je suis heureux d'avoir son opinion sur ce travail. Xavier d'Haultfœuille fût un de mes premiers professeurs d'économétrie à l'ENSAE, et a continué à suivre mes travaux après cela. J'ai toujours admiré sa capacité à appliquer sa rigueur et ses compétences techniques à des problèmes statistiques importants et concrets; son regard sur les défis statistiques propres à la mesure des inégalités est précieux. Enfin, c'est avec Muriel Roger que j'ai fait mes premiers pas dans la recherche, lors de mon stage de master. C'est elle qui m'a initié aux problématiques autour de la mesure des patrimoines. Je suis heureux de la voir de nouveau présente alors que je m'apprête à franchir cette nouvelle étape.

Cette thèse doit énormément au World Inequality Lab, et notamment à Lucas Chancel pour son travail de coordination. D'abord, pour avoir soutenu matériellement mes recherches durant les quatre dernières années. Ensuite, pour avoir largement contribué à la diffusion de ces travaux auprès de publics divers — non seulement universitaires, mais aussi les organisations internationales et les médias grand public. Enfin, et surtout, pour m'avoir fourni un environnement intellectuel aussi riche, stimulant et motivant durant toutes ces années. Le travail de thèse peut souvent s'avérer solitaire. Le fait d'avoir pu la réaliser au sein du World Inequality Lab m'a donné le privilège de voir comment mes travaux contribuer de manière très directe et très concrète à un projet collectif d'envergure et de long terme.

J'ai eu la chance de partager mon environnement de travail avec d'extraordinaires collègues — et amis — qui ont su entretenir bonne humeur et bienveillance durant toutes ces années. Certains sont devenus mes co-auteurs — Amory Gethin, Marc Morgan, et Ignacio Flores. Tous ont contribué à faire des ces années une expérience enrichissante, aussi bien intellectuellement qu'humainement: Lydia Assouad, Luis Bauluz, Yajna Govind, Mark Jenmana, Clara Martínez-Toledano, Mathilde Muñoz, Alix Myczkowski, Theresa Neef, Tom Raster, Olivia Ronsain, Mauricio de Rosa, Aurélie Sotura, Li Yang. Je n'oublie mes premiers co-bureaux, avec qui j'ai fait mes premiers pas de thésard: Nathalie Ferrière, Crisitina Herghelegiu et Cyril Milhaud.

Cette thèse a bénéficié — directement ou indirectement — des nombreuses rencontres que j'ai pu faire et des nombreuses discussions que j'ai pu avoir. Je pense en particulier à — et en m'excusant par avance auprès de tous ceux que j'aurais oublié: Facundo Alvaredo, Matthias André, Yonatan Berman, Antoine Bozio, Stefano Filauro, Matthew Fisher-Post, Bertrand Garbinti, Jonathan Goupille-Lebret, Branko Milanovic, Salvatore Morelli, Filip Novokmet, Daniel Waldenstrom, Gabriel Zucman.

Enfin, je souhaite remercier ma famille — ma sœur, et mes parents, pour m'avoir encouragé et soutenu toutes ces années.

Résumé

Cette thèse couvre plusieurs sujets sur la répartition des revenus et des richesses. Dans le premier chapitre, nous développons une nouvelle méthode pour exploiter les tabulations de revenu et de richesse, telle que celle publiée par les autorités fiscales. Nous y définissons les *courbes de Pareto généralisées* comme la courbe des coefficients de Pareto inversés $b(p)$, où $b(p)$ est le rapport entre le revenu moyen ou la richesse au-dessus du rang p et le p-ième quantile $Q(p)$ (c'est-à-dire $b(p) = \mathbb{E}[X|X > Q(p)]/Q(p)$). Nous les utilisons pour caractériser des distributions entières, y compris les endroits comme le sommet où la lois de Pareto est une bonne description, et les endroits plus bas où elles ne le sont pas. Nous développons une méthode pour reconstruire de manière flexible l'ensemble de la distribution sur la base de données tabulées sur le revenu ou le patrimoine, qui produit courbes de Pareto généralisées lisses et réalistes. À l'aide de tableaux détaillés tirés de données fiscales quasi exhaustives, nous démontrons la précision de notre méthode, à la fois empiriquement et analytiquement. Elle donne de meilleurs résultats que les autres techniques d'interpolation couramment utilisées.

Dans le deuxième chapitre, nous présentons une nouvelle approche pour combiner les données d'enquête et les tabulations fiscales afin de corriger la sous-représentation des plus riches au sommet. Elle détermine de façon endogène un "point de fusion" entre les données avant de modifier les poids tout au long de la distribution et de remplacer les nouvelles observations au-delà du support original de l'enquête. Nous fournissons des simulations de la méthode et des applications aux données réelles. Les premières démontrent que notre méthode améliore la précision et la stabilité des estimations de la distribution, par rapport à d'autres méthodes de correction d'enquêtes utilisant des données externes, et même en présence d'hypothèses extrêmes. Les applications empiriques montrent que non seulement les niveaux d'inégalité des revenus peuvent changer, mais aussi les tendances. Étant donné que notre méthode préserve les distributions multivariées des variables d'enquête, elle fournit aux chercheurs un cadre plus représentatif pour explorer les dimensions socio-économiques des inégalités.

Dans le troisième chapitre, nous estimons la distribution du revenu national dans 38 pays européens entre 1980 et 2017 en combinant enquêtes, données fiscales et comptes nationaux. Nous développons une méthodologie cohérente combinant des méthodes d'apprentissage statistique, de calage non linéaire des enquêtes et la théorie des valeurs extrêmes afin de produire des estimations de l'inégalité des revenus avant et après impôt, comparables d'un pays à l'autre et conformes aux taux de croissance macroéconomiques. Nous constatons que les inégalités se sont creusées dans une majorité de pays européens, en particulier entre 1980 et 2000. Le 1% les plus riches en Europe a augmenté plus de deux fois plus vite que les 50% les plus pauvres et a capturé 18% de la croissance des revenus régionaux. Les inégalités restent toutefois plus faibles et ont beaucoup moins augmenté en Europe qu'aux États-Unis, malgré la persistance de fortes différences de revenus entre les pays européens.

Dans le quatrième chapitre, je décompose la dynamique de la distribution de la richesse à l'aide d'un modèle stochastique dynamique simple qui sépare les effets de la consommation, du revenu du travail, des taux de rendement, de la croissance, de la démographie et du patrimoine. À partir de deux théorèmes de calcul stochastique, je montre que ce modèle est identifié de manière non paramétrique et qu'il peut être estimé à partir de données en coupes répétées. Je l'estime à l'aide des comptes nationaux distributifs des États-Unis depuis 1962. Je trouve que, de l'augmentation de 15 pp. de la part de la richesse détenue par les 1% les plus riches observée depuis 1980, environ 7 pp. peut être attribuée à l'inégalité croissante des revenus du travail, 6 pp. à la hausse des rendements sur le capital (principalement sous forme de plus-values), et 2 pp. à la baisse de la croissance. En suivant les paramètres actuels, la part de la richesse des 1% les plus riches atteindrait sa valeur stationnaire d'environ 45% d'ici les années 2040, un niveau similaire à celui du début du XXe siècle. Ces conclusions s'appliquent à un large éventail de modèles de répartition de la richesse, indépendamment de la façon exacte dont ils modèlent, par exemple, de la consommation ou du marché du travail. J'utilise ensuite le modèle pour analyser l'effet d'un impôt progressif sur les patrimoines au sommet de la distribution.

Mots Clés: inégalités; revenu; patrimoine; enquêtes; données fiscales; comptes nationaux; loi de Pareto; modèles non-paramétriques; modèles stochastiques

Codes JEL: D31; C14; C22

Abstract

This thesis covers several topics on the distribution of income and wealth. In the first chapter, we develop a new methodology to exploit tabulations of income and wealth such as the one published by tax authorities. In it, we define *generalized Pareto curves* as the curve of inverted Pareto coefficients $b(p)$, where $b(p)$ is the ratio between average income or wealth above rank p and the p-th quantile $Q(p)$ (i.e. $b(p) = \mathbb{E}[X|X > Q(p)]/Q(p)$). We use them to characterize entire distributions, including places like the top where power laws are a good description, and places further down where they are not. We develop a method to flexibly recover the entire distribution based on tabulated income or wealth data which produces smooth and realistic shapes of generalized Pareto curves. Using detailed tabulations from quasi-exhaustive tax data, we demonstrate the precision of our method both empirically and analytically. It gives better results than the most commonly used interpolation techniques.

In the second chapter, we present a new approach to combine survey data with tax tabulations to correct for the underrepresentation of the rich at the top. It endogenously determines a "merging point" between the datasets before modifying weights along the entire distribution and replacing new observations beyond the survey's original support. We provide simulations of the method and applications to real data. The former demonstrate that our method improves the accuracy and precision of distributional estimates, even under extreme assumptions, and in comparison to other survey correction methods using external data. The empirical applications show that not only can income inequality levels change, but also trends. Given that our method preserves the multivariate distributions of survey variables, it provides a more representative framework for researchers to explore the socio-economic dimensions of inequality.

In the third chapter, we estimate the distribution of national income in thirty-eight European countries between 1980 and 2017 by combining surveys, tax data and national accounts. We develop a unified methodology combining machine learning, nonlinear survey calibration and extreme value theory in order to produce estimates of pre-tax and post-tax income inequality, comparable across countries and consistent with macroeconomic growth rates. We find that inequality has increased in a majority of European countries, especially between 1980 and 2000. The European top 1% grew more than two times faster than the bottom 50% and captured 18% of regional income growth. Inequalities, however, remain lower and have increased much less in Europe than in the US, despite the persistence of strong income differences between European countries.

In the fourth chapter, I decompose the dynamics of the wealth distribution using a simple dynamic stochastic model that separates the effects of consumption, labor income, rates of return, growth, demographics and inheritance. Based on two results of stochastic calculus, I show that this model is nonparametrically identified and can be estimated using only repeated cross-sections of the data. I estimate it using distributional national accounts for the United States since 1962. I find that, out of the 15 pp. increase in the top 1% wealth share observed since 1980, about 7 pp. can be attributed to rising labor income inequality, 6 pp. to rising returns on wealth (mostly in the form of capital gains), and 2 pp. to lower growth. Under current parameters, the top 1% wealth share would reach its steady-state value of roughly 45% by the 2040s, a level similar to that of the beginning of the 20th century. These conclusions apply to a wide class of models of the wealth distribution, regardless of the exact primitives they use to account for, say, consumption or the labor market. I then use the model to analyze the effect of progressive wealth taxation at the top of the distribution.

Keywords: inequality; income; wealth; survey; tax data; national accounts; power law; non-parametric statistics; stochastic models

JEL codes: D31; C14; C22

Table of Contents

Remerciements **2**

Résumé **3**

Abstract **5**

Table of Contents **7**

List of Tables **11**

List of Figures **13**

General Introduction **16**

1 Generalized Pareto Curves **22**

1.1 Generalized Pareto Curves 25

1.1.1 Definition and Properties 25

1.1.2 Pareto Curves and Power Laws 26

1.1.3 Pareto Curves in Practice 28

1.1.4 Processes Generating Nonconstant Pareto Curves 29

1.2 Generalized Pareto Interpolation 32

1.2.1 Interpolation of the Pareto Coefficients 33

1.2.2 Enforcing Admissibility Constraints 35

1.2.3 Extrapolation in the Last Bracket 37

1.3 Tests Using Income Data 39

1.3.1 Overview of the Most Common Interpolation Methods 40

1.3.2 Extrapolation methods 45

1.4 Estimation Error 47

1.4.1 Theoretical results 49

1.4.2 Applications 50

Bibliography 55

2 The Weight of the Rich 59

2.1 Literature on Survey Correction Methods . . . 63

2.1.1 Reweighting Observations . . . 63

2.1.2 Replacing Incomes . . . 65

2.1.3 Combined Reweighting and Replacing . . . 66

2.2 Theory and Methodology . . . 67

2.2.1 Univariate Setting . . . 67

2.2.2 Multivariate Setting . . . 73

2.2.3 Expanding the Support . . . 79

2.3 Applications . . . 81

2.3.1 Simulations . . . 81

2.3.2 Real Data . . . 83

Bibliography . . . 98

3 How Unequal is Europe? 102

3.1 Related Literature . . . 104

3.2 Conceptual Framework . . . 108

3.2.1 Macroeconomic Concepts . . . 108

3.2.2 Income Distribution Concepts . . . 110

3.3 Sources and Methodology . . . 112

3.3.1 National Accounts . . . 112

3.3.2 Survey Microdata . . . 113

3.3.3 Survey Tabulations . . . 114

3.3.4 Survey Corrections . . . 116

3.3.5 Missing Incomes . . . 119

3.3.6 External validation . . . 122

3.4 Results . . . 122

3.4.1 Inequality between European Countries . . . 122

3.4.2 Inequality within European Countries . . . 126

3.4.3 Inequality between European Citizens . . . 129

3.4.4 Redistribution in Europe . . . 131

3.4.5 Inequality in Europe vs. the United States . . . 134

Bibliography . . . 142

4 Modeling the Dynamics of Wealth 147

4.1 The Distribution of Wealth . . . 151

4.1.1 Empirical Facts about Wealth in the United States . . . 152

4.1.2 Mechanisms That Account for Wealth Inequality . . . 155

4.2 Theoretical Framework . . . 156
4.2.1 Income and Consumption . . . 157
4.2.2 Birth and Death . . . 158
4.2.3 Inheritance . . . 159
4.3 Data, Demography and Inheritance . . . 159
4.3.1 Demography . . . 159
4.3.2 Inheritance and the Estate Tax . . . 161
4.3.3 Income and Wealth . . . 165
4.4 Identification and Estimation . . . 169
4.4.1 Identification . . . 169
4.4.2 Estimation . . . 172
4.5 Results . . . 177
4.5.1 Past Evolution (1962–2014) . . . 177
4.5.2 Future Evolution (2015–2100) . . . 181
4.5.3 The Taxation of Wealth . . . 181
Bibliography . . . 194

General Conclusion **198**

A Appendix to “Generalized Pareto Curves” **201**
A.1 Generalized Pareto curves: Omitted Proofs . . . 201
A.1.1 Proof of proposition 1 . . . 201
A.1.2 Proof of proposition 2 . . . 202
A.1.3 Proof of proposition 3 . . . 202
A.1.4 Proof of proposition 4 . . . 203
A.2 Other Concepts of Local Pareto Coefficients . . . 204
A.3 Dynamic Model of Income Growth an Wealth Accumulation . . . 206
A.3.1 Proofs omitted from the main text . . . 206
A.3.2 Alternative Calibrations . . . 207
A.4 Detailed interpolation method . . . 210
A.5 Other comparisons of interpolation methods . . . 212
A.6 Error estimation . . . 216
A.6.1 Decomposition of the error . . . 216
A.6.2 Misspecification error . . . 216
A.6.3 Sampling error . . . 220
A.6.4 Comparing Misspecification with Sampling Error . . . 230
Bibliography . . . 232

B Appendix to "The Weight of the Rich" 234

B.1 Formal Biases and Adjustments . . . 234

B.1.1 Double-Biased Density Functions . . . 234

B.1.2 Adjustment Methods: Reweighting vs. Replacing . . . 236

B.2 Merging Point Below the Trustable Span . . . 239

B.3 Geometrical Interpretation of Calibration . . . 241

B.4 Further Monte Carlo Simulations . . . 242

B.5 Data Details and Supplementary Results . . . 245

B.5.1 Country Specific Income Concepts and Observational Units . . 245

B.5.2 Further Tables and Figures . . . 249

B.5.3 Detailed Distribution . . . 261

Bibliography . . . 264

C Appendix to "How Unequal is Europe?" 266

C.1 Detailed Methodology . . . 266

C.1.1 Machine Learning . . . 267

C.1.2 Calibration . . . 271

C.1.3 Extreme Value Theory . . . 274

C.1.4 Income Distribution by State in the US . . . 275

C.2 Methodology and Alternative Assumptions . . . 277

C.3 Coverage of Data Sources . . . 280

C.4 Average Incomes in Europe . . . 281

Bibliography . . . 282

D Appendix to "The Dynamics of Wealth" 285

D.1 Comparison to Synthetic Saving Rates . . . 285

D.2 Omitted proofs . . . 290

D.2.1 Application of Gyöngy's (1986) Theorem . . . 290

D.2.2 Steady-State Wealth Distribution With a Wealth Tax . . . 291

Bibliography . . . 292

List of Tables

1.1 Mean relative error for different interpolation methods 42
1.2 Mean relative error on the top 1% for different extrapolation methods, knowing the top 10% and the top 5% 47
1.3 Observed maximum error and theoretical upper bound 51
1.4 Optimal bracket choice for a typical distribution of income 51
1.5 Mean relative error using subsamples of the full population 53

2.1 Structure of Corrected Population: Latest Year 89

3.1 Theil index decomposition of between-region and within-region inequalities in Europe and the US 138

A.1 Mean relative error for different interpolation methods (fiscal income) 213
A.2 Mean relative error for different interpolation methods (pre-tax national income) 214
A.3 Mean relative error on the top 0.1% for different extrapolation methods, knowing the top 10% and the top 1% 215

B.1 From Post-Tax to Pre-Tax Income in Chilean Surveys 248
B.2 Structure of Corrected Population in Brazil, 2007-2015 258
B.3 Structure of Corrected Population in Chile, 2009-2015 259
B.4 Structure of Corrected Population in France, 2004-2014 259
B.5 Structure of Corrected Population in Norway, 2004-2014 260
B.6 Structure of Corrected Population in United Kingdom, 2005-2014 . . 260
B.7 Income Shares: Raw Survey and Corrected Survey 262

C.1 Mean relative error on the average by percentile when imputing pretax and post-tax income by adults from a different concept using the machine learning algorithm 271
C.2 Coverage of data sources 280

C.3 Average national incomes in Europe, 1980-2017 281

List of Figures

1.1 Generalized Pareto curves of pre-tax national income 28

1.2 Calibration of $\sigma(x)$ on the US Distribution of Labor Income 31

1.3 P75 threshold and top 25% share in the United States (1962–2014), estimated using all interpolation methods and a tabulation with $p = 10\%, 50\%, 90\%, 99\%$ 43

1.4 P75 threshold and top 25% share in the United States (2000-2014), estimated using interpolation methods M0 and M3, and a tabulation with $p = 10\%, 50\%, 90\%, 99\%$ 43

1.5 Generalized Pareto curves implied by the different interpolation methods for the United States distribution of income in 2010 44

1.6 Extrapolation with generalized Pareto distribution 46

1.7 Comparison of extrapolation methods in the United States for the top 1%, knowing the top 10% and the top 5% 48

2.1 A “True” and Biased Income Distribution 69

2.2 The Intuition Behind Reweighting 70

2.3 Choice of Merging Point when $\bar{y} \geq y_{\text{trust}}$ 72

2.4 Benchmark Experiment Set-Up 83

2.5 Benchmark Experiment Results 84

2.6 Merging Point in Five Countries, Latest year 87

2.7 Top 1% Shares Before and After Correction 91

2.8 Gini Coefficients, Before and After Correction 93

2.9 Average Incomes Before and After Correction 95

3.1 DINA methodology 121

3.2 Inequality Between Countries in Europe 123

3.3 Inequality Within Countries in European Regions 127

3.4 Inequality Within Selected European Countries 128

3.5 Evolution of Inequality among European Citizens 130

3.6 Between- versus within-country inequality in Europe, 1980–2017 . . . 132
3.7 Redistribution in Europe 133
3.8 Inequality in Europe and in the United States 137
3.9 Growth of the Bottom 50% in Europe and in the United States . . . 139

4.1 Private Wealth in the United States 153
4.2 Wealth Inequality: The Role of the Top 1% vs. the Bottom 99% . . . 154
4.3 Income Inequality: Labor and Total Income 155
4.4 Modeling of Inheritance 162
4.5 Estate Tax 164
4.6 The Impact of Capital Gains on National Income and Inequality . . . 167
4.7 Simplified Dynamics of Wealth Inequality in the Top Tail 171
4.8 Estimation of the Main Model 174
4.9 Past Evolution of Wealth: Data vs. Model 178
4.10 Past Evolution of Wealth: The Role of Labor and Capital Income . . 180
4.11 Past Evolution of Wealth: The Role of Growth and Capital Gains . . 182
4.12 Future Evolution of Wealth 183
4.13 Impact on the Tax Base of a Linear Wealth Tax over $50m 191

A.1 Different concepts of local Pareto exponent 206
A.2 Calibration of $\mu(x)$ on the US distribution of labor income 208
A.3 Calibration of $\sigma(x)$ on the US distribution of personal wealth 209
A.4 Calibration of $\mu(x)$ on the US distribution of personal wealth 210
A.5 Bounds on the misspecification error term for φ and φ' 218
A.6 Estimations of $\varphi'''(x)$ 219
A.7 Asymptotic mean absolute value of the sampling error with finite variance 229
A.8 Asymptotic mean absolute value of the sampling error with infinite variance 230
A.9 Actual error and estimated misspecification error 230

B.1 Correcting for Nonresponse by Reweighting 236
B.2 Correcting for Misreporting by Replacing 238
B.3 Choice of Merging Point when $\bar{y} < y_{\text{trust}}$ 239
B.4 Geometrical Interpretation of Linear Calibration 241
B.5 Experiment with more Misreporting 242
B.6 Experiment with Replacing the Top 5% 243
B.7 Experiment with Poor Tax Data 244
B.8 Merging Points in Norway, 2004–2013 250

B.9 Merging Points in France, 2004–2013 252
B.10 Merging Points in United Kingdom, 2005–2013 254
B.11 Merging Points in Brazil, 2007–2014 255
B.12 Merging Points in Chile, 2009–2013 257
B.13 Discrepancy of income across datasets in Chile and Brazil: 2015 . . . 263

C.1 Role of the Various Steps: From Raw Surveys to Corrected Surveys to DINA Income 277
C.2 Bottom 50% Income Shares in France: Comparison 277
C.3 Distribution of Production Taxes 278
C.4 Narrow vs. Broad Equal-split 278
C.5 Distribution of Government Expenditures 279

General Introduction

This PhD thesis covers several topics on the distribution of income and wealth. These topics include measurement issues (chapters 1 and 2), applied empirical work in the case of Europe (chapter 3), and finally a quantitative modeling of the dynamics of wealth inequality in the United States (chapter 4).

These four chapters follow an explicit logical progression. The first two chapters lay out some basic tools to improve the measurement of inequality. In the third chapter, these tools are put to use in the concrete case of the European income distribution. This allows the production of harmonized inequality statistics that we consider to be more comprehensive and more reliable than previous estimates. This type of data opens up new opportunities for economic research to improve our understanding of inequality: this is what the fourth chapter demonstrates, using data for the United States.

A Better Use of Available Data

This work addresses issues at the intersection of three key challenges for contemporary inequality research. The first one concerns the use of administrative data sources. For a long time, surveys were the main source of knowledge for applied economists. And they have, indeed, been an invaluable source of information. However, in recent years, more and more people have started to rely on administrative data. Administrative data has advantages — exhaustivity, lower measurement error — that no survey can match. The research on inequality is no exception to that trend. Following the revival of Kuznets's (1953) work by Piketty (2003) and Piketty and Saez (2003), a large number of researchers have used tax data to measure inequality in the long run (Atkinson, 2007; Atkinson and Piketty, 2010). Their findings showed that surveys, due to misreporting and nonresponse, have had a tendency to miss the richest households, thus underestimating inequality and overlooking important trends.

Yet, administrative data raises its own set of challenges. Some countries — notoriously the Nordics, but also France and the United States — have been providing tax data to researchers in a very detailed and usable form (microdata). In many other countries, however, all researchers have to work with are tax tabulations — a highly censored version of the tax data that only contains income amounts by income bracket. The statistical tools used to exploit this type of data have not evolved much since Kuznets's (1953) time, and often fail to properly exploit the information at our disposal. The lack of individual observations also makes it difficult harmonize income concepts and the statistical unit of analysis. As a result, the body of work on inequality measurement using tax data that has been accumulated raises various comparability issues, and renders international comparisons of inequality difficult. Chapters 1 and 2 provide methodological improvements for using tax data tabulations, and for combining survey and tax data so as to make inequality estimates more precise and more comparable. Chapter 3 applies these ideas to get new estimates of income inequality across Europe.

Filling the Gaps between Micro and Macro Data

The second challenge addressed by this thesis is the gap between macroeconomic and microeconomic estimates of income and wealth. Over the second half of the 20th century, economists — first in academia, then in official statistical institutes — have been developing an impressive statistical apparatus to track the evolution of income and wealth all over the world. This achievement, known as the national accounts, is not perfect, and national accountants are in fact constantly trying to improve it. But it represents the most complete attempt at defining income, wealth and their components in a meaningful and internationally agreed way.

One of the main blind spots of national accounting as it exists concerns the distribution of income and wealth: national accounts are solely concerned with aggregates. This did not have to be the case. In fact, the first national accounts — King's social tables compiled in the late 17th century — were technically distributional national accounts, and showed how aggregates were distributed across various social classes (see Piketty, Saez, and Zucman, 2018). Simon Kuznets, one of the fathers of national accounting, was also known for his work on the distribution of income (Kuznets, 1953). Yet in practice, the development of national accounts strayed away from distributional issues. Perhaps because their development took place right after a strong compression of the income distribution in industrialized economies, so that inequality could be viewed as a secondary issue. Or perhaps because of the tight link between national accounting and the emergence of the field of macroeconomics, before the arrival of heterogeneous agents models that would have made distributional estimates useful.

Whatever the reason, the field of inequality research developed somewhat independently from national accounting, in spite of their major conceptual overlap. Fast forward to the 21st century, and the statistical apparatus that tracks income and wealth at the aggregate level is largely distinct from the one we use to track the evolution of income and wealth at the individual level, sometimes leading to major inconsistencies.

In recent years, actors from academia and statistical institutes have recognized the need to address this issue. One of these initiatives is the distributional national accounts (DINA) project (Alvaredo et al., 2017). Chapter 3 of this thesis presents the first attempt to apply this framework to construct pan-European distributional national accounts since 1980. These new estimates are conceptually comparable to the work of Piketty, Saez, and Zucman (2018) in the United States, and enable better comparison of the income distribution on both sides of the Atlantic. They provide an interesting case study of how distributional national accounts can be applied across a region with heterogeneous institutional frameworks and data availability.

Connecting Theoretical Models of Inequality to the Data

A third challenge in the field of inequality is the connection between the data and theories of wealth inequality. Ultimately, estimates of the income and wealth distribution are there to improve our understanding of what shapes inequality. Chapter 4 of this thesis addresses this issue, using the DINA data from Piketty, Saez, and Zucman (2018).

Several models of the wealth distribution have been developed, and they have had many successes in replicating stylized facts about wealth inequality and its distribution. These models are often complex, microfounded structural models that are calibrated but not directly estimated on the data. Indeed, there is no explicit identification strategy that would allow for a direct estimation of these models, and a result it can be hard to understand the generality, robustness and limitations of the various mechanisms.

In chapter 4, I show how the detailed historical data on the distribution of income and wealth put together by Piketty, Saez, and Zucman (2018) can help us make progress on the issue. I explain how the key drivers of wealth inequality can be identified in a “semi-structural” way, which allows for a clean decomposition of the different mechanisms that have been suggested to account for inequality.

Outline and Summary

Chapter 1 was written with Juliette Fournier and Thomas Piketty, and is called “Generalized Pareto Curves: Theory and Applications.” Its goal is to develop a new methodology to exploit tabulations of income and wealth such as the one published by tax authorities. In it, we define *generalized Pareto curves* as the curve of inverted Pareto coefficients $b(p)$, where $b(p)$ is the ratio between average income or wealth above rank p and the p-th quantile $Q(p)$ (i.e. $b(p) = \mathbb{E}[X|X > Q(p)]/Q(p)$). We use them to characterize entire distributions, including places like the top where power laws are a good description, and places further down where they are not. We develop a method to flexibly recover the entire distribution based on tabulated income or wealth data which produces smooth and realistic shapes of generalized Pareto curves. Using detailed tabulations from quasi-exhaustive tax data, we demonstrate the precision of our method both empirically and analytically. It gives better results than the most commonly used interpolation techniques.

Chapter 2 was written with Marc Morgan and Ignacio Flores, and is called “The Weight of the Rich: Improving Surveys with Tax Data.” Household surveys fail to capture the top tail of income and wealth distributions, as evidenced by studies based on tax data. Yet to date there is no consensus on how to best reconcile both sources of information. This paper presents a novel method, rooted in calibration theory, which helps to solve the problem under reasonable assumptions. It has the advantage of endogenously determining a “merging point” between the datasets before modifying weights along the entire distribution and replacing new observations beyond the survey’s original support. We provide simulations of the method and applications to real data. The former demonstrate that our method improves the accuracy and precision of distributional estimates, even under extreme assumptions, and in comparison to other survey correction methods using external data. The empirical applications provide useful and coherent illustrations in a wide variety of contexts. Results show that not only can income inequality levels change, but also trends. Given that our method preserves the multivariate distributions of survey variables, it provides a more representative framework for researchers to explore the socio-economic dimensions of inequality, as well as to study other related topics, such as fiscal incidence.

Chapter 3 was written with Lucas Chancel and Amory Gethin, and is called “How

Unequal is Europe? Evidence from Distributional National Accounts." It estimates the distribution of the national income in thirty-eight European countries between 1980 and 2017 by combining surveys, tax data and national accounts. In it, we develop a unified methodology combining machine learning, nonlinear survey calibration and extreme value theory in order to produce estimates of pre-tax and post-tax income inequality, comparable across countries and consistent with macroeconomic growth rates. We find that inequality has increased in a majority of European countries, especially between 1980 and 2000. The European top 1% grew more than two times faster than the bottom 50% and captured 18% of regional income growth. Inequalities, however, remain lower and have increased much less in Europe than in the US, despite the persistence of strong income differences between European countries.

Chapter 4 is called "Modeling the Dynamics of Wealth Inequality in the United States, 1962–2100." In it, I decompose the dynamics of the wealth distribution using a simple dynamic stochastic model that separates the effects of consumption, labor income, rates of return, growth, demographics and inheritance. Based on two results of stochastic calculus, I show that this model is nonparametrically identified and can be estimated using only repeated cross-sections of the data. I estimate it using distributional national accounts for the United States since 1962. I find that, out of the 15 pp. increase in the top 1% wealth share observed since 1980, about 7 pp. can be attributed to rising labor income inequality, 6 pp. to rising returns on wealth (mostly in the form of capital gains), and 2 pp. to lower growth. Under current parameters, the top 1% wealth share would reach its steady-state value of roughly 45% by the 2040s, a level similar to that of the beginning of the 20th century. These conclusions apply to a wide class of models of the wealth distribution, regardless of the exact primitives they use to account for, say, consumption or the labor market. I then use the model to analyze the effect of progressive wealth taxation at the top of the distribution.

Chapter 1

Generalized Pareto Curves: Theory and Applications

It has long been known that the upper tail of the distribution of income and wealth can be approximated by a Pareto distribution, or power law (Pareto, 1896). This fact has been widely used in the empirical literature on inequality to overcome certain limitations of the data. In particular, Pareto interpolation methods have been used by Kuznets (1953), Atkinson and Harrison (1978), Piketty (2001, 2003), Piketty and Saez (2003) and the subsequent literature exploiting historical tax tabulations to construct long-run series on income and wealth inequality. The widespread applicability of this functional form is often justified using models where income and wealth evolves according to random multiplicative shocks (Champernowne, 1953; Simon, 1955; Wold and Whittle, 1957). Recent contributions have shown how such models can account for both the levels and the changes in inequality (Nirei, 2009; Benhabib, Bisin, and Zhu, 2011; Piketty and Zucman, 2015; Jones and Kim, 2017; Jones, 2015; Benhabib and Bisin, 2016; Gabaix et al., 2016).

But while the Pareto approximation is acceptable for some purposes, it is not entirely correct, not even at the top. As a result, empirical methods that strictly rely on it can miss important features of the distribution (Jenkins, 2016; Atkinson, 2017). If we want to better exploit the data at our disposal, and also to better understand the economic mechanisms giving rise to the observed distributions of income and wealth, we need to move beyond standard Pareto distributions.

In this paper, we develop the flexible notion of *generalized Pareto curve* in order to characterize and estimate income and wealth distributions. A generalized Pareto curve is defined as the curve of inverted Pareto coefficients $b(p)$, where $0 \leq p < 1$

is the rank, and $b(p)$ is the ratio between average income or wealth above rank p and the p-th quantile $Q(p)$ (i.e. $b(p) = \mathbb{E}[X|X > Q(p)]/Q(p)$). If the tail follows a standard Pareto distribution, the coefficient $b(p)$ is constant. For example, if $b(p) = 2$ at the top of the wealth distribution, then the average wealth of individuals above €1 million is €2 million, the average wealth of individuals above €10 million is €20 million, and so on. In practice, we find that $b(p)$ does vary within the upper tail of observed income and wealth distributions (including within the top 10% or the top 1%), but that the curves $b(p)$ are relatively similar (typically U-shaped).

Our contribution is twofold. First, we start by showing that these generalized Pareto curves have direct connections to a more general theory of power laws known as Karamata's (1930) theory of *regular variations*. Therefore, they constitute a practical tool to study power laws in a more general sense, and move away from certain parametric assumptions that characterized earlier work. While we confirm that distributions of income and wealth are indeed power laws in Karamata's (1930) sense, we also see clear deviations from the strict Pareto distribution: we find that the distribution of income is more skewed toward the very top than what the standard Pareto model implies, especially in the United States. We further explain how Karamata's (1930) power laws and the generalized Pareto curves we observe in practice arise from straightforward generalizations of the models of income and wealth accumulation that are used to explain the emergence of standard Pareto laws. These generalizations are consistent with findings found elsewhere in the literature on the nature individual income processes, which can explain why and how we should expect Pareto coefficient to diverge from strict Paretian behavior.

Then, we exploit this framework to develop an improved methodological approach for the estimation of income and wealth distribution using tax data, which is often available solely in the form of tabulations with a finite number of inverted Pareto coefficients $b_1, \dots, b_K$ and thresholds $q_1, \dots, q_K$ observed for ranks $p_1, \dots, p_K$. We call it *generalized Pareto interpolation*. Existing methods typically rely on diverse Paretian assumptions (or even less realistic ones) that, by construction, blur or even erase deviations from the standard Pareto distribution. We show that taking into account how the Pareto coefficient $b(p)$ varies can dramatically improve the way we produce statistics on income and wealth inequality, especially with few data points. By using quasi-exhaustive annual micro files of income tax returns available in the United States and France over the 1962–2014 period (a time of rapid and large transformation of the distribution of income, particularly in the United States), we demonstrate the precision of the method. That is, based on the information for a small number of ranks (e.g. $p_1 = 10\%$, $p_2 = 50\%$, $p_3 = 90\%$, $p_4 = 99\%$), we can recover the entire distribution with remarkable precision. The method gives good results both for the top and for the bottom of the distribution, and generates a consistent and smooth distribution with a continuous density. In fact, we find that the precision of the method is such that it is often preferable to use tabulations based on exhaustive data rather than individual data from a non-exhaustive subsample of the population, even for subsamples considered very large by statistical standards. For example, a subsample of 100 000 observations can typically lead to a mean relative error of about 3% on the top 5% share, while a tabulation based on exhaustive data that includes the percentile ranks $p = 10\%, 50\%, 90\%$ and 99% gives a mean relative error of less than 0.5%. For the top 0.1% share, the same error can reach 20% with the same subsample, while the same tabulation yields an error below 4%.

We believe that the methodology developed in this paper can help researchers avoid excessive reliance on restrictive assumptions when using tabulated data, which is still commonplace in some areas of research.[1] To that end, we developed an R package, named `gpinter`, that implements the methods described in this article and make them easily available to researchers. We also provide a web interface built on top of this package, available at `http://wid.world/gpinter`, to estimate and manipulate distributions of income and wealth on the basis of simple tabulated data files (such as those provided by tax administrations and statistical institutes) and generalized Pareto interpolation methods.[2] These tools have successfully been used to estimate series of the income distribution in the Middle-East (Alvaredo, Assouad, and Piketty, 2017), Poland (Bukowski and Novokmet, 2017), Brazil (Morgan, 2017), India (Chancel and Piketty, 2017), Russia (Novokmet, Piketty, and Zucman, 2017), Ivory Coast (Czajka, 2017), China (Piketty, Yang, and Zucman, 2017), France (Garbinti, Goupille-Lebret, and Piketty, 2016), and India (Chancel and Piketty, 2017). And we plan to use them to keep expanding the World Inequality Database (`wid.world`). But the method is not limited to the production of specific inequality statistics: it outputs a complete and consistent distribution which, depending on what is most practical, can be characterized by its density, its cumulative distribution function, its quantile function or its Lorenz curve. As such, it offers readily available tools for using tabulated data in a variety of contexts (see for example Bierbrauer and Boyer (2017) in the field of optimal taxation).

[1] That is especially true in economic history, or when studying inequality is less developed countries. For example, the World Bank's PovcalNet or the World Panel Income Distribution (Lakner and Milanovic, 2016) take this form.

[2] R is maintained by the R Core Team (2016). The web interface uses `shiny` (Chang et al., 2017).

1.1 Generalized Pareto Curves

1.1.1 Definition and Properties

We characterize the distribution of income or wealth by a random variable X with cumulative distribution function (CDF) F. We assume that X is integrable (i.e. $\mathbb{E}[|X|] < +\infty$) and that F is differentiable over a domain $D = [a, +\infty[$ or $D = \mathbb{R}$. We note f the probability density function (PDF) and Q the quantile function. Our definition of the inverted Pareto coefficient follows the one first given by Fournier (2015).

Definition 1 (Inverted Pareto coefficient)**.** *For any income level $x > 0$, the inverted Pareto coefficient is $b^*(x) = \mathbb{E}[X|X > x]$, or:*

$$b^*(x) = \frac{1}{(1 - F(x))x} \int_x^{+\infty} z f(z) \, \mathrm{d}z$$

We can express it as a function of the fractile p with $p = F(x)$ and $b(p) = b^*(x)$:

$$b(p) = \frac{1}{(1 - p)Q(p)} \int_p^1 Q(u) \, \mathrm{d}u$$

If X follows a Pareto distribution with coefficient α and lower bound $\bar{x}$, so that $F(x) = 1 - (\bar{x}/x)^\alpha$, then $b(p) = \alpha/(\alpha - 1)$ is constant (a property also known as van der Wijk's (1939) law), and the top $100 \times (1-p)\%$ share is an increasing function of b and is equal to $(1-p)^{1/b}$. Otherwise, $b(p)$ will vary. We can view the inverted Pareto coefficient as an indicator of the tail's fatness, or similarly an indicator inequality at the top. It also naturally appears in some economic contexts, such as optimal taxation formulas (Saez, 2001). We favor looking at them as a function of the fractile p rather than the income x, because it avoids differences due to scaling, and make them more easily comparable over time and between countries. We call *generalized Pareto curve* the function $b : p \mapsto b(p)$ defined over $[\bar{p}, 1[$ with $\bar{p} = F(\bar{x})$.[3]

Proposition 1. *If X satisfies the properties stated above, then b is differentiable and for all $p \in [\bar{p}, 1[$, $1 - b(p) + (1 - p)b'(p) \leq 0$ and $b(p) \geq 1$.*

The proof of that proposition — as well as all the others in this section — are available in appendix A.1. The definition of $b(p)$ directly imply $b(p) \geq 1$. The fact that the quantile function is increasing implies $1 - b(p) + (1 - p)b'(p) \leq 0$. Conversely,

[3] We solely consider inverted Pareto coefficient above a strictly positive threshold $\bar{x} > 0$, because they have a singularity at zero and a less clear meaning below that.

for $0 \leq \bar{p} < 1$ and $\bar{x} > 0$, any function $b : [\bar{p}, 1[\to \mathbb{R}$ that satisfies property 1 uniquely defines the top $(1-\bar{p})$ fractiles of a distribution with $\bar{p} = F(\bar{x})$.

Proposition 2. *If X is defined for $x > \bar{x}$ by $F(\bar{x}) = \bar{p}$ and the generalized Pareto curve $b : [\bar{p}, 1[\to \mathbb{R}$, then for $p \geq \bar{p}$, the p-th quantile is:*

$$Q(p) = \bar{x}\frac{(1-\bar{p})b(\bar{p})}{(1-p)b(p)} \exp\left(-\int_{\bar{p}}^{p} \frac{1}{(1-u)b(u)}\,\mathrm{d}u\right)$$

The coefficient defined in 1 is only one of several "local" notion Pareto coefficient that may be defined using a similar logic. In appendix A.2, we show this definition fits into this larger family of Pareto coefficients.

1.1.2 Pareto Curves and Power Laws

For a strict power law (i.e. a Pareto distribution), the Pareto curve is constant. But strict power laws rarely exist in practice, so that we may want to characterize the Pareto curve when power law behavior is only approximate. Approximate power laws are traditionally defined based on Karamata's (1930) theory of *slowly varying* functions. In informal terms, we call a function slowly varying if, when multiplied by power law, it behaves asymptotically like a constant under integration.[4]

Definition 2 (Asymptotic power law)**.** *We say that X is an asymptotic power law if for some $\alpha > 0$, $1 - F(x) = L(x)x^{-\alpha}$, where $L :]0, +\infty[\to]0, +\infty[$ is a slowly varying function, which means that for all $\lambda > 0$, $\lim_{x\to+\infty} \frac{L(\lambda x)}{L(x)} = 1$.*

Definition 2 corresponds to the broadest notion of power laws. We call them "asymptotic" power laws to distinguish them from "strict" power laws (i.e. Pareto distributions). Strict power laws are characterized by their *scale invariance*, meaning that for all $\lambda > 0$, $1 - F(\lambda x) = \lambda^{-\alpha}(1 - F(x))$. The requirement that L is slowly varying in definition 2 means that $1 - F$ must be asymptotically scale invariant. That includes in particular situations where $1 - F$ is equivalent to a power law (i.e. $1 - F(x) \sim Cx^{-\alpha}$ for some $C > 0$). But we could also set, for example, $L(x) \propto (\log x)^{\beta}$ with any $\beta \in \mathbb{R}$.

We will in general restrict ourselves to situations where $\alpha > 1$ to ensure that the means are finite.[5] With $\alpha > 1$, there is a strong link between generalized Pareto

[4] See Bingham, Goldie, and Teugels (1989) for a full account of this theory.

[5] Hence, we exclude edge cases were the inverted Pareto coefficients are finite, but converge to $+\infty$ as $p \to 1$ (for example $b(p) = 3 - \log(1-p)$). Technically, they correspond to a power law, with $\alpha = 1$, but unlike a strict Pareto distribution with $\alpha = 1$, they have a finite mean. In practice,

curve and asymptotic power laws.

Proposition 3. *Let $\alpha > 1$. X is an asymptotic power law with Pareto coefficient α, if and only if* $\lim_{p\to 1} b(p) = \frac{\alpha}{\alpha-1}$.

Proposition 3 generalizes van der Wijk's (1939) characterization of Pareto distributions to asymptotic power laws. Because $\alpha > 1 \Leftrightarrow \alpha/(\alpha-1) > 1$, a distribution is an asymptotic power law if and only if its asymptotic inverted Pareto coefficient is strictly above one. It will tend toward infinity when α approaches one, and to one when α approaches infinity. This behavior is in contrast with distributions with a thinner tail, whose complementary CDF is said to be *rapidly varying.*

Proposition 4. $1 - F(x)$ *is rapidly varying (of index $-\infty$), meaning that for all $\lambda > 1$,* $\lim_{x\to +\infty} \frac{1-F(\lambda x)}{1-F(x)} = 0$ *if and only if* $\lim_{p\to 1} b(p) = 1$.

Distributions concerned by proposition 4 include the exponential, the normal or the log-normal. More broadly, it includes any distribution that converges to zero faster than any power law (i.e. $1 - F(x) = o(x^{-\alpha})$ for all $\alpha > 0$). For all those distributions, the generalized Pareto curve will eventually converge to one. Looking at the Pareto curve near $p = 1$ can therefore help discriminate fat-tailed distributions from others.

Propositions 3 and 4 imply that probability distributions may be divided into three categories, based on the behavior of their generalized Pareto curve. First, power laws, for which $b(p)$ converges to a constant strictly greater than one. Second, "thin-tailed" distributions, for which $b(p)$ converges to one. The third category includes distributions with an erratic behavior in the tail, for which $b(p)$ may oscillate at an increasingly fast rate without converging toward anything.[6] That last category does not include any standard parametric family of distributions, and its members can essentially be considered pathological. If we exclude it, we are left with a straightforward dichotomy between power laws, and thin tails.

When $\lim_{p\to 1} b(p) > 1$, so that X is an asymptotic power law, the generalized Pareto curve can further be used to observe *how* the distribution converges. If $b(p)$ increases near $p = 1$, the tail is getting fatter at higher income levels. But if $b(p)$ decreases, it is getting thinner.

With a strict power law, so that $b(p)$ is constant, the level of inequality stays the same as we move up through the distribution. The share of the top 10% among the whole population is the same as the share of the top 1% among the top 10%

Pareto coefficients for the distribution of income or wealth are clearly above one, so there is no reason to believe that such cases are empirically relevant.

[6] For example $b(p) = 3 + \sin(\log(1 - p))$.

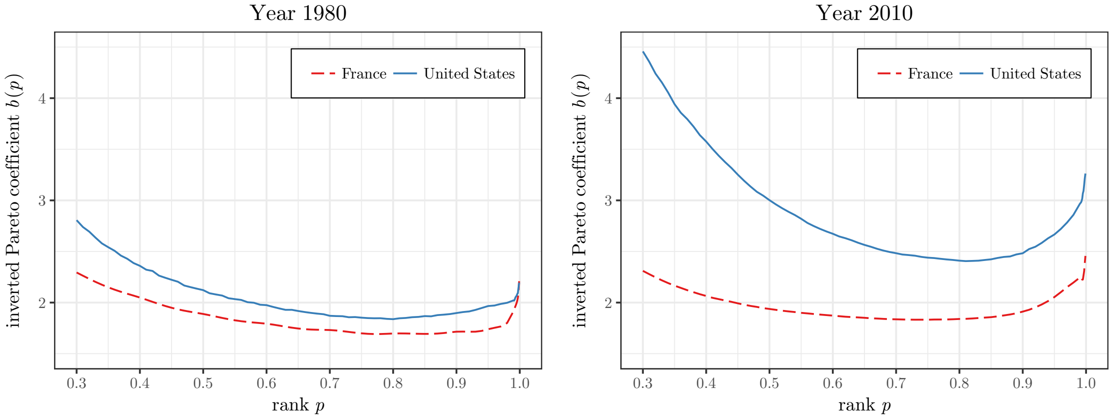
or the share of the top 0.1% among the top 1%. This property is often called the "fractal" nature of inequality. Deviations from a constant $b(p)$ indicate deviations from this rule: if $b(p)$ is increasing for $p > 0.9$, the top 0.1% gets a larger fraction of the income of the top 1% than the top 1% does for the top 10%, so that the top 1% is more unequal than the top 10%.

1.1.3 Pareto Curves in Practice

We now consider a sample $(X_1, \dots, X_n)$ of n iid. copies of X. We write $X_{(r)}$ the r-th order statistic (i.e. the r-th largest value). The natural estimator of the inverted Pareto coefficient may be written:[7]

$$\hat{b}_n(p) = \frac{1}{(n - \lfloor np \rfloor)X_{(\lfloor np \rfloor + 1)}} \sum_{k=\lfloor np \rfloor + 1}^{n} X_{(k)}$$

Figure 1.1 depicts the empirical Pareto curves for the distribution of pre-tax national income in France and in the United States in 1980 and 2010, based on quasi-exhaustive income tax data. The curve has changed a lot more in the United States than in France, which reflects the well-known increase in inequality that the United States has experienced over the period. In 2010, the inverted Pareto coefficients are much higher in the United States than in France, which means that the tail is fatter, and the income distribution more unequal.



Sources: Piketty, Saez, and Zucman (2016) (United States), Garbinti, Goupille-Lebret, and Piketty (2016) (France).

Figure 1.1: Generalized Pareto curves of pre-tax national income

[7] Note that for $(n-1)/n \le p < 1$, we have $\hat{b}_n(p) = 1$ regardless of the distribution of X. This speaks to the impossibility of directly estimating asymptotic quantities from a finite sample. However, with fiscal data, for which samples are extremely large, we need not be concerned by the problem until extremely narrow top income groups.

In both countries, $b(p)$ does appear to converge toward a value strictly above one, which confirms that the distribution of income is an asymptotic power law. However, the coefficients vary significantly, even within the top decile, so that the strict Pareto assumption will miss important patterns in the distribution. Because $b(p)$ rises within the top 10% of the distribution, inequality in both France and the United States is in fact even more skewed toward the very top than what the standard Pareto model suggests. And the amount by which inverted Pareto coefficients vary is not negligible. For the United States, in 2010, at its lowest point (near $p = 80\%$), $b(p)$ is around 2.4. If it were a strict Pareto distribution, it would correspond to the top 1% owning 15% of the income. But the asymptotic value is closer to 3.3, which would mean a top 1% share of 25%.

Though empirical evidence leads us to reject the strict Pareto assumption, we can notice that the generalized Pareto curves are U-shaped. We observe that fact for all countries and time periods for which we have sufficient data.

1.1.4 Processes Generating Nonconstant Pareto Curves

The emergence of the Pareto distribution for the distribution of income and wealth is generally explained by models in which random multiplicative shocks accumulate over time (Gabaix, 2009). While these models have been used to justify the use of the standard Pareto distribution, we show below that it can be extended to justify the type of varying $b(p)$ that we observe in practice.

The key feature explaining the Pareto shape is scale invariance: the evolution of individual incomes is subject to random multiplicative shocks that are the same regardless of where people are in the distribution. We can model this in continuous time using a stochastic differential equation:

$$\frac{\mathrm{d}X_t}{X_t} = \mu \,\mathrm{d}t + \sigma \,\mathrm{d}W_t \tag{1.1}$$

where X_t is the value of income at the date t, and W_t is a Wiener process (i.e. a Brownian motion). It means that the rate of growth of income ($\mathrm{d}X_t/X_t$) over a small time period $[t, t + \mathrm{d}t]$ is random and independent from X_t with a constant mean $\mu \,\mathrm{d}t$ and a constant variance $\sigma^2 \,\mathrm{d}t$. If relation (1.1) holds exactly throughout the entire distribution, then the process does not converge and income follows a log-normal distribution. However, if we add some friction that prevents income from becoming too small, then we can get a stationary distribution. To that end, Gabaix (1999) suggested the introduction of a reflecting barrier at a positive income level.

An alternative approach is to make the variance of *relative* income growth go to infinity at low income levels, so that the variance of *absolute* income growth remains well above zero (Saichev, Malevergne, and Sornette, 2010, p. 17).

While the focus of these models has been to explain standard Pareto behavior, a natural extension can justify the shapes of the Pareto curves that we observe in practice. Let us generalize the process (1.2) as:

$$\frac{\mathrm{d}X_t}{X_t} = \mu(X_t)\,\mathrm{d}t + \sigma(X_t)\,\mathrm{d}W_t \tag{1.2}$$

That is, we allow both the mean and the variance of shocks to change with income. Then we can state the following result:

Theorem 1. *Let X_t follow the stochastic differential equation (1.2) with a stationary distribution $\mathcal{D}$. If both $\mu(x)$ and $\sigma^2(x)$ converge toward a constant, $\mathcal{D}$ is a power law in the sense of definition 2.*

Theorem 1 makes the connection between asymptotic power laws and the asymptotic behavior of the process generating them. The Pareto distribution arises because of scale invariance (above a certain threshold) in the stochastic process that describes the evolution of income or wealth. But if the scale invariance doesn't hold exactly but only asymptotically, then instead of a Pareto distribution we get an asymptotic power law. We can specify more precisely the shape of the stationary distribution:

Theorem 2. *Assume that $\mu(x)$ and $\sigma^2(x)$ converge toward a constant: $\mu(x) \to \mu$, $\sigma^2(x) \to \sigma^2$. Define:*

$$\zeta(x) = 1 - \frac{2\mu(x)}{\sigma^2(x)} \qquad \text{and} \qquad \zeta = \lim_{x\to+\infty} \zeta(x) = 1 - \frac{2\mu}{\sigma^2}$$

Let f be the density of the stationary distribution, and F its CDF. We have:

$$f(x) \propto x^{-\zeta-1} \exp\left(-\log(\sigma^2(x)) - \int_1^x \frac{\zeta(t)-\zeta}{t}\,\mathrm{d}t\right)$$

and:

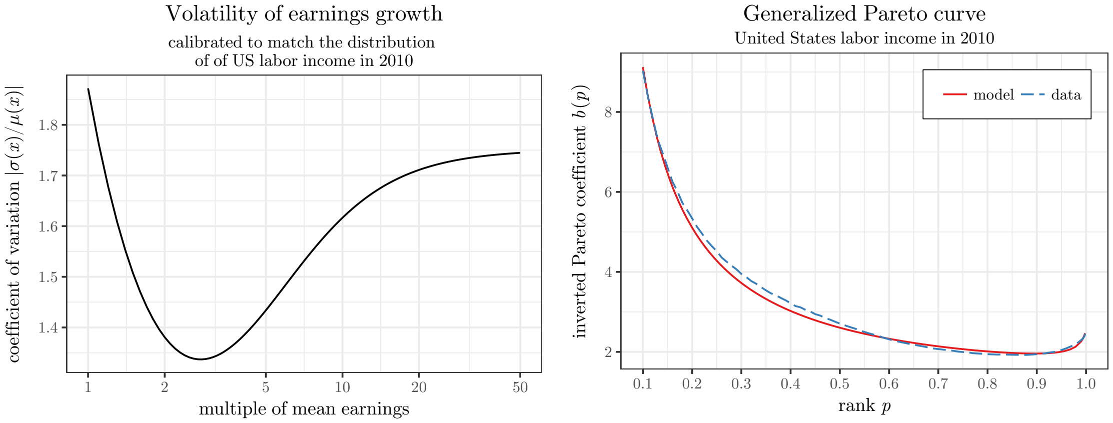
$$1 - F(x) = L(x)x^{-\zeta}$$

where L is a slowly varying function. Therefore, the stationary distribution has an asymptotic inverted Pareto coefficient equal to $1 - \sigma^2/(2\mu)$.

The asymptotic behavior of the process determines the asymptotic value of the Pareto coefficient. The characteristics of the process in the lower part of the distribution explain the rest of the Pareto curve. To explore this issue in more details, we can perform the following calibration exercise. Assume that average income growth is constant but that the variance of income has the following functional form for $\sigma(x)$, with $c_1, c_2, c_3, c_4 > 0$:

$$\sigma(x) = \sqrt{\frac{c_1 + c_2 x^2}{x^2} + \frac{c_3 x^2}{1 + c_4 x^2}} \tag{1.3}$$

The first term, $(c_1 + c_2 x^2)/x^2$, ensures that the variance goes to infinity as x goes to zero. The second term, $c_3 x^2/(1 + c_4 x^2)$, allows the variance to increase for high income. For $x \to +\infty$, we get $\sigma(x) \to \sqrt{1 + c_2 + c_3/c_4}$, so according to theorem 1 the stationary distribution is a power law. This corresponds to a U-shaped variance, with a high relative volatility of earnings at the bottom and at the top, and a lower one for the middle of the distribution. This profile of variance is in fact strongly suggested by empirical work on panel data using either surveys (Chauvel and Hartung, 2014; Hardy and Ziliak, 2014; Bania and Leete, 2009) or administrative data (Guvenen et al., 2015). Hardy and Ziliak (2014) describe it as the "wild ride" at the top and the bottom of the distribution. We also include a reflecting barrier at zero to prevent incomes from becoming negative, which helps makes the process stationary.



Volatility of earnings growth
calibrated to match the distribution
of of US labor income in 2010

Generalized Pareto curve
United States labor income in 2010

Model calibrated to match the US distribution of labor income in 2010 ($c_1 = 2.341, c_2 = 1.104, c_3 = 0.061, c_4 = 0.031$). The coefficient of variation corresponds to the standard deviation divided by the absolute value of the mean growth of non-reflected units.

Figure 1.2: Calibration of $\sigma(x)$ on the US Distribution of Labor Income

We calibrate formula (1.3) so that the Pareto curve of the stationary distribution matches actual data. Figure 1.2 shows the results for the United States labor income in 2010. The volatility of earnings growth has indeed a U-shaped profile. At the very top of the distribution, the volatility of earnings shocks is about 30% higher than at its lowest point, which occurs around the 90% percentile. Overall, this model is able to match most of the distribution of income, as shown by the two similar Pareto curves in figure 1.2. We can achieve similar results by adjusting the mean of shocks rather than the variance, or when looking at the case of wealth, and we present those results in appendix A.3. Modest breaks in scale invariance can therefore explain the variation of Pareto coefficients we observe in practice. The nee to break with scale invariance is consistent with other findings in the litterature: as Gabaix et al. (2016) explains, such deviation from scale invariance are also necessary to explain the pace of increase of inequality.

1.2 Generalized Pareto Interpolation

The tabulations of income or wealth such as those provided by tax authorities and national statistical institutes typically take the form of K fractiles $0 \leq p_1 < \cdots < p_K < 1$ of the population, alongside their income quantiles $q_1 < \cdots < q_K$ and the income share of each bracket $[p_k, p_{k+1}]$.[8] The interpolation method that we now present uses the way inverted Pareto coefficients vary smoothly to estimate a complete distribution based solely on that information: we call it *generalized Pareto interpolation.*

The first goal of the method is to be as flexible as we are allowed to be: that is, we do not force the estimated distribution into a predetermined shape. We stress that a fully nonparametric approach is not possible here due to the lack of a suitable asymptotic framework.[9] But we can still get a lot more flexibility than a strict Pareto model by introducing a large enough number of parameters. The second goal is to generate a solution with desirable properties. Indeed the interpolation problem is technically ill-posed as it has an infinite number of candidate solutions. Our method overcomes that issue by looking for a “regular” curve of Pareto coefficients.

Our method combines three components, which solve different aspects of the problem. First, we interpolate the generalized Pareto curve in a way that maximizes its smoothness while satisfying two sets of constraints: those related to the quantiles, and those related to the means. Second, we enforce if necessary the constraint that the quantile function is increasing by finding an admissible solution that is as close as possible to the original one. Finally, we deal separately with last bracket, for which the interpolation is not possible due to the lack of an endpoint in the interval.

[8] That last element may take diverse forms (top income shares, bottom income shares, average income in the brackets, average income above the bracket, etc.), all of which are just different ways of presenting the same information.

[9] The number of brackets would have to go to infinity, which is not the setting we are interested in.

For the exposition of the method, we will set aside sampling related issues, and treat empirical quantities as equivalent to their theoretical counterpart. But we come back to that issue in section 1.4.

1.2.1 Interpolation of the Pareto Coefficients

The tabulations let us compute $b(p_1), \dots, b(p_K)$ directly. But interpolating the curve $b(p)$ based solely on those points offers no guarantee that the resulting function will be consistent with the input data on quantiles. To that end, the interpolation needs to be constrained. To do so in a computationally efficient and analytically tractable way, we start from the following function:

$$\forall x \geq 0 \qquad \varphi(x) = -\log \int_{1-\mathrm{e}^{-x}}^{1} Q(p)\,\mathrm{d}p$$

which is essentially a transform of the Lorenz curve:

$$\varphi(x) = -\log((1 - L(p))\mathbb{E}[X])$$

with $p = 1 - \mathrm{e}^{-x}$. The value of φ at each point $x_k = -\log(1 - p_k)$ can therefore be estimated directly from the data in the tabulation. Moreover:

$$\forall x \geq 0 \qquad \varphi'(x) = \mathrm{e}^{\varphi(x)-x} Q(1 - \mathrm{e}^{-x}) = 1/b(1 - \mathrm{e}^{-x})$$

which means that the generalized Pareto coefficient $b(p)$ is equal to $1/\varphi'(x)$. Hence, the value of $\varphi'(x_k)$ for $k \in \{1, \dots, K\}$ is also given by the tabulation.

Because of the bijection between $(p, b(p), Q(p))$ and $(x, \varphi(x), \varphi'(x))$, the problem of interpolating $b(p)$ in a way that is consistent with $Q(p)$ is identical to that of interpolating the function φ, whose value and first derivative are known at each point x_k.

We assume that we know a set of points $\{(x_k, y_k, s_k), 1 \leq k \leq K\}$ that correspond to the values of $\{(x_k, \varphi(x_k), \varphi'(x_k)), 1 \leq k \leq K\}$, and we seek a sufficiently smooth function $\hat{\varphi}$ such that:

$$\forall k \in \{1, \dots, K\} \qquad \hat{\varphi}(x_k) = \varphi(x_k) = y_k \qquad \hat{\varphi}'(x_k) = \varphi'(x_k) = s_k \tag{1.4}$$

By sufficiently smooth, we mean that φ should be at least twice continuously differentiable. That requirement is necessary for the estimated Pareto curve (and by extension the quantile function) to be once continuously differentiable, or, put differently, not to exhibit any asperity at the fractiles included in the tabulation.

Our interpolation method relies on splines, meaning piecewise polynomials defined on each interval $[x_k, x_{k+1}]$. Although cubic splines (i.e. polynomials of degree 3) are the most common, they do not offer enough degrees of freedom to satisfy both the constraints given by (1.4) and the requirement that φ is twice continuously differentiable. We use quintic splines (i.e. polynomials of degree 5) to get more flexibility. To construct them, we start from the following set of polynomials for $x \in [0, 1]$:

$$\begin{aligned} h_{00}(x) &= 1 - 10x^3 + 15x^4 - 6x^5 & h_{01}(x) &= 10x^3 - 15x^4 + 6x^5 \\ h_{10}(x) &= x - 6x^3 + 8x^4 - 3x^5 & h_{11}(x) &= -4x^3 + 7x^4 - 3x^5 \\ h_{20}(x) &= \tfrac{1}{2}x^2 - \tfrac{3}{2}x^3 + \tfrac{3}{2}x^4 - \tfrac{1}{2}x^5 & h_{21}(x) &= \tfrac{1}{2}x^3 - x^4 + \tfrac{1}{2}x^5 \end{aligned}$$

which were designed so that $h_{ij}^{(k)}(\ell) = 1$ if $(i, j) = (k, \ell)$, and 0 otherwise. They are analogous to the basis of cubic Hermite splines (e.g. McLeod and Baart, 1998, p. 328), but for the set of polynomials of degree up to five. Then, for $k \in \{1, \dots, K-1\}$ and $x \in [x_k, x_{k+1}]$, we set:

$$\begin{aligned} \hat{\varphi}_k(x) = {} & y_k h_{00}\left(\tfrac{x-x_k}{x_{k+1}-x_k}\right) + y_{k+1} h_{01}\left(\tfrac{x-x_k}{x_{k+1}-x_k}\right) \\ & + s_k (x_{k+1} - x_k) h_{10}\left(\tfrac{x-x_k}{x_{k+1}-x_k}\right) + s_{k+1}(x_{k+1} - x_k) h_{11}\left(\tfrac{x-x_k}{x_{k+1}-x_k}\right) \\ & + a_k (x_{k+1} - x_k)^2 h_{20}\left(\tfrac{x-x_k}{x_{k+1}-x_k}\right) + a_{k+1}(x_{k+1} - x_k)^2 h_{21}\left(\tfrac{x-x_k}{x_{k+1}-x_k}\right) \end{aligned}$$

for some $a_k, a_{k+1} \in \mathbb{R}$, and $\hat{\varphi}(x) = \hat{\varphi}_k(x)$ for $x \in [x_k, x_{k+1}]$. By construction, we have $\hat{\varphi}(x_k) = y_k$, $\hat{\varphi}(x_k) = y_{k+1}$, $\hat{\varphi}'(x_k) = s_k$, $\hat{\varphi}'(x_{k+1}) = s_{k+1}$, $\hat{\varphi}''(x_k) = a_k$ and $\hat{\varphi}''(x_{k+1}) = a_{k+1}$. Hence, $\hat{\varphi}$ satisfies all the constraints and regularity requirements of the problem.

To pick appropriate values for $a_1, \dots, a_k$, we follow the usual approach of imposing additional regularity conditions at the jointures. We have a system of $K-2$ equations, linear in $a_1, \dots, a_k$, defined by:

$$\forall k \in \{2, \dots, K-1\} \qquad \hat{\varphi}_{k-1}'''(x_k) = \hat{\varphi}_k'''(x_k)$$

Two additional equations are required for that system to have a unique solution. One solution is to use predetermined values for a_1 and a_K (known as the "clamped spline"). Another, known as the "natural spline", sets:

$$\hat{\varphi}_1'''(x_1) = 0 \qquad \text{and} \qquad \hat{\varphi}_{K-1}'''(x_K) = 0$$

Both approaches are equivalent to the minimization of an irregularity criterion (e.g. Lyche and Mørken, 2002):

$$\min_{a_1,\dots,a_K} \int_{x_1}^{x_k} \{\hat{\varphi}'''(x)\}^2 \, \mathrm{d}x$$

subject to fixed values for a_1 and a_K (clamped spline) or not (natural spline). Hence, both methods can be understood as a way to minimize the curvature of $\hat{\varphi}'$, and therefore find a regular $b(p)$. That is, by construction, the method aims at finding the most "regular" generalized Pareto curve that satisfies the constraints of the problem.

We adopt a hybrid approach, in which a_1 is determined through $\hat{\varphi}_1'''(x_1) = 0$, but where a_K is estimated separately using the two-points finite difference:

$$a_K = \frac{s_K - s_{K-1}}{x_K - x_{K-1}}$$

Because the function is close to linear near x_K, it yields results that are generally similar to traditional natural splines. But that estimation of $\varphi''(x_K)$ is also more robust, so we get more satisfactory results when the data exhibit potentially troublesome features.

The vector $\boldsymbol{a} = [a_1 \quad \cdots \quad a_K]'$ is the solution of a linear system of equation $\boldsymbol{X}\boldsymbol{a} = \boldsymbol{v}$, where $\boldsymbol{X}$ depends solely on the $x_1, \dots, x_K$, and $\boldsymbol{b}$ is linear in $y_1, \dots, y_K$ and $s_1, \dots, s_K$. Therefore, we find the right parameters for the spline by numerically solving a linear system of equation. We provide the detailed expressions of $\boldsymbol{X}$ and $\boldsymbol{b}$ in appendix A.4.

1.2.2 Enforcing Admissibility Constraints

The interpolation method presented above does not guarantee that the estimated generalized Pareto curve will satisfy property 1 — or equivalently that the quantile will be an increasing function. In most situations, that constraint need not be enforced, because it is not binding: the estimated function spontaneously satisfy it. But it may occasionally not be the case, so that estimates of quantiles of averages at different points of the distribution may be mutually inconsistent. To solve that problem, we present an *ex post* adjustment procedure which constrains appropriately the interpolated function.

We can express the quantile as a function of φ:

$$\forall x \geq 0 \qquad Q(1 - \mathrm{e}^{-x}) = \mathrm{e}^{x - \varphi(x)} \varphi'(x)$$

Therefore:

$$\forall x \geq 0 \qquad Q'(1-\mathrm{e}^{-x}) = \mathrm{e}^{2x-\varphi(x)}[\varphi''(x) + \varphi'(x)(1-\varphi'(x))]$$

So the estimated quantile function is increasing if and only if:

$$\forall x \geq 0 \qquad \Phi(x) = \hat{\varphi}''(x) + \hat{\varphi}'(x)(1-\hat{\varphi}'(x)) \geq 0 \tag{1.5}$$

The polynomial Φ (of degree 8) needs to be positive. There are no simple necessary and sufficient conditions on the parameters of the spline that can ensure such a constraint. However, it is possible to derive conditions that are only sufficient, but general enough to be used in practice. We use conditions based on the Bernstein representation of polynomials, as derived by Cargo and Shisha (1966):

Theorem 3 (Cargo and Shisha, 1966). *Let $P(x) = c_0 + c_1 x_1 + \cdots + c_n x^n$ be a polynomial of degree $n \geq 0$ with real coefficients. Then:*

$$\forall x \in [0,1] \qquad \min_{0\leq i\leq n} b_i \leq P(x) \leq \max_{0\leq i\leq n} b_i$$

where:

$$b_i = \sum_{r=0}^{n} c_r \binom{i}{r} \bigg/ \binom{n}{r}$$

To ensure that the quantile is increasing over $[x_k, x_{k+1}]$ $(1 \leq k < K)$, it is therefore enough to enforce the constraint that $b_i \geq 0$ for all $0 \leq i \leq 8$, where b_i is defined as in theorem 3 with respect to the polynomial $x \mapsto \Phi(x_k + x(x_{k+1} - x_k))$. Those 9 conditions are all explicit quadratic forms in $(y_k, y_{k+1}, s_k, s_{k+1}, a_k, a_{k+1})$, so we can compute them and their derivative easily.

To proceed, we start from the unconstrained estimate from the previous section. We set $a_k = -s_k(1-s_k)$ for each $1 \leq k \leq K$ if $a_k + s_k(1-s_k) < 0$, which ensures that condition (1.5) is satisfied at least at the interpolation points. Then, over each segment $[x_k, x_{k+1}]$, we check whether the condition $\Phi(x) \geq 0$ is satisfied for $x \in [x_k, x_{k+1}]$ using the theorem 3, or more directly by calculating the values of Φ over a tight enough grid of $[x_k, x_{k+1}]$. If so, we move on to next segment. If not, we consider $L \geq 1$ additional points $(x_1^*, \dots, x_L^*)$ such that $x_k < x_1^* < \cdots < x_L^* < x_{k+1}$, and we redefine the function $\hat{\varphi}_k$ over $[x_k, x_{k+1}]$ as:

$$\tilde{\varphi}_k(x) = \begin{cases} \varphi_0^*(x) & \text{if} \quad x_k \le x < x_1^* \\ \varphi_\ell^*(x) & \text{if} \quad x_\ell^* \le x < x_{\ell+1}^* \\ \varphi_L^*(x) & \text{if} \quad x_L^* \le x < x_{k+1} \end{cases}$$

where the φ_ℓ^* $(0 \le \ell \le L)$ are quintic splines such that for all $1 \le \ell < L$:

$$\begin{aligned} \varphi_0^*(x_k) &= y_k & (\varphi_0^*)'(x_k) &= s_k & (\varphi_0^*)''(x_k) &= a_k \\ \varphi_L^*(x_{k+1}) &= y_{k+1} & (\varphi_L^*)'(x_{k+1}) &= s_{k+1} & (\varphi_L^*)''(x_{k+1}) &= a_{k+1} \\ \varphi_\ell^*(x_\ell^*) &= y_\ell^* & (\varphi_\ell^*)'(x_\ell^*) &= s_\ell^* & (\varphi_\ell^*)''(x_\ell^*) &= a_\ell^* \\ \varphi_\ell^*(x_{\ell+1}^*) &= y_{\ell+1}^* & (\varphi_\ell^*)'(x_{\ell+1}^*) &= s_{\ell+1}^* & (\varphi_\ell^*)''(x_{\ell+1}^*) &= a_{\ell+1}^* \end{aligned}$$

and $y_\ell^*, s_\ell^*, a_\ell^*$ $(1 \le \ell \le L)$ are parameters to be adjusted. In simpler terms, we divided the original spline into several smaller ones, thus creating additional parameters that can be adjusted to enforce the constraint. We set the parameters $y_\ell^*, s_\ell^*, a_\ell^*$ $(1 \le \ell \le L)$ by minimizing the L^2 norm between the constrained and the unconstrained estimate, subject to the $9 \times (L + 1)$ conditions that $b_i^\ell \ge 0$ for all $0 \le i \le 8$ and $0 \le \ell \le L$:

$$\min_{\substack{y_\ell^*, s_\ell^*, a_\ell^* \\ 1 \le \ell \le L}} \int_{x_k}^{x_{k+1}} \{\hat{\varphi}_k(x) - \tilde{\varphi}_k(x)\}^2 \,\mathrm{d}x \qquad \text{st.} \qquad b_i^\ell \ge 0 \quad (0 \le i \le 8 \text{ and } 0 \le \ell \le L)$$

where the b_i^ℓ are defined as in theorem 3 for each spline ℓ. The objective function and the constraints all have explicit analytical expressions, and so does their gradients. We solve the problem with standard numerical methods for nonlinear constrained optimization.[10,11]

1.2.3 Extrapolation in the Last Bracket

The interpolation procedure only applies to fractiles between p_1 and p_K, but we generally also want an estimate of the distribution outside of this range, especially for $p > p_K$.[12] Because there is no direct estimate of the asymptotic Pareto coefficient $\lim_{p\to 1} b(p)$, it is not possible to interpolate as we did for the rest of the distribution:

[10] For example, standard sequential quadratic programming (Kraft, 1994) or augmented Lagrangian methods (Conn, Gould, and Toint, 1991; Birgin and Martìnez, 2008). See NLopt for details and open source implementations of such algorithms: `http://ab-initio.mit.edu/wiki/index.php/NLopt_Algorithms`.

[11] Adding one point at the middle of the interval is usually enough to enforce the constraint, but more points may be added if convergence fails.

[12] It is always possible to set $p_1 = 0$ if the distribution has a finite lower bound.

we need to extrapolate it.

The extrapolation in the last bracket should satisfy the constraints imposed by the tabulation (on the quantile and the mean). In accordance with the principle of a regular Pareto curve, it should also ensure derivability of the quantile function at the juncture. To do so, we use the information contained in the four values (x_K, y_K, s_K, a_K) of the interpolation function at the last point. Hence, we need an appropriate functional form for the last bracket with enough degrees of freedom to satisfy all the constraints. To that end, we turn to the generalized Pareto distribution.

Definition 3 (Generalized Pareto distribution)**.** *Let $\mu \in \mathbb{R}$, $\sigma \in]0, +\infty[$ and $\xi \in \mathbb{R}$. X follows a generalized Pareto distribution if for all $x \geq \mu$ $(\xi \geq 0)$ or $\mu \leq x \leq \mu - \sigma/\xi$ $(\xi < 0)$:*

$$\mathbb{P}\{X \leq x\} = \text{GPD}_{\mu,\sigma,\xi}(x) = \begin{cases} 1 - \left(1 + \xi \frac{x-\mu}{\sigma}\right)^{-1/\xi} & \text{for } \xi \neq 0 \\ 1 - e^{-(x-\mu)/\sigma} & \text{for } \xi = 0 \end{cases}$$

μ is called the location parameter, σ the scale parameter and ξ the shape parameter.

The generalized Pareto distribution is a fairly general family which includes as special cases the strict Pareto distribution ($\xi > 0$ and $\mu = \sigma/\xi$), the (shifted) exponential distribution ($\xi = 0$) and the uniform distribution ($\xi = -1$). It was popularized as a model of the tail of other distributions in extreme value theory by Pickands (1975) and Balkema and Haan (1974), who showed that for a large class of distributions (which includes all power laws in the sense of definition 2), the tail converges towards a generalized Pareto distribution.

If $X \sim \text{GPD}(\mu, \sigma, \xi)$, the generalized Pareto curve of X is:

$$b(p) = 1 + \frac{\xi\sigma}{(1-\xi)[\sigma + (1-p)^\xi(\mu\xi - \sigma)]}$$

We will focus on cases where $0 < \xi < 1$, so that the distribution is a power law at the limit ($\xi > 0$), but its mean remains finite ($\xi < 1$). When $\xi\mu = \sigma$, the generalized Pareto curve is constant, and the distribution is a strict power law with Pareto coefficient $b = 1/(1-\xi)$. That value also corresponds in all cases to the asymptotic coefficient $\lim_{p\to 1} b(p) = 1/(1-\xi)$. But there are several ways for the distribution to converge toward a power law, depending on the sign of $\mu\xi - \sigma$. When $\mu\xi - \sigma > 0$, $b(p)$ converges from below, increasing as $p \to 1$, so that the distribution gets more unequal in higher brackets. Conversely, when $\mu\xi - \sigma < 0$, $b(p)$ converges from above, and decreases as $p \to 1$, so that the distribution is more equal in higher brackets.

The generalized Pareto distribution can match a wide diversity of profiles for the behavior of $b(p)$, while offering the right number of degrees of freedom for our purpose. In the context of our method, the value of its parameters is not of direct interest. In particular, the setting does not allow for a particularly accurate estimation of the asymptotic Pareto coefficient, and we do not focus on providing such an estimate. However, we can use it to find a reasonable functional form that makes an efficient use of the information at our disposal on the mean, the quantile and its derivative at the last threshold. The generalized Pareto distribution offers a way to extrapolate the coefficients $b(p)$ in a way that is consistent with all the input data and preserves the regularity of the Pareto curve.

We assume that, for $p > p_K$, the distribution follows a generalized Pareto distribution with parameters (μ, σ, ξ), which means that for $q > q_K$ the CDF is:

$$F(q) = p_K + (1 - p_K)\mathrm{GPD}_{\mu,\sigma,\xi}(q)$$

For the CDF to remain continuous and differentiable, we need $\mu = q_K$ and $\sigma = (1 - p_K)/F'(q_K)$, where $F'(q_K)$ comes from the interpolation method of section 1.2.1. Finally, for the Pareto curve to remain continuous, we need $b(p_K)$ equal to $1 + \sigma/(\mu(1 - \xi))$, which gives the value of ξ. That is, if we set the parameters (μ, σ, ξ) equal to:

$$\begin{aligned}
\mu &= s_K \mathrm{e}^{x_K - y_K} \\
\sigma &= (1 - p_K)(a_K + s_K(1 - s_K))\mathrm{e}^{2x_K - y_K} \\
\xi &= 1 - \frac{(1 - p_K)\sigma}{\mathrm{e}^{-y_K} - (1 - p_K)\mu}
\end{aligned}$$

then the resulting distribution will have a continuously differentiable quantile function, and will match the quantiles and the means in the tabulation.

1.3 Tests Using Income Data from the United States and France, 1962–2014

We test the quality of our interpolation method using income tax data for the United States (1962–2014) and France (1994–2012).[13] They correspond to cases for which we have detailed tabulations of the distribution of pre-tax national income based on quasi-exhaustive individual tax data (Piketty, Saez, and Zucman, 2016;

[13] More precisely, the years 1962, 1964 and 1966–2014 for the United States.

Garbinti, Goupille-Lebret, and Piketty, 2016), so that we can know quantiles or shares exactly.[14] We compare the size of the error in generalized Pareto interpolation with alternatives most commonly found in the literature.

1.3.1 Overview of the Most Common Interpolation Methods

Method 1: constant Pareto coefficient That method was used by Piketty (2001) and Piketty and Saez (2003), and relies on the property that, for a Pareto distribution, the inverted Pareto coefficient $b(p)$ remains constant. We set $b(p) = b = \mathbb{E}[X|X > q_k]/q_k$ for all $p \geq p_k$. The p-th quantile becomes $q = q_k \left(\frac{1-p}{1-p_k}\right)^{-1/\alpha}$ with $\alpha = b/(b-1)$. By definition, $\mathbb{E}[X|X > q] = bq$ which gives the p-th top average and top share.

Method 2: log-linear interpolation The log-linear interpolation method was introduced by Pareto (1896), Kuznets (1953), and Feenberg and Poterba (1992). It uses solely threshold information, and relies on the property of Pareto distributions that $\log(1 - F(x)) = \log(c) - \alpha \log(x)$. We assume that this relation holds exactly within the bracket $[p_k, p_{k+1}]$, and set $\alpha_k = -\frac{\log((1-p_{k+1})/(1-p_k))}{\log(q_{k+1}/q_k)}$. The value of the p-th quantile is again $q = q_k \left(\frac{1-p}{1-p_k}\right)^{-1/\alpha_k}$ and the top averages and top shares can be obtained by integration of the quantile function. For $p > p_K$, we extrapolate using the value α_K of the Pareto coefficient in the last bracket.

Method 3: mean-split histogram The mean-split histogram uses information on both the means and the thresholds, but uses a very simple functional form, so that the solution can be expressed analytically. Inside the bracket $[q_k, q_{k+1}]$, the density takes two values:

$$f(x) = \begin{cases} f_k^- & \text{if} \quad q_k \leq x < \mu_k \\ f_k^+ & \text{if} \quad \mu_k \leq x < q_{k+1} \end{cases}$$

[14] We use pre-tax national income as our income concept of reference. It was defined by Alvaredo, Atkinson, et al. (2016) to be consistent with the internationally agreed definition of net national income in the system of national accounts. Even though they are mostly based on individual tax data, estimates of pre-tax national income do involves a few corrections and imputations, which may affect the results. That is why we also report similar computations in appendix using fiscal income, which is less comparable and less economically meaningful, but doesn't suffer from such problems.

where μ_k is the mean inside the bracket.[15] To meet the requirement on the mean and the thresholds, we set:

$$f_k^- = \frac{(p_{k+1} - p_k)(q_{k+1} - \mu_k)}{(q_{k+1} - q_k)(\mu_k - q_k)} \qquad \text{and} \qquad f_k^+ = \frac{(p_{k+1} - p_k)(\mu_k - q_k)}{(q_{k+1} - q_k)(q_{k+1} - \mu_k)}$$

The means-split histogram does not apply beyond the last threshold of the tabulation.

Comparison Methods 1 and 2 make a fairly inefficient use of the information included in the original tabulation: method 1 discards the data on quantiles and averages at the higher end of the bracket, while method 2 discards the information on averages. As a consequence, none of these methods can guarantee that the output will be consistent with the input. The method 3 does offer such a guarantee, but with a very simple — and quite unrealistic — functional form.

Our generalized Pareto interpolation method makes use of all the information in the tabulation, so that its output is guaranteed to be consistent with its input. Moreover, contrary to all other methods, it leads a continuous density, hence a smooth quantile and a smooth Pareto curve. None of the other methods can satisfy this requirement, and their output exhibit stark irregularities at the beginning and the end of the brackets in the tabulation in input.

Application to France and the United States Using the individual income tax data, we compute our own tabulations in each year. We include four percentiles in the tabulation: $p_1 = 0.1$, $p_2 = 0.5$, $p_3 = 0.9$ and $p_4 = 0.99$.

We interpolate each of those tabulations with the three methods above, labelled "M1", "M2" and "M3" in what follows.[16] We also interpolate them with our new generalized Pareto interpolation approach (labeled "M0"). We compare the values that we get with each method for the top shares and the quantiles at percentiles 30%, 75% and 95% with the value that we get directly from the individual data. (We divide all quantiles by the average to get rid of scaling effects due to inflation

[15] The breakpoint of the interval $[q_k, q_{k+1}]$ could be different from μ_k, but not all values between q_k and q_{k+1} will work if we want to make sure that $f_k^- > 0$ and $f_k^+ > 0$. The breakpoint q^* must be between q_k and $2\mu_k - q_k$ if $\mu_k < (q_k + q_{k+1})/2$, and between $2\mu_k - q_{k+1}$ and q_{k+1} otherwise. Choosing $q^* = \mu_k$ ensures that the condition is always satisfied.

[16] We also provide extended tables in appendix with a fourth method, which is much more rarely used.

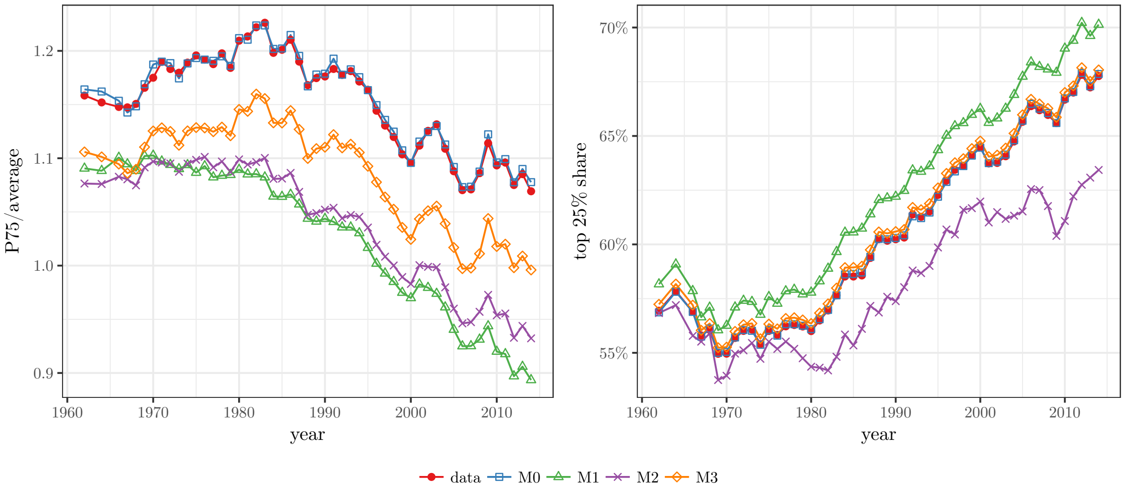
Table 1.1: Mean relative error for different interpolation methods

		mean percentage gap between estimated and observed values			
		M0	M1	M2	M3
United States (1962–2014)	Top 70% share	0.059% (ref.)	2.3% (×38)	6.4% (×109)	0.054% (×0.92)
	Top 25% share	0.093% (ref.)	3% (×32)	3.8% (×41)	0.54% (×5.8)
	Top 5% share	0.058% (ref.)	0.84% (×14)	4.4% (×76)	0.83% (×14)
	P30/average	0.43% (ref.)	55% (×125)	29% (×67)	1.4% (×3.3)
	P75/average	0.32% (ref.)	11% (×35)	9.9% (×31)	5.8% (×18)
	P95/average	0.3% (ref.)	4.4% (×15)	3.6% (×12)	1.3% (×4.5)
France (1994–2012)	Top 70% share	0.55% (ref.)	4.2% (×7.7)	7.3% (×13)	0.14% (×0.25)
	Top 25% share	0.75% (ref.)	1.8% (×2.4)	4.9% (×6.5)	0.37% (×0.49)
	Top 5% share	0.29% (ref.)	1.1% (×3.9)	8.9% (×31)	0.49% (×1.7)
	P30/average	1.5% (ref.)	59% (×40)	38% (×26)	2.6% (×1.8)
	P75/average	1% (ref.)	5.2% (×5.1)	5.4% (×5.3)	4.7% (×4.6)
	P95/average	0.58% (ref.)	5.6% (×9.6)	3.2% (×5.5)	1.8% (×3.2)

Pre-tax national income. Sources: author's calculation from Piketty, Saez, and Zucman (2016) (United States) and Garbinti, Goupille-Lebret, and Piketty (2016) (France). The different interpolation methods are labeled as follows. M0: generalized Pareto interpolation. M1: constant Pareto coefficient. M2: log-linear interpolation. M3: mean-split histogram. We applied them to a tabulation which includes the percentiles $p = 10\%$, $p = 50\%$, $p = 90\%$, and $p = 99\%$. We included the relative increase in the error compared to generalized Pareto interpolation in parentheses. We report the mean relative error, namely:

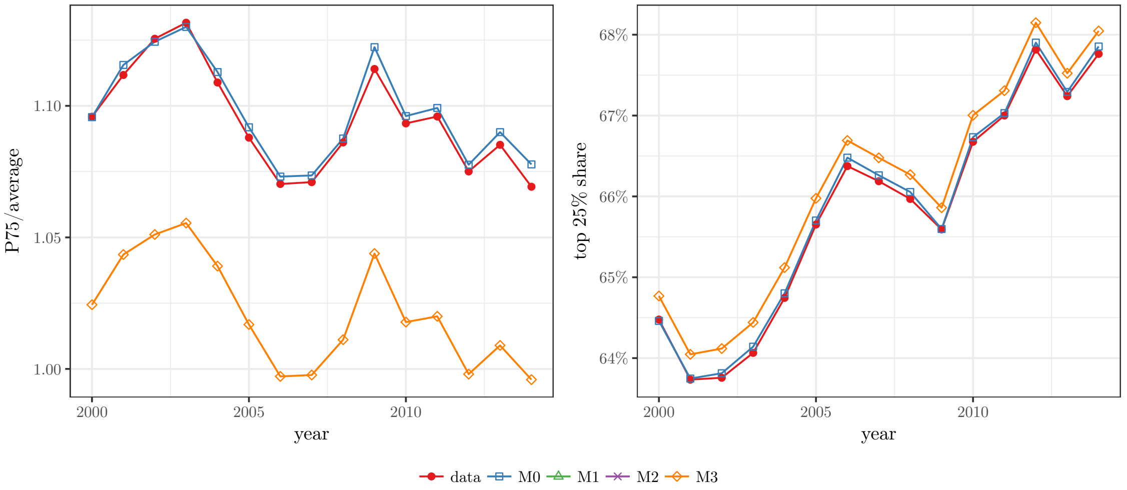
$$\frac{1}{\text{number of years}} \sum_{t=\text{first year}}^{\text{last year}} \left| \frac{\hat{y}_t - y_t}{y_t} \right|$$

where y is the quantity of interest (income threshold or top share), and $\hat{y}$ is its estimate using one of the interpolation methods. We calculated the results over the years 1962, 1964 and 1966–2014 in the United States, and years 1994–2012 in France.



Pre-tax national income. Sources: author's computation from Piketty, Saez, and Zucman (2016). M0: generalized Pareto interpolation. M1: constant Pareto coefficient. M2: log-linear interpolation. M3: mean-split histogram.

Figure 1.3: P75 threshold and top 25% share in the United States (1962–2014), estimated using all interpolation methods and a tabulation with $p = 10\%, 50\%, 90\%, 99\%$



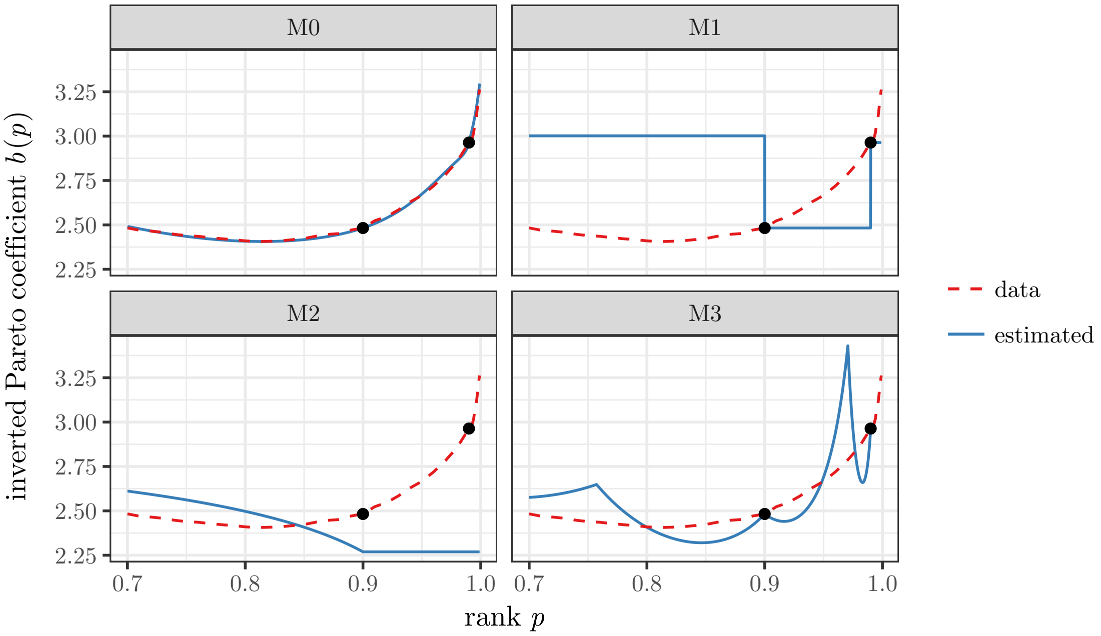
Pre-tax national income. Sources: author's computation from Piketty, Saez, and Zucman (2016). M0: generalized Pareto interpolation. M3: mean-split histogram.

Figure 1.4: P75 threshold and top 25% share in the United States (2000-2014), estimated using interpolation methods M0 and M3, and a tabulation with $p = 10\%, 50\%, 90\%, 99\%$

and average income growth.) We report the mean relative error in table 1.1:

$$\text{MRE} = \frac{1}{\text{number of years}} \sum_{t=\text{first year}}^{\text{last year}} \left| \frac{\hat{y}_t - y_t}{y_t} \right|$$

where y is the quantity of interest (income threshold or top share), and $\hat{y}$ is its estimate using one of the interpolation methods.



Pre-tax national income. Sources: author's computation from Piketty, Saez, and Zucman (2016). M0: generalized Pareto interpolation. M1: constant Pareto coefficient. M2: log-linear interpolation. M3: mean-split histogram.

Figure 1.5: Generalized Pareto curves implied by the different interpolation methods for the United States distribution of income in 2010

The two standard Pareto interpolation methods (M1 and M2) are the ones that perform worst. M1 is better at estimating shares, while M2 is somewhat better at estimating quantiles. That shows the importance not to dismiss any information included in the tabulation, as exhibited by the good performance of the mean-split histogram (M3), particularly at the bottom of the distribution.

Our generalized Pareto interpolation method vastly outperforms the standard Pareto interpolation methods (M1 and M2). It is also better than the mean-split histogram (M3), except in the bottom of the distribution where both methods work well (but standard Pareto methods M1 and M2 fail badly).

Figure 1.3 shows how the use of different interpolation methods affects the estimation of the top 25% share and associated income threshold. Although all methods roughly respect the overall trend, they can miss the level by a significant margin. The generalized Pareto interpolation estimates the threshold much better than either M1, M2, or M3.

For the estimation of the top 25% share, M3 performs fairly well, unlike M1 and M2. To get a more detailed view, we therefore focus on a more recent period (2000–2014) and display only M0 and M3, as in figure 1.4. We can see that M3 has, in that case, a tendency to overestimate the top 25% by a small yet persistent amount. In comparison, M4 produces a curve almost identical to the real one.

We can also directly compare the generalized Pareto curves generated by each method, as in figure 1.5. Our method, M0, reproduces the inverted Pareto coefficients $b(p)$ very faithfully, including above the last threshold (see section 1.3.2). All the other methods give much worse results. Method M1 leads to discontinuous curve, which in fact may not even define a consistent probability distribution. The M2 method fails to account for the rise of $b(p)$ at the top. Finally, the M3 leads to an extremely irregular shape due to the use a piecewise uniform distribution to approximate power law behavior.

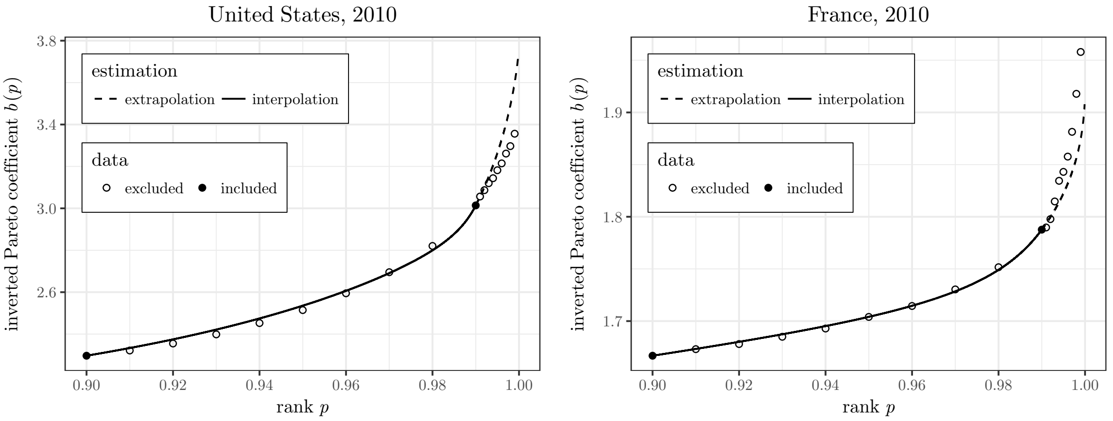
Overall, the generalized Pareto interpolation method performs well. In most cases, it gives results that are several times better than methods commonly used in the literature. And it does so while ensuring a smoothness of the resulting estimate that no other method can provide. Moreover, it works well for the whole distribution, not just the top (like M1 and M2) or the bottom (like M3).

1.3.2 Extrapolation methods

Of the interpolation methods previously described, only M1 and M2 can be used to extrapolate the tabulation beyond the last threshold. Both assume a standard Pareto distribution. Method M1 estimates $b(p)$ at the last fractile p_K, and assumes a Pareto law with $\alpha = b(p_K)/(b(p_K) - 1)$ after that. Method M2 estimates a Pareto coefficient based on the last two thresholds, so in effect it assumes a standard Pareto distribution immediately after the second to last threshold.

The assumption that $b(p)$ becomes approximately constant for p close to 1, however, is not confirmed by the data. Figure 1.6 demonstrate this for France and the United States in 2010. The profile of $b(p)$ is not constant for $p \approx 1$. On the contrary, it increases faster than for the rest of the distribution.

In section 1.2.3 we presented an extrapolation method based on the generalized



Fiscal income. Sources: author's computation from Piketty, Saez, and Zucman (2016) (for the United States) and Garbinti, Goupille-Lebret, and Piketty (2016) (for France).

Figure 1.6: Extrapolation with generalized Pareto distribution

Pareto distribution that had the advantage of preserving the smoothness of the Pareto curve, use all the information from the tabulation, and allow for a nonconstant profile of generalized Pareto coefficients near the top. As figure 1.6 shows, this method leads to a more realistic shape of the Pareto curve.

Table 1.2 compares the performance of the new method with the other ones, as we did in the previous section. Here, the tabulation in input includes $p = 90\%$ but stops at $p = 95\%$, and we seek estimates for $p = 99\%$.[17,18] Method M2 is the most imprecise. Method M1 works quite well in comparison. But our new method M0 gives even more precise results. This because it can correctly capture the tendency of $b(p)$ to keep on rising at the top of the distribution.

Figure 1.7 compares the extrapolation methods over time in the United States. We can see M1 overestimates the threshold by about as much as M2 underestimates it, while M0 is much closer to reality and makes no systematic error. For the top share, M1 is much better than M2. But it slightly underestimates the top share because it fails to account for the rising profile of inverted Pareto coefficients at the top, which is why our method M0 works even better.

[17] Here, we use fiscal income instead of pre-tax national income to avoid disturbances created at the top by the imputation of some sources of income in pre-tax national income.

[18] We provide in appendix an alternative tabulation which stops at the top 1% and where we seek the top 0.1%. The performances of M0 and M1 are closer but M0 remains preferable.

Table 1.2: Mean relative error on the top 1% for different extrapolation methods, knowing the top 10% and the top 5%

		mean percentage gap between estimated and observed values		
		M0	M1	M2
United States (1962–2014)	Top 1% share	0.78% (ref.)	5.2% (×6.7)	40% (×52)
	P99/average	1.8% (ref.)	8.4% (×4.7)	13% (×7.2)
France (1994–2012)	Top 1% share	0.44% (ref.)	2% (×4.6)	11% (×25)
	P99/average	0.98% (ref.)	2.5% (×2.5)	2.4% (×2.4)

Fiscal income. Sources: author's calculation from Piketty, Saez, and Zucman (2016) (United States) and Garbinti, Goupille-Lebret, and Piketty (2016) (France). The different extrapolation methods are labeled as follows. M0: generalized Pareto distribution. M1: constant Pareto coefficient. M2: log-linear interpolation. We applied them to a tabulation which includes the percentiles $p = 90\%$, and $p = 95\%$. We included the relative increase in the error compared to generalized Pareto interpolation in parentheses. We report the mean relative error, namely:

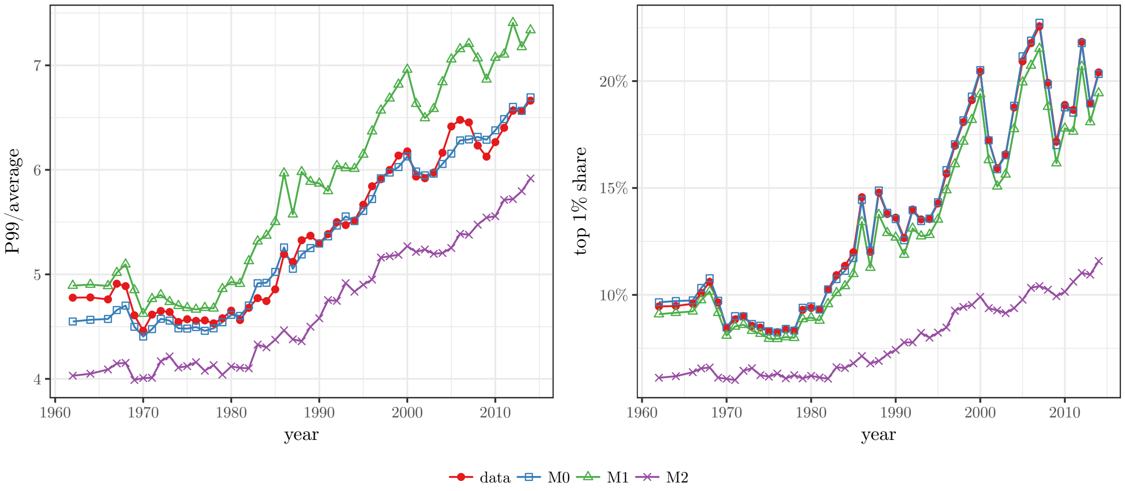
$$\frac{1}{\text{number of years}} \sum_{t=\text{first year}}^{\text{last year}} \left| \frac{\hat{y}_t - y_t}{y_t} \right|$$

where y is the quantity of interest (income threshold or top share), and $\hat{y}$ is its estimate using one of the interpolation methods. We calculated the results over the years 1962, 1964 and 1966–2014 in the United States, and years 1994–2012 in France.

1.4 Estimation Error

The previous section calculated empirically the precision of our new interpolation method. We did so by systematically comparing estimated values with real ones coming from individual tax data. But whenever we have access to individual data, we do not in fact need to perform any interpolation. So the main concern about the previous section is its general validity. To what extent can its results be extended to different tabulations, with different brackets, corresponding to a different distribution? Is it possible to get estimates of the error in the general case? How many brackets do we need to reach a given precision level, and how should they be distributed?

To the best of our knowledge, none of these issues have been tackled directly in the previous literature. The main difficulty is that most of the error is not due to mere sampling variability (although part of it is), which we can assess using standard



Fiscal income. Sources: author's computation from Piketty, Saez, and Zucman (2016).

Figure 1.7: Comparison of extrapolation methods in the United States for the top 1%, knowing the top 10% and the top 5%

methods. It comes mostly from the discrepancy between the functional forms used in the interpolation, and the true form of the distribution. Put differently, it corresponds to a "model misspecification" error, which is harder to evaluate. But the generalized Pareto interpolation method does offer some solutions to that problem. We can isolate the features of the distribution that determine the error, and based on that provide approximations of it.

In this section, we remain concerned with the same definition of the error as in the previous one. Namely, we consider the difference between the estimate of a quantity by interpolation (e.g. shares or thresholds) and the same quantity defined over the true population of interest. This is in contrast with a different notion of error common in statistics: the difference between an empirical estimate and the value of an underlying statistical model. If sample size were infinite — so that sampling variability would vanish — both errors would be identical. But despite the large samples that characterize tax data, sampling issues cannot be entirely discarded. Indeed, because income and wealth distributions are fat-tailed, the law of large numbers may operate very slowly, so that both types of errors remain different even with millions of observations (Taleb and Douady, 2015).

We consider our notion of the error to be more appropriate in the context of the methods we are studying. Indeed, concerns for the distribution of income and wealth only arise to the extent that it affects actual the actual population, not a model of it. Moreover, this allows us to remain agnostic as to the "true" model for the distribution of income.

In order to get tractable analytical results, we also focus on the unconstrained interpolation procedure of section 1.2.1, and thus leave aside the monotonicity constraint of the quantile. That has very little impact on the results in practice since the constraint is rarely binding, and when it is the adjustments are small.[19]

1.4.1 Theoretical results

Let n be the size of the population (from which the tabulated data come). Recall that $x = -\log(1-p)$. Let $e_n(x)$ be the estimation error on $\varphi_n(x)$, and similarly $e'_n(x)$ the estimation error on $\varphi'_n(x)$. If we know both those errors then we can retrieve the error on any quantity of interest (quantiles, top shares, Pareto coefficients, etc.) by applying the appropriate transforms. Our first result decompose the error between two components. Like all the theorems of this section, we give only the main results. Details and proofs are in appendix.

Theorem 4. *We can write $e_n(x) = u(x) + v_n(x)$ and $e'_n(x) = u'(x) + v'_n(x)$ where $u(x), u'(x)$ are deterministic, and $v_n(x), v'_n(x)$ are random variables that converge almost surely to zero when $n \to +\infty$.*

We call the first terms $u(x)$ and $u'(x)$ the "misspecification" error. They correspond to the difference between the functional forms that we use in the interpolation, and the true functional forms of the underlying distribution. Even if the population size was infinite, so that sampling variability was absent, they would still remain nonzero. We can give the following representation for that error.

Theorem 5. *$u(x)$ and $u'(x)$ can be written as a scalar product between two functions ε and φ''':*

$$u(x) = \int_{x_1}^{x_K} \varepsilon(x,t)\varphi'''(t)\,\mathrm{d}t \qquad \text{and} \qquad u'(x) = \int_{x_1}^{x_K} \frac{\partial\varepsilon}{\partial x}(x,t)\varphi'''(t)\,\mathrm{d}t$$

where $\varepsilon(x,t)$ is entirely determined by $x_1, \dots, x_K$.

The function $\varepsilon(x,t)$ is entirely determined by the known values $x_1, \dots, x_K$, so we can calculate it directly. Its precise definition is given in appendix. The other function, φ''', depends on the quantity we are trying to estimate, so we do not know it exactly. The issue is common in nonparametric statistics, and complicates the application of the formula.[20] But if we look at the value of φ''' in situations where we have enough

[19] For example, the monotonicity constraint is not binding in any of the tabulations interpolated in the previous section.

[20] For example, the asymptotic mean integrated squared error of a kernel estimator depends on the second derivative of the density (Scott, 1992, p. 131).

data to estimate it directly, we can still derive good approximations and rules of thumb that apply more generally.

We call $v_n(x)$ and $v_n'(x)$ the "sampling error". Even if the true underlying distribution matched the functional used for the interpolation, so that there would be no misspecification error, they would remain nonzero. We can give asymptotic approximation of their distribution for large n. We do not only cover the finite variance case ($\mathbb{E}[X^2] < +\infty$), but also the infinite variance case ($\mathbb{E}[X^2] = +\infty$), which leads to results that are less standard. Infinite variance is very common when dealing with distributions of income and wealth.

Theorem 6. *$v_n(x)$ and $v_n'(x)$ converge jointly in distribution at speed $1/r_n$:*

$$r_n \begin{bmatrix} v_n(x) \\ v_n'(x) \end{bmatrix} \xrightarrow{\mathcal{D}} \mathcal{J}$$

If $\mathbb{E}[X^2] < +\infty$, then $r_n = \sqrt{n}$ and $\mathcal{J}$ is a bivariate normal distribution. If $\mathbb{E}[X^2] = +\infty$ and $1 - F(x) \sim Cx^{-2}$, then $r_n = (n/\log n)^{1/2}$ and $\mathcal{J}$ is a bivariate normal distribution. If $\mathbb{E}[X^2] = +\infty$ and $1 - F(x) \sim Cx^{-\alpha}$ $(1 < \alpha < 2)$, then $r_n = n^{1-1/\alpha}$ and $\mathcal{J} \overset{\mathcal{D}}{=} (\gamma_1 Y, \gamma_2 Y)$ where Y follows a maximally skewed stable distribution with stability parameter α.

Again, we provide more detailed expressions of the asymptotic distributions in appendix alongside the proof of the result. More importantly, we also show that in practice, we always have $v_n(x) \ll u(x)$ and $v_n'(x) \ll u'(x)$, regardless of the precise characteristics of the underlying distribution. This means that sampling variability is negligible compared to the misspecification error. Therefore, we will apply the result of this section assuming $e_n(x) \approx u(x)$ and $e_n'(x) \approx u'(x)$.

1.4.2 Applications

1.4.2.1 Estimation of Error Bounds

Given that thee sampling error is negligible, theorem 5 may be used to get bounds on the error in the general case. As an example, imagine that in both France and the United States, we have access to individual data for the most recent ten years, but that we only have access to tabulated data for the years before that. This, in fact, is what happens in the United States (before 1962) and France (before 1970). We can use the envelope of $|\varphi'''|$ over the ten years with individual data as a reasonable upper bound of it for the rest of the period. We write $|\varphi'''(x)| \leq M(x)$ for all x. Using the triangular inequality, we get $e_n(x) \leq \int_{x_1}^{x_K} |\varepsilon(x,t)| M(t) \, \mathrm{d}t$.

Table 1.3: Observed maximum error and theoretical upper bound

		maximum absolute error on φ		maximum absolute error on φ'		maximum relative error on top shares	
		actual	bound	actual	bound	actual	bound
United States (1962–2004)	$p = 30\%$	0.0014	0.0030	0.0074	0.0100	0.14%	0.30%
	$p = 75\%$	0.0023	0.0137	0.0048	0.0088	0.23%	1.37%
	$p = 95\%$	0.0020	0.0059	0.0044	0.0077	0.20%	0.59%
France (1994–2002)	$p = 30\%$	0.0054	0.0097	0.0038	0.0231	0.54%	0.97%
	$p = 75\%$	0.0080	0.0208	0.0033	0.0076	0.80%	2.08%
	$p = 95\%$	0.0040	0.0088	0.0060	0.0109	0.40%	0.88%

Table 1.3 compares the bound on the error calculated as such with reality. The estimates are conservative by construction due to the use of an upper bound for φ''' and the triangular inequality in the integral. We indeed observe that the theoretical bound is always higher than the actual maximum observed error. Yet in general, the bound that we calculate gives an approximate idea of the error we may expect in each case.

1.4.2.2 Optimal Choice of Brackets

We now consider the inverse problem: namely, how many brackets do we need to achieve a given precision level, and how should they be placed? Based on theorem 5, we can answer that question for any given φ''' by solving an optimization program. Hence, if we pick a functional form for φ''' which is typical of what we observe, we get the solution of the problem for the typical income distribution.

Table 1.4: Optimal bracket choice for a typical distribution of income

	3 brackets	4 brackets	5 brackets	6 brackets	7 brackets
optimal placement of thresholds	10.0% 68.7% 95.2% 99.9%	10.0% 53.4% 83.4% 97.1% 99.9%	10.0% 43.0% 70.4% 89.3% 98.0% 99.9%	10.0% 36.8% 60.7% 80.2% 93.1% 98.6% 99.9%	10.0% 32.6% 53.3% 71.8% 86.2% 95.4% 98.9% 99.9%
maximum relative error on top shares	0.91%	0.32%	0.14%	0.08%	0.05%

We assume that we want our tabulation to span from the 10% to the 99.9% percentiles, so we set $p_1 = 0.1$ and $p_K = 0.999$. We pick the median profile of φ''' estimated over all available years for France and the United States. For a given number K of thresholds, we solve the optimization problem:[21]

$$\min_{p_2,\dots,p_{K-1}} \left\{ \max_{t\in[x_1,x_K]} \int_{x_1}^{x_K} \varepsilon(x,t)\varphi'''(t)\,\mathrm{d}t \right\} \quad \text{st.} \quad p_1 < p_2 < \cdots < p_{K-1} < p_K$$

where as usual $x_k = -\log(1-p_k)$ for $1 \leq k \leq K$.

Table 1.4 shows that a important concentration of brackets near the top is desirable, but that we also need quite a few to cover the bottom. Half of the brackets should cover the top 20%, most of which should be within just the top 10%. The rest should be used to cover the bottom 80% of the distribution. We can also see that a relatively small number of well-placed brackets can achieve remarkable precision: only six are necessary to achieve a maximal relative error of less than 0.1%.

1.4.2.3 Comparison with Partial Samples

We have seen that generalized Pareto interpolation can be quite precise, but how does it compare to the use of a subsample of individual data? The question may be of practical interest when researchers have access to both exhaustive data in tabulated form, or a partial sample of individual data. Such a sample could either be a survey, or a subsample of administrative data.

We may address that question using an example and Monte-Carlo simulations. Take the 2010 distribution of pre-tax national income in the United States. We can estimate that distribution and use it to simulate a sample of size $N = 10^8$ (the same order of magnitude as the population of the United States).

Then, we create subsamples of size $n \leq N$ by drawing without replacement from the large population previously generated.[22] In the case of surveys, we ignore nonresponse and no misreporting, a simplification which favors the survey in the comparison. For each of those subsamples, we estimate the quantiles and top shares at different points of the distribution, and compare it to the same values in the original sample of size N. Table 1.5 shows the results for different values of n. We see that even for large samples ($n = 10^5$, $n = 10^6$, $n = 10^7$), the case for using tabulations of exhaustive data rather than subsamples to estimates quantities such as the top 1% or 0.1% share remains strong. Indeed, even with $n = 10^6$ observations, the typical error on

[21] We solve the problem using the derivative-free Nelder-Mead algorithm.

[22] This survey design is called simple random sampling.

Table 1.5: Mean relative error using subsamples of the full population

	mean percentage gap between estimated and observed values for a survey with simple random sampling and sample size n					
	$n = 10^3$	$n = 10^4$	$n = 10^5$	$n = 10^6$	$n = 10^7$	$n = 10^8$
Top 70% share	0.42%	0.20%	0.10%	0.04%	0.01%	0.00%
Top 50% share	1.26%	0.63%	0.32%	0.13%	0.04%	0.00%
Top 25% share	4.00%	2.04%	1.05%	0.44%	0.15%	0.00%
Top 10% share	9.29%	4.80%	2.50%	1.05%	0.35%	0.00%
Top 5% share	14.32%	7.48%	3.94%	1.65%	0.55%	0.00%
Top 1% share	29.13%	16.01%	8.57%	3.61%	1.21%	0.00%
Top 0.1% share	52.94%	35.23%	19.91%	8.57%	2.89%	0.00%
P30 threshold	4.67%	1.44%	0.45%	0.15%	0.04%	0.00%
P50 threshold	3.29%	1.03%	0.33%	0.10%	0.03%	0.00%
P75 threshold	2.92%	0.91%	0.31%	0.10%	0.03%	0.00%
P90 threshold	3.91%	1.21%	0.39%	0.12%	0.04%	0.00%
P95 threshold	5.86%	1.76%	0.59%	0.18%	0.06%	0.00%
P99 threshold	14.39%	4.79%	1.42%	0.46%	0.14%	0.00%
P99.9 threshold	44.31%	16.29%	5.47%	1.70%	0.49%	0.00%

Original sample of size $N = 10^8$ simulated using the distribution of 2010 pre-tax national income in the United States. Source: author's computations from Piketty, Saez, and Zucman (2016).

the top 1% share is larger than what we get in table 1.4, even with few thresholds. In practice, the thresholds may not be positioned in an optimal way as in table 1.4, so may also want to compare the results with table 1.1. The differences in the orders of magnitude are large enough so that the implications of that comparison hold.

Concluding comments

In this paper, we introduce the concept of generalized Pareto curve to characterize, visualize and estimate distributions of income or wealth. We show strong connections between those curves and the theory of asymptotic power laws, which makes them a natural tool for analyzing them.

Based on quasi-exhaustive individual tax data, we reveal some stylized facts about the distribution of income that lets us move beyond the standard Pareto assumption. We find that although generalized Pareto curves can vary a lot over time and between countries, they tend to stay U-shaped.

Then we develop a method to interpolate tabulated data on income or wealth — as is typically available from tax authorities and statistical institutes — that can correctly reproduce the subtleties of generalized Pareto curves. In particular, the method guarantees the smoothness of the estimated distribution, and work well over most of the distribution, not just the very top. We show that method to be several times more precise than the alternatives most commonly used in the literature. In fact, it can often be more precise than using non-exhaustive individual data. Moreover, we can derive formulas for the error term that let us approximately bound the error of our estimates, and determine the number of optimally placed brackets that is necessary to achieve a given precision.

Finally, we show how our finding can be connected to the existing literature on the income and wealth distribution that emphasizes the role of random growth in explaining power law behavior. The typical shape of Pareto curves that we observe may be explained by a simple and natural deviation from standard random growth models, which is also backed by theoretical models and empirical studies on panel data. Namely, the very top experience higher growth and/or more risk, meaning the processes that generate income and wealth are not fully scale invariant.

We believe that more empirical work — especially a careful use of administrative data sources — is necessary to study those dynamics in a fully satisfying way. We hope that the interpolation method presented in this paper will allow future researchers make progress in that direction. To that end, we made the methods presented in this paper available as a R package named `gpinter`, and also in the form of an online interface that can be used without any installation or knowledge of any programming language. Both are available at `http://wid.world/gpinter`.

Bibliography

Alvaredo, Facundo, Lydia Assouad, and Thomas Piketty (2017). “Measuring Inequality in the Middle East 1990-2016: The World’s Most Unequal Region?”

Alvaredo, Facundo, A. B. Atkinson, et al. (Dec. 2016). “Distributional National Accounts (DINA) Guidelines: Concepts and Methods used in WID.world”. In: *WID.world Working Paper Series*.

Atkinson, A. B. (2017). “Pareto and the Upper Tail of the Income Distribution in the UK: 1799 to the Present”. In: *Economica* 84.334, pp. 129–156. URL: `http://dx.doi.org/10.1111/ecca.12214`.

Atkinson, A. B. and A. J. Harrison (1978). *Distribution of Personal Wealth in Britan*. Cambridge University Press.

Balkema, A. A. and L. de Haan (Oct. 1974). “Residual Life Time at Great Age”. In: *The Annals of Probability* 2.5, pp. 792–804. URL: `http://dx.doi.org/10.1214/aop/1176996548`.

Bania, Neil and Laura Leete (2009). “Monthly household income volatility in the U.S., 1991/92 vs. 2002/03”. In: *Economics Bulletin* 29.3, pp. 2100–2112.

Benhabib, Jess and Alberto Bisin (2016). “Skewed Wealth Distributions: Theory and Empirics”. In: *NBER Working Paper Series* 21924, p. 37. URL: `http://www.nber.org/papers/w21924`.

Benhabib, Jess, Alberto Bisin, and Shenghao Zhu (2011). “The Distribution of Wealth and Fiscal Policy in Economies With Finitely Lived Agents”. In: *Econometrica* 79.1, pp. 123–157. URL: `http://dx.doi.org/10.3982/ECTA8416`.

Bierbrauer, Felix J. and Pierre C. Boyer (2017). “Politically Feasible Reforms of Non-Linear Tax Systems”.

Bingham, N. H., C. M. Goldie, and J. L. Teugels (1989). *Regular Variation*. Encyclopedia of Mathematics and its Applications. Cambridge University Press.

Birgin, E. G. and J. M. Martìnez (Apr. 2008). “Improving Ultimate Convergence of an Augmented Lagrangian Method”. In: *Optimization Methods Software* 23.2, pp. 177–195. URL: `http://dx.doi.org/10.1080/10556780701577730`.

Bukowski, Pawel and Filip Novokmet (2017). “Inequality in Poland: Estimating the whole distribution by g-percentile, 1983-2015”.

Cargo, G. T. and O. Shisha (Jan. 1966). “The Bernstein Form of a Polynomial”. In: *Journal of Research of the National Bureau of Standards* 60B.1, pp. 79–81.

Champernowne, D. G. (1953). “A Model of Income Distribution”. In: *The Economic Journal* 63.250, pp. 318–351. URL: `http://www.jstor.org/stable/2227127`.

Chancel, Lucas and Thomas Piketty (2017). "Indian income inequality, 1922-2014: From British Raj to Billionaire Raj?" URL: `http://wid.world/document/chancelpiketty2017widworld/`.

Chang, Winston et al. (2017). *shiny: Web Application Framework for R.* R package version 1.0.3. URL: `https://CRAN.R-project.org/package=shiny`.

Chauvel, Louis and Anne Hartung (2014). "Dynamics of Income Volatility in the US and in Europe, 1971-2007: The Increasing Lower Middle Class Instability". In: *Paper Prepared for the IARIW 33rd General Conference.*

Conn, Andrew R., Nicholas I. M. Gould, and Philippe L. Toint (1991). "A Globally Convergent Augmented Langrangian Algorithm for Optimization with General Constraints and Simple Bounds". In: *SIAM Journal on Numerical Analysis* 28.2, pp. 545–572. URL: `http://www.jstor.org/stable/2157828`.

Czajka, Léo (2017). "Income Inequality in Côte d'Ivoire: 1985-2014". In: *WID.world Working Paper* July.

Feenberg, Daniel and James Poterba (Dec. 1992). "Income Inequality and the Incomes of Very High Income Taxpayers: Evidence from Tax Returns". In: Working Paper Series 4229. URL: `http://www.nber.org/papers/w4229`.

Fournier, Juliette (2015). "Generalized Pareto curves: Theory and application using income and inheritance tabulations for France 1901-2012". MA thesis. Paris School of Economics.

Gabaix, Xavier (1999). "Zipf's Law for Cities: An Explanation". In: *The Quarterly Journal of Economics* 114.3, p. 739.

– (2009). "Power Laws in Economics and Finance". In: *Annual Review of Economics* 1.1, pp. 255–294.

Gabaix, Xavier et al. (2016). "The Dynamics of Inequality". In: *Econometrica* 84.6, pp. 2071–2111. URL: `http://dx.doi.org/10.3982/ECTA13569`.

Garbinti, Bertrand, Jonathan Goupille-Lebret, and Thomas Piketty (2016). "Income Inequality in France, 1900-2014: Evidence from Distributional National Accounts (DINA)". In: WID.world Working Paper. URL: `http://wid.world/document/b-garbinti-j-goupille-and-t-piketty-inequality-dynamics-in-france-1900-2014-evidence-from-distributional-national-accounts-2016/`.

Guvenen, Fatih et al. (Jan. 2015). "What Do Data on Millions of U.S. Workers Reveal about Life-Cycle Earnings Risk?" In: *NBER Working Papers.* URL: `https://ideas.repec.org/p/nbr/nberwo/20913.html`.

Hardy, Bradley and James P Ziliak (Jan. 2014). "Decomposing Trends in Income Volatility: the "Wild Ride" at the Top and Bottom". In: *Economic Inquiry* 52.1, pp. 459–476.

Jenkins, Stephen P (2016). “Pareto distributions, top incomes, and recent trends in UK income inequality”. URL: http://sticerd.lse.ac.uk/dps/pep/pep30.pdf.

Jones, Charles I. (2015). “Pareto and Piketty: The Macroeconomics of Top Income and Wealth Inequality”. In: *Journal of Economic Perspectives* 29.1, pp. 29–46. URL: http://dx.doi.org/10.1257/jep.29.L29.

Jones, Charles I. and Jihee Kim (2017). “A Schumpeterian Model of Top Income Inequality”. URL: http://search.proquest.com/docview/1629323393?accountid=17248.

Karamata, J. (1930). “Sur un mode de croissance regulière des fonctions”. In: *Mathematica* 4.

Kraft, Dieter (Sept. 1994). “Algorithm 733: TOMP — Fortran Modules for Optimal Control Calculations”. In: *ACM Transactions on Mathematical Software* 20.3, pp. 262–281. URL: http://doi.acm.org/10.1145/192115.192124.

Kuznets, Simon (1953). *Shares of Upper Income Groups in Income and Savings.* NBER. URL: http://www.nber.org/books/kuzn53-1.

Lakner, Christoph and Branko Milanovic (2016). “Global Income Distribution: From the Fall of the Berlin Wall to the Great Recession”. In: *The World Bank Economic Review* 30.2, pp. 203–232. URL: https://academic.oup.com/wber/article-lookup/doi/10.1093/wber/lhv039.

Lyche, Tom and Knut Mørken (2002). “Spline Methods”. URL: http://folk.uio.no/in329/komp.html.

McLeod, R.J.Y. and M.L. Baart (1998). *Geometry and Interpolation of Curves and Surfaces.* Cambridge University Press.

Morgan, Marc (2017). “Extreme and Persistent Inequality: New Evidence for Brazil Combining National Accounts , Surveys and Fiscal Data, 2001-2015”. URL: http://wid.world/wp-content/uploads/2017/09/Morgan2017BrazilDINA.pdf.

Nirei, Makoto (2009). “Pareto Distributions in Economic Growth Models”.

Novokmet, Filip, Thomas Piketty, and Gabriel Zucman (2017). “From Soviets to Oligarchs: Inequality and Property in Russia 1905-2016”. URL: http://gabriel-zucman.eu/files/NPZ2017.pdf.

Pareto, Vilfredo (1896). *Cours d’économie politique.*

Pickands, James (Jan. 1975). “Statistical Inference Using Extreme Order Statistics”. In: *The Annals of Statistics* 3.1, pp. 119–131. URL: http://dx.doi.org/10.1214/aos/1176343003.

Piketty, Thomas (Sept. 2001). *Les hauts revenus en France au XXème siècle.* Grasset.

– (2003). “Income Inequality in France, 1901–1998”. In: *Journal of Political Economy* 111.5, pp. 1004–1042.

Piketty, Thomas and Emmanuel Saez (2003). “Income Inequality in the United States, 1913-1998”. In: *The Quarterly Journal of Economics* 118.1, pp. 1–39.

Piketty, Thomas, Emmanuel Saez, and Gabriel Zucman (Dec. 2016). “Distributional National Accounts: Methods and Estimates for the United States”. In: Working Paper Series 22945. URL: `http://www.nber.org/papers/w22945`.

Piketty, Thomas, Li Yang, and Gabriel Zucman (2017). “Capital Accumulation, Private Property and Rising Inequality in China, 1978-2015”. URL: `http://www.nber.org/papers/w23368.pdf`.

Piketty, Thomas and Gabriel Zucman (2015). “Wealth and Inheritance in the Long Run”. In: *Handbook of Income Distribution.* Vol. 2. Handbook of Income Distribution. Elsevier, pp. 1303–1368.

R Core Team (2016). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing. Vienna, Austria. URL: `https://www.R-project.org/`.

Saez, Emmanuel (2001). “Using Elasticities to Derive Optimal Income Tax Rates”. In: *The Review of Economic Studies* 68.1, pp. 205–229. URL: `https://academic.oup.com/restud/article-lookup/doi/10.1111/1467-937X.00166`.

Saichev, Alexander, Yannick Malevergne, and Didier Sornette (2010). *Theory of Zipf’s Law and Beyond.* Berlin, Heidelberg: Springer Berlin Heidelberg. URL: `http://dx.doi.org/10.1007/978-3-642-02946-2`.

Scott, David W. (1992). *Multivariate Density Estimation.* John Wiley & Sons, Inc. URL: `http://dx.doi.org/10.1002/9780470316849`.

Simon, Herbert (1955). “On a Class of Skew Distribution Functions”. In: *Biometrika* 42.3-4, pp. 425–440. URL: `http://biomet.oxfordjournals.org.libproxy.lib.unc.edu/content/42/3-4/425.full.pdf+html?sid=3d2e52aa-23bb-4338-83be-d71ec5197171`.

Taleb, N. N. and R. Douady (July 2015). “On the super-additivity and estimation biases of quantile contributions”. In: *Physica A Statistical Mechanics and its Applications* 429, pp. 252–260. eprint: `1405.1791`.

van der Wijk, J. (1939). *Inkomens- En Vermogensverdeling.*

Wold, H. O. A. and P. Whittle (1957). “A Model Explaining the Pareto Distribution of Wealth”. In: *Econometrica* 25.4, pp. 591–595.

Chapter 2

The Weight of the Rich: Improving Surveys with Tax Data

For a long time, most of what we knew about the distribution of income, wealth and their covariates came from surveys, in which randomly chosen households are asked to fill a questionnaire. Household surveys have been an invaluable tool for tracking the evolution of society. But in recent years, the research community has grown increasingly concerned with their limitations. In particular, surveys have struggled to keep track of the evolution of the top tail of the distribution, due mainly to heterogeneous response rates, misreporting and small sample bias, which distort all sorts of distributional estimates. These biases end up affecting the way public policy is designed and evaluated.

For this reason, researchers have increasingly been turning to a different source to study inequality: tax data. The idea is not new; we can trace it back to the seminal work of Kuznets (1953), or even Pareto (1896). More recently, Piketty and Saez (2003) and Piketty (2003) applied their method to more recent data for France and the United States. This work was extended to more countries by many researchers whose contributions were collected in two volumes by Atkinson and Piketty (2007, 2010) and served as the basis for the World Inequality Database (`http://wid.world`).

But tax data have their own limitations. They usually only cover the top of the distribution and include at best a limited set of covariates. They do not capture well informal and tax-exempt income. They are often not available as microdata but rather as tabulations summarizing the distribution, which limits their use. The statistical unit that they use (individuals or households) depends on the local legislation and may not be comparable from one country to the next. This is why many indicators, such as poverty rates or gender gaps, have to be calculated from surveys. The use of different — and sometimes contradictory — sources to compute statistics can make it hard to build consistent and accurate narratives on distributional matters. This explains the ongoing effort to combine the different data sources at our disposal in a way that exploits their strengths and corrects their weaknesses.

The Distributional National Accounts (DINA) project is a prominent example of this effort. Its guidelines (Alvaredo et al., 2017) emphasize the need to look at the entire distribution, harmonize concepts, and where possible decompose the distribution according to socio-demographic characteristics. Piketty, Saez, and Zucman (2018) for the United States, and Garbinti, Goupille-Lebret, and Piketty (2018) for France have used both survey and tax data to construct distributional statistics that account for all of the income recorded in national accounts. The resulting dataset not only allowed them to reassess the evolution of income concentration statistics, but also to study subjects such as: gender gaps, growth incidence curves or the distributive impact of fiscal policy. But these examples depend in large part on the existence of reliable administrative microdata accessible to researchers, to which information from surveys can be added to account for the limited sources of income not covered in the tax data.

In many countries, both developed and less developed, such direct access is quite rare. Instead, tabulations of fiscal income, containing information on the number and declared income of individuals by income bracket, are more commonly available. The population coverage in the tabulations is often substantially less than the total adult population, and the difference varies with the country studied. Furthermore, in contexts of high informality, which is the case for many developing countries, even if tax declarations had full population coverage, they could not be assumed to be reliable across the whole distribution. In such cases it is better to proceed the other way round: rather than incorporating survey information into the tax data, we need to incorporate tax information into the survey data.

There has been a number of suggested approaches to deal with the problem of merging tax and survey data, yet the literature has largely failed to converge towards a standard. Crucially, most of the existing approaches directly adjust the income or wealth distributions, overlooking the goal of preserving the survey's representativeness in terms of covariates, while relying on arbitrary assumptions in the process. In this paper, we develop a methodology that has significant advantages over previous ones, and which should cover most practical cases within a single, united framework. Our method avoids relying, to the extent possible, on *ad hoc* assumptions and parameters.

We present a data-driven way to determine the point in the survey data where the under-coverage of income starts. This is our “merging point” — the point in the distribution were survey data and tax data are merged. We perform necessary adjustments in a way that minimize distortions from the original survey, and preserve desirable properties, such as the continuity of the density function. Rather than directly making assumptions on complex summary statistics such as quantiles or bracket averages, our method makes assumptions that are easily interpretable at the level of observations. The algorithm acknowledges the presence of covariates, so that we ensure the representativeness of the survey in terms of income while maintaining — and possibly improving — its representativeness in terms of age, gender, or any other dimension along the distribution. As a result, we can preserve the richness of information in surveys, both in terms of covariates and household structure.

Our method proceeds in three steps, the first aimed at selecting the merging point between the datasets, and the other two aimed at correcting for the two main types of error in surveys: non-sampling error and sampling error. Non-sampling error refers to issues that cannot easily be solved with a larger sample size, and typically arise from unobserved heterogeneous response rates. In the second step, we correct for these issues using a reweighting procedure rooted in survey calibration theory (Deville and Särndal, 1992). In doing so, we address a longstanding inconsistency between the empirical literature on top incomes in surveys, and the established practice of most survey producers. Indeed, since Deming and Stephan (1940) introduced their raking algorithm, statistical institutes have regularly reweighted their surveys to match known demographic totals from census data. Yet the literature on income has mostly relied on adjusting the value of observations, rather than their weight, to enforce consistency between tax and survey data. We argue that the theoretical foundations of such approaches are less explicit and harder to justify.

This initial correction step addresses non-sampling error, but it is limited in its ability to correct for sampling error, meaning a lack of precision due to limited sample size.[1] A clear example is the maximum income, which is almost always lower in surveys than in tax data, something no amount of reweighting can do anything about. Top income shares of small income groups are also strongly downward biased in small samples (Taleb and Douady, 2015), so inequality will be underestimated even if all the non-sampling error has been corrected. To overcome this problem, we supplement the survey calibration with a further step, in which we replace observations at the top

[1] Calibration methods can, to some extent, correct for sampling error. But their ability to do so only holds asymptotically (Deville and Särndal, 1992), so it does not apply to narrow income groups at the top of the distribution.

by a distribution generated from the tax data, and match the survey covariates to it. The algorithm for doing so preserves the distribution of covariates in the original survey, their correlation with income, and the household structure, regardless of the statistical unit in the tax data. The result is a dataset where sampling variability of income at the top has been mostly eliminated, and whose covariates have the same statistical properties as the reweighted survey. Because we preserve the nature of the original microdata, we can use the output to experiment with different statistical units, equivalence scales, calculate complex indicators, and perform decompositions along any dimension included in the survey.

In order to illustrate how the method operates in practice, we run two different types of applications. First, since the true distribution of income and wealth is always unknown, we simulate artificial populations that are drawn from parametric distributions. These include behavioral assumptions that define two main sources of bias, namely heterogeneous response rates and misreporting. Using these biases, we simulate a large number of consecutive surveys and then apply our correction method using synthetic tax tabulations. We use these experiments to assess the accuracy and precision of the resulting estimates and to compare them to those derived from both the raw sample and the most common alternative methods using external data — namely methods that directly replace survey incomes with tax incomes for the same quantiles in the distribution. We demonstrate that our method is superior to available options, not only because it relies on reasonable assumptions that enable the use of resulting micro-datasets — unlike the "replacing" alternative — but also because it produces estimates that are consistently closer to true values with lower variance.

In our second application, we apply our method to real data from five countries: France, U.K., Norway, Brazil and Chile. Our case studies are chosen to showcase the wide applicability of the method to both developed countries and less-developed countries. The method makes upward revisions to inequality estimates in all cases, with varying degrees of magnitude, depending on the quality of the underlying data and the level of inequality in each country. It can also produce differing inequality trends. Moreover, our empirical results support the findings of our simulations concerning the difference between our method and the replacing alternative.

For practical use, we have developed a Stata command that applies our method. The program works with several input types, income concepts and statistical units, ensuring flexibility for users. Our method may therefore easily be used by researchers interested in analyzing the different dimensions of inequality.[2] The main goal of this paper is to describe the theoretical and practical details behind this readily usable method, as well as its advantages with respect to existing approaches.

The remainder of the paper is structured as follows. In section 2.1 we relate our paper to the existing literature. In section 2.2 we lay out the theoretical framework of our method. This is followed by applications to simulated distributions and practical applications to specific countries in section 3, before concluding.

2.1 Literature on Survey Correction Methods

Numerous studies have sought to adjust survey data primarily to improve the latter's representativeness and/or produce a more accurate distribution of income. In some instances this has been achieved with the aid of external administrative data. We identify three distinguishable methodological strands present in this literature. The first strand opts to reweight survey observations. The second strand replaces the income value of observations with a value typically drawn from a parametric distribution or an external data source. Finally, a third strand identifies the need to employ a hybrid procedure by combining reweighting and replacing.

2.1.1 Reweighting Observations

The studies that focus on reweighting usually formalize the bias as nonresponse. Many papers in this literature estimate a parametric model of nonresponse to adjust survey weights, but do not use direct data on the distribution of income. Korinek, Mistiaen, and Ravallion (2006) make this type of adjustment using nonresponse rates across geographic areas and the characteristics of respondents within regions. This type of approach can be sensitive to the degree of geographic aggregation used calculating response rates. This is an issue explored in more detail by Hlasny and Verme (2017; 2018) for the US and European case respectively, using similar probabilistic models. Depending on the nature of the survey data, greater or less geographic dis-aggregation on nonresponse rates can be more appropriate to the adjustment at hand.

Crucially, these models do not use direct information on the income distribution — often due to lack of availability. Instead, they have to infer relationships between

[2] The package to download is `bfmcorr` for the correction method, which includes two sub-commands: `postbfm` for the post-estimation output and `bfmtoy` for parametric simulations. The command and its sub-commands come with a full set of user instructions.

individual nonresponse and individual characteristics based on aggregate relationships between average nonresponse and an average of certain characteristics. This make these methods susceptible to the pitfalls of ecological inference. In particular, to the extent that nonresponse bias is a strongly nonlinear function of income (which we observe in practice), the relationship between income and nonresponse will look very different at the aggregate and the individual level. Our proposal instead makes use of direct administrative data to determine the relationship between income and nonresponse.

There are a few studies in this literature that combine surveys with external sources to measure inequality. An example of this is the case study of Argentina in Alvaredo (2011), in which the corrected Gini coefficient is estimated by assuming that the top of the survey distribution (top 1% or top 0.1%) completely misses the richest individuals that are represented in tax data. This accounts for the bias of nonresponse and corrects the distribution via an implicit reweighting procedure. The specific form of the nonresponse bias that is assumed tacitly is, nonetheless, a rather restrictive one. Indeed, the correction implies a deterministic nonresponse rate equal to 1 above a previously selected quantile and 0 under it. Furthermore, in both of the empirical applications (the US and Argentina) the threshold beyond which the tax data is used is chosen arbitrarily.[3] Our method on the other hand tries at best to avoid arbitrary choices on the portion of the survey distribution to be corrected or on the form of the bias implied by the correction.

To our knowledge the paper that comes closest to proposing an approach that resembles the one we propose here, in terms of criteria and methodology, is Medeiros, Castro Galvão, and Azevedo Nazareno (2018) applied to Brazilian data. That is, it is the only study that combines tabulated tax data with survey micro-data by explicitly reweighting survey observations. More specifically, the authors apply a Pareto distribution to incomes from the tax tabulation to correct the top of the income distribution calculated from the census. Their method involves re-weighting the census population by income intervals above a specified merging point, which is determined from the comparison of the median total income reported in each quantile of the tax data and in the Census (0.5% of the adult population sorted by income).

However, important differences remain. Contrary to our method, the choice of the merging point is not endogenous, but chosen by the authors as the most relevant point

[3] In any case, the goal of the paper is not to tackle the nonresponse or misreporting biases directly, but to provide a simple estimation of a corrected Gini coefficient.

beyond which the tax data presents a more concentrated distribution. Thus, multiple points can be used, and indeed the authors test two. Our method endogenously determines a single merging point based on a more comprehensive treatment of the form of the non-response bias. Importantly, our approach preserves the continuity of the density of income — something that only a specific choice of the merging point can ensure. To guide their choice of the merging point, Medeiros, Castro Galvão, and Azevedo Nazareno (2018) look at the rank at which income in the tax data exceed that of the survey. Yet from the perspective of correcting for non-response, such a point does not have any well-defined interpretation.

Moreover, while they increase the weight of observations above the merging point, they do not reduce the weight of individuals below this point, such that the corrected population ends up being larger than the original official population. The authors do not provide a way to ensure the representativeness of characteristics other than income after the adjustment either — their purpose is to remedy the underestimation of top incomes in surveys, without a unified calibration framework. Moreover, their method does not remedy the lack of precision at the top of the distribution arising from sampling limitations, resulting in downward biased income shares of small income groups, especially in small samples. In contrast, our method addresses all of these issues.

2.1.2 Replacing Incomes

The general feature of the "replacing" approach is that it involves the direct replacing of survey incomes with incomes from tax data. Although there is no unified theory or explicit justification behind the applications of this adjustment procedure, most of these methods share some defining characteristics. In practice, they generally adjust distributions by replacing cell-means in the survey distribution of income with those from the tax distribution for the same sized cells (i.e. fractiles) of equivalent rank in the population. The size of the cells varies by study (Burkhauser, Hérault, et al., 2016; Piketty, Yang, and Zucman, 2017; Chancel and Piketty, 2017; Czajka, 2017). Furthermore, the overall size of the population group whose income is to be adjusted is sometimes chosen arbitrarily, such as the top 20% in the distribution (Piketty, Yang, and Zucman, 2017), the top 10% (Burkhauser, Hérault, et al., 2016; Chancel and Piketty, 2017), the top 1% (Burkhauser, Hahn, and Wilkins, 2016; Alvaredo, 2011), or the top 0.5% of survey observations (DWP, 2015).

This decision can be made less arbitrary using the comparison of threshold or average incomes by fractile in the two distributions. The size of the group is then chosen as the point in the distribution where the two quantile functions cross (e.g. Czajka (2017)). As we have noted earlier, this point is not really meaningful from a statistical viewpoint. In fact, under the most natural assumption (increasing nonresponse profile) it should not even exist, because the quantile functions do not intersect.

Other non-arbitrary choices take the minimum income level that requires mandatory tax filing (Diaz-Bazan, 2015). While these try to keep the use of survey data to measure the top of the income distribution to a strict minimum, they assume that the entire tax distribution is reliable. We argue however that not all the income in tax data should be considered reliable given the difference between declarable income thresholds and taxable income thresholds. The quality of tax data generally increases with income in a manner that is often not well defined, and given this uncertainty it makes sense to limit their use to the portion that is absolutely necessary.

In certain cases, the survey distribution stops being reliable before the tax data can be trusted. This happens in particular in countries where only a small part of the population file a tax return. The fact that quantiles from both sources do not cross is often viewed as evidence of this problem. In such cases, from the point at which we stop trusting the survey to the point at which we start trusting the tax data, one option is to rescale upwards the income values from the survey distribution. This can be done using various profiles of rescaling coefficients (usually linear) (Chancel and Piketty, 2017; Piketty, Yang, and Zucman, 2017; Novokmet, Piketty, and Zucman, 2018). This procedure ensures at least that the quantile function is continuous. These rescaling methods can be seen as an extension of the general replacing methods.

Replacing survey-respondents' declared income has been viewed as adjusting for the misreporting bias in surveys (Burkhauser, Hérault, et al., 2018; Jenkins, 2017). In Appendix B.1 we formalise the existence of this bias both when it operates alone and when it operates in the presence of non-response. We compare existing replacing methods to our own method, and we explain why they only correct for misreporting under very strong and unrealistic assumptions — namely that the income rank in the survey distribution and in the benchmark distribution are the same, and that underreporting is a deterministic function of this rank.

2.1.3 Combined Reweighting and Replacing

Some voices stress the need to combine the aforementioned correction approaches. Bourguignon (2018), while reviewing the typical adjustment methods employed, correctly highlights that any method must dwell on three important parameters:

the amount of income to be assigned to the top, the size of this top group, and the share of the population added to the top in the survey. The definition of these three parameters implies a correction procedure combining reweighting and replacing methods. His analysis goes on to study the ways in which these choices impact the adjustments made to the original distribution. However, this analysis does not shed light on *how* to make these choices. Moreover, in reviewing multiple correction methods and applying them to Mexican survey data (including the combined case, where all three parameters mentioned take non-zero values), he only considers the situation "where nothing is known about the distribution of the missing income, unlike when tax records or tabulations are available" (Bourguignon, 2018). This is in contrast to our approach for correcting survey microdata, which combines the two previous methods, but which explicitly merges tax data with surveys to produce more realistic distributions of income.

In summary, contrary to existing methods, our method uses external tax data, endogenously finds a non-arbitrary merging point, and preserves the multivariate distribution of covariates and population totals. Moreover, it is grounded on a more solid theoretical framework, which we now turn to explain in the following section.

2.2 Theory and Methodology

To describe our method and the theory behind it, we part from the simple univariate setting, where we adjust the weight of observations in the survey at different income levels. The second section explains how to use the theory of survey calibration to handle more complex multivariate settings. Finally, the third section explains how we address the problem of sampling error, which reweighting has only a limited ability to address.

2.2.1 Univariate Setting

In this section we first explain the intuition behind the correction before presenting how we choose the merging point between the two distributions.

2.2.1.1 Intuition

Let X and Y be two real random variables. We will use Y to represent the true income distribution, part of which we assume is recorded in the tax data.[4] And

[4] In reality, part of the true income may also be missing from the tax data due to non-taxable income not reported on the declaration and tax evasion. The extent of these omissions vary by

we will use X to represent the income distribution recorded in the survey. Each random variable has a probability density function (PDF) f_Y and f_X, a cumulative probability function (CDF) F_Y and F_X, and a quantile function Q_Y and Q_X.

Let $\theta(y) = f_X(y)/f_Y(y)$ be the ratio of the survey density to the true density at the income level y. This represents the number of people within an infinitesimal bracket $[y, y + \mathrm{d}y]$ according to the the survey, relative to the actual number of people in the bracket. If $\theta(y) < 1$, then people with income y are underrepresented in the survey. Conversely, if $\theta > 1$, then they are overrepresented.

The value of $\theta(y)$ may be interpreted as a relative probability. Indeed, let D be a binary random variable that denotes participation to the survey: if an observation is included in the sample, then $D = 1$, otherwise $D = 0$. Then Bayes' formula implies:

$$\theta(y) = \frac{f_X(y)}{f_Y(y)} = \frac{1}{f_Y(y)} \times f_Y(y) \frac{\mathbb{P}\{D = 1|Y = y\}}{\mathbb{P}\{D = 1\}} = \frac{\mathbb{P}\{D = 1|Y = y\}}{\mathbb{P}\{D = 1\}} \tag{2.1}$$

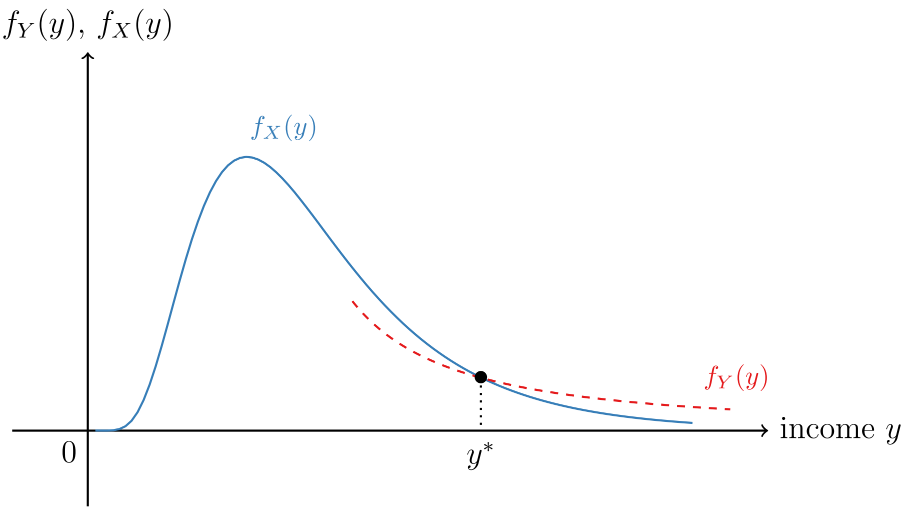
If everyone has the same probability of response, then $\mathbb{P}\{D = 1|Y = y\} = \mathbb{P}\{D = 1\}$, and $\theta(y) = 1$. Hence $f_X(y) = f_Y(y)$ and the survey is unbiased. What matters for the bias is the probability of response at a given income level relative to the average response rate, which is why we have the constraint $\mathbb{E}[\theta(Y)] = 1$. Intuitively, if some people are underrepresented in the survey, then mechanically others have to be overrepresented, since the sum of weights must ultimately sum to the population size.

This basic constraint has important consequences for how we think about the adjustment of distributions. Any modification of one part of the distribution is bound to have repercussions on the rest. In particular, it makes little sense to assume that the survey is not representative of the rich, and at the same time that it is representative of the non-rich.

Figure 2.1 represents the situation graphically, in the more common case where $\theta(y)$ is lower for top incomes. We show a truncated version of f_Y since tax data often only cover a limited part of the whole distribution. The fact that the dashed red line $f_Y(y)$ is above the solid blue line $f_X(y)$ mean that top incomes are underrepresented. Therefore, lower incomes must be overrepresented, which is what we see below the point y^*. This pivotal value is unique assuming that θ is monotone. The appropriate correction procedure here would be to increase the value of the density above it, and decrease its value below it. The intuition behind reweighting is that we have to multiply the survey density f_X by a factor $1/\theta(y)$ to make it equal to the true

country, and their treatment are beyond the scope of this paper.

Figure 2.1: A "True" and Biased Income Distribution



The solid blue line represents the survey density f_X. The dashed red line represents the tax data density f_Y, which is only observed at the top. For high incomes, the survey density is lower than the tax data density, which means that high incomes are underrepresented. If some individuals are underrepresented, then other have to be overrepresented: they correspond to people below the point y^*.

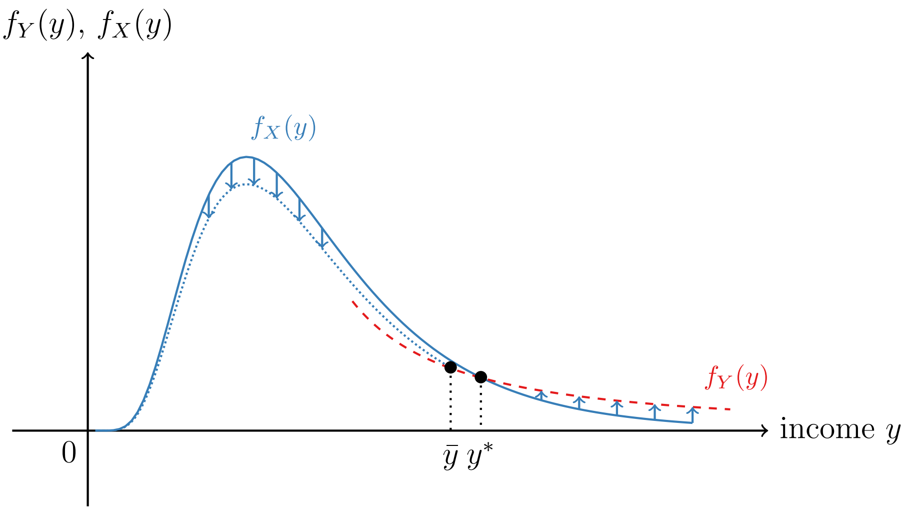
density f_Y. In practice, this means multiplying the weight of any observation Y_i by $1/\theta(Y_i)$.

When we observe both f_Y and f_X, we can directly estimate θ nonparametrically. But because we do not observe the true density over the entire support, we have to make an assumption on the shape of θ for values not covered by the tax data. We will assume a constant value. Behind this assumption, there are both theoretical motivations that we develop in section 2.2.2, and empirical evidence that we present in section 2.3. Intuitively, it means that there is no problem of representativeness within the bottom of the distribution, so that the overrepresentation of the non-rich is only the counterpart of the underrepresentation of the rich. We can therefore write the complete profile of θ as:

$$\theta(y) = \begin{cases} \bar{\theta} & \text{if } y < \bar{y} \\ f_X(y)/f_Y(y) & \text{if } y \geq \bar{y} \end{cases} \tag{2.2}$$

We call $\bar{y}$ the *merging point*. It is the value at which we merge observations from the tax data into the survey. A naive choice would be to use the tax data as soon as they become available, but this will often lead to poor results. This is because the point from which the tax data become reliable is not necessarily sharp and well-defined, so in practice it will be better to start using the tax data only when it becomes clearly

Figure 2.2: The Intuition Behind Reweighting



The solid blue line represents the survey density f_X. The dashed red line represents the tax data density f_Y. Above the merging point $\bar{y}$, the reweighted survey data have the same distribution as the tax data (dashed red line). Below the merging point, the density has been uniformly lowered so that it still integrates to one, creating the dotted blue line.

necessary. The proper choice of that point is an important aspect of the method on which we return to in section 2.2.1.2. For now we will take it as given, and only assume that it is below the pivotal point y^* of figure 2.1. Figure 2.2 shows how the reweighting using (2.2) operates.

Let $\tilde{f}_X$ be the reweighted survey, i.e. $\tilde{f}_X(y) = f_X(y)/\theta(y)$. By construction, we have $\tilde{f}_X(y) = f_Y(y)$ for $y \geq \bar{y}$. As indicated by upward arrows on the right of figure 2.2, the density has been increased for $y > y^*$. Since densities must integrate to one, values for $y < y^*$ have to be lowered. The uniform reweighting below $\bar{y}$ creates the dotted blue line.

2.2.1.2 Choice of the Merging Point

For many countries, tax data only covers the top of the distribution. We use the term *trustable span* to name the interval over which the tax data may be considered reliable. It takes the form $[y_{\text{trust}}, +\infty[$. This interval is determined by country specific tax legislation. It relies on the portion of the distribution covered in the data (declarations) or just on the portion of the tax population that pays income tax (taxpayers).

We do not usually wish to use the tax data over the entire trustable span. First, because the beginning of the trustable span is not always sharp — the reliability of the tax data increases with income in a way that is not well-defined, therefore it is more prudent to restrict their use to the minimum that is necessary. Second, once we are past the point where there is clear evidence of a bias, we prefer to avoid distorting the survey in unnecessary ways.

We suggest a simple, data-driven way for choosing the merging point point with desirable properties. In particular, we seek to approximately preserve the continuity of the underlying density function after reweighting. We start from the typical case where $\bar{y}$ is inside the trustable span $[y_{\text{trust}}, +\infty[$. In Appendix B.2 we consider cases where the trustable span may be too small to observe an overlap between the densities.

Assume that the bias function $\theta(y)$ follows the form in (2.2). We introduce a second function, the cumulative bias, defined as:

$$\Theta(y) = \frac{F_X(y)}{F_Y(y)} \tag{2.3}$$

In figure 2.3, we examine the shape of $\theta(y)$ and $\Theta(y)$ in relation to the density functions presented in figure 2.2. We have the relationship $\Theta(y)F_Y(y) = \int_{-\infty}^{y} \theta(t) f_Y(t)\,\mathrm{d}t$. Given (2.2), for $y < \bar{y}$, $\Theta(y) = \bar{\theta}$. As figure 2.3 shows, we should expect the merging point $\bar{y}$ to be the highest value y such that $\Theta(y) = \theta(y)$.

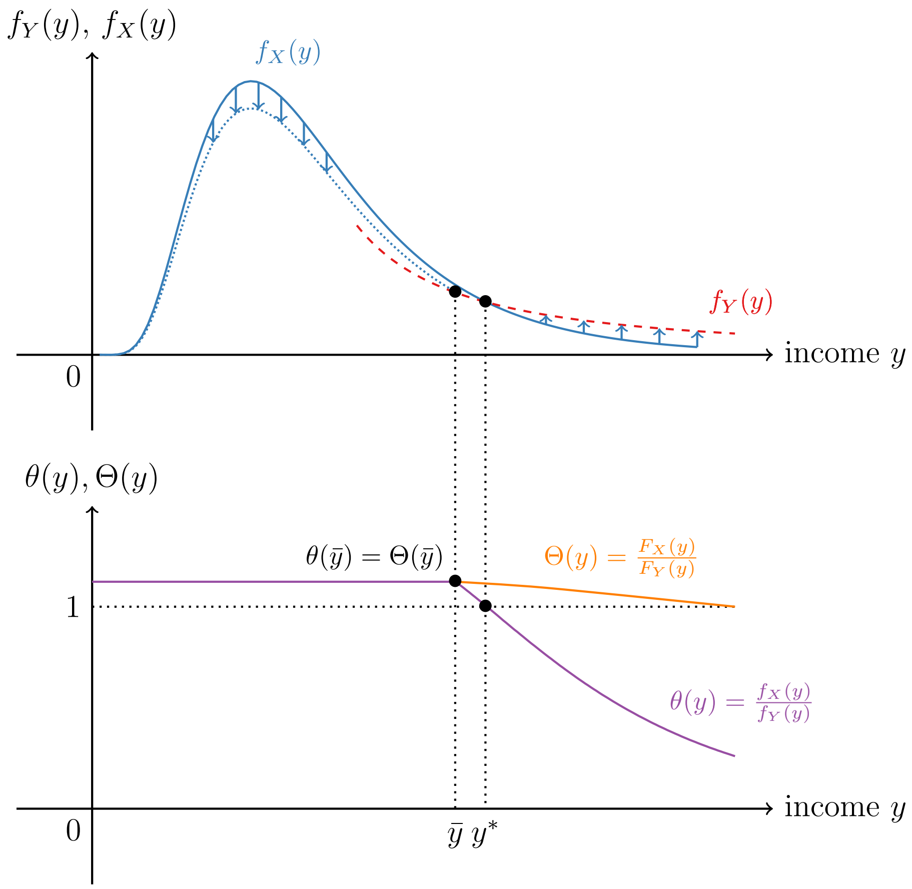
We can contrast this choice of merging point with the one implicitly chosen in at least some replacing approaches: the point at which the quantile functions of the survey and the tax data cross.[5] This is equivalent to setting equal densities (i.e. $\theta(y) = 1$) until this merging point, which will in general be lower than ours. At that point, there is a discontinuity in $\theta(y)$ which jumps above one, and then progressively decreases toward zero. As a result, the people just above the merging point are implicitly assumed to be overrepresented compared to those below, even though they are richer. This discontinuity and lack of monotonicity of θ is hard to justify, and our choice of merging point avoids it.

We can estimate both $\theta(y)$ and $\Theta(y)$ over the trustable span of the tax data. To determine the merging point in practice, we look for the moment when the empirical curves for $\Theta(y)$ and $\theta(y)$ cross, and discard the tax data below this point. This choice is the only one that can ensure that the profile of $\theta(y)$, and by extension the income density function, remains continuous.

The estimation of $\Theta(y)$ poses no difficulty as it suffices to replace the CDFs by their

[5] Appendix B.1.2 presents a theoretical comparison of both procedures.

Figure 2.3: Choice of Merging Point when $\bar{y} \geq y_{\text{trust}}$


$f_Y(y)$, $f_X(y)$
$f_X(y)$
$f_Y(y)$
income y
0
$\theta(y), \Theta(y)$
$\theta(\bar{y}) = \Theta(\bar{y})$
$\Theta(y) = \frac{F_X(y)}{F_Y(y)}$
1
$\theta(y) = \frac{f_X(y)}{f_Y(y)}$
income y
0
$\bar{y}$ y^*

empirical counterpart in (2.3) to get the estimate $\hat{\Theta}_k$. For $\theta(y)$, however, we have to estimate densities. We define m bins using fractiles of the distribution (from 0% to 99%, then 99.1% to 99.9%, then 99.91% to 99.99% and 99.991% to 99.999%). We approximate the densities using histogram functions over these bins. This gives a first estimate for each bin that we call $(\tilde{\theta}_k)_{1\leq k\leq m}$. The resulting estimate is fairly noisy, so we get a second, more stable one named $(\hat{\theta}_k)_{1\leq k\leq m}$ using an antitonic (monotonically decreasing) regression (Brunk, 1955; Ayer et al., 1955; Eeden, 1958). That is, we solve:

$$\min_{\hat{\theta}_1,\dots,\hat{\theta}_m} \sum_{k=1}^{m} (\hat{\theta}_k - \tilde{\theta}_k)^2 \qquad \text{s.t.} \qquad \forall k \in \{2,\dots,m\} \quad \hat{\theta}_{k-1} \geq \hat{\theta}_k$$

We solve the problem above using the Pool Adjacent Violators Algorithm (Ayer et al., 1955). The main feature of this approach is that we force $(\hat{\theta}_k)_{1\leq k\leq m}$ to be decreasing. This turns out to be enough to smooth the estimate so that we can work with it, without the need introduce additional regularity requirements. We use as the merging point bracket the lowest value of k such that $\hat{\theta}_k < \hat{\Theta}_k$.

2.2.2 Multivariate Setting

The previous subsection presented the main idea of the method. But while this intuition works well in the univariate case, the introduction of other dimensions from the survey (gender, age, income composition, etc.) complicates the problem significantly. Indeed, it is not enough for the survey to be solely representative in terms of total income, we also need to preserve (or possibly enforce) representativeness in terms of these other variables. This subsection thus explains how we adapt our method to the survey-calibration framework, mainly to address two types of representational issues.[6] First, if the survey is already assumed to be representative at the aggregate level in terms of age or gender (i.e., because it has already been adjusted to fit census data), then we should aim to preserve such features. Second, when the adjustment is made using income alone (i.e. the univariate case), it corrects weights based on the observed probability of response conditional on income, ignoring interactions between total income and other characteristics, which are sometimes reported in tax data.[7] We start by presenting the theory in its general setting below, before explaining how to apply it to the problems at hand.

2.2.2.1 Calibration

Problem Survey calibration considers the following problem. We have a survey sample of size n. Each observation is a k-dimensional vector $\boldsymbol{x}_i = (x_{1i}, \dots, x_{ki})'$. The sample can be written $(\boldsymbol{x}_1, \dots, \boldsymbol{x}_n)$, and the corresponding survey weights are $(d_1, \dots, d_n)$. We know from a higher-quality external source the true population totals of the variables $x_{1i}, \dots, x_{ki}$ as the vector $\boldsymbol{t}$. We seek a new set of weights, $(w_1, \dots, w_n)$, such that the totals in the survey match their true value, i.e. $\sum_{i=1}^n w_i \boldsymbol{x}_i = \boldsymbol{t}$.

This problem will in general have an infinity of solutions, therefore survey calibration introduces a regularization criterion to select the preferred solution out of all the different possibilities. The idea is to minimize distortions from the original survey data, so we consider:

$$\min_{w_1,\dots,w_n} \sum_{i=1}^n \frac{(w_i - d_i)^2}{d_i} \qquad \text{s.t.} \qquad \sum_{i=1}^n w_i \boldsymbol{x}_i = \boldsymbol{t} \tag{2.4}$$

[6] Survey calibration was introduced with the raking procedure of Deming and Stephan (1940). Deville and Särndal (1992) provided major improvements. While statistical institutes routinely use calibration methods with respect to age and gender variables, they are not yet traditionally used for income variables.

[7] For instance, if rich elderly persons are more likely to respond to surveys (say, because they have more free time) than younger rich people, then the univariate adjustment will produce an accurate income distribution without solving the over-representation of older people. A similar rationale can be applied to the issue of income composition.

That is, we minimize the χ^2 distance between the original and the calibrated weights, under the constraint on population totals: this is called linear calibration. While alternative distances are sometimes used, linear calibration is advantageous in terms of analytical and computational tractability.

Solution Solving the problem (2.4) leads to:

$$\frac{w_i}{d_i} = 1 + \boldsymbol{\beta}\boldsymbol{x}_i \tag{2.5}$$

where $\boldsymbol{\beta}$ is a vector of Lagrange multipliers determined from the constraints as:

$$\boldsymbol{\beta} = \boldsymbol{T}^{-1}\left(\boldsymbol{t} - \sum_{i=1}^{n} d_i \boldsymbol{x}_i\right) \qquad \text{with} \qquad \boldsymbol{T} = \sum_{i=1}^{n} d_i \boldsymbol{x}_i \boldsymbol{x}_i'$$

where the matrix $\boldsymbol{T}$ is invertible as long as there are no collinear variables in the $\boldsymbol{x}_i$ (meaning neither redundancy nor incompatibility of the constraints).[8] One undesirable feature of linear calibration is that it may lead to weights below one or even negative, which prevents their interpretation as an inverse probability and is incompatible with several statistical procedures. Therefore, in practice, we enforce the constraints $w_i \geq 1$ for all i using an standard iterative method described in Singh and Mohl (1996, method 5). This is known as truncated linear calibration.

Interpretation This procedure can be interpreted in terms of a nonresponse model.[9] In this context, the survey weights are the inverse of the probability of inclusion in the survey sample. This probability of inclusion is the product of two components. The first one depends on whether a unit is selected for the survey, regardless of whether that unit accepts to answer or not. We note $D_i = 1$ if unit i is selected, and $D_i = 0$ otherwise. The value $\delta_i = 1/\mathbb{P}\{D_i = 1\}$ is called the design weight. The design weight in constructed by the survey producer and therefore known exactly. The second component depends on whether a unit contacted for the survey accepts to answer or not. We note $R_i = 1$ if unit i accepts to participate in the survey, and $R_i = 0$ otherwise. The value $\rho_i = 1/\mathbb{P}\{R_i = 1\}$ is called the response weight. Since both D_i and R_i must be equal to 1 for a unit to be observed, the final weight is the product of these two components $\delta_i\rho_i$.

Nonresponse is unknown so it has to be estimated using certain assumptions. The simplest one is that ρ_i is the same for all units, therefore all weights are up-scaled by

[8] In practice, we use the Moore–Penrose generalized inverse to circumvent the collinearity problem.

[9] For a geometric interpretation of linear calibration see Appendix B.3.

the same factor so that their sum matches the population of interest. More complex models use information usually available to the survey producer, that is, basic socio-demographic variables which we will write $\boldsymbol{U}_i$. The survey producer models nonresponse as a function of these variables: $\rho_i = \phi(\boldsymbol{U}_i)$. The survey producer provides weights equal to $\delta_i \phi(\boldsymbol{U}_i)$. If nonresponse is also a function of income, which is not observed by the survey producer, then the estimated nonresponse will fail to accurately reflect true nonresponse, leading to biased estimates of the income distribution. Using the tax data $\boldsymbol{Y}_i$, we can estimate a new model that takes income into account: $\psi(\boldsymbol{U}_i, \boldsymbol{Y}_i)$. The final weight becomes:

$$\begin{aligned} w_i &= \frac{1}{\mathbb{P}\{D_i = 1\}} \frac{1}{\mathbb{P}\{R_i = 1\}} \\ &= \frac{1}{\mathbb{P}\{D_i = 1\}} \psi(\boldsymbol{U}_i, \boldsymbol{Y}_i) \\ &= \delta_i \phi(\boldsymbol{U}_i) \times \frac{\psi(\boldsymbol{U}_i, \boldsymbol{Y}_i)}{\phi(\boldsymbol{U}_i)} \\ &= d_i \times \frac{\psi(\boldsymbol{U}_i, \boldsymbol{Y}_i)}{\phi(\boldsymbol{U}_i)} \end{aligned} \tag{2.6}$$

Comparing equation (2.5) with (2.6), we see that the calibration problem suggests both a functional form and an estimation method for $\psi(\boldsymbol{U}_i, \boldsymbol{Y}_i)/\phi(\boldsymbol{U}_i)$. This functional form assumes nonresponse profiles that are as uniform (thus non-distortive) as possible, and only modify the underlying distribution if it is necessary to do so. The preference for non-distortive functional forms can also help justify the use of a constant reweighting profile below the merging point in section 2.2.1.1.

Application to Income Data The calibration problem is presented so as to enforce the aggregate value of variables. In order to use it to enforce the distribution of a variable, we have to discretize this distribution. In the case of income tax data, the income distribution may be presented in various tabulated forms, and we use the generalized Pareto interpolation method of Blanchet, Fournier, and Piketty (2017) to turn it into a continuous distribution.[10] We output the distribution discretized over a narrow grid made up of all percentiles from 0% to 99%, 99.1% to 99.9%, 99.91% to 99.99% and 99.991% to 99.999%. We discard tax brackets below the merging point, whose choice is described in section 2.2.1.2. We then match the survey data to their corresponding tax bracket. In general, it is necessary to regroup certain tax brackets to make sure that we have at least one (and preferably more) observations in each bracket. Otherwise the calibration will not be possible. We automatically regroup

[10] See `wid.world/gpinter` for an online interface and a R package to apply the method.

brackets to have a partition of the income distribution at the top such that each bracket has at least 5 survey observations.

Our correction procedure also tries to constrains the number of times the weights are expanded or reduced to avoid disproportionate adjustments to single observations already in the dataset. Consequently we introduce the condition that brackets with a $\theta(y)$ outside the boundary defined by $1/a \leq \theta(y) \leq a$ are automatically grouped into larger brackets. The default limit we choose is $a = 5$. Thus, in this case, no observation would have their weight multiplied by more than 5 times or less than 0.2 times.[11]

Assume that we eventually get m brackets, with the k-th bracket covering a fraction p_k of the population. We create dummy variables $b_1, \dots, b_m$ for each income bracket. If the total population is N and the sample-size is n, then the calibrated weights should satisfy:

$$\forall k \in \{1, \dots, m\} \qquad \sum_{i=1}^{n} w_i b_{ik} = N p_k$$

Since these equations are expressed as totals of variables, they can directly enter the calibration problem (2.4). In practice, we are enforcing the income distribution through a histogram approximation of it.

The flexibility of the calibration procedure lets us put additional constraints in the calibration problem. In particular, if the survey is already assumed to be representative in terms of age or gender, then their distribution can be kept constant during the procedure. Hence we correct for the income distribution while maintaining the representativeness of the survey along the other dimensions. Additional constraints are also possible, if external information on other variables is available (see section 2.2.2.2).

For all the observations below the merging point, the dummy variables $b_1, \dots, b_m$ are all equal to zero, so the weight adjustment only depends on a constant and possibly other calibration variables such as age and gender, but not income. This matches the uniform adjustment profile (2.2) at the bottom of the distribution that we present in section 2.2.1.1. The calibration, by construction, avoids distorting the bottom of the distribution because it is not necessary to enforce the constraints of the calibration problem.

[11] Some observations may still fall outside of these constraints if covariates are present, but in practice only to a limited extent.

2.2.2.2 Extensions

The calibration framework is generic enough to incorporate information into the survey in different forms. While the most standard problem is to directly correct the income distribution using the income concept of interest, more complicated settings can sometimes occur. The flexibility of the calibration framework makes it generally possible to deal with these settings without resorting to additional *ad hoc* assumptions. We discuss below three common cases.

Using Population Characteristics by Income Tax data sometimes provides information on the population characteristics by income level, typically, the gender composition. This can tell us how the interaction between income and other characteristics impacts the bias, so it can be useful to include this information in the survey.

Assume that we have m income tax brackets that contain a share $p_1, \dots, p_m$ of the overall population N. For each of them, we know the share $\boldsymbol{s} = (s_1, \dots, s_m)$ of people with a given characteristic, such as belonging to a certain gender or age group. Let v_i be the variable equal to 1 if unit i belongs to that group in the survey, and 0 otherwise. Let b_{ik} be the variable equal to 1 if unit i in the survey is in income bracket k, and 0 otherwise.

To make sure that the survey reproduces the information in the tax data, we add the following constraints to the calibration problem (2.4):

$$\forall k \in \{1, \dots, m\} \qquad \sum_{i=1}^{n} w_i b_{ik} v_i = N s_k p_k$$

Using Income Composition Another source of information that is commonly available in tax data is the composition of income within brackets. Using that information is useful if we assume that the bias may be different for people that derive their income from, say, capital rather than labor.

Assume that we have m income brackets. For each of them, we know the share $\boldsymbol{s} = (s_1, \dots, s_m)$ of capital income. In the survey, total income is recorded as y_i and capital income as c_i. Let b_{ik} be a variable equal to 1 if unit i in the survey is in income bracket k. In order to enforce the constraint that the share of capital income within each bracket is the same as in the tax data, it suffice to enforce the constraints:

$$\forall k \in \{1, \dots, m\} \qquad \sum_{i=1}^{n} w_i b_{ik}(c_i - s_k y_i) = 0$$

Indeed, the first part of the sum is $\sum_{i=1}^{n} w_i b_{ik} c_i$, which is the total capital income of the bracket. In the second part we have the total income of the bracket $\sum_{i=1}^{n} w_i b_{ik} y_i$, multiplied by the capital share s_k. This constraint can be expressed as a total of the variable $b_{ik}(c_i - s_k y_i)$. We can see that units will see their weight decrease or increase depending on whether their capital share is below or above the average of the bracket they belong to.

Using several income concepts Until now we have considered the case where the income recorded in tax data more or less matches the income concept of interest, which is the income likely to drive the bias. Yet sometimes only part of this income is recorded in the tax data. For example, in developing countries, only income from the formal sector may be recorded in the tax data, and there is a sizable informal sector only present in the survey data, which is widely spread across the distribution (as in Czajka (2017)).

In such cases, it would be problematic to directly apply the calibration method described previously. Indeed, since the adjustment factor of the weights would only depend on formal sector income, two people with the same income, one working in the formal sector and the other in the informal sector, would see their weight adjusted very differently. As a result, there would be almost no correction for the income distribution of the informal sector.

The solution to that problem is to use Deville's (2000) generalized calibration approach. The standard calibration approach formulated in (2.4) does not specify on what variable the weight adjustment factors should depend. In the solution of the problem, they depend directly on the variables used in the constraint. This is because the method always favors the least distorting adjustments, so it only uses the variables most directly related to the constraints.

If we have some prior knowledge of what the bias should depend on, then we can use generalized calibration to specify these variables *ex ante*. We still use $\boldsymbol{x}_i$ to denote the k calibration variables for which we know the true population totals $\boldsymbol{t}$. In the example, it would include formal sector income in addition to basic socio-demographic characteristics. We also define $\boldsymbol{z}_i$, a vector of instrumental calibration variables with the same size as $\boldsymbol{x}_i$. They may include variables in $\boldsymbol{x}_i$ (e.g. socio-demographic variables) but more importantly also some variables imperfectly correlated with the

$\boldsymbol{x}_i$, in the example the sum of formal and informal sector income. We write the calibration problem as finding $w_1, \dots, w_n$ such that:

$$\sum_{i=1}^{n} w_i \boldsymbol{x}_i = \boldsymbol{t} \qquad \text{and} \qquad \forall i \in \{1, \dots, n\} \qquad \frac{w_i}{d_i} = 1 + \boldsymbol{\beta} \boldsymbol{z}_i \tag{2.7}$$

When $\boldsymbol{x}_i = \boldsymbol{z}_i$, the problem (2.7) is equivalent to (2.4). The solution of (2.7) given by Deville (2000) is similar to that of (2.5):

$$\boldsymbol{\beta} = \boldsymbol{T}^{-1} \left(\boldsymbol{t} - \sum_{i=1}^{n} d_i \boldsymbol{x}_i \right) \qquad \text{with} \qquad \boldsymbol{T} = \sum_{i=1}^{n} d_i \boldsymbol{z}_i \boldsymbol{x}_i'$$

While we may view the standard calibration as performing a projection of the variable of interest y_i onto the calibration variables $\boldsymbol{x}_i$ using an OLS regression, the generalized calibration performs that same projection using an IV regression with $\boldsymbol{z}_i$ as a vector of instruments for $\boldsymbol{x}_i$. For this to work properly, we need $\boldsymbol{z}_i$ to be sufficiently correlated with $\boldsymbol{x}_i$, otherwise we face a weak instrument problem similar to that of traditional IV regressions (Lesage, Haziza, and D'Haultfoeuille, 2018). This is not a major concern in the example since the sum of formal and informal income is strongly correlated with formal income by construction.

2.2.3 Expanding the Support

After applying the methods of the previous sections, the survey should be statistically indistinguishable from the tax data. However, the precision that we get at the top of the income distribution may still be insufficient for some purposes. Indeed, the number of observations in the survey is still significantly lower than what we would get in theory from administrative microdata. The extent to which this represents a problem varies. If we use survey weights to, say, run regressions and get better estimates of average partial effects in presence of unmodeled heterogeneity of treatment effects (Solon, Haider, and Wooldridge, 2015), then the reweighting step is enough. But problems may arise if we wish to produce indicators of inequality, especially the ones that focus on the top of the distribution, like top income shares. The combination of a low number of observations with fat-tailed distributions can create small sample biases for the quantiles and top shares (Okolewski and Rychlik, 2001; Taleb and Douady, 2015), and skewed distributions of the sample mean (Fleming, 2007). In most cases, we would underestimate levels of inequality.

Unlike problems caused by, say, heterogeneous response rates, these biases are part of *sampling error*. They do not reflect fundamental issues with the validity of the survey, but arise purely out of its limited sample size. The calibration method (section 2.2.2) does, to some extent, reduce sampling error. Yet it only does so under asymptotic conditions (Deville and Särndal, 1992) that cannot hold for narrow groups at the top of the income distribution. For this reason, we prefer to consider that the role of survey calibration in our methodology is to deal with *non-sampling error*. We use a different approach to deal with sampling error.

In particular, we aim to solve the case where tax statistics include a positive number of income-declarations beyond the survey's support. That is, we need to account for individuals declaring higher income than the richest persons in the surveys, which cannot be solved by re-weighting observations. To do so, we start from the original tax tabulations, which were created from the entire population of taxpayers and should therefore be free of sampling error. We use it alongside a generalized Pareto interpolation to estimate a continuous income distribution (Blanchet, Fournier, and Piketty, 2017) that reproduces the features of the tax data with high precision. We then statistically match the information in the calibrated survey data with the tax data by preserving the rank of each observation.

More precisely: we inflate the number of data points in the survey by making k_i duplicates of each observation i. We attribute to each new observation the weight $q_i = w_i/k_i$, where w_i is the calibrated weight from the previous step. We choose $k_i = [\pi \times w_i]$ where $[x]$ is x rounded to the nearest integer. Therefore all new observations have an approximately equal weight close to $1/\pi$. The size of the new dataset, made out of the duplicated observations, can be made arbitrarily high by adjusting π, yet any linear weighted statistic will be the same over both datasets.

Let M be the number of observations in the new dataset. The weights are assumed to sum to the population size N. We will associate to each of them a small share $[0, q_{j_1}/N], [q_{j_1}/N, (q_{j_1} + q_{j_2})/N], \dots, [\sum_{k=1}^{M} q_{j_k}/N, 1]$ of the true population. If we attribute to each observation the average income of their population share in the tax data, then by construction the income distribution of the newly created survey will be the same as in the tax data. We rank observations in increasing order by income to preserve the joint distribution between income and the covariates in the survey.

From an intuitive perspective, this process can be described as replacing the income of observations beyond the merging point with the income of observations with equivalent weight and rank in the tax distribution. This step ensures that the we reproduce exactly the income distribution from tax data, preserve the survey's covariate distribution (including the household structure), and limit distortions in the relationship between income and covariates from survey data.

2.3 Applications

In section 2.3.1, we run controlled experiments with parametric distributions, using the Monte-Carlo approach, in order to assess the accuracy of estimates produced after applying both our adjustment method and the common replacing alternative found in the literature.[12] In section 2.3.2, we illustrate how the method operates with actual household surveys and tax statistics, applying it to data from five countries (France, U.K., Norway, Brazil and Chile). Our chosen case studies showcase the wide applicability of the method to both developed countries and less-developed ones — the latter's data tending to be more challenging.

2.3.1 Simulations

Our experiments start with the simulation of a 'true' distribution with several million individuals, which follow a parametric distribution. We emulate a typical tax-tabulation, which summarizes information on the richest fractiles of that same distribution in different intervals. We then draw a number of pseudo-random samples from the original distribution, simulating surveys to a given share of the population each time, which we adjust following both our method and the replacing method common in the literature.

All samples are biased by definition, including both the misreporting and non-response biases. The former is defined by a probability of misreporting that is assumed to be flat for most of the distribution and assumed to increase linearly with rank only at the top. The distribution of misreported income is also defined parametrically, in such a way to ensure a prevalence of under-reporting over over-reporting, and we misreported incomes are drawn randomly from that distribution. Response rates are also assumed to be flat for most of the distribution and they only fall — linearly with rank — at the top. In what follows, we comment on what we consider to be our benchmark experiment. Yet, other experiments were conducted, using different sets of parameters and assumptions (Appendix B.4). These include alternative assumptions for each bias, variations in the replacing procedure, the size of the replaced population and the coverage of the simulated tax data. However, despite different — and sometimes extreme — assumptions, these experiments consistently

[12] We choose the replacing alternative as it is the most prevalent one which utilises external data to correct surveys.

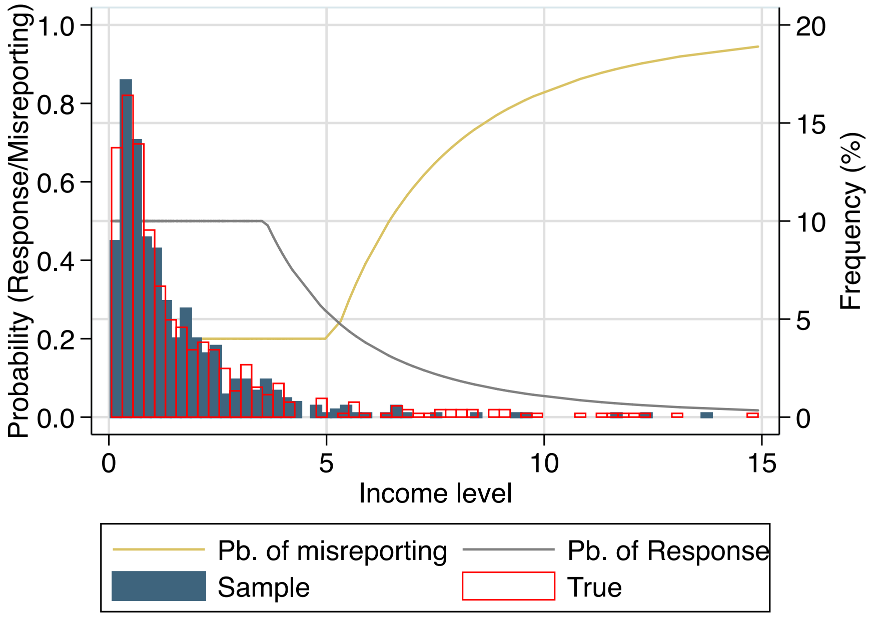
demonstrate that our algorithm is adaptive and capable of implementing adjustments that push surveys closer to the true distribution when its right tail is biased.[13]

In our benchmark experiment, we study a population of 9 million individuals that are randomly drawn from a standard normal distribution. We use the exponential function of sampled values so that the distribution fits a lognormal distribution. We select a thousand random subsamples from it, whose size correspond to 1% of the total population. The expected response rate, conditional of being sampled, is 50% for most of the population; it then decreases from percentile 90 (P90) onward and tends to 0 for the richest individual, resulting in a general response rate of 47.5%. The probability of misreporting is 20% until percentile 95 (P95); it then increases, approaching 100% at the very top. The probability of misreporting is close to 22% on average and the distribution of misreported income is also a standard lognormal. In practice, all individuals in the simulated distribution have the same probability of being 'surveyed' (1 in 100), yet individuals have their own likelihood of answering the survey and if they do answer, their response can be either accurate or misreported. Hence, in such context, although the surveyed sample is 1% of the population, only close to 0.5% of the population effectively reports income. Figure 2.4 graphically depicts the set-up of our benchmark experiment, for one of the random samples. We apply both our adjustment method, described in the previous section, and the alternative replacing procedure. The latter corresponds to the most common form that is found in the literature, which consists in replacing the top 1% of the survey distribution with that from tax data.

Figure 2.5 compares the accuracy of distributional estimates that result from the raw simulated survey to those resulting from the application of both our method and replacing. It displays errors with respect to true values for a series of estimates. Kernel densities provide a visual appreciation of the set of measurements that are found for all the 1000 iterations. The true values are: an average close to 1.6, a Gini coefficient close to 0.52, a top 1% share close to 9.3% and a top 10% share close to 39%. It appears quite clearly that our method's estimates tend to be more accurate than others in all cases, as they are systematically closer to the true estimates and they are visibly less variant. Although both adjustment methods operate differently, in purely distributional terms they both reproduce the information of the top 1% that is found in tax data. That is, after applying the adjustment, the average income

[13] All our experiments were conducted using the `bfmtoy` command that comes with the `bfmcorr` Stata package. Not only was it coded to be able to reproduce our experiments, but it also provides a tool for researchers to simulate artificial distributions and easily change all the parameters involved to test survey-adjustment methods.

Figure 2.4: Benchmark Experiment Set-Up

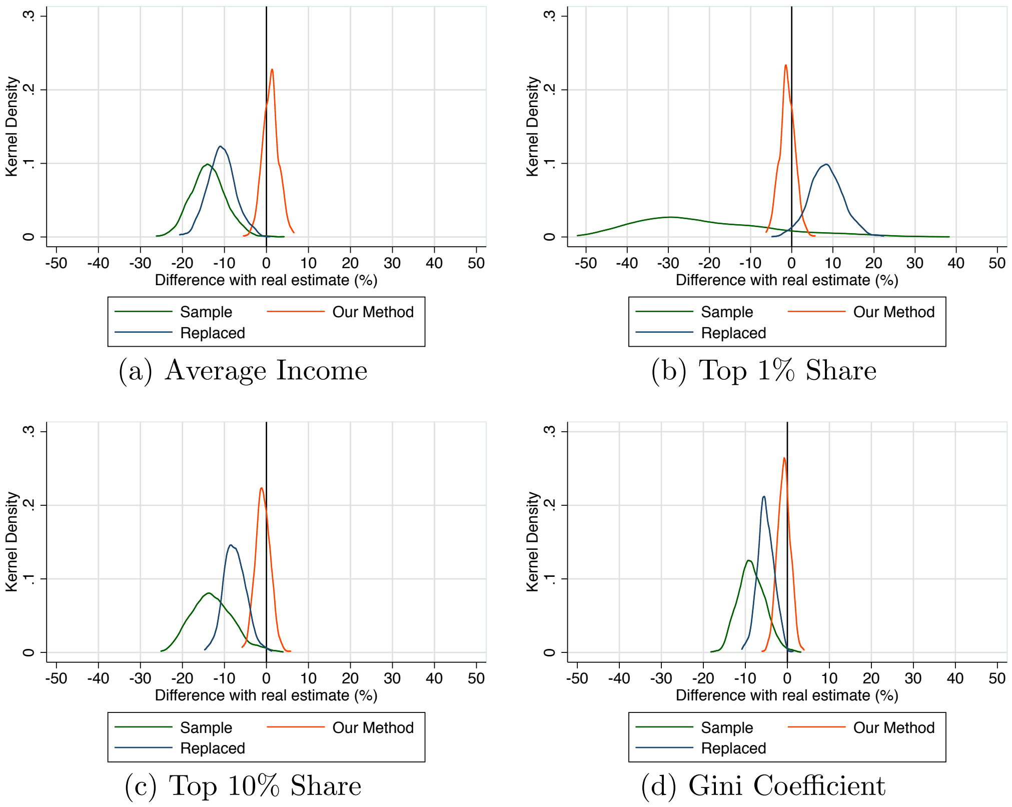


of the top percentile should be equivalent in both cases. However, the same is not necessarily true for the rest of the distribution, and thus not for average income either. Indeed, figure 2.5a shows that even if the average income gets closer to the true value with replacing, it still remains underestimated by a tenth of the true value on average, instead of 15% in the raw survey. The lower total income is thus what explains that in figure 2.5b, the top 1% shares seem to be systematically overestimated with replacing because the numerator of the top share is the same in both, but the denominator is underestimated in replacing. In the case of the top 10%, the error goes on the opposite direction (figure 2.5c). This is because an arbitrary correction of the top 1% is not enough to adjust for a distribution where the top decile is affected both by higher non-response and misreporting. When we focus on a synthetic indicator of inequality, such as the Gini coefficient, we find a similar hierarchy of estimates (figure 2.5d). It is also worth noticing that the raw sample estimate of the top 1% share is considerably less precise than any of the other estimates. This is due to a large extent to the small sample bias referred to by Taleb and Douady (2015), which is amplified by both nonresponse and misreporting.

2.3.2 Real Data

Our method can be replicated for all countries with the requisite data, namely, survey micro-data covering the entire population and tax data covering at least a fraction of

Figure 2.5: Benchmark Experiment Results



(a) Average Income

(b) Top 1% Share

(c) Top 10% Share

(d) Gini Coefficient

it.[14] We experiment with five real distributions, three European countries — making use of the common survey framework applied to them — and two less-developed Latin American countries, which can be imagined to present more of a challenge regarding data quality and scope.

2.3.2.1 Definitions and Data

A crucial preliminary step in the analysis is to reconcile both the definition of income and the unit of observation in national surveys with the ones that are used in tax declarations. Our algorithm functions under the supposition that these definitions have been made consistent in the two datasets by researchers. For France, Norway and the U.K., our analysis broadly covers the years 2004-2014. For Brazil, we cover 2007-2015 and for Chile we include the years 2009, 2011, 2013 and 2015. Consistent with the calibration procedure explained in section 2.2.2 we preserve the representativeness — not only of income — but also of other variables for which the

[14] In the case where users only avail of tabulated survey data our method will still perform the correction, using percentile bracket-information from the synthetic micro-files produced by the *gpinter* program (see `wid.world/gpinter`).

survey is assumed to be already representative, namely gender and age variables.[15]

Income Concept Given that we seek to approximate the benchmark distribution, our method is by definition anchored to the income concept that is used in the tax tabulations, which in all of our case studies is pre-tax income. However, countries differ in the income concept included in their respective surveys. Brazil's PNAD reports individuals' pre-tax income, while Chile's CASEN gives after-tax income. Thus, for Chile we require to impute taxes paid to arrive at gross income. Appendix B.5.1 explains how this imputation is done, as well as the construction of income units in surveys and their approximation with tax data in all countries. For European countries we work with gross incomes (pre-tax and employee contributions deducted at source) from the SILC database.[16] France is the exception since incomes reported in the tax files are net of employee contributions deducted at source. For this reason we use the concept of net income in SILC for France that deducts social contributions levied at source.

The tax data we use is presented in tabulated form, containing at the very least, the number of income recipients by given income intervals and the total or average income declared within each interval. For France, we use the tabulated tax statistics produced by Garbinti, Goupille-Lebret, and Piketty (2016) from the ministry of finance's tax microdata. The data cover all tax units (*foyers fiscaux* – singles or married couples), with about 50% of these subject to positive income tax. For the U.K. we use tax tabulations from the Survey of Personal Incomes (SPI) available from the Office of National Statistics. The underlying data covers about 80-90% of tax units (individuals) aged 15+, with about 60% subject to positive income tax. For Norway, we use tax data from Statistics Norway, which covers 100% of tax units (individuals) aged 17 and over, of which roughly 90% have positive income tax payments. For Brazil we use tax data from the personal income tax declarations (DIPRF tables), which covers about 20% of the adult population, with about 14% subject to the personal income tax on taxable income. For Chile we exploit income tax data from the *Global Complementario* and *Impesto Único de Segunda Categoría* (IGC and IUSC tabulations), which covers 70% of the adult population, with about 15-20% subject to the personal income tax on taxable income. For all cases, we take the proportion of population with positive tax payments as the "trustable span" of the tax data. The intuition for this choice is that individuals subject to income tax are less likely to misreport their income compared individuals who declare but are

[15] We do so using the command `holdmargins`. See the instructions to `bfmcorr` in Stata.

[16] In all countries, gross income is after employer social contributions.

under the tax-paying threshold.

Observational Unit Concerning the observational units, we anchor the definition to the official tax unit in each country. In all of our country cases declarations are made at the individual level, except in France and Brazil, where declarations are jointly filed by married couples (in the case of the latter, at their own discretion). However, for France we make use of the individually-declared fiscal income files produced by Garbinti, Goupille-Lebret, and Piketty (2016) from the administratice microdata. Therefore for all countries, we define the unit of analysis across datasets as individual income, including for Brazil, where the joint income of couples is equally split between the component members (see Appendix B.5.1 and Morgan (2018) for further details).

2.3.2.2 Empirical Bias and Corrected Population

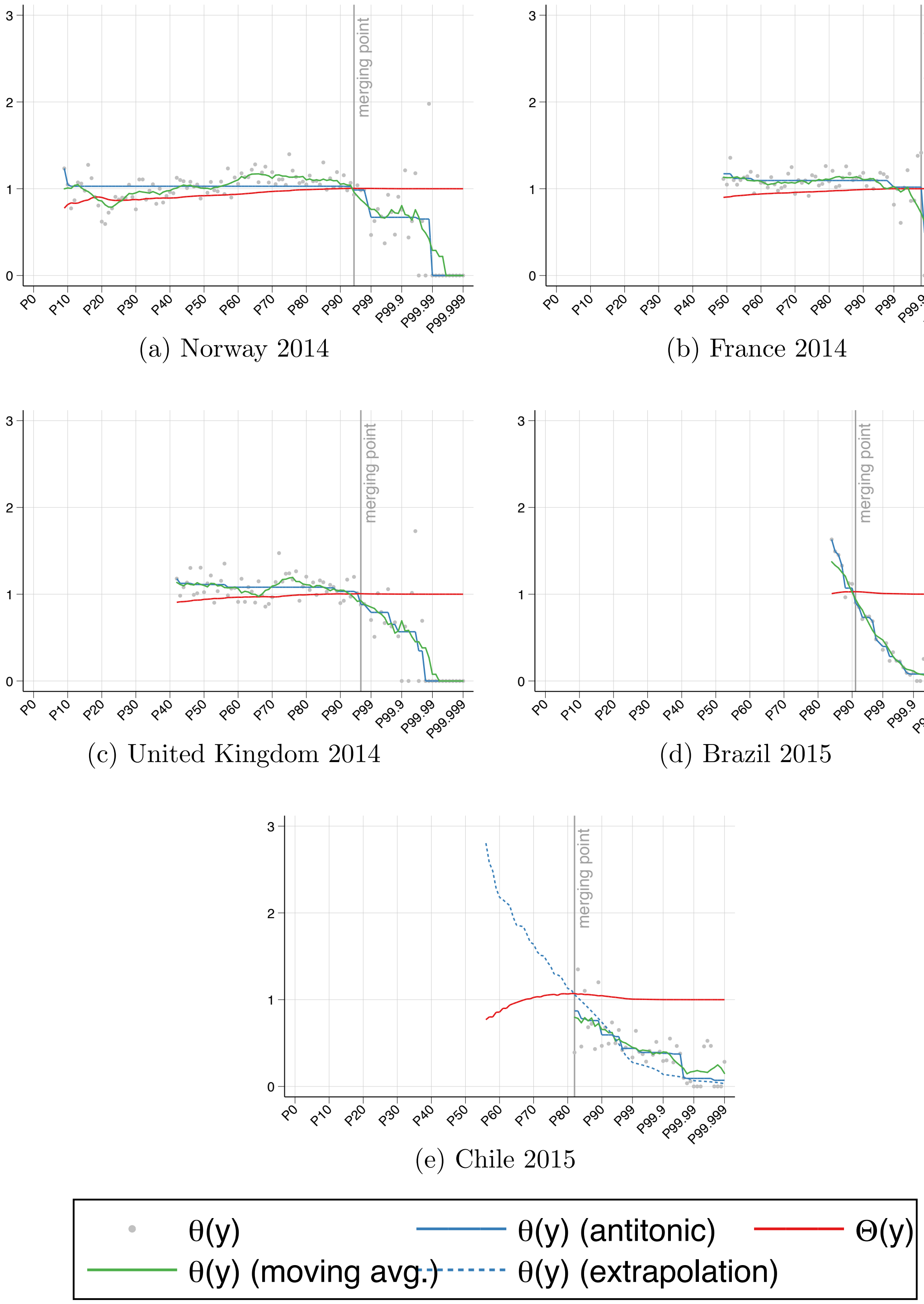
The Shape of the Bias Our method finds the merging point between surveys and tax data by comparing the population densities at specified income levels, as explained in section 2.2.1.2. To do so we first interpolate the fiscal incomes in the tabulation using the generalized Pareto interpolation (`https://wid.world/gpinter`) developed by Blanchet, Fournier, and Piketty (2017), which allows for the expansion of the tabulated income values into 127 intervals.[17] Using the thresholds of these intervals we can construct our key statistics: the frequency ($\theta(y)$) and cumulative frequency ($\Theta(y)$) of individuals along the income distribution.

Figure 2.6 presents depictions of the shape of the empirical bias within the tax data's "trustable span" for all countries for the latest available year. First of all, the shape of the bias we measure from the data is very similar to the one we present in the theoretical formalization, depicted in figures 2.3 and B.3. In particular, we always observe a convex shape in the top tail, to the right of the merging point. It thus appears that surveys tend to increasingly underestimate the frequency of incomes beyond a certain point in the distribution.

For the more developed countries (Norway, France and the United Kingdom), the

[17] These comprise of 100 percentiles from P0 to P100, where the top percentile (P99–100) is split into 10 deciles (P99.0, P99.1, . . . , P99.9-100), the top decile of the top percentile (P99.9–100) being split into ten deciles itself (P99.90, P99.91, . . . , P99.99-100), and so forth until P99.999. This interpolation technique, contrary to the standard Pareto interpolation, allows us to recover the income distribution without the need for parametric approximations. It estimates a full set of Pareto coefficients by using a given number of empirical thresholds provided by tabulated data. As such the Pareto distribution is given a flexible form, which overcomes the constancy condition of standard power laws, and produces smoother and more precise estimates of the distribution.

Figure 2.6: Merging Point in Five Countries, Latest year



(a) Norway 2014

(b) France 2014

(c) United Kingdom 2014

(d) Brazil 2015

(e) Chile 2015

Notes: the figures depict the estimated bias in the survey relative to the tax data. Grey dots are, for each quantile of the fiscal income distribution, the ratio of income density in the survey over that of tax data. The green line is the centered average of $\theta(y)$ at each quantile and eight neighboring estimates. The blue line is the result of an *antitonic* regression applied to $\theta(y)$. It is constrained to be decreasing as it is used to find a single merging point. The blue dotted line, which only appears in figure 2.6e,is an extrapolation of the trend described by $\theta(y)$ based on a *ridge* regression (see Appendix B.2). The red line is the ratio of the cumulative densities. For details refer to section 2.2.1.2.

shape of the empirical bias $\theta(y)$ can be observed for a more comprehensive share of the population, due to the greater population coverage in tax data. This enables us to empirically test our theoretical expectations on the specific behavior of the bias to the left of the merging point. We indeed observe on the left side of figures 2.6a, 2.6b and 2.6c, a general stability in the relative rate of response, with averages trending above 1. The extent and quality of tax data below the merging point in less-developed countries is such that we cannot observe the same trends.[18]

The merging points found by our algorithm vary by country and by year, again revealing differences in data quality and coverage between them. The Chilean case (figure 2.6e) provides an example of our program needing to extrapolate the shape of the bias to find the merging point (see Appendix B.2 for more details of this procedure). For this case we rely on parameters observed for Brazil (specifically, values for the elasticity of response to income) above its trustable span as inputs for the Chilean extrapolation.[19] The fit with the existing data seems to work quite well. The empirical bias that is observed in previous years for all countries is presented in Appendix B.5.2.1.

Corrected Population Our program then adjusts the individual weights of survey respondents in line with information from tax data, as described in section 2.1. We provide some summary statistics of the population we correct in table 2.1, again using the last available year for each country as illustrations (see Appendix B.5.2.2 for other years). According to the comparison of surveys with tax records, a varying proportion of the total population is adjusted at the top of the survey distribution in each country (column [4] of table 2.1), ranging from 5.9% in Chile to 0.05% in France for their most recent years.[20] This is derived from the comparison of the share of the population above the merging point in the two datasets. Since we use incomes in tax data as the benchmark for the top of the distribution, the share of the population above the merging point in tax data is directly related to the merging point. The share of the population above this point in surveys is always lower, indicating under-coverage of top incomes. But in both cases, the

[18] Tax enforcement issues affecting this portion of the distribution could be at play here, as well as the sharp difference in incomes between the top and the rest in these countries leading to higher inequality levels than developed countries.

[19] The value of the baseline elasticity of response to income, γ_1^*, extracted from the Brazilian data is -0.99.

[20] Across years there is less variation in this share, with Norway and particularly France being relative exceptions. In the French case, we believe the significant break in the series is due to the use of register data in SILC alongside the household survey from 2008. Despite the SILC survey making use of register data, the goal is not to over-sample the top of the distribution, but rather to improve the precision of responses.

Table 2.1: Structure of Corrected Population: Latest Year

Country	Population over Merging Point (% total population)		Corrected population		
	Tax data	Survey	Total	Share inside survey support	Share outside survey support
	[2]	[3]	[4] = [2] − [3]	[5]	[6]
Chile	17.0%	11.1%	5.9%	99.99%	0.01%
Brazil	8.0%	5.3%	2.7%	99.0%	1.0%
UK	3.0%	2.5%	0.5%	93.6%	6.4%
Norway	5.0%	4.6%	0.4%	96.0%	4.0%
France	0.1%	0.05%	0.05%	99.0%	1.0%

Notes: The table orders countries by the size of the corrected population. Column [2] shows the proportion of the population that is above this merging point in the tax data. Column [3] shows the proportion that is above the merging point in survey data. The difference between the two is the proportion of the survey population that is corrected (Column [4]). As explained in the text, we adjust survey weights below the merging point by the same proportion. The corrected proportion above the merging point can be decomposed into the share of the corrected population that is inside the survey support (up to the survey's maximum income) and the share that is outside the support (observations with income above the survey's maximum). Brazil and Chile refer to 2015, while all the European countries refer to 2014.

overwhelming majority of the adjustment (over 90%) can be seen to come from inside the survey support, rather than outside the survey's original support, suggesting that non-sampling issues related to heterogeneous response rates matter more than problems related to under-sampling for the size of the corrected population.

In general, this step of the algorithm is a useful guide to assess the income coverage of surveys across countries. For instance, it appears on the basis of our analysis that the Brazilian surveys do a better job at capturing gross income, given the lower share of the underrepresented population, than the Chilean household surveys. Moreover, comparing France and the UK, it seems that sampling error is greater in the UK surveys, given the higher share of the population beyond the survey's maximum income that needs to be added. Non-sampling error itself is greatest in Chile, derived from the share of the corrected population found inside the survey's support.

2.3.2.3 Results

We now turn to unveil how different our merged distributions are with respect to the raw survey distributions and other corrected distributions based on the most common replacing method found in the literature that utilises external data. The latter corresponds to the procedure reproduced in the simulation in section 2.3.1, whereby the top 1% of the survey distribution is directly replaced by the top 1% of the tax distribution. We present results on top 1% income shares, Gini coefficients and average incomes.[21]

Top Income Shares In line with the improved income coverage that are method produces — by more accurately including upper incomes — estimates of the income concentrated at the top of the distribution are revised upwards in all countries. The size of the adjustment, however, varies by country. Figure 2.7 depicts this for the Top 1% share.[22] Brazil has the most extensive correction, with a top 1% share that increases by about 10 percentage points every year (figure 2.7d). Conversely, France and Norway experience relatively smaller adjustments, starting from relatively lower levels of inequality. In addition, Brazil offers the clearest illustration of the distinct trends in inequality that can emerge after making a correction to the survey's representation of income. While the raw survey depicts falling top income shares, the corrected survey distribution shows slightly increasing top shares. Distinct trends are also visible, albeit for shorter periods of time, in the other countries.

The quality of both surveys and tax statistics may have a substantial impact on the size of the adjustment. For instance, in the case of France, several improvements were made to the survey's methodology starting in 2008. In particular, the matching of individuals across survey and register data allowed for the use of tax data as an external source to assess individual income without recourse to self-reporting. This testifies to the more accurate reporting of income in subsequent years, even though the gap in shares does not fully disappear in all years. Although this incorporation of register data remedies problems of misreporting and item non-response (failure to answer certain income questions), it cannot itself get around unit non-response (failure to answer the entire survey), or issues of under-sampling.

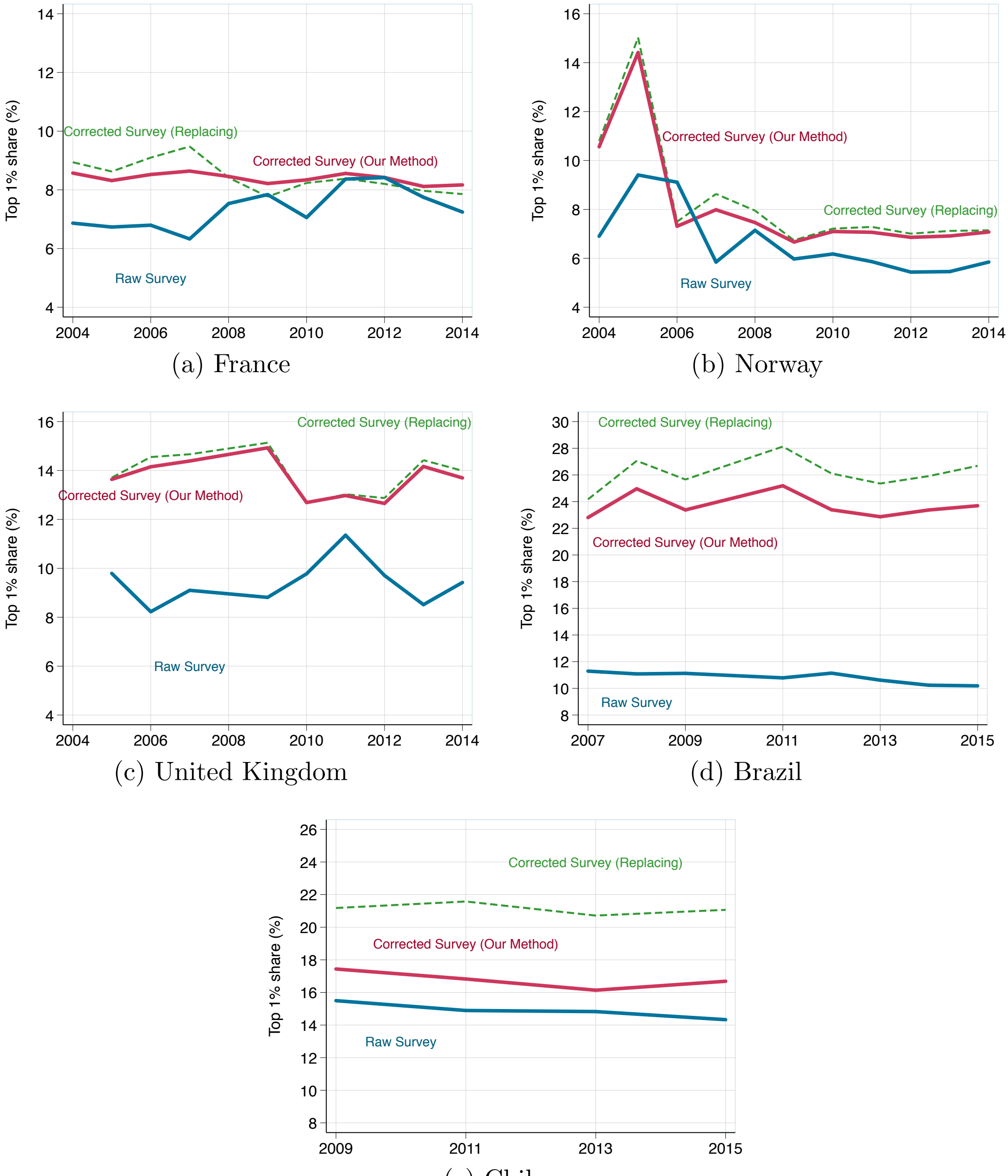
Moreover, when we compare the size of the adjustment in Chile and Brazil (figures 2.7d and 2.7e respectively), two highly unequal Latin American countries, the latter has a considerably higher adjustment. One of the reasons that could be behind this phenomenon is the fact that capital income, especially dividends, is better recorded in Brazilian tax statistics. Indeed, the Brazilian tax agency has relatively good means to verify the accuracy of capital income declarations (Morgan, 2018), while Chilean tax authorities are generally constrained by bank secrecy (Fairfield and Jorratt De Luis, 2016). In this case, the limited quality of Chilean tax statistics explains the smaller correction.[23]

[21] Appendix B.5.3 presents results for other income groups in the distribution.

[22] The one exception to this upward correction is Norway in 2006 (see figure 2.7b). However, this is likely due to a change in the local tax legislation affecting the distribution of business profits (Alstadsæter et al., 2016), as we explain in the text.

[23] There is also a considerable difference between these countries' tax systems and their respective

Figure 2.7: Top 1% Shares Before and After Correction



(a) France

(b) Norway

(c) United Kingdom

(d) Brazil

(e) Chile

Notes: the corrected survey using the replacing method directly replaces the survey distribution above P99 with the distribution above P99 from the tax distribution.

Following the same rationale, the inclusion or exclusion of some types of income in a given dataset can also affect the size of the correction. In the case of Norway, tax incentives started favoring the retention of corporate profits inside corporations after 2005, with the creation of a permanent dividends tax in 2006. This resulted in less dividend payments, and thus less income to be registered as personal income in tax data. The reform also gave strong incentives for higher-than-normal dividend payouts in 2005, which contributed to the sharp increase in top shares observed for this year (Aaberge and Atkinson, 2010; Alstadsæter et al., 2016). In figure 2.7b, it can be clearly perceived that the size of the adjustment appears to drop durably after this year. Additionally, it should be noticed that the Norwegian survey appears to be rather insensitive to this change, implying that dividends where badly represented before 2005. Other potential explanations for the difference in the size of adjustments could have to do with behavioural differences between populations across countries related to response rates and reporting accuracy.

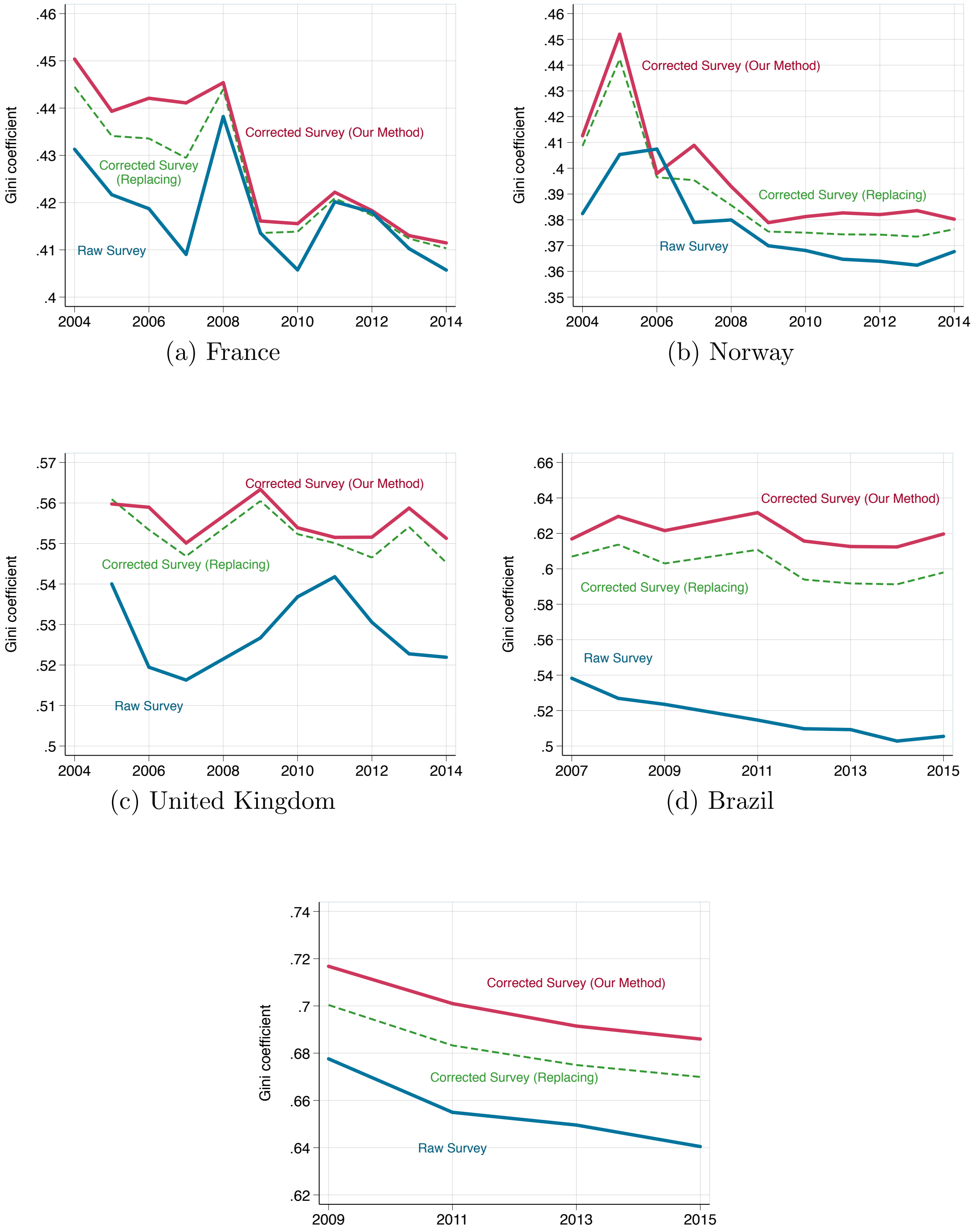
The extent of the adjustment, by definition, depends directly on the shape of the bias that is observed in figure 2.6. Both the steepness of $\theta(y)$, when it is to the right side of the merging point, and the size of the corrected population (column 4 in table 2.1) are decisive factors for the size of such an increase. Another way to think about the size of the corrected population is to look at the size of the area between $\theta(y)$ and 1, to the right side of the merging point.

Finally, comparing between correction methods, we can observe — in line with our simulations — that the top 1% share is generally higher in the replacing scenario than in our method due to the fact that while the level of numerator incomes is equivalent in both settings, average incomes (the denominator) is underestimated in the former scenario, as we show further below.

Gini Coefficients Figure 2.8 shows the time series of the Gini coefficients before and after the correction for all available years. Overall, we find a similar hierarchy of estimates, mirroring our simulations in the previous section — inequality is corrected upwards, more so in countries whose raw survey is not already matched with any administrative source, and to different degrees depending on the year, thus producing distinct trends. This is further evidence that surveys need to be adjusted if they are to better represent the income distribution, in the same manner as they are currently

incentives. In Chile most dividends received by individuals are taxed, while in Brazil they are not. This, in addition to the fact that Chilean realized capital gains are mostly un-taxed, provokes incentives towards the artificial retention of profits that are not as present in Brazil. This is why, in Chile, the imputation of undistributed profits to the distribution of personal income appears to be necessary when making international comparisons (Flores et al., n.d.).

Figure 2.8: Gini Coefficients, Before and After Correction



(a) France

(b) Norway

(c) United Kingdom

(d) Brazil

(e) Chile

Notes: the corrected survey using the replacing method directly replaces the survey distribution above P99 with the distribution above P99 from the tax distribution.

calibrated to better represent the distribution of various demographic variables.

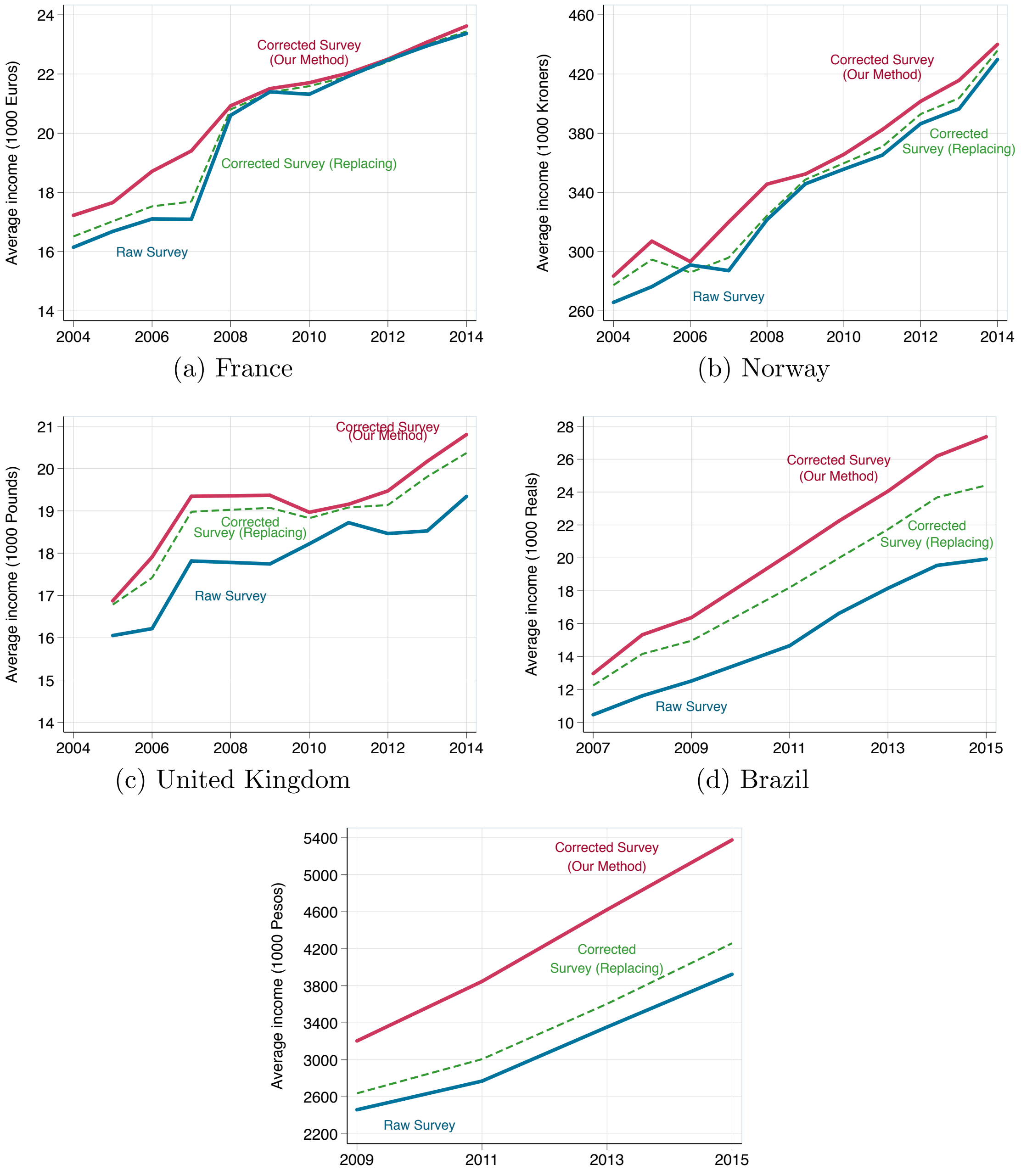
Again consistent with our simulations, the replacing procedure seems to undershoot inequality levels compared to our method, which more accurately accounts for higher non-response and misreporting at the top. An arbitrary correction of the top 1% is not enough to adjust the under-coverage of income coming from these errors. This is especially the case where the corrected population is larger than the arbitrarily chosen fractile, such as in Brazil and Chile (see figure 2.6 and table 2.1).

Average Incomes As alluded to before, the average income of the top percentile using both correction methods is the same, which is higher than the level observed in the raw surveys. However, the crucial difference between the two methods is that the average incomes for the other groups in the population are not equal. In our method, the weight of persons with lower incomes are reduced, while the replacing method keeps the same average income for the bottom income groups. This subsequently produces differences in the overall average income of the population in both cases. Figure 2.9 depicts that our method increases the average income in the surveys in all countries, although with highly varying degrees of magnitude. In the lower-income countries, which have the highest corrections — Brazil and Chile — average incomes increase broadly by 30-50%, with the gap increasing over time. The higher income countries in Europe experience lower corrections to their average incomes, with the orders of magnitude between them reproducing the rank of countries by size of correction in table 2.1 — the U.K. experiences a larger correction than Norway, which experiences a larger correction than France. Visibly, in figure 2.9a the gap between the average in the raw data and corrected data is reduced from 2008 on-wards on account of the reduction in the size of the survey bias coming from the methodological novelties (see table B.4 for further details).

The result for the replacing method goes in line with expectations. It is higher than the raw survey result, as more income is given to the top of the distribution, but it is also consistently lower than the average our method produces, since it does not reduce the weight of individuals with lower incomes. This is an inconsistency coming from its own rationale, as explained in Appendix B.1.2, which our method explicitly overcomes.

The relative underestimation of incomes is further evident in figure B.13, which shows income coverage across datasets in the two countries with the largest corrected populations. The corrected survey income total from our method, which is already higher than the total from replacing as figure 2.9 testifies, is closer to a broadly

Figure 2.9: Average Incomes Before and After Correction



(a) France

(b) Norway

(c) United Kingdom

(d) Brazil

(e) Chile

Notes: the corrected survey using the replacing method directly replaces the survey distribution above P99 with the distribution above P99 from the tax distribution. Average incomes are rescaled accordingly.

equivalent income total from national accounts in both Brazil and Chile.

Conclusion

The main objective of this paper is to provide a more rigorous methodological tool that enables researchers to combine income or wealth surveys with administrative data in a simple and consistent manner. We present a new methodology on the combination of such sources, which incorporates a clearer formal understanding of the potential biases at play and a solution to remedy them. The result of our calibration-inspired approach, we argue, should be a more representative dataset that can serve as a basis to study the different dimensions of social inequality. Our algorithm is built in such way that it automatically generates, from raw surveys and tax data, an adjusted micro-dataset including new modified weights and new observations, while preserving the consistency of other pre-existing socio-demographic variables, at both the individual and aggregate level.

Our paper can thus be viewed as an attempt to improve survey representativeness by taking the income (or wealth) distribution into account. While it is common to adjust survey weights in accordance to external information on the distribution of basic socio-demographic variables, our paper motivates the use of auxiliary administrative data sources on the distribution of income to further improve the representativeness of the population.

Our procedure has several advantages compared to available options to correct surveys. First, it is based on a solid and intuitive theoretical framework. Second, our method avoids *a priori* assumptions on the size of the population to be corrected. Instead, it offers a clear procedure to find the merging point between datasets non-arbitrarily. Third, the algorithm can be applied to a wide variety of countries, both developed and less developed, since it accounts for different levels of data coverage. Fourth, our method respects original individual self-reported profiles and socio-demographic totals for variables other than income. We thus preserve the internal consistency of surveys, while better approximating the external consistency of its income distribution. Although we preserve socio-demographic totals for variables other than income, our method allows for their conditional distribution to vary upon the addition of new income information. However, our method also accommodates the input of distributional information of other variables (age, sex, income type, etc.) if they are available in the tax data. As such, one may also calibrate and correct the survey on covariates of income, in addition to income itself, if reliable statistics exist on their interaction. Finally, it should be clear that this method can serve multiple research objectives — from single-country and cross-country empirical analyses using income statistics as well as their covariates, to fiscal incidence analysis.

To the extent of harmonizing our correction procedure among different countries, we stress the importance of analyzing the underlying data in each case. For this, our method provides useful tools to researchers wishing to assess the population coverage of surveys conditional on income. Figure 2.6 and table 2.1 are examples of the type of information directly computed by our algorithm, which is made available to users as a program on Stata. With standard survey and tax data at hand, researchers can perform our correction procedure with relative ease, as long as the income concepts are/can be made comparable across the datasets.

Our practical applications show the accuracy and scope of the method. The Monte Carlo simulations reveal that our method produces results — on average incomes and inequality indicators — that are closer to values from the true distribution with lower variance, compared to the drawn sample and the common "replacing" alternative employed in the literature. This is because the structure of our method's correction takes seriously the nature of the potential biases at play. Finally, when applied to real data, our approach is shown to be robust to different contexts, with the size of adjustments depending on data quality and inequality levels by country. The wider the gap between survey and administrative data and higher the level of inequality in the country, the greater the correction is likely to be. Our empirical results are consistent with experiments we run with simulated data. Overall, we claim that our method is accurate, robust and pragmatic in unifying the strengths of separate datasets on the distribution of income/wealth and their covariates into one source of information.

Bibliography

Aaberge, Rolf and A. B. Atkinson (2010). “Top incomes in Norway”. In: *Top incomes: a global perspective* 2.

Alstadsæter, Annette et al. (2016). *Accounting for business income in measuring top income shares: Integrated accrual approach using individual and firm data from Norway.* Tech. rep. National Bureau of Economic Research.

Alvaredo, Facundo (2011). “A note on the relationship between top income shares and the Gini coefficient”. In: *Economics Letters* 110.3, pp. 274–277. URL: `http://dx.doi.org/10.1016/j.econlet.2010.10.008`.

Alvaredo, Facundo et al. (2017). “Distributional National Accounts (DINA) Guidelines: Concepts and Methods used in WID.world”.

Atkinson, A. B. and Thomas Piketty (2007). *Top incomes over the twentieth century: a contrast between continental European and English-speaking countries.* Oxford University Press, p. 585. URL: `https://global.oup.com/academic/product/top-incomes-over-the-twentieth-century-9780199286881?lang=en&cc=fr`.

– (2010). *Top incomes: a global perspective.* Oxford University Press, p. 776. URL: `https://global.oup.com/academic/product/top-incomes-9780199286898?cc=fr&lang=en&#`.

Ayer, Miriam et al. (1955). “An Empirical Distribution Function for Sampling with Incomplete Information”. In: *Ann. Math. Statist.* 26.4, pp. 641–647. URL: `https://doi.org/10.1214/aoms/1177728423`.

Blanchet, Thomas, Juliette Fournier, and Thomas Piketty (2017). “Generalized Pareto Curves: Theory and Applications”.

Bourguignon, François (2018). “Simple adjustments of observed distributions for missing income and missing people”. In: *The Journal of Economic Inequality*, pp. 1–18.

Brunk, H D (1955). “Maximum Likelihood Estimates of Monotone Parameters”. In: *Ann. Math. Statist.* 26.4, pp. 607–616. URL: `https://doi.org/10.1214/aoms/1177728420`.

Burkhauser, Richard V, Markus H Hahn, and Roger Wilkins (2016). “Top Incomes and Inequality in Australia: Reconciling Recent Estimates from Household Survey and Tax Return Data”.

Burkhauser, Richard V, Nicolas Hérault, et al. (2016). “What has Been Happening to UK Income Inequality Since the Mid-1990s? Answers from Reconciled and Combined Household Survey and Tax Return Data”. URL: `http://www.nber.org/papers/w21991`.

– (2018). “Survey Under-Coverage of Top Incomes and Estimation of Inequality: What is the Role of the UK’s SPI Adjustment?” In: *Fiscal Studies* 39.2, pp. 213–240.

Chancel, Lucas and Thomas Piketty (2017). “Indian income inequality, 1922-2014: From British Raj to Billionaire Raj?” URL: `http://wid.world/document/chancelpiketty2017widworld/`.

Czajka, Léo (2017). “Income Inequality in Côte d’Ivoire: 1985-2014”. In: *WID.world Working Paper* July.

Deming, W. Edwards and Frederick F. Stephan (1940). “On a Least Squares Adjustment of a Sampled Frequency Table When the Expected Marginal Totals are Known”. In: *The Annals of Mathematical Statistics* 11.4, pp. 427–444. URL: `https://projecteuclid.org/euclid.aoms/1177731829`.

Deville, Jean-Claude (2000). “Generalized calibration and application to weighting for non-response”. In: *COMPSTAT: Proceedings in Computational Statistics 14th Symposium held in Utrecht, The Netherlands, 2000.* Ed. by Jelke G Bethlehem and Peter G M van der Heijden. Heidelberg: Physica-Verlag HD, pp. 65–76. URL: `https://doi.org/10.1007/978-3-642-57678-2_6`.

Deville, Jean-Claude and Carl-Erik Särndal (1992). “Calibration Estimators in Survey Sampling”. In: *Journal of the American Statistical Association* 87.418, pp. 376–382.

Diaz-Bazan, Tania (2015). “Measuring Inequality from Top to Bottom”. In: *Policy Research Working Paper* 7237.

DWP (2015). “Households Below Average Income: An analysis of the income distribution 1994/95 – 2013/14”. URL: `https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/437246/households-below-average-income-1994-95-to-2013-14.pdf`.

Eeden, Constance van (1958). “Testing and Estimating Ordered Parameters of Probability Distributions”. PhD thesis. University of Amsterdam.

Fairfield, Tasha and Michel Jorratt De Luis (2016). “Top Income Shares, Business Profits, and Effective Tax Rates in Contemporary Chile”. In: *Review of Income and Wealth* 62, S120–S144.

Fleming, Kirk G (2007). “We’re Skewed—The Bias in Small Samples from Skewed Distributions”. In: *Casualty Actuarial Society Forum* 2.2, pp. 179–183.

Flores, Ignacio et al. (n.d.). “Top Incomes in Chile: A Historical Perspective on Income Inequality, 1964–2017”. In: *Review of Income and Wealth* 0.0 (). eprint: `https://onlinelibrary.wiley.com/doi/pdf/10.1111/roiw.12441`. URL: `https://onlinelibrary.wiley.com/doi/abs/10.1111/roiw.12441`. Forthcoming.

Garbinti, Bertrand, Jonathan Goupille-Lebret, and Thomas Piketty (2016). “Income Inequality in France, 1900-2014: Evidence from Distributional National Accounts (DINA)”. URL: http://piketty.pse.ens.fr/filles/GGP2016DINA.pdf.

– (2018). “Income inequality in France, 1900–2014: Evidence from Distributional National Accounts (DINA)”. In: *Journal of Public Economics* 162, pp. 63–77.

Hlasny, Vladimir and Paolo Verme (2017). “The impact of top incomes biases on the measurement of inequality in the United States”.

– (2018). “Top Incomes and Inequality Measurement: A Comparative Analysis of Correction Methods Using the EU SILC Data Vladimir”. In: *Econometrics* 6.30, pp. 1–38.

Jenkins, Stephen P (2017). “Pareto Models, Top Incomes and Recent Trends in UK Income Inequality”. In: *Economica* 84.334, pp. 261–289.

Korinek, Anton, Johan A. Mistiaen, and Martin Ravallion (2006). “Survey nonresponse and the distribution of income”. In: *Journal of Economic Inequality* 4.1, pp. 33–55.

Kuznets, Simon (1953). *Shares of Upper Income Groups in Income and Savings.* NBER. URL: http://www.jstor.org/stable/10.2307/2343040?origin=crossref.

Lesage, Éric, David Haziza, and Xavier D’Haultfoeuille (2018). “A cautionary tale on instrument vector calibration for the treatment of unit nonresponse in surveys”.

Medeiros, Marcelo, Juliana de Castro Galvão, and Luìsa de Azevedo Nazareno (2018). “Correcting the Underestimation of Top Incomes: Combining Data from Income Tax Reports and the Brazilian 2010 Census”. In: *Social Indicators Research* 135.1, pp. 233–244.

Morgan, Marc (2018). “Essays on Income Distribution: Methodological, Historical and Institutional Perspectives with Applications to the Case of Brazil (1926–2016)”. PhD Dissertation in Economics. Paris: Paris School of Economics & EHESS.

Novokmet, Filip, Thomas Piketty, and Gabriel Zucman (2018). “From Soviets to oligarchs: inequality and property in Russia 1905-2016”. In: *The Journal of Economic Inequality* 16.2, pp. 189–223.

Okolewski, Andrzej and Tomasz Rychlik (2001). “Sharp distribution-free bounds on the bias in estimating quantiles via order statistics”. In: *Statistics and Probability Letters* 52.2, pp. 207–213.

Pareto, Vilfredo (1896). *Écrits sur la courbe de la répartition de la richesse.*

Piketty, Thomas (2003). “Income Inequality in France, 1901–1998”. In: *Journal of Political Economy* 111.5, pp. 1004–1042. URL: http://www.journals.uchicago.edu/doi/10.1086/376955.

Piketty, Thomas and Emmanuel Saez (2003). “Income Inequality in the United States, 1913–1998”. In: *Quarterly Journal of Economics* CXVIII.1.

Piketty, Thomas, Emmanuel Saez, and Gabriel Zucman (2018). “Distributional National Accounts: Methods and Estimates for the United States”. In: *Quarterly Journal of Economics* 133.May, pp. 553–609.

Piketty, Thomas, Li Yang, and Gabriel Zucman (2017). “Capital Accumulation, Private Property and Rising Inequality in China, 1978-2015”. URL: `http://www.nber.org/papers/w23368.pdf`.

Singh, A C and C A Mohl (1996). “Understanding Calibration Estimators in Survey Sampling”. In: *Survey Methodology* 22.2, pp. 107–115.

Solon, Gary, Steven J. Haider, and Jeffrey M. Wooldridge (2015). “What Are We Weighting For?” In: *Journal of Human Resources* 50.2, pp. 301–316. URL: `http://jhr.uwpress.org/lookup/doi/10.3368/jhr.50.2.301`.

Taleb, Nassim Nicholas and Raphael Douady (2015). “On the super-additivity and estimation biases of quantile contributions”. In: *Physica A: Statistical Mechanics and its Applications* 429, pp. 252–260. URL: `http://dx.doi.org/10.1016/j.physa.2015.02.038`.

Chapter 3

How Unequal is Europe? Evidence from Distributional National Accounts

Despite the relevance of Europe as an economic and political entity, it is remarkably hard to know how growth has been shared over the past few decades across its population. This difficulty is not the result of a lack of data *per se.* In fact, there is a fair amount of data available, at least since the 1980s. The problem is that these data are scattered across a variety of sources, taking several forms, using diverse concepts and different methodologies. So we find ourselves with a disparate set of indicators that are not always comparable, are hard to aggregate, provide uneven coverage, and can tell conflicting stories.

As a result, the literature has struggled to answer simple questions such as: which income groups in which countries have benefited the most from European growth? How is European inequality affected by taxes and transfers? Is Europe as a whole more or less equal than the United States? This paper addresses these problems by constructing distributional national accounts for 38 European countries since 1980. While we still face considerable challenges in the construction of good estimates of the income distribution in some countries, we believe that our new series present major improvements over existing ones.

First, our estimates combine virtually all the existing data on the income distribution of European countries in a consistent way. That includes, first and foremost, surveys, national accounts, and tax data. It also includes additional databases on social contribution schedules, social benefits by function, and government spending on health that have been compiled by several institutions over the years (OECD, Eurostat, WHO). Our methodology exploits the strengths of each data source to correct for the weaknesses of the others. It avoids the kind of systematic errors that would arise from the comparison of different income concepts, different statistical units and different methodologies. As such, our estimates are meant to reflect the best of our current knowledge on what has been the evolution of inequality in Europe.

Second, in line with the logic of distributional national accounts (DINA), we distribute the entirety of national income. This includes money that never explicitly shows up on anyone's bank account, such as imputed rents, production taxes or the retained earnings of corporations, yet can account for a significant share of the income recorded in national accounts and official publications of macroeconomic growth. Therefore, our results are consistent with macroeconomic totals, and provide a comprehensive picture of how income accrues to individuals, both before and after government redistribution. Using a broad definition of income makes our results less sensitive to various legislative changes, and therefore more comparable both over time and between countries.

Third, rather than focusing on a handful of indicators, we cover the entire distribution from the bottom to the top 0.001% — which we can capture thanks to tax data. Therefore, we can aggregate our distributions at different regional levels, and analyze the structure of inequality in great details. We can, furthermore, use our estimates to compute any set of synthetic indicators in a consistent way, such as top and bottom income shares, poverty rates or Gini coefficients.

Our results are as follows. In terms of inequalities between countries, we do not observe a clear pattern of convergence in average income levels since the early 1980s. Per adult income in Eastern Europe was about 35% lower than the European average in 2017. This was the same value as in the early 1980s, before the fall of the USSR. In Southern European countries, per adult average incomes have been declining relatively to the continental average since the 1990s and were 10% below the average in 2017. Northern European countries were 25% richer than the average in the mid-1990s and ended up 50% richer.

Inequalities have been increasing in nearly all European countries, both at the bottom and at the top of the distribution. Nearly all European countries failed to reach the United Nations Sustainable Development Goals inequality target over the 1980-2017 period, which seeks to ensure that the bottom 40% of the population grows faster than the average. Since the 2000s, European countries have been relatively more successful at ensuring that bottom income groups secure a fair share of growth, but the majority of countries still failed to achieve the UN objective.

As a result of a limited convergence process and rising inequality within countries, Europeans are more unequal today than they were four decades ago. Between 1980 and 2017, the top 1% grew more than two times faster and captured as much growth as the bottom 50%. The share of national income captured by the richest 10% Europeans increased from 30% to 36% between 1980 and 2017.

Despite rising inequality in Europe and in the European Union, European countries have been much more successful at promoting inclusive growth than the United States. This is largely because European countries succeeded in generating much higher growth rates for low-income groups. The average pre-tax income of the poorest half of the European adult population was 35% higher in 2017 than in 1980, while it was essentially the same as in 1980 for the poorest 50% of US citizens. Consequently, Europe was much less unequal than the US, despite higher inequalities between European countries than between US states.

In the online appendix (`https://wid.world/europe2019`), we provide detailed information on data sources, methodological steps and key results for all the countries and European regions covered in this paper. Detailed inequality series covering the distributions of pre-tax and post-tax incomes can be downloaded from the website of the World Inequality Database (`https://wid.world`).

3.1 Related Literature

This paper contributes to the growing literature combining distributional data with national accounts to measure income and wealth inequalities. Following the seminal contributions of Piketty (2003) and Piketty and Saez (2003), who used income tax tabulations to study the evolution of top incomes in France and the United States in the course of the twentieth century, a new body of research has combined income tax returns and Pareto interpolation techniques to compute estimates of top income shares in a number of countries (see Atkinson and Piketty, 2007; Atkinson and Piketty, 2010 for a global perspective). This area of study has provided a number of insights into the long-run evolution of inequality. However, top income shares tend to rely on country-specific definitions of taxable income and tax units, and only cover a small fraction of the population (generally the top 10% or top 1%). Fiscal income also diverges from the national income, due to the existence of tax exempt income components, and is therefore inconsistent with macroeconomic growth figures.

The increasing availability of tax data has also shed light on the limitations of household surveys, which are traditionally used by statistical institutes and researchers to measure the distribution of income. Surveys remain an invaluable source of information to measure income inequality. However, they tend to underestimate the incomes of top earners, because of small sample sizes (Taleb and Douady, 2015), and because the rich are less likely to answer surveys (Korinek, Mistiaen, and Ravallion, 2006) and more likely to underreport their income (e.g. Cristia and Schwabish, 2009; Paulus, 2015; Angel, Heuberger, and Lamei, 2017). These issues can have serious consequences on estimates of income inequality, both in terms of comparisons between countries and comparisons over time. This is particularly problematic in Europe and the United States, where the rise of income inequality in past decades has been concentrated at the very top of the distribution (Atkinson and Piketty, 2010; Alvaredo, Chancel, et al., 2018).

Recent studies have attempted to overcome these issues by combining surveys or tax data with national accounts to produce more reliable measures of the distribution of income. Statistical institutes and international organizations have increasingly recognized the need to bridge the micro-macro gap. Since 2011, an expert group on the Distribution of National Accounts mandated by the OECD has been working on methods to allocate gross disposable household income to income quintiles (Fesseau and Mattonetti, 2013; Zwijnenburg, Bournot, and Giovannelli, 2019). In a similar fashion, experimental statistics on the distribution of personal income and wealth have been recently published by Eurostat (2018), Statistics Netherlands (2014), Statistics Canada (2019) and the Australian Bureau of Statistics (2019). These exercises have improved upon traditional survey-based estimates, but do not make systematic use of tax data and are restricted to the household sector. This can make estimates of inequality sensitive to the tax base in ways that are not economically meaningful, since firms can have differential incentives to distribute dividends or accumulate retained earnings depending on local tax legislation. Piketty, Saez, and Zucman (2018) were among the first to allocate all components of the US national income to individuals based on tax microdata and explicit assumptions about the distribution of tax exempt income. Several research works have since then followed a similar methodology to extend the distributional national accounts (DINA) approach to other countries.[1] This is the framework that we adopt in this paper: that is, we combine data from surveys, income tax returns and national accounts to estimate the distribution of the national income in thirty-eight European countries between

[1] A comprehensive discussion of the DINA methodology is presented in Alvaredo, Atkinson, et al. (2016). Recent studies following the DINA approach include Morgan (2017) for Brazil and Jenmana (2018) for Thailand.

1980 and 2017. With the exception of France, where extensive work has now been conducted on the distributions of both pre-tax income (Garbinti, Goupille-Lebret, and Piketty, 2018) and post-tax income (Bozio et al., 2018), our study is to the best of our knowledge the first to estimate DINA series for European countries.

This paper directly contributes to the existing literature on the evolution of income inequality in European economies and in Europe as a whole. It has generally been acknowledged that Europe has not been spared by the rise in income disparities visible in the developed world since the beginning of the 1980s (OECD, 2008; Atkinson and Piketty, 2010). However, because a variety of sources and methodologies have been used to measure inequality in Europe, it remains remarkably difficult to study how growth has been shared across its population. Given the relevance of Europe as an increasingly integrated world region, there is a need to go beyond country-specific studies and study the European-wide distribution of income. Recent contributions such as Filauro (2018) and Brandolini and Rosolia (2019) have made advances in tackling this question by using harmonised data from the European Statistics on Income and Living Conditions (EU-SILC) to study income inequality in the European Union as a whole, but they do not address the potential under-representation of top incomes in surveys and only cover the 2004-2016 period. The comparison of the long-run distribution of economic growth across European countries is another area of study where much remains to be done. The effort made by the Luxembourg Income Study to harmonize surveys for a number of Western European countries has been hugely helpful in improving the comparability of pre-2000 inequality statistics in Europe, but surveys (because of sampling issues and misreporting at the top of the distribution) can reveal inequality trajectories which are inconsistent with those suggested by top income shares. The same limitations apply to Eastern Europe: the historical survey tabulations studied by Milanovic (1998), the EU-SILC surveys now conducted in new EU member countries and the top income shares recently estimated from income tax data (e.g. Novokmet, Piketty, and Zucman, 2018; Bukowski and Novokmet, 2019) are based on different income concepts and are therefore hard to compare.

Another question which has received much attention in recent years is that of the comparison between Europe and the United States. While it is acknowledged that post-tax income inequality is greater in the US than in most European countries today, it remains unclear whether this was also the case in past decades and whether this gap is due to differences in pre-tax inequality or to differences in the fiscal incidence of government redistribution. By following the distributional national accounts methodology, Bozio et al. (2018) find that government redistribution reduces inequality less in France than in the United States. This result contrasts with other existing studies (e.g. Jesuit and Mahler, 2010; Immervoll and Richardson, 2011; Guillaud, Olckers, and Zemmour, 2019) which rely on household surveys and restrict their analysis to direct taxes and transfers. Whether the US are more unequal than Europe as a whole also remains an open question. Seminal work on the distribution of income in the EU-15 (Atkinson, 1996) or the Eurozone (Beblo and Knaus, 2001) suggested that income inequality was higher in the US, but recent studies extending the analysis to new, poorer Eastern European member states have found mixed results (e.g. Brandolini, 2006; Dauderstädt and Keltek, 2011; Salverda, 2017; Filauro and Parolin, 2018). One of the limits of existing studies is that they are based on surveys. This may bias comparisons of income inequality between European countries and between Europe and the US if surveys capture more accurately top incomes in some countries than in others. The top-coding of incomes in the public-use samples of the US Current Population Survey, for instance, contrasts with the use of administrative data to fill in survey income components in several European countries, leading to important differences in the quality of survey-based inequality estimates.[2]

This article differs from existing studies on the distribution of income in Europe in a number of ways. First, we go beyond the available survey microdata by collecting and harmonizing a rich dataset of historical survey tabulations. This allows us to go back in time and consistently study the long-run evolution of income inequality in the large majority of European countries from the 1980s until today. Secondly, we use all available studies on the evolution of top income shares, as well as previously unused tax data sources, to correct for the under-representation of high-income earners. Thirdly, we allocate all components of the national income to individuals, including tax exempt income, production taxes and collective government expenditure. This allows us to analyse the distribution of macroeconomic growth in Europe and the effects of different forms of redistribution on inequality.

Methodologically, our approach also departs from existing distributional national accounts studies in the way we combine different available data sources. Piketty, Saez, and Zucman (2018) and Garbinti, Goupille-Lebret, and Piketty (2018) start

[2] Atkinson, Piketty, and Saez (2011) show for example that the CPS top 1 percent share effectively misses 10.4 points of the surge of the top 1 percent income share relative to income tax data in the United States between 1976 and 2006. In the Luxembourg Income Study — one of the most widely used source for comparative work on inequality — the top 1% share of household disposable income is 7% for the US in 2016. Using tax data, Piketty, Saez, and Zucman (2018) find a share of more than 15% for a comparable income concept. By contrast, as we show in the online appendix, the evolution of top incomes is relatively well approximated by EU-SILC data in Nordic countries.

with tax data, to which they progressively add information from surveys and national accounts. This "top-down" approach exploits all the richness of the tax microdata and yields extremely detailed and precise estimates. However, while this type of work can be and should be extended to as many European countries as possible, there are many countries and time periods for which tax microdata are simply not available. This justifies our "bottom-up" approach, which starts from surveys and gradually incorporates information from top income shares and unreported national income components. As such, we view our methodology as well-suited to estimating the distribution of the national income in countries gathering a mix of survey microdata, tabulated tax returns and a variety of heterogeneous historical data sources.[3]

3.2 Conceptual Framework

We study the distribution of the national incomes of thirty-eight European countries, spanning from Portugal to Cyprus and from Iceland to Malta, between 1980 and 2017. Our geographical area of interest includes the twenty-eight members of the European Union, five candidate countries (Bosnia and Herzegovina, Serbia, Montenegro, Macedonia, Albania), and five countries which are not part of the EU but have maintained tight relationships with it (Iceland, Norway, Switzerland, Kosovo and Moldova).

We follows as closely as possible the principles of the DINA guidelines (Alvaredo, Atkinson, et al., 2016), which we briefly outline below. This allows us to be comparable with existing studies, including Piketty, Saez, and Zucman (2018) for the United States.

3.2.1 Macroeconomic Concepts

Net National Income Our preferred measure to compare income levels between countries and over time is the net national income. It is equal to gross domestic product (GDP) net of capital depreciation, plus net foreign income received from abroad. While GDP figures are most often discussed by academics and the general

[3] In a similar fashion, Piketty, Saez, and Zucman (2019) have recently proposed a simplified method for recovering estimates of top pre-tax national income shares based on the fiscal income shares of Piketty and Saez (2003) and very basic assumptions on the distribution of untaxed labour and capital income components. Our methodology can also be viewed as a "simplified" approach to produce DINA estimates, but we stress that the type of data at our diposal differs, and therefore so does our methodology. As we show in section 3.3, we are able to reproduce very closely the results of Garbinti, Goupille-Lebret, and Piketty (2018) for France by combining their top fiscal income shares with available surveys and national accounts data.

public, we believe national income to be more meaningful, since capital depreciation is not earned by anyone, while foreign incomes are, on the contrary, received or paid by residents of a given country. While GDP and national income usually follow each other, there are countries where they can diverge. In particular, GDP can be sensitive to assumptions about the localization of production — a notion that can become murky in our globalized age. In countries such as Ireland or Luxembourg, GDP growth in recent years has been coupled with large outflows capital income, a phenomenon usually attributed to tax avoidance by multinational corporations.[4] Because it is an indicator of income rather than production, national income is less sensitive to such issues.[5]

From Survey and Taxable Income to National Income The national income is the sum of the primary incomes of households, corporations, non-profit institutions serving households and the general government. Household income includes the compensation of employees, mixed income and property income, which are generally — though imperfectly — covered by household surveys and tax data. It also includes the imputed rents of owner-occupied dwellings, which are much less often available from traditional sources but nonetheless represent a substantial share of the capital income of households. The primary incomes of other institutional sectors can amount to a fifth of the national income, but do not appear in either surveys or tax data (see figure 3.1a). These mainly consist in the taxes on production received by the general government (net of subsidies) and the retained earnings of financial and non-financial corporations. Taxes on production are a separate component of national income and their distribution can follow several conventions, which we address below. Retained earnings correspond to profits that are kept within the company rather than distributed to shareholders as dividends. This income ultimately increases the wealth of shareholders and therefore represents a source of income to them.[6]

[4] For example, Ireland officially estimated its real GDP growth in 2015 to be +26%. This number stirred controversy, as it is believed to be the sole result of a few large multinational corporations relocating their intangible assets in Ireland for tax purposes.

[5] Net foreign incomes compensate any change in GDP caused by different assumptions about the localization of production.

[6] Several papers have documented the impact of including retained earnings in the United States (Piketty, Saez, and Zucman, 2018), Canada (Wolfson, Veall, and Brooks, 2016), and Chile (Fairfield and Jorratt De Luis, 2016; Atria et al., 2018). In Norway, Alstadsæter et al. (2017) showed that the choice to keep profits within a company or to distribute them is highly dependent on tax incentives, and therefore that failing to include them in estimates of inequality makes top income shares and their composition artificially volatile. Previous work would sometimes include capital gains in their income definition, which indirectly accounts for this type of income. Yet this constitutes a poor proxy, because capital gains are recorded upon realization, rather than when they accrue to individuals. And whether capital gains are realized or not depends on their value and on tax incentives. Therefore, attributing retained earnings to individuals directly is more reliable,

3.2.2 Income Distribution Concepts

The DINA framework acknowledges three levels of distribution, called *factor income*, *pre-tax income* and *post-tax income*. Factor income is the income that accrues to individuals as a result of their labor or their capital, before any type of redistribution, be it through social insurance or social assistance schemes. Pre-tax income corresponds to income after the operation of the social insurance system (pension and unemployment), but before other types of redistribution. It is closest — though better harmonized and conceptually broader — to the "taxable income" in most countries. Finally, post-tax income accounts for all the redistribution of income operated by the government.

In this paper, we will mostly focus on pre-tax and post-tax income. Factor income is harder to compute given the type of data at our disposal, and also less meaningful. Indeed, many retirees have near zero factor income by construction, so that measures of factor income inequality are highly sensitive to the population structure of the countries.

Factor Income On the labor side, factor income includes the entire compensation that firms pay to their employees, including social contributions paid by employee or employers, and mixed income. On the capital side, it includes the property income distributed to households, the imputed rents for owner-occupiers, and the primary income of the corporate sector (i.e. undistributed profits). We attribute undistributed profits belong to the owners of the corresponding corporations, since it increases the value of their shares, and therefore their wealth.[7] Factor income also includes the primary income of the government, which essentially corresponds to taxes on products and production, minus the interests that the government pays on its debt. Following the DINA standard, we assume that these taxes are paid proportionally to income, but we also experiment with alternative assumptions. And we distribute interest payments of the government proportionally to income.[8]

From Factor to Pre-tax Income Pre-tax income correspond to factor income, to which we add social insurance benefits in the form of unemployment and pension,

more meaningful, more consistent with macroeconomic measures of income, and more comparable across countries.

[7] Their inclusion can be viewed as a way to some capture capital gains as they accrue to individuals rather than upon realization.

[8] Interest payments on government debt have no aggregate effect on national income because it represents a transfer from the government to households, but it does have a second-order distributional effect because ownership of government bonds is usually more concentrated than income.

and from which we remove the social contributions that pay for them. Note that for pre-tax income to sum up to national income, it is important to remove the same amount of social contributions as the amount of social benefits that we distribute. This way, we both avoid double counting and ensure that we look at the redistribution from social insurance in a way that is budget-balanced. In doing so, we observe significant heterogeneity between countries. In most countries, social contributions exceed pension and unemployment benefits, because social contributions also pay for health or family-related benefits that we classify as non insurance-based redistribution. Therefore, we only deduct a fraction of social contributions from pre-tax income. But in some countries, like Denmark, social contributions are virtually non-existent. In these cases, we have to assume that social insurance is financed by the income tax, and therefore deduct a fraction of the income tax from factor income to get to pre-tax income.

From Pre-tax to Post-tax income To move from pre-tax to post tax income, we first remove all taxes and social contributions that remain to be paid by individuals. This includes the taxes on products and production that we previously added, and also the corporate tax that was added through undistributed profits. Then we add all types of government transfers, and government consumption. We distribute all of government consumption proportionally to income, with the exception of public health expenditures. We use the proportionality assumption for simplicity, transparency and comparability with earlier work on distributional national accounts, in particular in the United States (Piketty, Saez, and Zucman, 2018). But we consider that it is important to make an exception for health spending. Indeed, while many European countries have public health insurance systems, the United States have a mostly private one, with some public programs such as Medicaid and Medicare, which are explicitly distributed to their recipients in the United States distributional national accounts. Therefore, distributing health spending proportionally to income could understate the amount of redistribution that European countries engage in. For other types of spending (education, military, police, etc.), we experiment with alternative assumptions, but we still use the proportionality assumption as our benchmark. We distribute the net saving of the government (the discrepancy between what the government collects in taxes and what its pays as transfer, consumption or interest) proportionaly to the income of individuals so that post-tax national income matches national income.

Unit of Analysis In our benchmark series, the statistical unit is the adult individual (defined as being 20 or older) and income is split equally among spouses, in line with other existing DINA studies (e.g. Piketty, Saez, and Zucman, 2018; Garbinti, Goupille-Lebret, and Piketty, 2018).[9]

3.3 Sources and Methodology

This section describes the main steps followed to estimate the distribution of the national incomes of European countries. We refer to the appendix C.1 for technical details on the methodology, and to the online appendix for a more detailed account of data sources, methodological steps and robustness checks. In broad strokes, our methodology starts from a variety of household surveys. We harmonized them and correct them using tax data. Finally, we account for the various parts of national of income that are absent from the usual sources. Add the European level, figure C.1 in appendix show the role that the various steps play in the final series. Most of the difference between raw survey estimates and our series come from the inclusion of tax data.

3.3.1 National Accounts

Main Aggregates For total national national income, we use series compiled by the World Inequality Database based on data from national statistical institutes, macroeconomic tables from the United Nations System of National Accounts and other historical sources (see Blanchet and Chancel, 2016). For the various components of national income, we collect national accounts data from the Eurostat, the OECD, and the UN. We use Eurostat and the OECD in priority, as they tend to have the most most reliable data, but their coverage is less extensive than the UN.[10] We provide a detailed view of the coverage that these data provide in our extended appendix.[11]

[9] We also compute additional series in which income is split between all adult household members, not just members of a couple (i.e. a "broad" rather than a "narrow" equal-split). The difference is not entirely negligible in certain Southern and Eastern European countries. Until now DINA studies have has a tendency to use the narrow equal-split developed countries (e.g. Piketty, Saez, and Zucman, 2018; Garbinti, Goupille-Lebret, and Piketty, 2018) and the broad equal-split in less developed ones (e.g. Novokmet, Piketty, and Zucman, 2018; Piketty, Yang, and Zucman, 2019). We focus on the narrow equal-split in our benchmark for comparability with the United States, but also shed some light on the issue by providing both concepts. See figure C.4 in appendix.

[10] We link together the various series, rescaling older and lower-quality series to match the newer and higher-quality ones in their latest year of overlap to avoid any structural break.

[11] Using these sources, we have a sufficiently detailed decomposition of national income that covers nearly 100% of the continent national income up until 1995. Before that, coverage becomes increasingly sparser: we have the full decomposition for about 50% of national income in 1990, decreasing to 20% at the very beginning of our series. We impute missing series by retropolating them using exponential smoothing with a coefficient of 0.9. As a last resort, we rely of regional

Additional Sources In a few cases, we need to rely on additional sources to perform decomposition of national income that are needed for our series and more precise than what is available through standard data portals. First, when we distribute the retained earnings of the corporate sector, we have to separate the share that is owned by private citizens from the share that belong to the governments. To do so, we use the fractions of equities owned by the households sector in the financial balance sheets available from the OECD. We also need to separate the social benefits that correspond to pension and unemployment from the other types of social benefits in order to calculate pre-tax income.[12] To do so, we rely on the OECD social expenditure database, which breaks down social benefits by function in great details since 1980. Finally, we need to separate health expenditures from the rest in the individual consumption of the government. For that, we extract health government spending from the System of Health Accounts, a database that emerged from a joint work between the OECD, Eurostat and the WHO.

3.3.2 Survey Microdata

Sources We collect and harmonize household survey data from several international and country-specific datasets. Our most important source of survey data is the European Union Statistics on Income and Living Conditions (EU-SILC), which have been conducted on a yearly basis since 2004 in thirty-two countries. We complement EU-SILC by its predecessor, the European Community Household Panel (ECHP), which covers the 1994-2001 period for thirteen countries in Western Europe. Our second most important source of survey data is the Luxembourg Income Study (LIS) provides access to harmonized survey microdata covering twenty-six countries over the 1975–2014 period. Most Western European countries are covered from 1985 until today, and several countries from Eastern Europe have been surveyed since the 1990s.

Imputations When we have access to survey microdata, we can usually estimate income concepts that are close to our concepts of interest (pre-tax and post-tax income) with only a few components of income that remain to be added separately (see section 3.3.5). A significant exception concerns social contributions in EU-SILC: while both employer and employee social contributions are recorded, employee contributions

averages.

[12] The DINA guidelines (Alvaredo, Atkinson, et al., 2016) recommend using the distinction between social insurance benefits (D621 + D622) and social assistance benefits in cash (D622). Unfortunately that level of details is not commonly available in the national accounts of most countries, which only report the aggregate item D62. This is why we rely on alternative sources.

are combined with income and wealth taxes. We use the social contribution schedules published in the OECD Tax Database to impute employee social contributions separately. Before 2007, employer contributions may also not be recorded despite having information on income before taxes and employee contributions. In such cases, we also impute employer contributions based on schedules from the OECD Tax Database. Beside that, measures of income before and after taxes and transfers have been recorded consistently as part of EU-SILC. The Luxembourg Income Study also produces some historical data on pre-tax income, in many cases by imputing direct taxes and social contributions as part of their harmonization effort. As a result, we have survey microdata on both pre-tax income and post-tax income in almost all countries since 2007, and for over a longer time period for a number of Western European countries (Germany, the United Kingdom, Switzerland, and Nordic countries).

3.3.3 Survey Tabulations

Sources We complement the survey microdata with a number of tabulations available from the World Bank's PovcalNet portal, the World Income Inequality Database (WIID) and other sources. PovcalNet provides pre-calculated survey distributions by percentile of post-tax income or consumption per capita. The WIID gathers inequality estimates obtained from various studies, and gives information on the share of income received by each decile or quintile of the population. Finally, we collect historical survey data on post-tax income inequality in former communist Eastern European countries provided by Milanovic (1998), as well as formerly unused tabulations covering Yugoslav republics from the 1970s to 1989.[13] In all cases, we use generalized Pareto interpolation (Blanchet, Fournier, and Piketty, 2017) to recover complete distributions from the tabulations. A detailed breakdown of available survey data sources by country is available in the online appendix.

Harmonization Contrary to microdata, tabulations only provide distributions covering specific welfare concepts and equivalence scales. The majority of tabulations recorded in PovcalNet and WIID correspond to post-tax income, while cases in which we only observe consumption are limited to a handful of Eastern European countries (Moldova, Kosovo, Montenegro).[14] The equivalence scales available are more diverse,

[13] We are grateful to Branko Milanovic for providing us with these tables.

[14] The only exceptions correspond to a handful of Eastern European countries at the beginning of the period (Bosnia and Herzegovina, Moldova, Montenegro) for which we have no other source available. In these cases we use the survey distribution of pre-tax income as a proxy for the "true" pre-tax income.

including households, adults, individuals, the OECD modified equivalence scale or the square root scale.[15] For these data sources, as well as for survey microdata where information on taxes and transfers is incomplete, we have to develop a strategy to transform the distribution of the observed "source concept" (e.g. consumption per capita or post-tax income among households) into an imputed distribution measured in a "target concept" (pre-tax or post-tax income per adult).

The key idea behind our harmonization procedure is that, while the different income or consumption concepts that we observe are different, they are also related. Using all the cases where the income distribution is simultaneously observed for two different concepts, we can map the way they tend to relate to one another, and use that to convert any source concept to our concept of interest. In practice, we formalize this idea by writing the average income of each percentile for the distribution of interest as a function of all the percentiles of the distribution from which we wish to impute, and also as a function of various auxiliary variables that may potentially account for that relationship (average income, population and household structure, marginal tax rates, social expenditures). Finding that function amounts to a regression problem, albeit a high-dimensional, non-parametric one. To avoid making *ad hoc* restrictions, we rely on a recent advances in non-parametric high dimensional statistics, also known as machine learning. We use XGBoost (Chen and Guestrin, 2016), a state-of-the-art implementation of a standard, robust and high performing algorithm called *boosted regression trees*. We provide a detailed view of the method and the results in the extended appendix.

We stress that this approach is not perfect: the relationships between the different concepts are not deterministic, so that these imputations involve their share of uncertainty. However, the existing literature has often chosen to ignore these issue altogether, and directly combined, say, income and consumption data (e.g. Lakner and Milanovic, 2016). We feel that our approach is preferable, because it corrects at least for what can be corrected. We further provide in our online appendix prediction intervals to give some idea of the amount of precision that the method achieves. Note that in practice, the output of the harmonization procedure is straightforward and intuitive: it mostly adjust the levels of the different series, but does note introduce any trend that was not already in the data.[16]

[15] When computing inequality estimates with the OECD modified equivalence scale, the first adult in the household is given a weight of 1, other adults are given a weight of 0.5, and children are given a weight of 0.3 each. The square root scale divides total income by the square root of the size of the household.

[16] Before the 2000s, only post-tax inequality estimates are available for many countries. In these countries, the trends for post-tax and pre-tax inequality estimates are thus implicitly assumed to

3.3.4 Survey Corrections

Survey data are known to often miss the very rich. For our purpose it is important to distinguish two reasons for that: non-sampling and sampling error. Sampling error refers to problems that arise purely out of the limited sample size of survey data. Low sample sizes affect the variance of estimates, but they may also create biases, especially when measuring inequality at the top of the distribution. Non-sampling error refers to the systematic biases that affect survey estimates in a way that is not directly affected by the sample size. These mostly include people refusing to answer surveys and misreporting their income in ways that are not observed, and therefore not corrected, by the survey producers. Estimates based on raw survey data do not account for any of these biases and therefore tend to underestimate incomes at the top end.

Non-sampling Error We correct survey data for non-sampling error using known top income shares estimated from administrative data. Following contributions by Piketty (2001) for France and Piketty and Saez (2003) for the United States, several authors have been using tax data to study top income inequality in the long run. Most of these studies have been published in two collective volumes (Atkinson and Piketty, 2007; Atkinson and Piketty, 2010), and their results have been compiled in the World Inequality Database.[17] In general, tax data is only reliable for the top of the distribution, and this is why these series do not cover anything below the top 10%. Researchers estimate the share of top income groups by dividing their income in the tax data by a corresponding measure of total income in the national accounts. At the time of writing, data series were available for nineteen European countries, providing information on the share of income received by various groups within the top 10%.

We complete this database by gathering and harmonizing a collection of formerly unused tabulated tax returns covering Austria (2008–2015), East Germany (1970–1988), Estonia (2002–2017), Iceland (1990–2016), Italy (2009–2016), Luxembourg (2010, 2012), Portugal (2005–2016), Romania (2013) and Serbia (2017). We use these tabulations to directly add new top income shares to our database. We provide a

be similar. We view this a reasonable approximation given that the main determinant of post-tax inequality is pre-tax inequality, both between countries and over time (e.g. Guillaud, Olckers, and Zemmour, 2019). This is all the more true that our "pre-tax" income concept includes pension and unemployment, which are the most important forms of government redistribution. In the United States, trends for post-tax and pre-tax income inequality are very similar, with the minor exception of the role played by government health spending (Medicare and Medicaid) (see Piketty, Saez, and Zucman, 2018) which are separately taken into account in our methodology.

[17] See `http://wid.world`.

detailed account of the computations for each country in the online appendix. In most cases, we directly correct the surveys with the tax data using the method of Blanchet, Flores, and Morgan (2018) rather than using a total income estimate from the national accounts. Direct correction of survey data is a more flexible and practical approach, at least for the recent period, and is now being preferred in the latest work on inequality (e.g. Piketty, Yang, and Zucman, 2019; Morgan, 2017; Bukowski and Novokmet, 2017). When extending existing series using that method, as in Italy or Portugal, our results are consistent with the work that was done previously, thus confirming the consistency and reliability of both approaches. Our results also reveal that the underestimation of top incomes varies a lot across surveys and is typically higher in Eastern European countries. This points to the importance of correcting surveys with tax data to make comparisons between countries more reliable.

We correct the survey data using standard survey calibration methods. The principle of survey calibration is to reweight observations in the survey in the least distortive way so as to match some external information. Statistical institutes already routinely apply these methods to ensure survey representativity in terms of age or gender. We directly extend them to also ensure representativity in terms of income. The applicability of these methods to correct for the underrepresentation of the rich in surveys has been discussed at length by Blanchet, Flores, and Morgan (2018).

One difficulty is that our external source of information consist in top income shares. Because top income shares are a non-linear statistic, they cannot directly be used in standard calibration procedures. We tackle that issue using suggestions from Lesage (2009). They involve linearizing top income shares statistics by calculating their influence function, and introducing a nuisance parameter. We discuss that methodology in details in our extended appendix. In concrete terms, we increase weight at the top of the distribution so that survey top incomes match their value observed in the tax data.

One advantage of calibration procedures is that they allow to perform survey correction with a income tax data concept that may differ from the income concept of interest — either in terms of income definition or statistical unit. We always match concepts to the best of our ability between the tax data and the survey data to perform the correction. Then we use income concepts that are better defined and more economically meaningful to produce our inequality series. Confronting tax data and survey data as such is a very powerful way to harmonized income tax statistics between countries.[18]

[18] For older time periods from which we cannot perform that exercise directly due to lack of

When we do not directly observe tax data in a country, we still perform a correction based on the profile of nonresponse that we observe in other countries. This is only the case for a few small countries — Albania, Bosnia and Herzegovina, Bulgaria, Cyprus, Kosovo, Latvia, Lithuania, Macedonia, Malta, Moldova, Montenegro and Slovakia. To capture statistical regularities, we estimate the nonresponse profile as a function of the distribution of income in the uncorrected survey using the same machine learning algorithm as in the previous section. We stress that this remains a rough approximation and that in our view the proper estimation of top income inequality requires access to tax data. Fortunately, our tax data covers the majority of European countries and of the European population, so that the impact of these corrections on our results is very limited.

Sampling Error The sample size of surveys varies a lot and can sometimes be quite low: this, in itself, can seriously affect estimates of inequality at the top and, in general, will underestimate it (Taleb and Douady, 2015). Correcting sampling error requires some sort of statistical modeling. We use methods coming from extreme value theory, which is routinely used in actuarial sciences to estimate the probability of occurrence of very rare events, but can similarly be used to estimate the distribution of income at the very top.

The main tenet of extreme value theory can be understood in analogy to the central limit theorem. According to the central limit theorem, under some regularity assumptions, but regardless of the exact distribution of iid. variables $X_1, \dots, X_n$, the distribution of the sum $\sum_{i=1}^{n} X_i$ as n goes to infinity will belong to a tightly parametrized family of distributions (a Gaussian one). Similarly, under mild regularity assumptions, the distribution of the largest value of the sample $\max(X_1, \dots, X_n)$ as n goes to infinity will belong to a certain parametric family. The same holds for the second-largest value, the third-largest value, and so on. As a result, the top k largest values will approximately follow a distribution known as the generalized Pareto distribution. That result is known as the Pickands–Balkema–de Haan theorem (e.g. Ferreira and Haan, 2006).

The generalized Pareto distribution therefore more or less provides a universal approximation of the distribution of the tails of distributions. It includes the Pareto or the exponential distribution as a special case. We use it to model the top 10% of income distributions. Because the likelihood surface of the generalized Pareto distribution is very flat, maximum likelihood estimation often gives poor results

proper survey microdata, we retropolate the correction on the income tax series that is done over the more recent period.

unless the sample size is very large. The standard method of moments also fails if the distribution has infinite variance, which can often occur with income distributions. We use a simple and robust alternative known as probability-weighted moments (Hosking and Wallis, 1987). We provide technical details for the method in appendix. Note that by construction, this adjustment has absolutely no impact on the top 10% income share, it only refines the income distribution within the top 10%.

3.3.5 Missing Incomes

Once we have harmonized and corrected our survey data using tax data, we find ourselves with more precise and comparable inequality series. But those series do yet account for all of national income because they lack some components from the household sector (imputed rents), the corporate sector (undistributed profits) and the government sector (taxes on products and government spending).[19]

Imputed Rents We extract the total value of imputed rents from the national accounts. To distribute them, we rely on (calibrated) EU-SILC data that does record imputed rents (although they are not included in the headline inequality figures). We perform a simple statistical matching procedure using income as a continuous variable to add imputed rents, which we describe in the appendix. The imputed rents total is rescaled to match national accounts. The method preserves the joint distribution of income and imputed rents in EU-SILC, the distribution of imputed rents in EU-SILC, the distribution of income in the original data, and the imputed rents total in the national accounts.

Undistributed Profits We distribute the private share of undistributed profits to individuals proportionally the ownership of corporate stock. This includes both private and public stocks that are held directly or indirectly through mutual funds and private pension plans. However, we exclude sole proprietorship, since in the national accounts they are not an entity separate from the household to which they belong.

The distribution of stock ownership comes from the Household Finance and Consumption Survey (HFCS), the pan-European wealth survey of the European Central Bank. We calibrate that survey on the top income shares as we do for other surveys to make it representative in terms of income and get consistent results. The HFCS only started around 2013, so before that year we keep the distribution of retained earnings

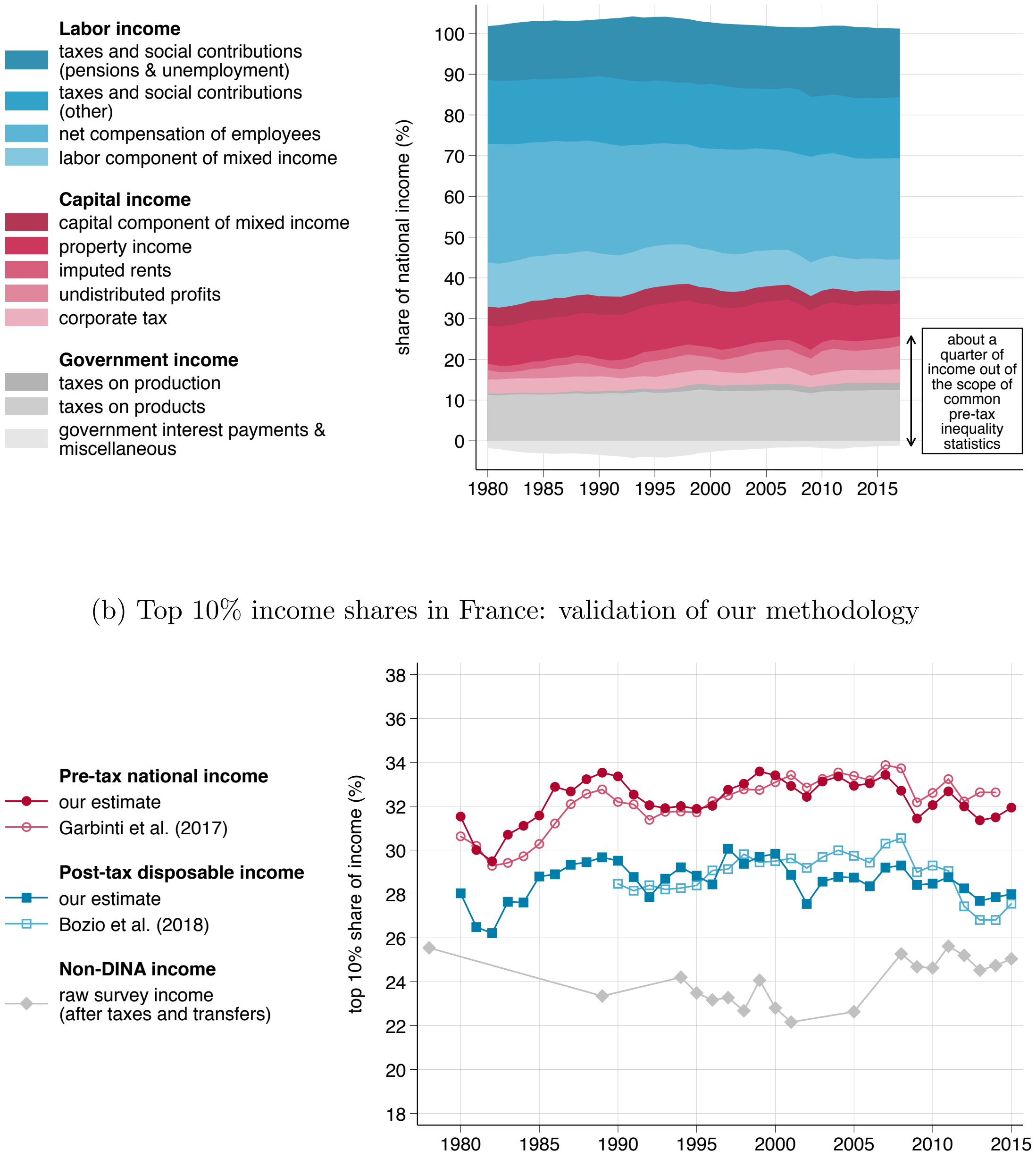
[19] Other missing items (taxes on production, government surplus, etc.) are smaller and less important because we distribute them proportionally.

constant and only change the amount of retained earnings to be distributed: this constitutes a reasonable approximation because stock ownership is always already highly concentrated, so that the main impact of retained earnings on inequality comes from changes in their average amount rather than changes in the inequality of stock ownership. After 2013, we use the wave that is closest to the year under consideration.

By distributing retained earnings proportionally to stock ownership, we assume that profits are similar in every company. To the extent can this assumption bias our results? If richer people keep more money in their companies, then we will underestimate inequality. A good point of reference if the study of Alstadsæter et al. (2017) in Norway. To our knowledge, this is the only study that analyzed the role of undistributed business income on inequality while being to match exactly businesses to their owners. Our approach yields similar results, which give us confidence in its validity.

Taxes on Products In our baseline estimates, we follow the standard DINA guidelines and distributes taxes on products and production proportionally to pre-tax income. We also experiment with an alternative assumption, namely that people pay taxes on products proportionally to their consumption. To that end, we rely on the Household Budget Surveys (HBS) from Eurostat to get the distribution of consumption and its dependency to income. We use the same statistical matching procedure as before to attribute a consumption to people alongside the income distribution, and attribute the taxes on products proportionally to it. As we show in appendix (table C.3), this lowers pre-tax income inequality somewhat, but does not change the trend. (Post-tax income inequality is not affected by production taxes.)

Government Expenditures In post-tax income, we directly distribute health expenditures lump-sum and other expenditures proportionally.[20] In appendix (see table C.5), we experiment with alternative assumption (full proportional allocation and full lump-sum). This changes levels of post-tax inequality, but the trends are similar.



(a) Composition of European national income, 1980-2017

(b) Top 10% income shares in France: validation of our methodology

The figure compares the evolution of income inequality as measured by raw surveys, by our methodology, and by complete DINA studies in France.

Figure 3.1: DINA methodology

3.3.6 External validation

Many existing DINA studies (Piketty, Saez, and Zucman, 2018; Garbinti, Goupille-Lebret, and Piketty, 2018; Bozio et al., 2018) rely on detailed tax microdata and microsimulation models. In comparison, our methodology relies on much sparser data. To what extent do we get comparable results despite having less data? Figure 3.1b compares our results with those of Garbinti, Goupille-Lebret, and Piketty (2018) and Bozio et al. (2018) for pre-tax and post-tax income inequality in France. As we can see, there is a strong agreement between both methods (results for the bottom 50% are along the same lines, see table C.2 in appendix.) Both find a overall tendency that is quite at odds with what the raw data (in grey) suggest. This gives us confidence that our method gives estimates that are comparable to more detailed DINA studies.

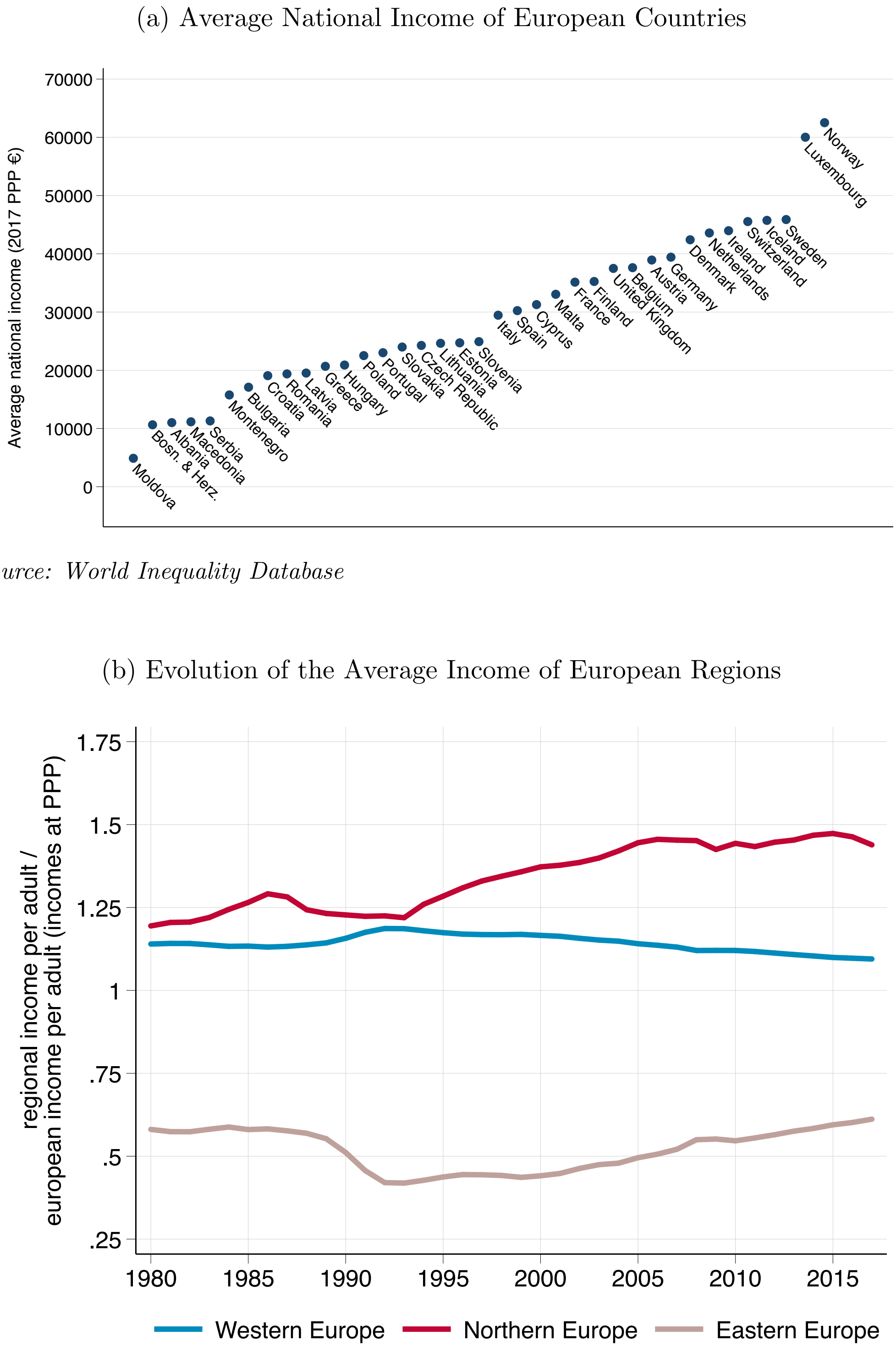
Note that we obtain these results in spite of the fact that our data for France are not of an especially high quality. The SILC statistics for France are a transcription of a survey (called SRCV) which is used for its extensive set of questions on material poverty, but is not considered the best survey for income inequality. For that purpose, the French statistical institute relies on another survey, called ERFS. But that survey is not part of any international scheme, such as EU-SILC, nor is it available through portals such as the Luxembourg Income Study. Therefore, we do not include it in our estimations. Before SILC is available, we rely on France's Household Budget Survey, which has been made available through LIS. While France's HBS is a key source for consumption data, it is not viewed a the best source for income data either. Therefore, there is no reason to think that our methodology would work better for France than other countries just because of the quality of the data in input.

3.4 Results

3.4.1 Inequality between European Countries

Before looking at inequalities within European countries and within wider regional entities (such as the European Union), it is worth having in mind how differences in countries' average national incomes have evolved between 1980 and 2017. As new countries joined the EU and further political integration was enhanced by policy makers in the 1990s and 2000s, convergence in standards of living gradually became part of the European Union agenda, along with the harmonization of economic

[20] To extent that the health risk profile is the same for everyone, and that health spending is actuarially fair, distributing health expenditures lump-sum properly captures the insurance value of government spending on health.



(a) Average National Income of European Countries

Source: World Inequality Database

(b) Evolution of the Average Income of European Regions

Source: authors' computations using data from the World Inequality Database. *Interpretation*: between 1980 and 2017, the average income of a Western European citizen remained about 20% higher than that of an average European.

Figure 3.2: Inequality Between Countries in Europe

policies. One of the explicit objectives of European integration, in particular, was to reduce average income gaps between EU Member States. The Lisbon Treaty, one of the legal basis of the EU, states that "the EU shall promote economic, social and territorial cohesion, and solidarity among Member States."[21]

In 2017, we can see important differences in standards of living between European countries were visible, but relatively homogeneous levels among the largest member states of the European Union (figure 3.2a). In most of the Balkan countries, per adult national incomes were below €15,000, while Southern and other Eastern European countries earned between €15,000 and €30,000. In most other EU countries, incomes ranged between €30,000 and €45,000. Luxembourg and Norway, finally, stood out with average national incomes higher than €60,000. Based on these differences as well as geographical proximity, we propose to divide Europe into three broad regions in the rest of this paper: Northern Europe, Western Europe, and Eastern Europe. Northern Europe includes Nordic countries spanning from Denmark to Iceland. Eastern Europe includes other countries located east from Austria, and Western Europe encompasses the remaining countries (see table C.3 for a full list of these countries and the evolution of their national incomes per adult).

Regional growth trajectories in the past forty years do not show a rapid equalization of absolute income levels (figure 3.2b). In Eastern Europe, sustained economic growth since the early 2000s has succeeded in bringing back the income levels that existed during the communist era and which dramatically fell after the dislocation of the USSR, but Eastern European citizens still earn about 40% less than the average European. Meanwhile, Western European nations have steadily been characterized by national incomes higher by 15–20% on average, Scandinavian countries have consolidated their positions at the top of the European distribution, experiencing high growth rates since the mid-1990s.

Looking more precisely at country-specific trajectories reveals a relatively complex picture, with no sign of long-run monotonic convergence. Between 1980 and 1989, national income growth was slightly higher in countries with standards of living closer to the European average — such as Finland, the United Kingdom, Slovenia and Sweden — while the rest of Europe saw annual growth rates of about 1% throughout the decade. Following the disintegration of the Soviet Union, former communist countries characterized by lower standards of living than Western European nations

[21] Article 3 of the Lisbon Treaty. Inequality reduction between Member States is also made clear in the Treaty on the Functioning of the European Union. Article 174, for instance, states that "the Union shall aim at reducing disparities between the levels of development of the various regions and the backwardness of the least favored regions."

experienced strong recessions, mirrored by negative growth rates (-1.7% per year on average) for the poorest 10% countries of the old continent. The 2000–2007 period, on the other hand, came with restored stability and revived economic growth for Central and Eastern European countries, which led to a moderate reduction in between-country inequalities. Finally, the fact that Spain, Italy, Portugal and Greece were strongly affected by the 2007–2008 crisis translated into negative growth rates at the middle of the European distribution during this period.

The analysis of income inequalities between countries therefore points to the importance of both macro-historical events and country-specific trajectories. The economic downturns in Eastern Europe which followed the collapse of the USSR, as well as macroeconomic imbalances exacerbated in Western Europe since 2008 have strongly affected national and regional growth trajectories. Yet, because European countries have been affected very differently by these crises, their overall effect on income differences between nations has remained unclear.

Did European integration contribute to decreasing inequality between member States? Unsurprisingly, European integration in itself has been associated with a gradual *widening* of income differences between EU members. This is the mechanical consequence of an integration process in which new member States have increasingly more diverse income levels. The integration of Spain and Portugal in 1986, both slightly poorer than EU-10 members, as well as the inclusion of Sweden and Finland in 1995 led to a slight increase in between-country inequalities at the EU level. As former communist countries joined the European community in 2004, 2007 and 2013, these differences became even wider. Thanks to new access to the common market, technological catch-up, economic reforms and EU cohesion policies, however, it is expected that new Member States catch up with the rest of the EU. Income growth rates of Eastern European countries which joined the EU after 2004 grew at a much faster rate than EU-15 countries.

This picture should, however, be interpreted cautiously. First, despite significantly higher growth rates, income levels in Eastern European countries remain significantly below that of EU-15 countries and at a relatively similar level to that of the early 1980s, before the collapse of Eastern European economies.[22] Second, since 2008, the growth differential between EU-15 and Eastern European Union countries is partly due to sluggish post-crisis growth in the EU-15. A large part of the high Eastern European growth is also related to economic recovery after the collapse

[22] In 2017, the average income of Eastern European Union countries was equal to 62% of EU-15 average income. This value was 54% in 1980.

of Eastern European economies in the early 1990s (up to the late 2000s, non-EU Eastern countries also caught up rapidly with EU-15 members).

3.4.2 Inequality within European Countries

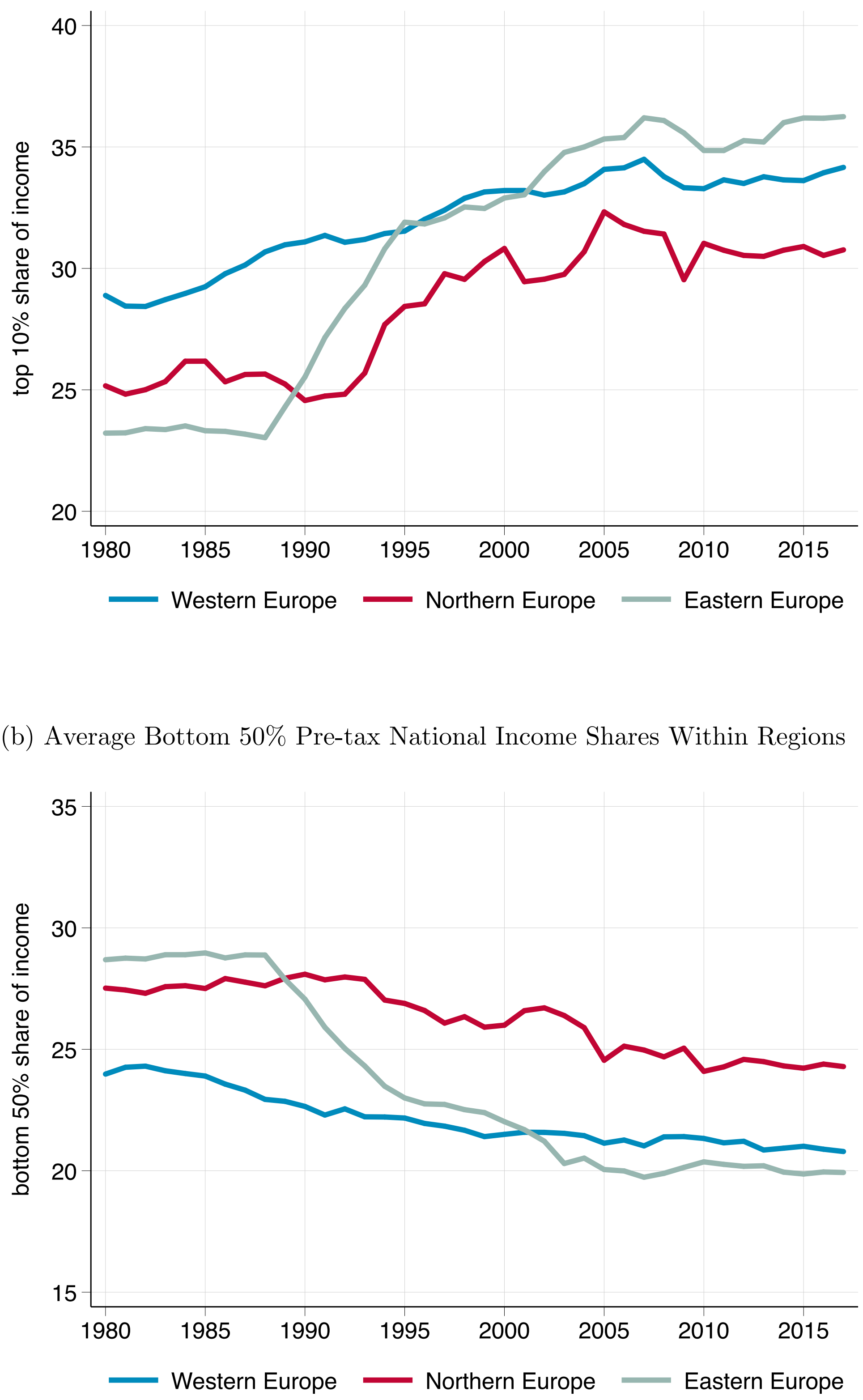
We now turn to the analysis of income inequalities within European countries. How did European countries perform in curbing inequality and promoting inclusive growth over the past decades? Beyond country-specific trajectories and short-run dynamics, it is possible to identify a set of stylized facts.

First, in a large majority countries where data is available since 1980, top earners have captured an increasing share of national income. If one looks precisely at average inequality levels among European regions, differences in trajectories between our three regions of interest are identifiable (figures 3.3a and 3.3b). In Northern Europe, inequality has increased during the 1990s. In Western Europe, the increase has been more linear. But Eastern Europe is the area where inequalities have risen the most, especially at the top of the distribution during the 1990s and the early 2000s, as Eastern European countries transitioned from communism to capitalism.[23] Today, pre-tax income inequality remains, on average, slightly lower in Northern Europe than in other regions of the continent, even if these differences should not be exaggerated.

While common trends are visible in broad European regions, there are also country-specific trajectories (figure 3.4). Germany, France and the United Kingdom, who together represent 80% of the adult population of Western Europe in 2017, all witnessed increasing inequalities at the top of the distribution. In the United Kingdom, the top 10% share increased from 1980 to the 2007-2008 crisis, while it mainly rose in Germany in the 2000s and remained more stable in France over the period. In Northern Europe, income inequality increased mainly during the 1990s.

Eastern Europe, finally, is clearly the region where inequalities within countries have risen most. Poland, the Czech Republic, Hungary, Romania and Bulgaria all went through important political and structural economic changes in the 1990s as they transitioned to market economies. At the beginning of the 1980s, Eastern European countries were among the least unequal of the continent; by 2000, they had caught up with Southern European inequality levels. Poland is the country where income

[23] It is important to stress here that we focus solely on monetary income inequality, which was unusually low in Russia and Eastern Europe under communism. Other forms of inequality prevalent at the time, in terms of access to public services or consumption of other forms of in-kind benefits, may have enabled local elites to enjoy much higher standards of living than what their income levels suggest.

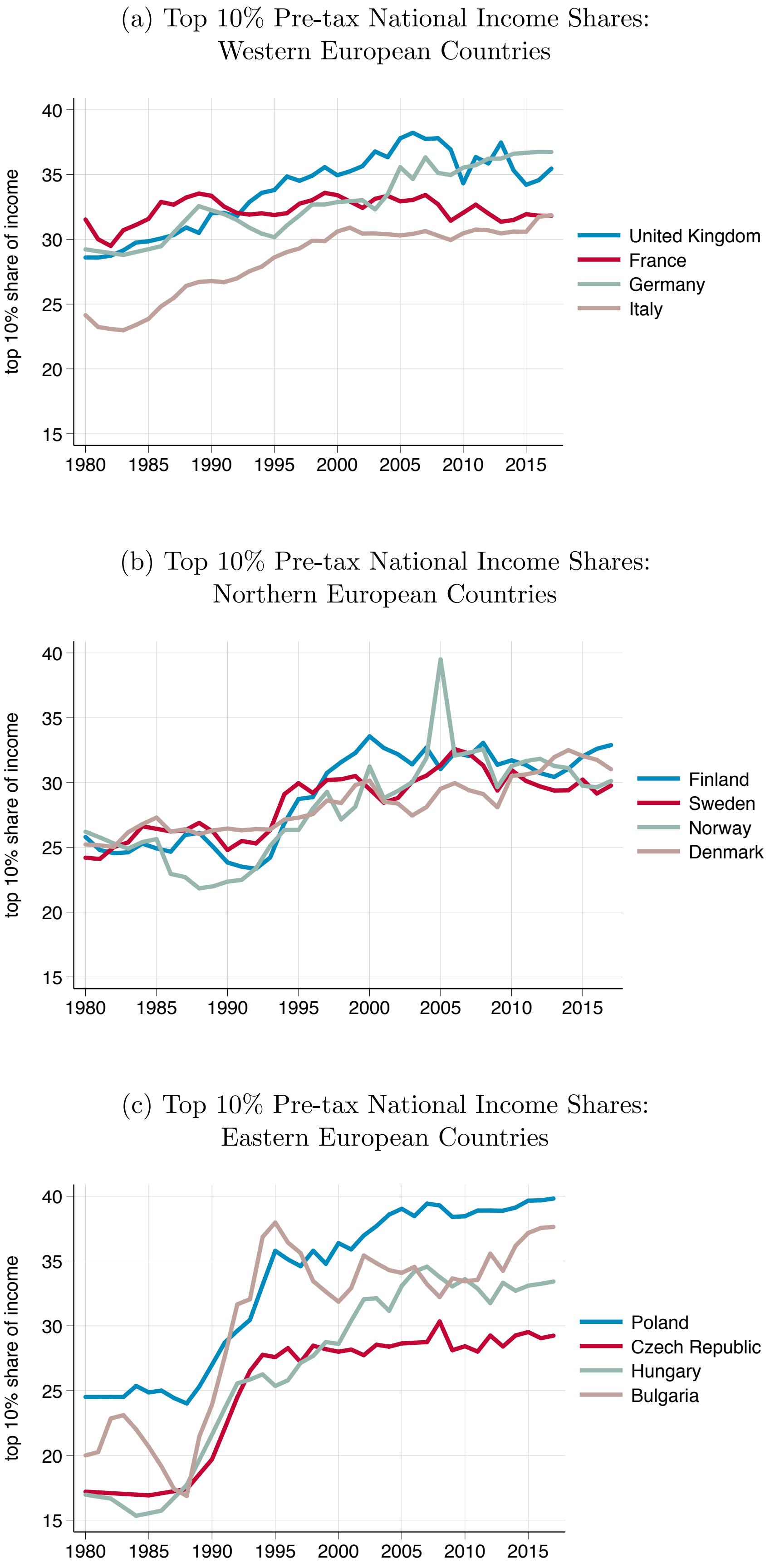


(a) Average Top 10% Pre-tax National Income Shares Within Regions

(b) Average Bottom 50% Pre-tax National Income Shares Within Regions

Source: authors' computations combining surveys, tax data and national accounts. Figures correspond to population-weighed country averages in the regions considered.

Figure 3.3: Inequality Within Countries in European Regions



(a) Top 10% Pre-tax National Income Shares: Western European Countries

(b) Top 10% Pre-tax National Income Shares: Northern European Countries

(c) Top 10% Pre-tax National Income Shares: Eastern European Countries

Source: authors' computations combining surveys, tax data and national accounts.

Figure 3.4: Inequality Within Selected European Countries

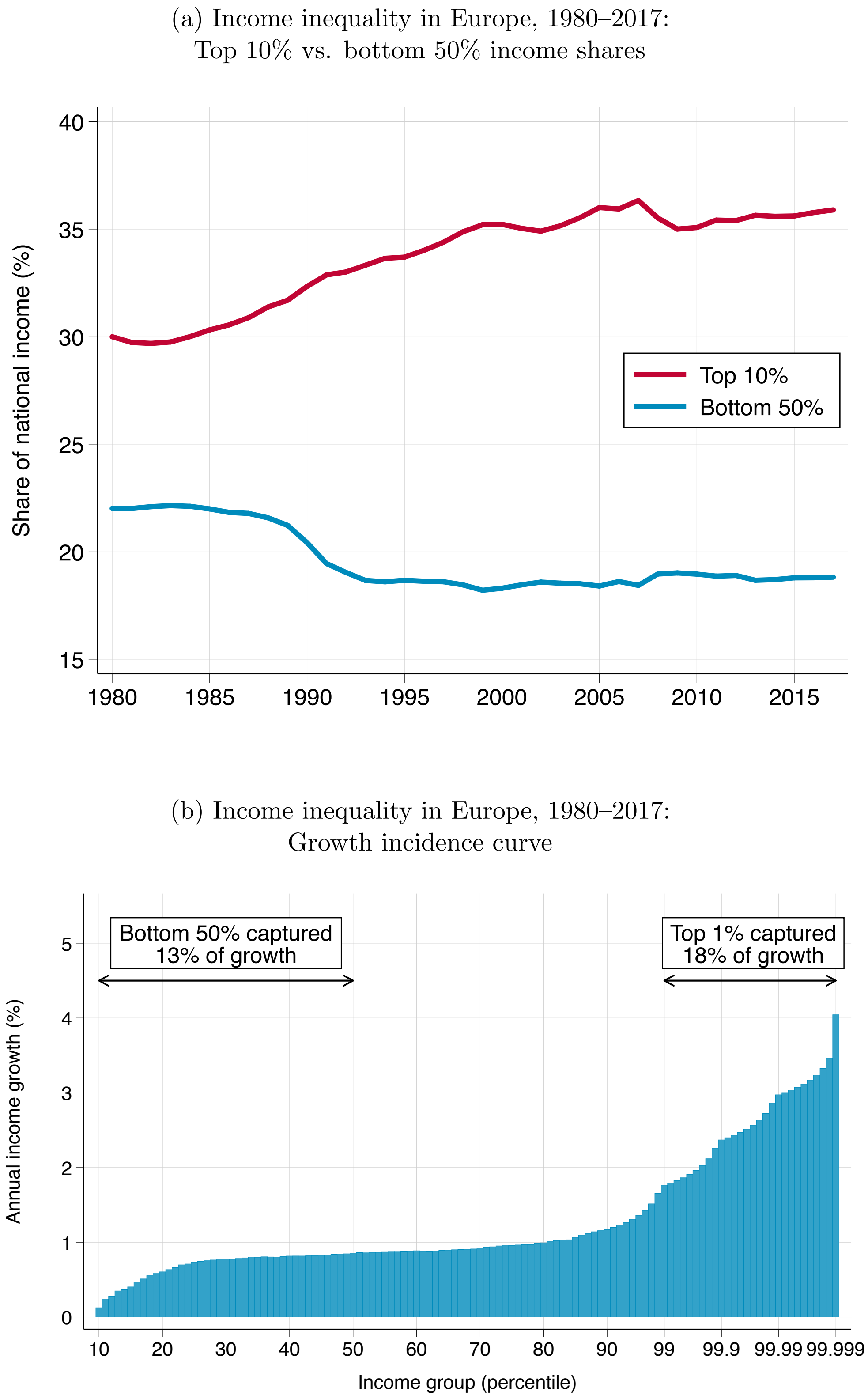
disparities rose most, in part because they continued to rise in the 2000s and 2010s while they more or less stabilized in the rest of the region. In 2017, top 10% Polish earners received nearly 40% of national income, more than any of their counterparts in other European countries.

3.4.3 Inequality between European Citizens

Having discussed the evolution of income inequality between and within European countries, we now look at income inequality in Europe as a whole. The level and evolution of inequality between European citizens depend upon three factors: the evolution of income inequalities between European countries, the evolution of inequalities within countries, and the relative weights of countries' populations. In this section, we measure European-wide income inequalities at purchasing power parities to account for differences in average costs of living between European countries. When comparing Europe to the US, however, we will adopt market exchange rates estimates to make results between the two regions more comparable — since PPP conversion factors exist for European countries but not for US states.

Income differences between European residents have increased in the past forty years (figure 3.5a). Top 10% earners in Europe received 30% of total regional income in 1980, while the bottom 50% received 22%. In 2017, by contrast, the top 10% share had risen to 35%, while 18% total income accrued to the poorest half of the population. In line with our previous findings, it appears that changes in the income distribution mostly occurred during the last two decades of the twentieth century. As top income inequality increased in most countries of Western Europe and Scandinavia between 1980 and 2000, the richest decile captured an increasing share of the continent's growth, before more of less stagnating since then. By contrast, the bottom 50% share decreased more suddenly in the early 1990s due to the combination of strong recessions and rising inequalities at the top in Eastern Europe. These two movements have driven most of variations in income inequality in Europe.

Long-run trends in Europe reveal that inequalities have mainly increased at the very top of the income distribution. Figure 3.5b plots the annualized growth rates of different income groups over the 1980–2017 period. In the past thirty-seven years, the poorest half of European residents saw their incomes increase by less than 1% annually. The "European middle class" only benefited slightly more from growth than these poorer groups: income earners between percentiles 50 and 90 saw their incomes increase by about 1% per year. As soon as one looks at groups within the top 10%, however, total growth rates are markedly higher. All income groups among



(a) Income inequality in Europe, 1980–2017:
Top 10% vs. bottom 50% income shares

(b) Income inequality in Europe, 1980–2017:
Growth incidence curve

Source: authors' computations combining surveys, tax data and national accounts.

Figure 3.5: Evolution of Inequality among European Citizens

top 0.1% earners saw their earnings grow by more than 2% per year during our period of interest, and even more for the top 0.001% of European.

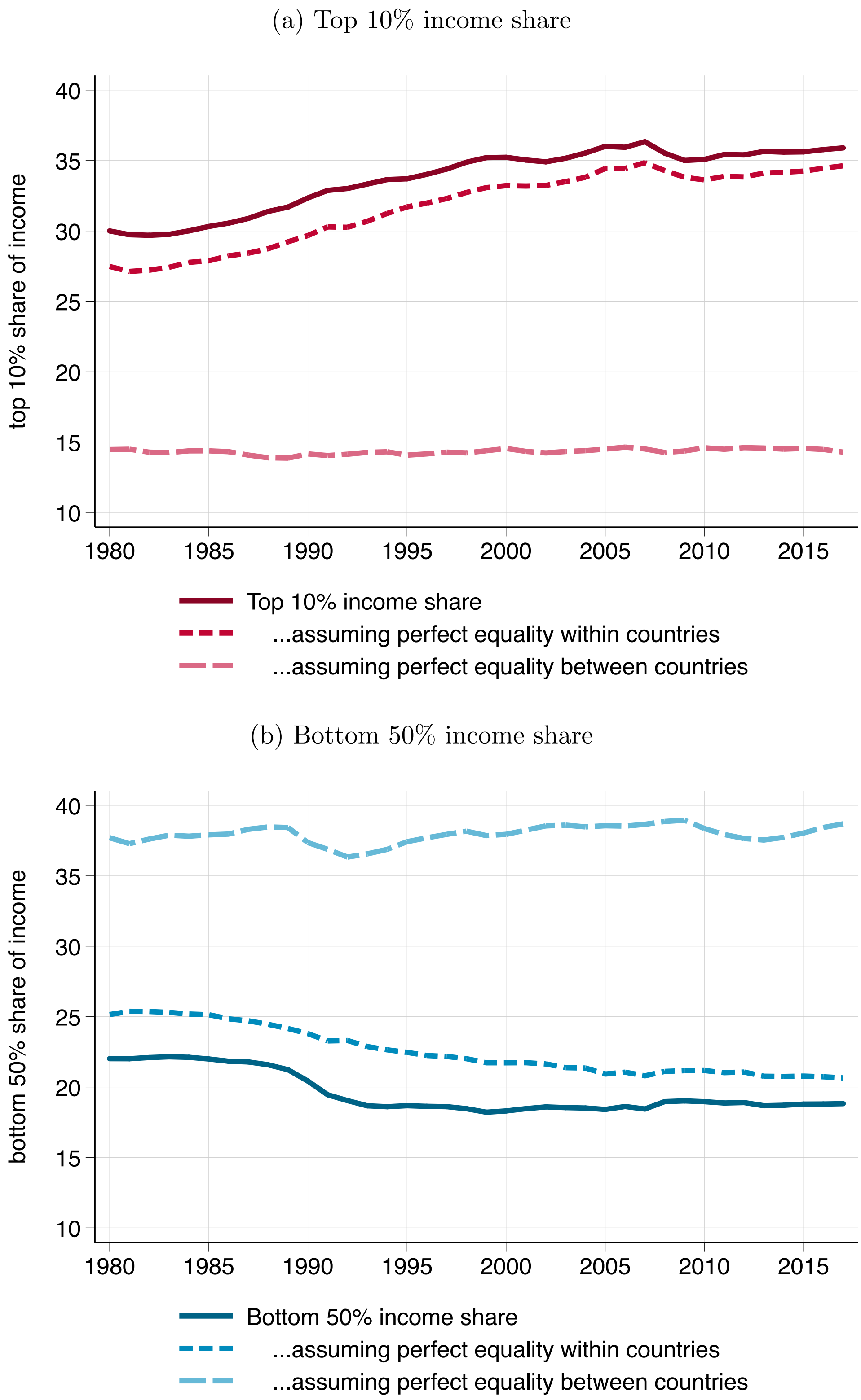
How important are between-country inequalities compared to within-country income differences in explaining these trends? Figure 3.6 shows the potential levels and dynamics of top 10% and bottom 50% income shares under different scenarios. Solid lines represent the true series, dotted lines correspond to the income disparities that would exist if there was no inequality between countries, and dashed lines correspond to those that would prevail if there were no inequalities within countries. Eradicating differences in countries' average national incomes would have a moderate effect on European inequalities: both the top 10% and the bottom 50% shares would change by a few percentage points in all years considered. If all Europeans were to earn the average national income of their country of residence, by contrast, differences in standards of living would be dramatically reduced. The top 10% share would have stagnated at about 15%, while bottom 50% earners would receive more than one third of total income in all years considered.

3.4.4 Redistribution in Europe

Until now, we have focused exclusively on the distribution of pre-tax income, that is the sum of all pre-tax personal income flows accruing to the owners of the production factors, before taking into account the operation of the tax and transfer system, but after taking into account the operation of the pension system. We will now look more precisely at the evolution of post-tax disposable income inequality in Europe and the extent of redistribution across European regions.

To that end, we will distinguish two concept of income after taxes and transfers. The first one is post-tax national income. Post-tax national income subtracts all taxes and contributions, and adds all transfers, including both cash transfers and government consumption, so that the total sums to national income. All explained is the methodological section, public health expenditures are distributed lump-sum, and other government consumption proportionally. Because the distribution of government expenditure raises more conceptual and methodological questions, we will also consider a narrower concept: post-tax disposable income. Post-tax disposable income only redistribute cash transfers, so that is does not sum up to national income.

In order to synthesize redistribution with a simple indicator, we propose to follow Bozio et al. (2018) and look at the percent reduction in the ratio of the top 10% to bottom 50% average incomes. This ratio is a simple and straightforward measure of

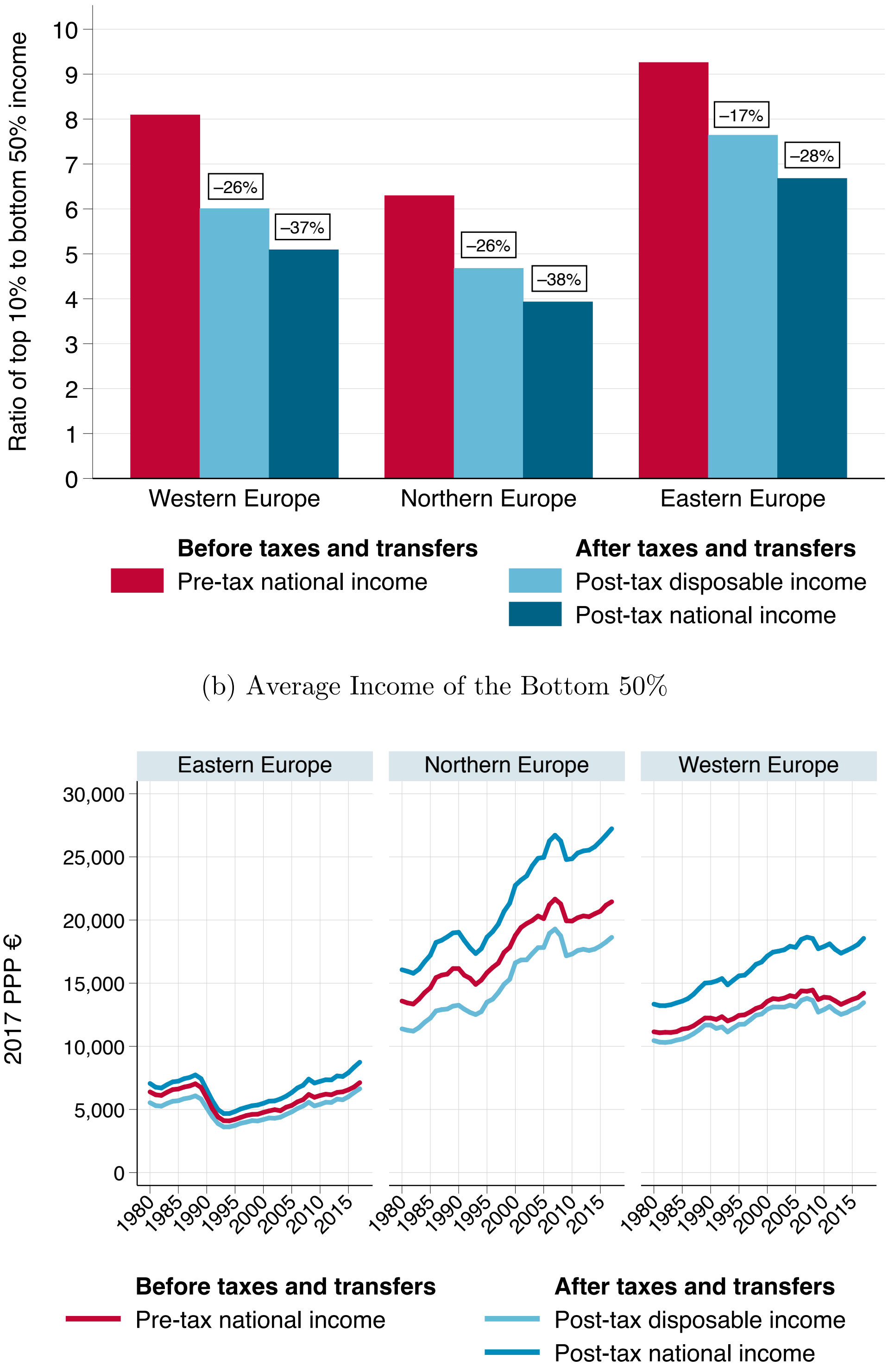


(a) Top 10% income share

(b) Bottom 50% income share

Source: authors' computations combining surveys, tax data and national accounts.

Figure 3.6: Between- versus within-country inequality in Europe, 1980–2017



(a) Ratio of Top 10% to Bottom 50% Average Income

(b) Average Income of the Bottom 50%

Source: authors' computations combining surveys, tax data and national accounts.

Figure 3.7: Redistribution in Europe

inequality, as it summarizes in a single number the gap between the earnings of the two sides of the income distribution. Looking at the extent to which fiscal systems reduce this gap can therefore inform directly on their redistributive effect.[24] Figure 3.7a compares average redistribution across European regions with this indicator. Most of the redistribution happens when moving from pre-tax to post-tax disposable income. Yet government expenditures also plays a sizable role, which is entirely driven by health expenditures. Western and Northern Europe have, according to that indicator, virtually identical levels of redistribution. But Eastern Europe redistributes significantly less.

In figure 3.7b, we can see the average income of the bottom 50%, both before and after taxes. Note that the post-tax *disposable* income of the bottom 50% is actually slightly lower than its pre-tax income: that is because some of the taxes paid by the bottom 50% finance government expenditures, which are not accounted for in disposable income. To properly capture the absolute increase in standard of livings when moving from pre-tax to post-tax income, we must look at post-tax *national* income, that incorporates all government spending.

3.4.5 Inequality in Europe vs. the United States

Income inequality in the US has increased dramatically in the past forty years, especially at the top of the distribution (Piketty, Saez, and Zucman, 2018). In this section, we seek to compare these dynamics to those observed in Europe. Europe and the United States are two large, integrated world regions, which share relatively high degrees of similarities in terms of levels of development, exposure to global markets or penetration of new technologies. Comparing the evolution of income inequality in these regions can thus provide insights into their different policy and economic trajectories since the 1980s. In particular, we will refine and expand on the recent work done in the World Inequality Report 2018 (Alvaredo, Chancel, et al., 2018) by focusing on two questions. Are income disparities in Europe larger than in the US? And what are the roles of between-country (between-states) and within-country (within-states) inequalities in explaining these differences? We explore these issues by comparing estimates from this paper with the US DINA estimates from Piketty, Saez, and Zucman (2018). We also explore the geography of inequality in Europe and the US by combining these estimates with state-specific top taxable income shares from Frank et al. (2015), survey distributions from the current population population survey (CPS), and state-level GDPs (see appendix C.1.4 for the detailed

[24] The results of this section are robust to the use of different groups for the top and the bottom of the distribution (e.g. top 1%/bottom 50%)

methodology). In what follows, when looking at inequality in Europe as a whole, we use market exchange rates estimates to measure differences in average income levels between European countries. This is to make the comparison between US states and European countries more meaningful. While purchasing power parity figures could be computed for European countries, there exist no conversion factor which would allow us to account for differences in average costs of living between US States.

Spatial inequalities have always been much smaller in the US than in Europe, at least since the mid-twentieth century.[25] In Europe, inequalities between countries have decreased slightly from 1950 to the beginning of the 1980s and have remained broadly stable since then: in 2017, the national income of top 10% European countries was 2.8 times higher than that of the bottom 50%. Spatial heterogeneity has never reached such levels in the US, where the top 10% to bottom 50% ratio has decreased from 2.5 at the beginning of the 1930s to 1.5 in 2017.[26]

These differences are apparent when comparing individual countries and states in recent years. The poorest European countries had national incomes per adult lower than the continental average by more than 50%, both in 1980 and in 2017. There was no such equivalent in the US, neither today nor thirty years ago. In 1980, poorest US states were characterized by standards of living lower than the national average by no more than 25%, and this figure did not exceed 40% in 2017. Similarly, the wealthiest countries of Europe have steadily remained richer than the average European by about 75%, compared to only 25% in the US. There were, both in 1980 and 2017, small US states who were significantly richer than the rest of the country: in 1980, residents of Alaska and Washington D.C. earned more than 300% of US national income. Beyond these exceptions, however, a vast majority of states have always had standards of living located between 70% and 120% of the national average.

There are at least two potential explanations for these differences. First, the United States has reached a significantly higher degree of economic integration than Europe, and have remained politically and institutionally stable for a much longer time. In this context, US states rapidly converged in their levels of development, especially at the beginning of the twentieth century (Barro and Sala-i-Martin, 1991; Barro and Sala-i-

[25] State domestic products provided by the Bureau of Economic Analysis go back as far as 1967. We extrapolate these series back to 1929 by using the growth rates in personal income per capita available from Barro and Sala-i-Martin (1992).

[26] The ratios of the top 10% to bottom 50% European states or US states adjust for population differences. That is, we split proportionally the population of states which are at the frontier between the top 10% and the bottom 90% of the continental population. This indicator is a simple measure of spatial inequality: it compares the average income of the "core" territories to that of the poorest states or countries gathering half of the total population.

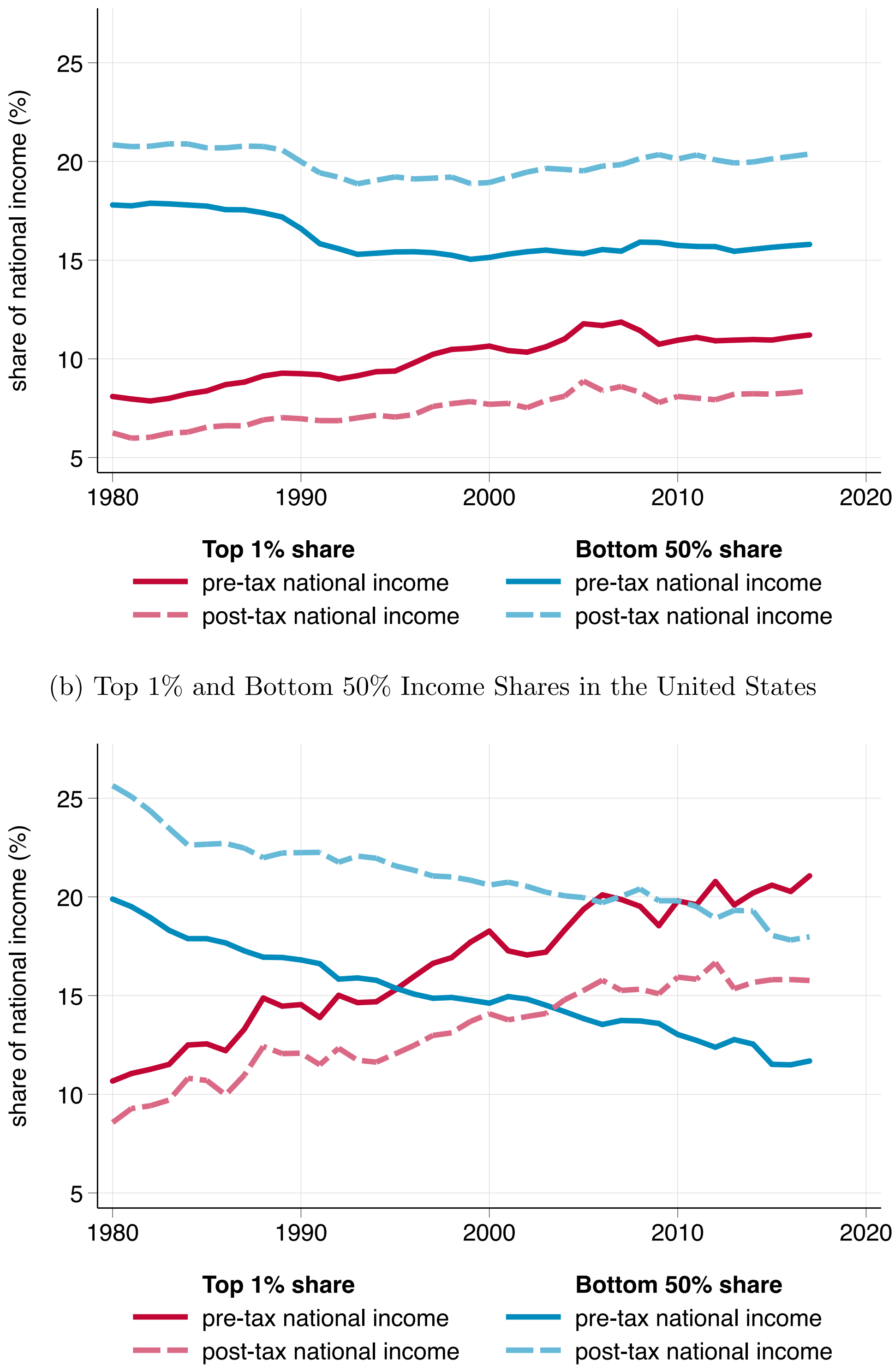
Martin, 1992). Accordingly, the persistence of high between-country inequalities in Europe can partially be explained by the multiplication of strong asymmetric shocks since the 1980s which have delayed potential convergence processes. The 1990s crises in Eastern Europe badly affected the poorest economies of the continent, just as the 2008 crisis hit only moderately richer European nations but led to stronger recessions in Southern and Baltic countries. That heterogeneity also has to do with a lack of political integration and coordinated policy responses among European countries and within the European Union. While the EU has decided to encourage the adhesion of future members with financial aid funds, it has only dedicated moderate sums to these programs.[27]

While geographical disparities are higher in Europe than in the US, inequalities within territories are higher and have grown much faster in the US. In 1980, US states were, on average, only slightly more unequal than Western European countries. Between 1980 and 2017, however, this gap grew significantly: while inequalities within European countries increased only moderately, they skyrocketed in most US states, with income shares for the top 10% reaching up to 60% in New York and Florida. The fact that inequalities increased only moderately in Europe, and mainly in Eastern European countries who "caught up" with their Western neighbors, announced a clear disconnection between the US and Europe. In 2017, top 10% shares in the most equal states of the US were close to those observed in the most unequal countries of Europe.

Spatial inequalities are therefore lower in the US than in Europe, while inequalities within European countries are lower than inequalities within US states. Adding up these two effects, are overall income differences wider in the US than in Europe as a whole? The answer is unequivocal: income inequality is substantially higher in the US than in Europe. In 2017, the top 1% in the US captured a share of national income twice as large as the poorest half of the population. In Europe, by contrast, the bottom 50% share was significantly larger than that of top 1% earners (figure 3.8). This was not always the case: in 1980, bottom 50% shares were

[27] Between 1991 and 2003, for instance, average transfers from West Germany to East Germany had amounted to some 4.5% of western GDP and 30% of eastern GDP, leading to rapid and significant regional convergence after reunification (see for example `http://ec.europa.eu/economy_finance/publications/pages/publication1437_en.pdf`). By contrast, the 2019 financial programming of the European Regional Development Fund, the main program for correcting imbalances between EU regions, is expected to amount to 31 billion euros, or less than 0.2% of total EU GDP (see `https://ec.europa.eu/budget/library/biblio/documents/2019/Programmes_performance_overview.pdf`). And when looking at net contributions to the EU budget, countries benefiting most from EU transfers (such as Bulgaria, Hungary of Lithuania) do not receive net income flows higher than 3% of their GDP, while the most important contributors (such as Germany or Sweden) give up less than 0.4% of their total annual production.

(a) Top 1% and Bottom 50% Income Shares in Europe



(b) Top 1% and Bottom 50% Income Shares in the United States

Source: Europe: authors' computations combining surveys, tax data and national accounts; United States: Piketty, Saez, and Zucman (2018)

Figure 3.8: Inequality in Europe and in the United States

Table 3.1: Theil index decomposition of between-region and within-region inequalities in Europe and the US

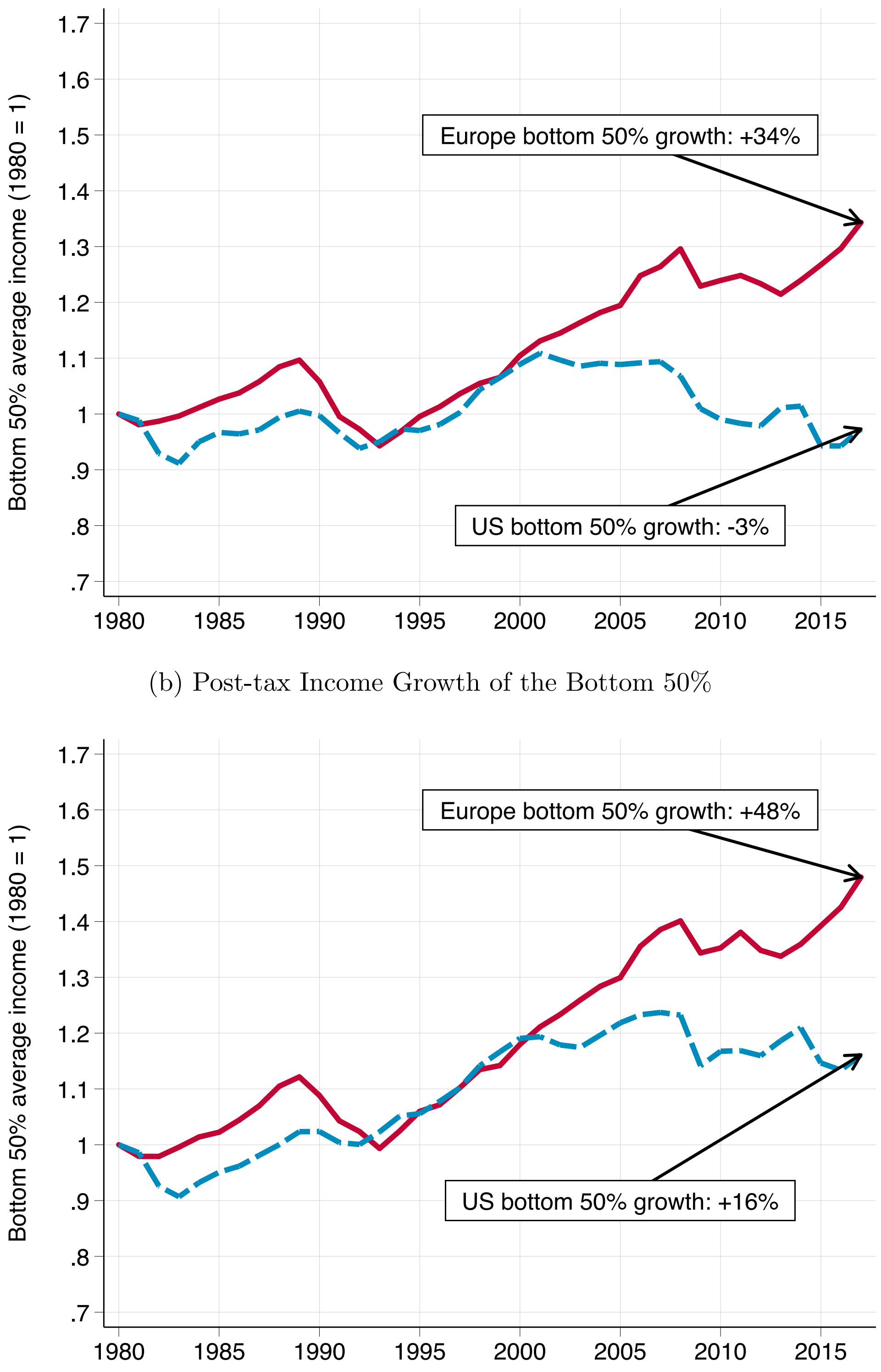
	Theil index	Within-group		Between-group	
		Value	% of total	Value	% of total
Europe					
1980	0.37	0.24	65.0 %	0.13	35.0 %
1990	0.43	0.29	67.4 %	0.14	32.6 %
2000	0.49	0.34	69.6 %	0.15	30.4 %
2007	0.52	0.39	74.8 %	0.13	25.2 %
2017	0.50	0.38	76.6 %	0.12	23.4 %
United States					
1980	0.45	0.44	96.7 %	0.01	3.3 %
1990	0.63	0.61	98.0 %	0.01	2.0 %
2000	0.85	0.84	98.5 %	0.01	1.5 %
2007	0.94	0.93	98.5 %	0.01	1.5 %
2017	1.00	0.98	98.3 %	0.02	1.7 %

Source: authors' computations combining surveys tax data and national accounts.

actually very similar between the two regions, amounting to about a fifth of national income. While income inequalities have increased in both Europe and the US, the trend has therefore been much steeper in the latter. In Europe, economic crises and rising income disparities in Eastern Europe contributed to moderately compressing the bottom 50% share at the beginning of the 1990s, while top income inequality increased slightly from the 1980s to the 2000s. Inequality dynamics in the US have been much more linear: in the past forty years, the top 1% share steadily surged from 11% to 20% and the bottom 50% share was nearly divided by two.

These differences appear even more striking if one compares the growth trajectories of the bottom 50% of the two regions (figure 3.9). Our estimates reveal that despite the fact that the average national income grew faster in the US than in Europe by 40% during this period, the poorest 50% experienced faster growth in Europe, both on a pre-tax and a post-tax basis. The pre-tax income of the bottom 50% stagnated in the US while it increased by 34% in Europe. The picture is not significantly different for post-tax incomes, which increase by 16% in the US compared to 48% in Europe.

Table 3.1 provides Theil decomposition of income inequality in Europe and the US between 1980 and 2017. In 1980, inequalities were slightly higher in the US than in Europe, if one considers the Theil index to be a broad measure of income concentration. This gap had widened considerably in 2017: the Theil index reached 1



(a) Pre-tax Income Growth of the Bottom 50%

(b) Post-tax Income Growth of the Bottom 50%

Source: Europe: authors' computations combining surveys, tax data and national accounts; United States: Piketty, Saez, and Zucman (2018)

Figure 3.9: Growth of the Bottom 50%
in Europe and in the United States

in the US, compared to only 0.5 in Europe. Furthermore, decomposition reveals that inequalities between countries explain a much larger share of income disparities in Europe than inequalities between states do in the US. At the beginning of our period of interest, about two thirds of income inequalities in Europe were explained by inequalities within countries. Due to rising income disparities in European nations, the share of income concentration explained by within-group inequalities increased to more than 75% in 2017. In the US, on the other hand, higher geographical integration and larger differences in standards of living within States have led between-group inequalities to remain of minor importance. Between 1980 and 2017, the share of overall US inequalities explained by within-state income differences remained above 95%.

Conclusion

We have developed a novel methodology combining surveys, tax data and national accounts in a consistent manner to produce pre-tax and post-tax income inequality statistics for all European countries covering the 1980-2017 period. Based on this methodology, we have documented the following results.

First, we do not observe a clear pattern of convergence in average income levels between countries since the early 1980s. Per adult income in Eastern Europe is about 35% lower than European average today. This is the same value as in the early 1980s, before the fall of the USSR. In Southern European countries, per adult average incomes have been declining relatively to the continental average since the 1990s and are now 10% below the average. Northern European countries were 25% richer than the average in the mid-1990s and are now 50% richer.

Personal income inequalities have been increasing in nearly all countries. Nearly all European countries failed to reach the United Nations Sustainable Development Goals inequality target over the 1980-2017 period, which seeks to ensure that the bottom 40% of the population grows faster than the average. Since the 2000s, European countries have been relatively more successful at ensuring that bottom income groups secure a fair share of growth, but the majority of countries still failed to achieve the UN objective.

As a result of a limited convergence process and rising inequality within countries, Europeans are more unequal today than four decades ago. Between 1980 and 2017, per adult average annual pre-tax income growth was below 1% for bottom 50% earners, while the top 0.1% grew at a rate higher than 2% per year. The top 1%

captured about as much growth as the bottom 50% of the population. The share of national income captured by the top 1% Europeans increased from less than 8% of national income to nearly 11% between 1980 and 2017.

Despite a rise of inequality in Europe and within the EU, European countries have been much more successful at containing rising inequalities than the US. This is largely because European countries succeeded in generating higher income growth rates for bottom earners than did the US. Average income of the poorest half of Europeans was 40% higher in 2017 than in 1980, while it was essentially the same as in 1980 (+3%) for the poorest 50% Americans. Consequently, Europe is much less unequal than the US, despite higher spatial inequalities in Europe than between US states.

To what extent did observed and perceived inequality dynamics in Europe contribute to current levels of resentment against national and European institutions? Which structural changes and set of policies enabled European countries to contain the surge of inequalities observed in the USA since 1980? This paper opens up many questions to which our inequality series will hopefully contribute to answering in future comparative research.

Bibliography

Alstadsæter, Annette et al. (2017). “Accounting for Business Income in Measuring Top Income Shares: Integrated Accrual Approach Using Individual and Firm Data from Norway”. URL: http://www.nber.org/papers/w22888.

Alvaredo, Facundo, Anthony B. Atkinson, et al. (2016). “Distributional National Accounts (DINA) Guidelines: Concepts and Methods used in WID.world”. In: *WID.working paper series 2016/1*. URL: http://wid.world/document/dinaguidelines-v1/.

Alvaredo, Facundo, Lucas Chancel, et al. (2018). *World Inequality Report*. Tech. rep. World Inequality Lab. URL: http://wir2018.wid.world/files/download/wir2018-full-report-english.pdf.

Angel, Stefan, Richard Heuberger, and Nadja Lamei (2017). “Differences Between Household Income from Surveys and Registers and How These Affect the Poverty Headcount: Evidence from the Austrian SILC”. In: *Social Indicators Research* 138.2, pp. 1–29.

Atkinson, Anthony B. (1996). “Income distribution in Europe and the United States”. In: *Oxford Review of Economic Policy* 12.1.

Atkinson, Anthony B. and Thomas Piketty (2007). *Top Incomes Over the Twentieth Century*. Oxford University Press.

– (2010). *Top Incomes - A Global Perspective*. Oxford University Press.

Atkinson, Anthony B., Thomas Piketty, and Emmanuel Saez (2011). “Top Incomes in the Long Run of History”. In: *Journal of Economic Literature* 49.1, pp. 3–71.

Atria, Jorge et al. (2018). “Top incomes in Chile: A Historical Perspective of Income Inequality (1964-2015)”. URL: http://ignacioflores.com/pdf/top-incomes-chile.pdf.

Australian Bureau of Statistics (2019). “Australian National Accounts: Distribution of Household Income, Consumption and Wealth, 2003-04 to 2017-18”. In: URL: https://www.abs.gov.au/ausstats/abs@.nsf/mf/5204.0.55.011.

Barro, Robert and Xavier Sala-i-Martin (1991). “Convergence across States and Regions”. In: *Center Discussion Paper No. 629*.

– (1992). “Convergence”. In: *Journal of Political Economy* 100.2, pp. 223–251.

Beblo, Miriam and Thomas Knaus (2001). “Measuring income inequality in Euroland”. In: *Review of Income and Wealth* 47.3.

Blanchet, Thomas and Lucas Chancel (2016). “National Accounts Series Methodology”. URL: http://wid.world/document/1676/.

Blanchet, Thomas, Ignacio Flores, and Marc Morgan (2018). “The Weight of the Rich: Improving Surveys Using Tax Data”.

Blanchet, Thomas, Juliette Fournier, and Thomas Piketty (2017). "Generalized Pareto Curves: Theory and Applications".

Bozio, Antoine et al. (2018). "Inequality and Redistribution in France, 1990-2018: Evidence from Post-Tax Distributional National Accounts (DINA)". In: *WID.world Working Paper 2018/10.*

Brandolini, Andrea (2006). "Measurement of income distribution in supranational entities: The case of the European Union". In: *LIS Working Paper Series, No. 452.*

Brandolini, Andrea and Alfonso Rosolia (2019). "The distribution of well-being among Europeans". In: *Banca d'Italia Occasional Papers Series 496.*

Bukowski, Pawel and Filip Novokmet (2017). "Top incomes during wars, communism and capitalism: Poland 1892-2015". In: *WID.world Working Paper Series 2017/22.*

– (2019). "Between Communism and Capitalism: Long-Term Inequality in Poland, 1892-2015". URL: `https://wid.world/document/between-communism-and-capitalism-long-term-inequality-in-poland-1892-2015/`.

Chen, Tianqi and Carlos Guestrin (2016). "XGBoost: A Scalable Tree Boosting System". URL: `http://dx.doi.org/10.1145/2939672.2939785`.

Cristia, Julian and Jonathan A. Schwabish (2009). "Measurement error in the SIPP: Evidence from administrative matched records". In: *Journal of Economic and Social Measurement* 34.1, pp. 1–17.

Dauderstädt, Michael and Cem Keltek (2011). "Immeasurable Inequality in the European Union". In: *Intereconomics* 46.1, pp. 44–51.

Eurostat (2018). "Income comparison: social surveys and national accounts". In: URL: `https://ec.europa.eu/eurostat/web/experimental-statistics/ic-social-surveys-and-national-accounts`.

Fairfield, Tasha and Michel Jorratt De Luis (2016). "Top Income Shares, Business Profits, and Effective Tax Rates in Contemporary Chile". In: *Review of Income and Wealth* 62.August, S120–S144.

Ferreira, Ana and Laurens de Haan (2006). *Extreme Value Theory: An Introduction.* Springer Series in Operations Research. Springer.

Fesseau, Maryse and Maria Liviana Mattonetti (2013). "Distributional Measures Across Household Groups in a National Accounts Framework: Results from an Experimental Cross-country Exercise on Household Income, Consumption and Saving". In: *OECD Statistics Working Papers 2013/4.*

Filauro, Stefano (2018). "The EU-wide income distribution: inequality levels and decompositions". In: *European Commission Working Paper.*

Filauro, Stefano and Zachary Parolin (2018). “Unequal unions? A comparative decomposition of income inequality in the European Union and United States”. In: *Journal of European Social Policy*.

Frank, Mark et al. (2015). “Frank-Sommeiller-Price Series for Top Income Shares by US States since 1917”. In: *WID.world Technical Note Series 2015/7*.

Garbinti, B., J. Goupille-Lebret, and Thomas Piketty (2018). “Income inequality in France, 1900-2014: Evidence from Distributional National Accounts (DINA)”. In: *Journal of Public Economics* 162.1, pp. 63–77.

Guillaud, Elvire, Matthew Olckers, and Michaël Zemmour (2019). “Four Levers of Redistribution: The Impact of Tax and Transfer Systems on Inequality Reduction”. In: *Review of Income and Wealth*.

Hosking, J R M and J R Wallis (1987). “Parameter and Quantile Estimation for the Generalized Pareto Distribution”. In: *Technometrics* 29.3, pp. 339–349. URL: `https://www.tandfonline.com/doi/abs/10.1080/00401706.1987.10488243`.

Immervoll, Herwig and Linda Richardson (2011). “Redistribution Policy and Inequality Reduction in OECD Countries: What has changed in two decades?” In: *OECD Social, Employment and Migration Working Papers No. 122*.

Jenmana, Thanasak (2018). “Democratisation and the Emergence of Class Conflicts Income Inequality in Thailand, 2001-2016”. URL: `https://wid.world/document/democratisation-and-the-emergence-of-class-conflicts-income-inequality-in-thailand-2001-2016-wid-world-working-paper-2018-15/`.

Jesuit, David K. and Vincent A. Mahler (2010). “Comparing Government Redistribution Across Countries: The Problem of Second-Order Effects”. In: *Social Science Quarterly* 91.5, pp. 1390–1404.

Korinek, Anton, Johan A. Mistiaen, and Martin Ravallion (2006). “Survey nonresponse and the distribution of income”. In: *Journal of Economic Inequality* 4.1, pp. 33–55.

Lakner, Christoph and Branko Milanovic (2016). “Global Income Distribution: From the Fall of the Berlin Wall to the Great Recession”. In: *The World Bank Economic Review* 30.2, pp. 203–232. URL: `https://academic.oup.com/wber/article-lookup/doi/10.1093/wber/lhv039`.

Lesage, Éric (2009). “Calage non linéaire”.

Milanovic, Branko (1998). *Income, Inequality, and Poverty during the Transition from Planned to Market Economy*. World Bank Regional and Sectoral Studies.

Morgan, Marc (2017). "Extreme and Persistent Inequality: New Evidence for Brazil Combining National Accounts , Surveys and Fiscal Data, 2001-2015". URL: `http://wid.world/wp-content/uploads/2017/09/Morgan2017BrazilDINA.pdf`.

Novokmet, Filip, Thomas Piketty, and Gabriel Zucman (2018). "From Soviets to oligarchs: inequality and property in Russia 1905-2016". In: *The Journal of Economic Inequality* 16.2, pp. 189–223.

OECD (2008). *Growing Unequal? Income Distribution and Poverty In OECD Countries*. OECD Publishing.

Paulus, Alari (2015). "Income underreporting based on income expenditure gaps: Survey vs tax records". URL: `http://hdl.handle.net/10419/126467`.

Piketty, Thomas (2001). *Les hauts revenus en France au XXe siècle*. Seuil.

– (2003). "Income Inequality in France, 1901–1998". In: *Journal of Political Economy* 111.5, pp. 1004–1042. URL: `http://www.journals.uchicago.edu/doi/10.1086/376955`.

Piketty, Thomas and Emmanuel Saez (2003). "Income Inequality in the United States, 1913–1998". In: *Quarterly Journal of Economics* CXVIII.1.

Piketty, Thomas, Emmanuel Saez, and Gabriel Zucman (2018). "Distributional National Accounts: Methods and Estimates for the United States". In: *Quarterly Journal of Economics* 133.2, pp. 553–609.

– (2019). "Simplified Distributional National Accounts". In: *AEA Papers and Proceedings* 109, pp. 289–295.

Piketty, Thomas, Li Yang, and Gabriel Zucman (2019). "Capital Accumulation, Private Property, and Rising Inequality in China, 1978-2015". In: *American Economic Review* 109.7, pp. 2469–2496.

Salverda, Wiemer (2017). "The European Union's income inequalities". URL: `https://www.nbp.pl/badania/seminaria/7iv2017-3.pdf`.

Statistics Canada (2019). "Distributions of Household Economic Accounts, estimates of asset, liability and net worth distributions, 2010 to 2018, technical methodology and quality report". In: URL: `https://www150.statcan.gc.ca/n1/pub/13-604-m/13-604-m2019001-eng.htm`.

Statistics Netherlands (2014). "Measuring Inequalities in the Dutch Household Sector". URL: `https://www.cbs.nl/en-gb/background/2014/19/measuring-inequalities-in-the-dutch-household-sector`.

Taleb, Nassim Nicholas and Raphael Douady (2015). "On the super-additivity and estimation biases of quantile contributions". In: *Physica A: Statistical Mechanics and its Applications* 429, pp. 252–260. URL: `http://dx.doi.org/10.1016/j.physa.2015.02.038`.

Wolfson, Michael, Mike Veall, and Neil Brooks (2016). “Piercing the Veil – Private Corporations and the Income of the Affluent”. In: *Canadian Tax Journal* 64.1, pp. 1–25.

Zwijnenburg, Jorrit, Sophie Bournot, and Federico Giovannelli (2019). “OECD Expert Group on Disparities in a National Accounts Framework: Results from the 2015 Exercise.” In: *OECD Statistics Working Papers 2019/76.*

Chapter 4

Modeling the Dynamics of Wealth Inequality in the United States, 1962–2100

This paper develops a new approach to address a basic question: what drives the evolution of the wealth distribution, and how does it react to economic, demographic or policy changes?

The economic literature has already made many contributions to our understanding of the wealth distribution (e.g. Wold and Whittle, 1957; Laitner, 1979; Vaughan, 1979; Benhabib, Bisin, and Zhu, 2011). It has been able to explain key stylized facts, in particular its Pareto-shaped tails. It has uncovered several plausible mechanisms that could explain the levels and changes in wealth inequality. It has emphasized, among others, the role of labor income inequality, unequal rates of return, taxation, demographics through children and the sharing of inheritance, and the spread between the rate of return on capital and the rate of economic growth (Stiglitz, 1969; Cowell, 1998; Favilukis, 2013; Piketty and Zucman, 2014; Böhl and Fischer, 2017; Hubmer and Smith, 2018).

But it remains difficult to untangle these effects and assess their respective importance. Theoretical models tend to focus on a limited set of mechanisms and make simplifying assumptions for the sake of tractability. While understandable, this limits our ability to connect these models to the data beyond the replication of the main stylized facts, and use them for policy purposes.

On the empirical side, most of the literature has been concerned with pure measurement issues (e.g. Saez and Zucman, 2016; Kopczuk, 2015; Bricker, Henriques, and

Hansen, 2018). Still, a few papers have tried to test certain mechanisms directly using reduced-form specifications. Acemoglu and Robinson (2015) and Góes (2016) have tested the impact of the difference between the rate of return on capital and the growth rate $(r - g)$ that was popularized by Piketty (2014), and found no supporting evidence. But these approaches face certain difficulties. Wealth inequality statistics are still in their infancy, with limited time and geographical coverage, and varying quality. This paucity of wealth inequality data makes it hard to get meaningful variation — let alone exogenous variation — that could be used to capture the effects at hand. This is all the more limiting that theory suggests such effects may be slow and take decades to materialize clearly (Gabaix et al., 2016). The studies have avoided the issue by relying on more widely available but fairly noisy proxies such as income inequality, which complicates the interpretation of the results. In a world with a widespread, cross-country dataset on wealth inequality spanning several centuries, it might be easy to just "let the data speak." But until then, pure reduced-form approaches face considerable challenges.

Another method involves the construction of "synthetic saving rates" (Saez and Zucman, 2016; Garbinti, Goupille-Lebret, and Piketty, 2016; Berman, Ben-Jacob, and Shapira, 2016). These synthetic saving rates are calculated so as to reconcile the wealth changes of the different parts of the distribution with the income of the corresponding groups. They are meant to capture many different effects including mobility, the inequality of savings, and their correlation with wealth. In essence, this is an accounting exercise that looks separately at the different parts of the distribution. That approach constitutes a practical middle ground between the theory and the data. Strictly speaking, however, synthetic saving rates can only be interpreted as structural parameters if we assume no mobility between groups, and homogeneous behavior within groups, which affects the domain of applicability ofthe method and the generality of its conclusions. In fact, Gomez (2016) shows that these synthetic saving rates can be decomposed into a "within" term and a "displacement" term which captures mobility. He shows that the displacement terms plays an important role in the dynamic of wealth inequality. This decomposition, however, requires access to panel data on wealth, which is quite rare.

Overall, the answer to many questions remains unclear because the empirical literature has been working with synthetic indicators that are hard to tie explicitly to the individual behavior of people. And, as a consequence, it has had difficulties connecting itself to the theory. This state of affairs is at least partly the result of data limitations. Ideally, to directly integrate theoretical models with the data, we would use a long-run, high-quality panel dataset of both income and wealth. Unfortunately, no such thing exists in most countries, including the United States.

In this paper, I suggest a way to resolve this divide between data and theory. Using only repeated cross-sections, and based on two results of stochastic calculus, I am able to nonparametrically identify and structurally estimate the key parameters that determine the dynamics of the wealth distribution. These parameters directly relate to the individual saving behavior of people, so they do not merely capture reduced form relationships between synthetic indicators of wealth inequality. But they remain very general and mostly agnostic as to the exact reason why people save, making the approach compatible with a wide class of models.

This framework is flexible enough to incorporate a realistic model of income, the tax system, inheritance and demographics. Yet it remains simple enough to allow for a transparent identification of the parameters. The model can reproduce the data, and be used run conditional forecasts and counterfactuals in which we change various economic parameters, such as the growth rate, labor income inequality, or rates of return. The model captures both the steady state and the transitional effects of the different shocks — an important feature given that some shocks can take a lot of time to noticeably affect the wealth distribution.

Models of the wealth distribution that can accurately reproduce its Pareto-shaped fat tails virtually all share the same core idea: that people accumulate wealth through a succession of random multiplicative shocks. These may be preference shocks, shocks to rates of return, to the number of children, and so on. What matters is that, as long as the mean and the variance of these shocks falls into the right range, the steady-state distribution will have a power-law tail (Kesten, 1973; Gabaix, 2009). That leads us to the key insight of the paper: although this process of multiplicative random shocks is very general, it actually makes some sharp predictions regarding the evolution of the wealth distribution. And we can exploit these predictions as a source of identification. To that end, it is important not to solely focus on the steady state. Indeed, there is an infinite number of ways in which we could calibrate the mean and variance of the aforementioned multiplicative shocks so as to reach any given steady-state level of inequality, making the model underidentified. But under these various calibrations, the distribution of wealth would change at widely different speeds. Therefore, as long as we observe the wealth distribution outside of its steady state — which is clearly the case in the United States since the 1960s — we can unambiguously identify the parameters of the underlying wealth accumulation process.

In practice, the approach of the paper is made tractable by the use of the continuous time formalism advocated for instance by Gabaix et al. (2016). This formalism provides access to two highly useful results. First, there is the Fokker-Planck equation, which explicitly relates the evolution of the wealth distribution to the parameters of the underlying accumulation process. Then, there is Gyöngy's (1986) theorem. The process of wealth accumulation is itself a function of stochastic processes for income and consumption that are potentially hard to model accurately. Gyöngy's (1986) theorem shows that, in order to properly model the marginal distribution of wealth, it is not necessary to fully model these processes: all we need to know is their mean and variance conditional on wealth. In essence, the mean and variance of savings conditional on wealth turn out to be "sufficient statistics" which entirely define the evolution of the wealth distribution. This considerably reduces the dimensionality of the problem, and makes the analysis much simpler. In the end, and despite the richness of the model, the estimation reduces to the estimation of a linear relationship between observable quantities: therefore, it provides a clear-cut and visual interpretation that can be used to discuss the quality of the fit, or the presence of structural changes.

I apply the method using the Distributional National Accounts (DINA) data from Piketty, Saez, and Zucman (2018), which I complement using the Survey of Consumer Finances (SCF) and various additional sources to account for demography and inheritance. Piketty, Saez, and Zucman (2018) provide public use samples of their data, available yearly since 1962. This data distributes all of national income and wealth to individuals, making it possible to track their distribution in a way that is consistent with macroeconomic totals, and over a long period that include some major economic changes. I find that, out of the 15 pp. increase in the top 1% wealth share observed since 1980, about 7 pp. can be attributed to rising labor income inequality, 6 pp. to rising returns on wealth (mostly in the form of capital gains), and 2 pp. to lower growth. Over the entire period, rich households appear to have been consuming, on average, a constant fraction of their wealth. At the same time they have seen their income rising, due to both higher labor income inequality and capital gains. Hence, they have been saving a higher fraction of their income, leading to an important accumulation of wealth at the top. Under current parameters, the top 1% wealth share would reach its steady-state value of roughly 45% by the 2040s, a level similar to that of the beginning of the 20th century.

This model of the wealth distribution has some practical applications, in particular for the theory of wealth taxation. Recent contributions have emphasized that the long-run elasticity of the capital stock is a sufficient statistic for optimal capital taxation (Saez and Stantcheva, 2018), yet little is known about its value. I use this paper's model to investigate the issue. It allows me to approach the problem in a way that combines insights from several recent contributions, in particular the role of mobility (Saez and Zucman, 2019), tax avoidance, and saving responses (Jakobsen et al., 2019). I develop a simple formula to estimate how the tax base would react to a wealth tax at the top at the steady-state. This formula suggests that the elasticity can be sizeable, but also that it is higher for small tax rates than for larger ones. As a result, revenue-maximizing tax rates may still be quite high.

The rest of the paper is organized as follows: in section 4.1, I review the main stylized facts about wealth inequality in the United States, and discuss the mechanisms that account for it. In section 4.2, I explain how I model the three components that shape the distribution of wealth: income and consumption, inheritance, and demography. In section 4.3, I explain what data I use and how I estimate demography and inheritance in the model. In section 4.4, I explain how to identify and estimate the main model. Section 4.5 discusses the results of the model with an application to wealth taxation, and section 4.5.3 concludes.

4.1 The Distribution of Wealth

Over the past few years, there has been a widespread regain of interest in the topic of wealth and its distribution. In the United States, we know the aggregate level of wealth from the official balance sheets compiled by the Federal Reserve. The distribution of that wealth, on the other hand, is a more complicated issue. Historically, there is no official source for the distribution of household wealth, nor is there any direct administrative data sources that we could use the calculate it. Economists, therefore, have had to devise several indirect methods to estimate wealth inequality.

There is the SCF, a triennial survey of household assets conducted by the Federal Reserve. It exists since 1949 (Schularick, Kuhn, and Steins, 2018), with publicly available data since 1962. The SCF serves as the basis for the recently published Distributional Financial Accounts (DFA) of the United States (Batty et al., 2019). However, it has only been conducted regularly and with a consistent methodology since 1989. While the SCF strongly oversamples the richest households, like all surveys it may suffer from some nonresponse and misreporting — and in fact it explicitly excludes extremely wealthy households from its sampling frame for confidentiality reasons. An alternative approach is the capitalization method (Saez and Zucman, 2016), which estimates wealth from administrative capital income tax data.[1] In this paper, I will primarily rely on estimates from the capitalization method as applied by Saez and Zucman (2016), but also use the SCF for some purposes. I will address divergences between the two sources when necessary.

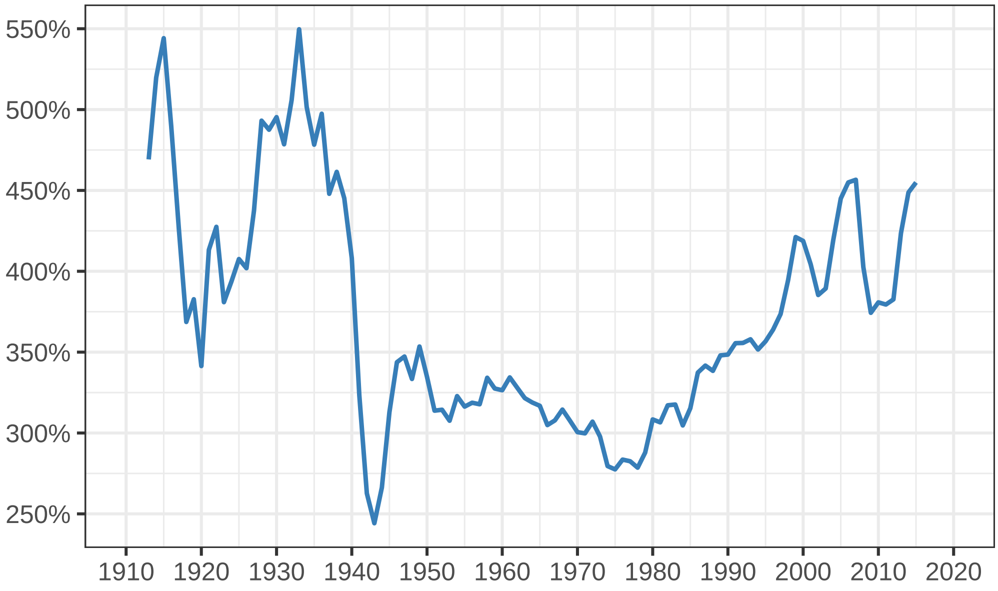
4.1.1 Empirical Facts about Wealth in the United States

Since the 1980s, household wealth in the United States has grown larger and more concentrated. From 1980 to 2015, the ratio private of private wealth to national income grew from 310% to 450% (figure 4.1a), reaching a level not seen since before the Second World War (Piketty and Zucman, 2014). Over the same period wealth inequality has also increased (figure 4.1b). Using the capitalization method, Saez and Zucman (2016) find that the top 1% owns 37% of private wealth in 2014, compared to 23% in 1980. Data from the SCF shows similar trends.

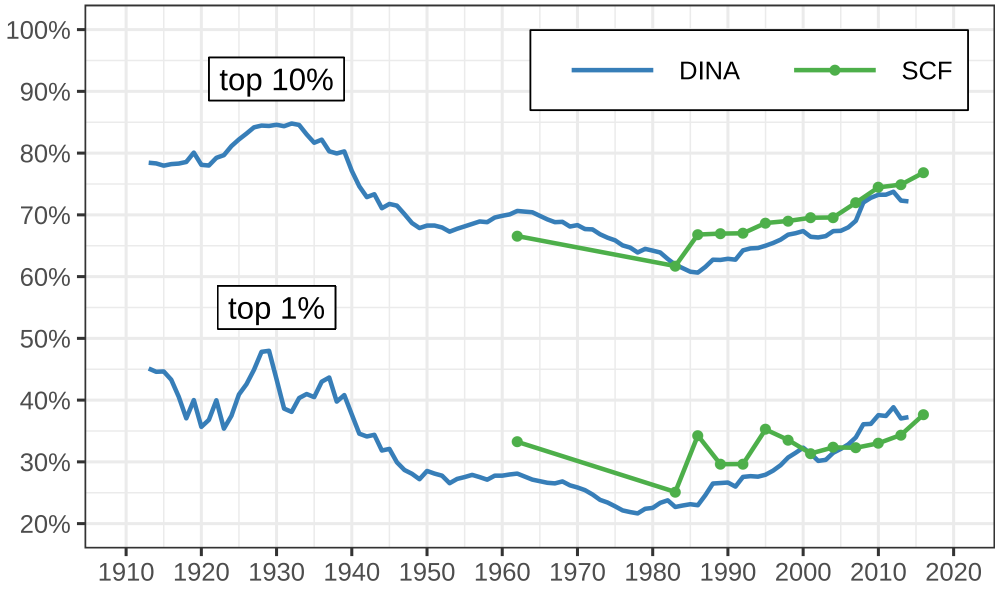
This rise in wealth inequality has been entirely driven by the top tail of the distribution. While there have been some changes for the bottom, notably a rise in the number of indebted households (Wolff, 2010), in practice these dynamics are not central to the topic of increasing inequality. As shown in figure 4.2a, the top 1% used to hold on average an amount of wealth equal to 70 times average national income in 1980. That amount now exceeds 150 times average national income. During the same period, the bottom 99% owned about 2.5 times the average national income in wealth, a value that remained relatively constant. As a result, if we were to hold constant the wealth/income ratio of the bottom 99% as in figure 4.1b, the evolution of the top 1% share would be very similar to what we observe in reality. Conversely, if we were to fix the wealth/income ratio of the top 1% at its 1980 level, inequality would not have risen at all.

The rise in wealth inequality has had consequences not only for wealth, but also for income. As shown in figure 4.3, the share of pre-tax national income owned by the top 1% nearly doubled since 1980, going from 11% to 20%. That increase can only be partially explained by the rise in labor income inequality. Since the early 2000s, the top 1% share of labor income has been mostly flat while the top 1% share of total income has kept on increasing. Similarly, up until 1980, income inequality

[1] A third approach is the estate multiplier method (Kopczuk and Saez, 2004), which estimates wealth from inheritance tax data. The estate multiplier stands out from other methods in that it does not find any increase in wealth inequality over the past decades. However, these estimates are usually considered unreliable for the recent period due to differential mortality and tax avoidance (Saez and Zucman, 2016; Kopczuk, 2015), and by now the estate tax has become too narrow to keep on applying the method.



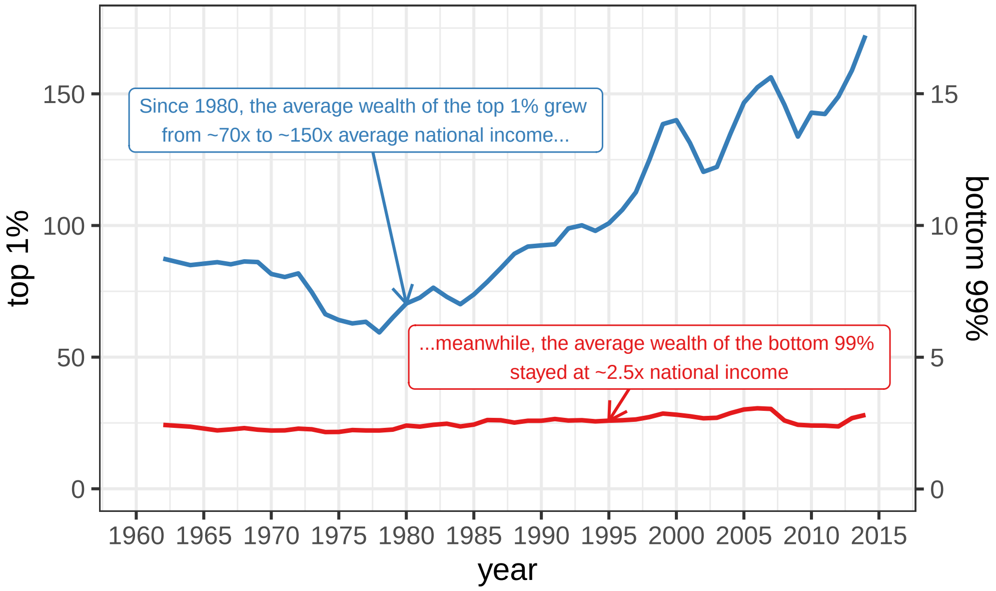
(a) Private Wealth to Income Ratio



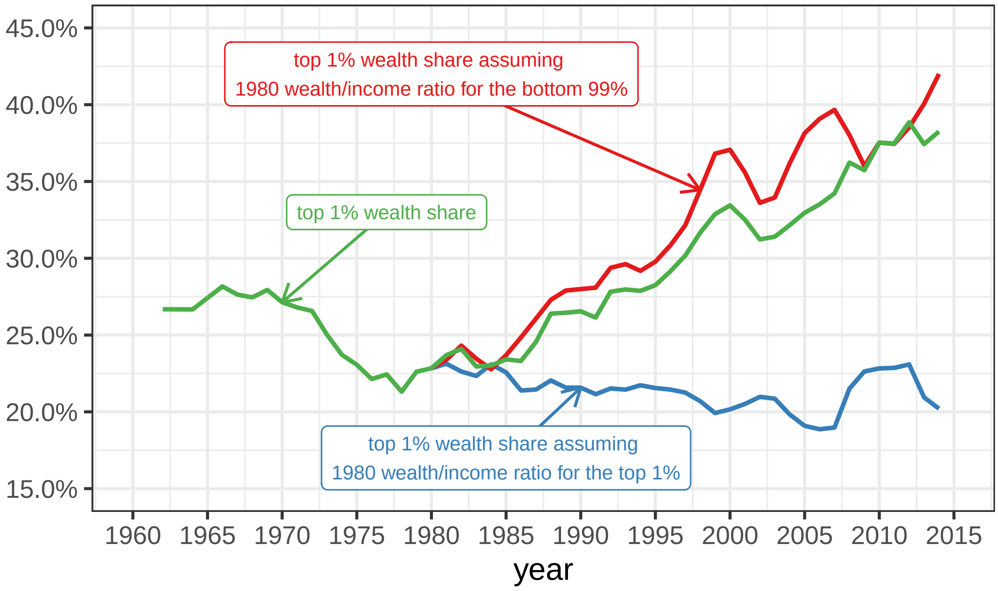
(b) Wealth inequality

Source: Figure 4.1a: DINA data from Piketty, Saez, and Zucman (2018). Figure 4.1b: author's computations using DINA data from Piketty, Saez, and Zucman (2018) and SCF data. *Note:* For the wealth-to-income ratios, the income concept for the denominator is net national income. For inequality data, the unit is the adult (20 or older) individual, and wealth is split equally between members of couples.

Figure 4.1: Private Wealth in the United States



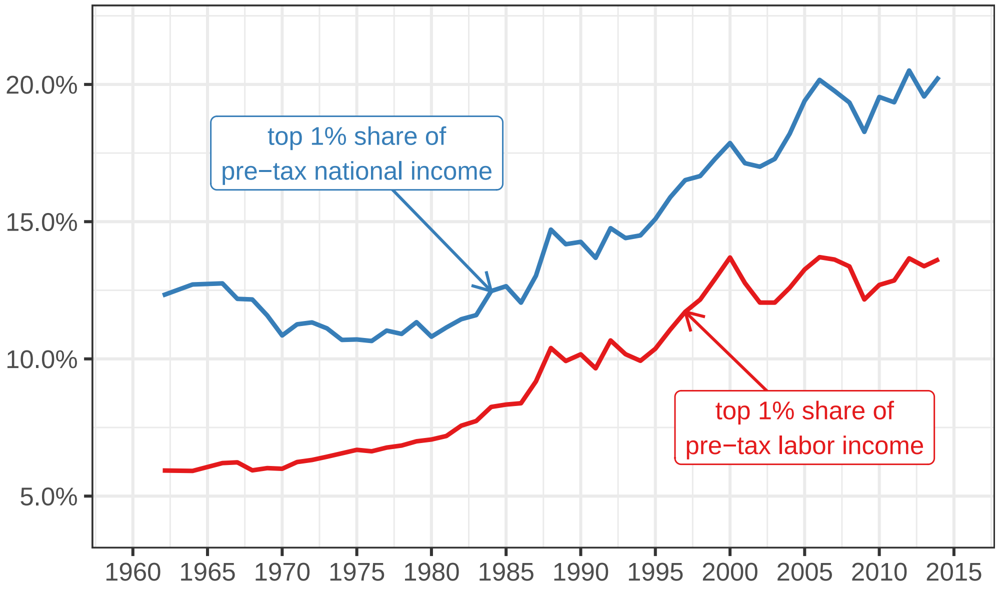
(a) Average Wealth/National Income



(b) Top 1% Share

Source: Author's computation using DINA data from Piketty, Saez, and Zucman (2018). *Note:* For the wealth-to-income ratios, the income concept for the denominator is net national income. For inequality data, the unit is the adult (20 or older) individual, and wealth is split equally between members of couples.

Figure 4.2: Wealth Inequality: The Role of the Top 1% vs. the Bottom 99%



Source: Author's computation using DINA data from Piketty, Saez, and Zucman (2018). *Note:* The unit is the adult (20 or older) individual, and income is split equally between members of couples.

Figure 4.3: Income Inequality: Labor and Total Income

was slightly decreasing even though labor income inequality was on the rise. These divergences can only be explained by changes in the distribution of capital income, which is directly related to the distribution of wealth.

4.1.2 Mechanisms That Account for Wealth Inequality

The standard models of savings that explain behavior for the bulk of the distribution do not, in general, account well for the shape of the distribution in the tail, which as we have seen in section 4.1.1 is what explains the increase in inequality. This is true of life-cycle models (Atkinson, 1971) and precautionary saving models (Carroll, 1998). The models that can realistically reproduce the distribution usually incorporate a taste for wealth, either directly (Carroll, 1998; Piketty and Zucman, 2014) or as a bequest motive (Benhabib, Bisin, and Zhu, 2011), and random shocks to preferences (Piketty and Zucman, 2014), number of children (Cowell, 1998), or rates of return (Benhabib, Bisin, and Zhu, 2011). The key feature of all these models is that wealth follows a transition equation of the form $w_{t+1} = a_t w_t + b_t$, where a_t and b_t are random. This type of multiplicative process with random shocks was studied by Kesten (1973), who showed that regardless of the exact distribution of a_t and b_t, w_t converges towards a distribution with a power-law tail.

The Kesten (1973) process justifies why, broadly speaking, power laws arise from multiplicative random shocks with frictions. However, the discrete time formalism of Kesten (1973) quickly gets intractable, so for more elaborate applications it is better to move to continuous time. In continuous time, we can model wealth accumulation as a stochastic differential equation (SDE). Like a deterministic differential equation, a SDE relates the current value of a variable to its immediate evolution (i.e. its derivative). But it is stochastic because it assumes that this relationship involves some randomness. Concretely, while a first-order ordinary differential equation for w_t may be written $\frac{\partial}{\partial t} w_t = \mu_t(w_t)$, a SDE formalizes the idea that $\frac{\partial}{\partial t} w_t = \mu_t(w_t) +$ "noise". The proper formalization of "noise" in continuous time is called a Wiener process. Traditionally, we write:

$$\mathrm{d}w_t = \mu_t(w_t)\,\mathrm{d}t + \sigma_t(w_t)\,\mathrm{d}B_t \tag{4.1}$$

to say the variance of the "noise" over a small amount of time $\mathrm{d}t$ is $\sigma_t^2(w_t)\,\mathrm{d}t$, so that the derivative of w_t is random with mean $\mu_t(w_t)$ and standard deviation $\sigma_t(w_t)$. The value $\mu_t(w_t)$ is called the drift, and $\sigma_t(w_t)$ the diffusion. Using $\mu_t(w_t) = a + bw_t$ and $\sigma_t^2(w_t) = c + dw_t^2$, we get a continuous-time analog to the Kesten (1973) process and, assuming proper parameter values, we converge to a power law. More generally, if we assume $\mu_t(w_t) \propto w_t$ and $\sigma_t(w_t) \propto w_t$ for high w_t, and that some friction prevents w_t from becoming too small, then we converge towards a power law (Gabaix, 2009). The continuous time framework allows us to abstract ourselves from short-term effects that are not relevant in practice but can seriously complicate the analysis.

While the evolution of w_t in equation (4.1) is random at the individual level, we can characterize the distribution of w_t at the aggregate level using the Fokker-Planck equation:

$$\frac{\partial}{\partial t} f_t(x) = -\frac{\partial}{\partial x}\left[\mu_t(x) f_t(x)\right] + \frac{1}{2}\frac{\partial^2}{\partial x^2}\left[\sigma_t^2(x) f_t(x)\right] \tag{4.2}$$

This is a deterministic partial differential equation that characterizes the evolution of the density f_t of w_t at time t, and which will be central the methodology of this paper because it lets us connect the way the wealth distribution evolves with the underlying parameters of the wealth accumulation process.

4.2 Theoretical Framework

Time is continuous, indexed by t. The distribution of wealth is driven by three factors: income and consumption, birth and death, and inheritance. We treat each of them in turn.

4.2.1 Income and Consumption

At each time t, the individual i holds W_{it} in wealth, consumes C_{it}, earns Z_{it} in labor income, and gets a rate of return r_{it} on their wealth (including capital gains, if any). At the individual level, wealth follows the differential equation:

$$\frac{\partial}{\partial t} W_{it} = Z_{it} + r_{it} W_{it} - C_{it}$$

Let $\bar{Y}_t$ be the average income (labor and capital), and $g_t \equiv \frac{\partial}{\partial t}\bar{Y}_t / \bar{Y}_t$ is the growth rate of average income. Define $w_{it} \equiv W_{it}/\bar{Y}_t$, $z_{it} \equiv Z_{it}/\bar{Y}_t$ and $c_{it} \equiv C_{it}/\bar{Y}_t$. To stationarize the dynamics of wealth, I will be working with these normalized quantities. The evolution of wealth becomes:

$$\frac{\partial}{\partial t} w_{it} = z_{it} + (r_{it} - g_t) w_{it} - c_{it}$$

Define $y_{it} \equiv z_{it} + (r_{it} - g_t) w_{it}$, so that $\frac{\partial}{\partial t} w_{it} = y_{it} - c_{it}$. I now introduce stochasticity to the income process and the consumption process. Assume, without much loss of generality, that over a small interval of time $[t, t + \mathrm{d}t]$, income (y_{it}) and consumption (c_{it}) are random with mean $\nu_{it}\,\mathrm{d}t$ and $\mu_{it}\,\mathrm{d}t$, and variance $\tau_{it}^2\,\mathrm{d}t$ and $\sigma_{it}^2\,\mathrm{d}t$ respectively (ν_{it}, μ_{it}, τ_{it}^2 and σ_{it}^2 being themselves random processes). Then wealth evolves according to the SDE:

$$\mathrm{d}w_{it} = [\nu_{it} - \mu_{it}]\,\mathrm{d}t + [\tau_{it}^2 + \sigma_{it}^2]^{1/2}\,\mathrm{d}B_{it}$$

where B_{it} is a Wiener process.[2] That SDE has stochastic coefficients, which prevents us from directly applying the Fokker-Planck equation (4.2). To avoid the need to explicitly model the income and consumption processes separately, I apply a result of stochastic calculus known as Gyöngy's (1986) theorem, which allows us to drastically reduce the dimensionality of the problem to solely focus on wealth.

Theorem 7 (Gyöngy, 1986)**.** *Let $\boldsymbol{X}_t$ be a n-dimensional stochastic process satisfying:*

$$\mathrm{d}\boldsymbol{X}_t = \boldsymbol{\alpha}_t\,\mathrm{d}t + \boldsymbol{\beta}_t\,\mathrm{d}\boldsymbol{B}_t$$

where $\boldsymbol{\alpha}_t$ and $\boldsymbol{\beta}_t$ are bounded and nonanticipative $n \times 1$ and $n \times m$ stochastic processes, respectively, $\boldsymbol{\beta}_t\boldsymbol{\beta}_t'$ is uniformly positive definite, and $\boldsymbol{B}_t$ is a m-dimensional Wiener

[2] This formulation implicitly assumes that income and consumption are uncorrelated conditional on wealth. But the analysis still holds if they are. Define $\rho_{it} \equiv \mathrm{Cov}(y_{it}, c_{it})$. Then the equation holds if we redefine σ_{it}^2 to include covariance as $\sigma_{it}^2 + 2\rho_{it}$.

process. Then there is a Markov process $\boldsymbol{Y}_t$ satisfying:

$$\mathrm{d}\boldsymbol{Y}_t = \boldsymbol{a}_t(\boldsymbol{Y}_t)\,\mathrm{d}t + \boldsymbol{b}_t(\boldsymbol{Y}_t)\,\mathrm{d}\boldsymbol{B}_t$$

where $\boldsymbol{X}_t$ and $\boldsymbol{Y}_t$ have the same marginal distributions for each t. We can construct $\boldsymbol{Y}_t$ by setting:

$$\boldsymbol{a}_t(\boldsymbol{y}) = \mathbb{E}[\boldsymbol{\alpha}_t|\boldsymbol{X}_t = \boldsymbol{y}] \qquad\qquad \boldsymbol{b}_t(\boldsymbol{y}) = \mathbb{E}[\boldsymbol{\beta}_t\boldsymbol{\beta}_t'|\boldsymbol{X}_t = \boldsymbol{y}]^{1/2}$$

Gyöngy's (1986) theorem implies that we can write:

$$\mathrm{d}w_{it} = [\nu_t(w_{it}) - \mu_t(w_{it})]\,\mathrm{d}t + [\tau_t^2(w_{it}) + \sigma_t^2(w_{it})]^{1/2}\,\mathrm{d}B_{it} \tag{4.3}$$

where $\nu_t(w)$, $\mu_t(w)$ are the means of income and consumption conditional on wealth, and $\tau_t^2(w)$, $\sigma_t^2(w)$ are the variances of income and consumption conditional on wealth.[3]

The Fokker-Planck equation associated to (4.3) and which describes the density of wealth f_t is:

$$\frac{\partial}{\partial t}\,f_t(w) = -\frac{\partial}{\partial w}\left[(\nu_t(w) - \mu_t(w))f_t(w)\right] + \frac{1}{2}\frac{\partial^2}{\partial w^2}\left[(\tau_t^2(w) + \sigma_t^2(w))f_t(w)\right] \tag{4.4}$$

4.2.2 Birth and Death

I extend the model above with a birth and death process. People die randomly according to year, age and sex-specific fertility rates. Let g_t be the density of wealth weighted by these mortality rates. Other people appear with a random initial endowment drawn from a distribution with density h.

Let β_t and δ_t be the overall birth and death rate. The total population N_t grows at a rate $n_t = \dot{N}_t/N_t = \beta_t - \delta_t$. Adding this process turns equation (4.4) into:

$$\begin{aligned}\frac{\partial}{\partial t}\,f_t(w) = & \underbrace{-\frac{\partial}{\partial w}\left[(\nu_t(w) - \mu_t(w))f_t(w)\right] + \frac{1}{2}\frac{\partial^2}{\partial w^2}\left[(\tau_t^2(w) + \sigma_t^2(w))f_t(w)\right]}_{\text{income and consumption}} \\ & + \underbrace{\beta_t h(w) - \delta_t g_t(w) - n_t f_t(w)}_{\text{birth and death}}\end{aligned}$$

[3] See appendix D.2.1 for details on how to arrive at that result.

4.2.3 Inheritance

The wealth of people who die gets redistributed to their spouse or children, after payment of the estate tax, if any. Contrary to income that can be viewed as a continuous flow, inheritance is punctual and introduces a discontinuity in the evolution of wealth. So I model it as a jump process.

The inheritance process is partially connected to the demographic process: it redistributes the wealth of people who die in a given year to their next of kin. The way that inheritance is redistributed depends on the estate tax and additional parameters that capture intergenerational mobility (i.e. do wealthier people inherit more?) I explain how I fully model the process in section 4.3.2. For now I take the joint distribution of inheritance and wealth as given.

With a probability $\pi_t(w)$, people see their wealth jump from w to $w + \lambda$ where λ is the amount of inheritance received, net of taxes. Let $s_t(\lambda|w)$ be the density of the value of the inheritance, conditional on the value of wealth, and conditional on receiving inheritance. We can model the jump process as a death with rate $\pi_t(w)$ and as an injection with rate $\int \pi_t(w-\lambda)f_t(w-\lambda)s_t(\lambda|w-\lambda)\,\mathrm{d}\lambda$:

$$\begin{aligned}\frac{\partial}{\partial t}f_t(w) = & \underbrace{-\frac{\partial}{\partial w}\big[(\nu_t(w)-\mu_t(w))f_t(w)\big]+\frac{1}{2}\frac{\partial^2}{\partial w^2}\big[(\tau_t^2(w)+\sigma_t^2(w))f_t(w)\big]}_{\text{income and consumption}} \\ & + \underbrace{\beta_t h(w) - \delta_t g_t(w) - n_t f_t(w)}_{\text{birth and death}} \\ & + \underbrace{\int \pi_t(w-\lambda)f_t(w-\lambda)s_t(\lambda|w-\lambda)\,\mathrm{d}\lambda - \pi_t(w)f_t(w)}_{\text{inheritance}}\end{aligned} \tag{4.5}$$

4.3 Data, Demography and Inheritance

4.3.1 Demography

I compute the entire demography of the United States from 1850 to 2100. Although the income and wealth data does not start until 1962, the model requires demographic data that starts much earlier. Indeed, I need to simulate how wealth gets transmitted from one generation to the next. Therefore, if a supercentenarian dies in the 1960s, I have to be able to simulate their entire life history to know how many live children they have, and how old they are. For all years and all ages, I estimate data on the population structure by age and sex, mortality (i.e. life tables), fertility (for both sexes) and intergenerational ties (age and sex of children). Sometimes, data is only available by age groups (e.g. of five years) or a subset of years (e.g. every ten years). Whenever necessary, I interpolate estimates with a monotonic cubic spline (Fritsch and Carlson, 1980) to get data for every single year and age.

Population by Age and Sex Before 1900, I directly estimate the population pyramid using decennial census microdata from the IPUMS USA database (Ruggles et al., 2019). From 1900 to 1932, I use the National Intercensal Tables from the United States Census Bureau. From 1933 to 2016, I use population estimates from the Human Mortality Database.[4] After 2016, I use the projections from the World Population Prospects (United Nations, 2017).

Life Tables Before 1900, I use the historical life tables from Haines (1998). From 1900 to 1932, I use the Human Life Table Database, and from 1933 to 2016, life tables from the Human Mortality Database. After 2016, I rely on projections from the World Population Prospects (United Nations, 2017). All the tables are broken down by sex.

Age-Specific Fertility Rates by Birth Order I estimate age-specific fertility rates by birth order, for both sexes. For women, they are directly available from 1933 to 2016 from the Human Fertility Database. From 1917 to 1932, I use data from the Human Fertility Collection. That same source provides fertility rates until going back to 1895–1899, but without the breakdown by birth order. Therefore, before 1917, I assume that the birth order composition remains constant. Before 1895, there is no age-specific data available, so I use the data on total fertility rate and rescale the age profile from 1895–1899 to that value.[5]

Unlike female fertility rates, male fertility rates are not a standard demographic indicator, so they are not directly available from any source. To estimate them, I combine the age-specific female fertility rates with the joint distribution of the age of opposite-sex couples since 1850 calculated using the decennial census microdata from the IPUMS USA database (Ruggles et al., 2019).

Age and Sex of Children I simulate the distribution of the number, age and sex of living children for in each year after 1962 (when income and wealth data starts), every age and both sexes, which allows me to realistically model how wealth gets transmitted from one generation to the next. To that end, I combine all of the data

[4] See `https://www.mortality.org/hmd/USA/DOCS/ref.pdf` for detailed primary sources.

[5] See Gapminder: `https://www.gapminder.org/news/children-per-women-since-1800-in-gapminder-world/`

above. I make every person have children randomly over their past lifetime according to year, age and sex-specific fertility rate. Because I have the breakdown by birth order, I can take into account how the decision to have another child depends on the number of children that one already has. Then, I make each child go through life and die at random according to their year, age and sex-specific mortality rate. As result, I can tie every individual in the database to fictitious descendants that are, on average, representative the true composition of descendants.

4.3.2 Inheritance and the Estate Tax

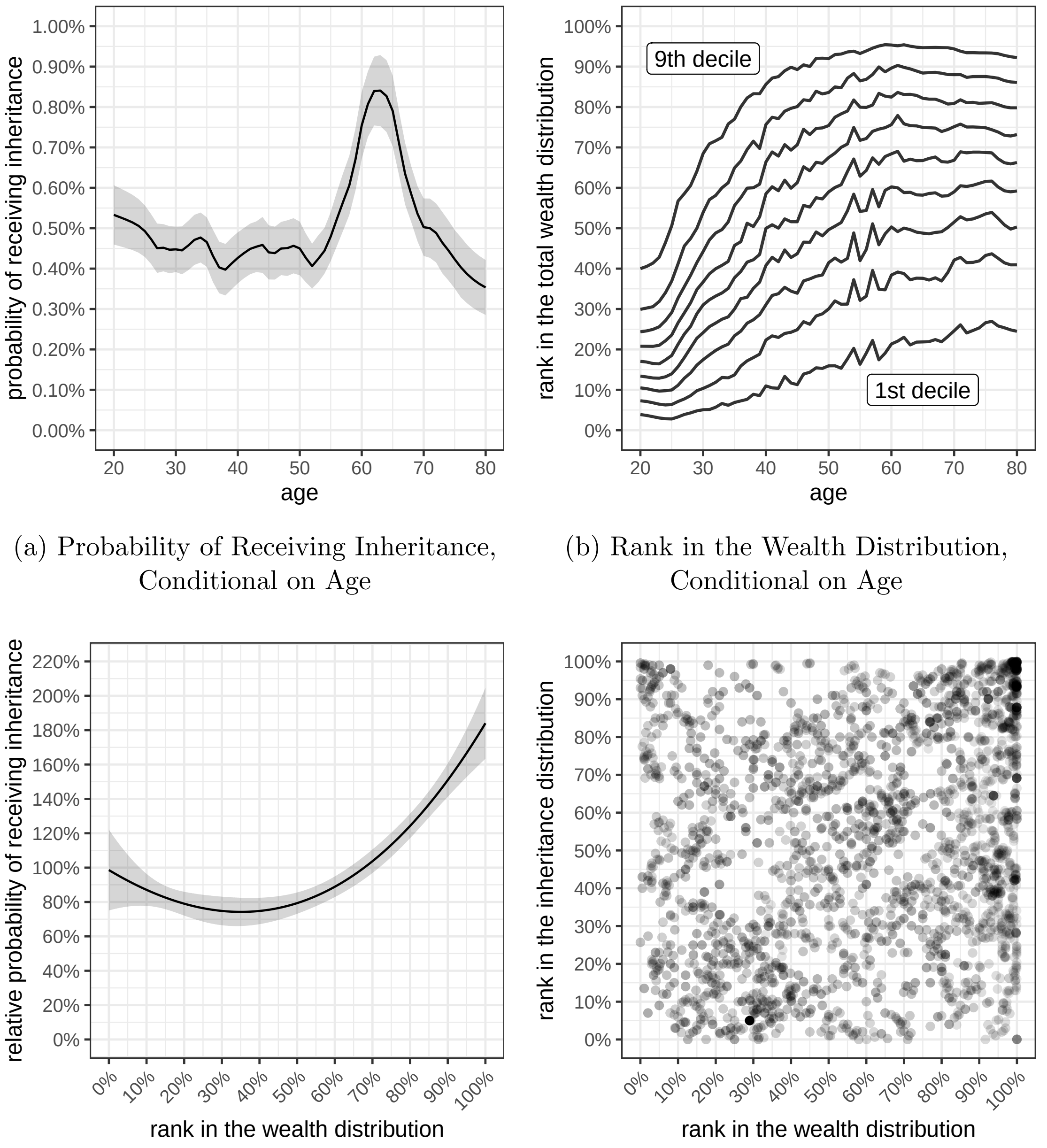
Part of the inheritance process is determined by the demographics and the distribution of wealth, while other parts have to be modeled separately. I assume that people die at random, conditional on their age and sex, so that the distribution of inheritances correspond to the distribution of wealth, weighted by mortality rates. I then assume that the wealth of decedents is either redistributed to their spouse (if any) or to their descendants (if they have no living spouse), after payment of the estate tax. The age and sex of decedents are given by the demography (see section 4.3.1). I assume that inheritance is split equally between children, as is the norm in the United States (Menchik, 1980).

While the demographic aspect of inheritance is endogenously determined by demography, I still need to model separately how wealth gets distributed for a given age and sex. This captures intergenerational wealth mobility in the sense that wealthier people might also have wealthier parents and thus inherit more. There are two aspects to this question: the extensive margin (how likely are you to receive inheritance in a given year?) and the intensive margin (how much inheritance do you receive?) To address this question, I use data from the SCF, which has been recording inheritance consistently since 1989. Note that because the probability of receiving inheritance in a given year is very low overall (about half a percent, see figure 4.4a), I have to pool all the 1989–2016 waves in order to get sufficient sample sizes.

Extensive Margin Let $D_i = 1$ if individual i receives inheritance, and $D_i = 0$ otherwise. Let A_i be their age, and W_i their wealth. Assume that:

$$\mathbb{P}\{D_i = 1 | A_i = a, W_i = w\} = \mathbb{P}\{D_i = 1 | A_i = a\} \phi(F_{A_i=a}(w)) \tag{4.6}$$

where $F_{A_i=a}$ is the cumulative distribution function (CDF) of wealth conditional on age, and $\int_0^1 \phi(r) \, \mathrm{d}r = 1$. By construction, the expected value of the right-hand side of (4.6) conditional on age is equal to $\mathbb{P}\{D_i = 1 | A_i = a\}$ so that the



(a) Probability of Receiving Inheritance, Conditional on Age

(b) Rank in the Wealth Distribution, Conditional on Age

(c) Relative Probability of Inheritance, Conditional on Age

(d) Joint Ranks in the Wealth and the Inheritance Distribution, Conditional on Age

Source: Author's computation using the SCF (1989–2016). Gray ribbons correspond to the 95% confidence intervals. In figure 4.4d, opacity is proportional to the weight of observations.

Figure 4.4: Modeling of Inheritance

specification makes probabilistic sense.[6] Note that $F_{A_i=a}(w)$ is the rank of w in the wealth distribution (conditional on age), which is how we can make the formula (4.6) consistent regardless of the shape of the wealth distribution.

The value of $\mathbb{P}\{D_i = 1|A_i = a\}$ is determined by demography, so we only need to estimate ϕ. I start by calculating a rank in the wealth distribution conditional on age by running nonparametric quantile regression of wealth on age for every percentile (see figure 4.4b). I then regress the dummy D_i for having received inheritance on that rank, multiplied by $\mathbb{P}\{D_i = 1|A_i = a\}$. I use ordinary least squares (OLS) and a cubic polynomial with coefficients constrained so that its integral over $[0, 1]$ equals one (see figure 4.4c). As we can see, even after partialling out the effect of age, wealthier people still experience a higher probability of receiving inheritance. I use that polynomial as my estimate of ϕ.

Intensive Margin I account for the intensive margin by modeling the joint distribution of the ranks in the wealth distribution and the inheritance distribution (i.e. the copula), conditional on age and on having received inheritance. I take the subsample of inheritance receivers and calculate their rank in the wealth and the inheritance distribution using nonparametric quantile regression as I did for the extensive margin.

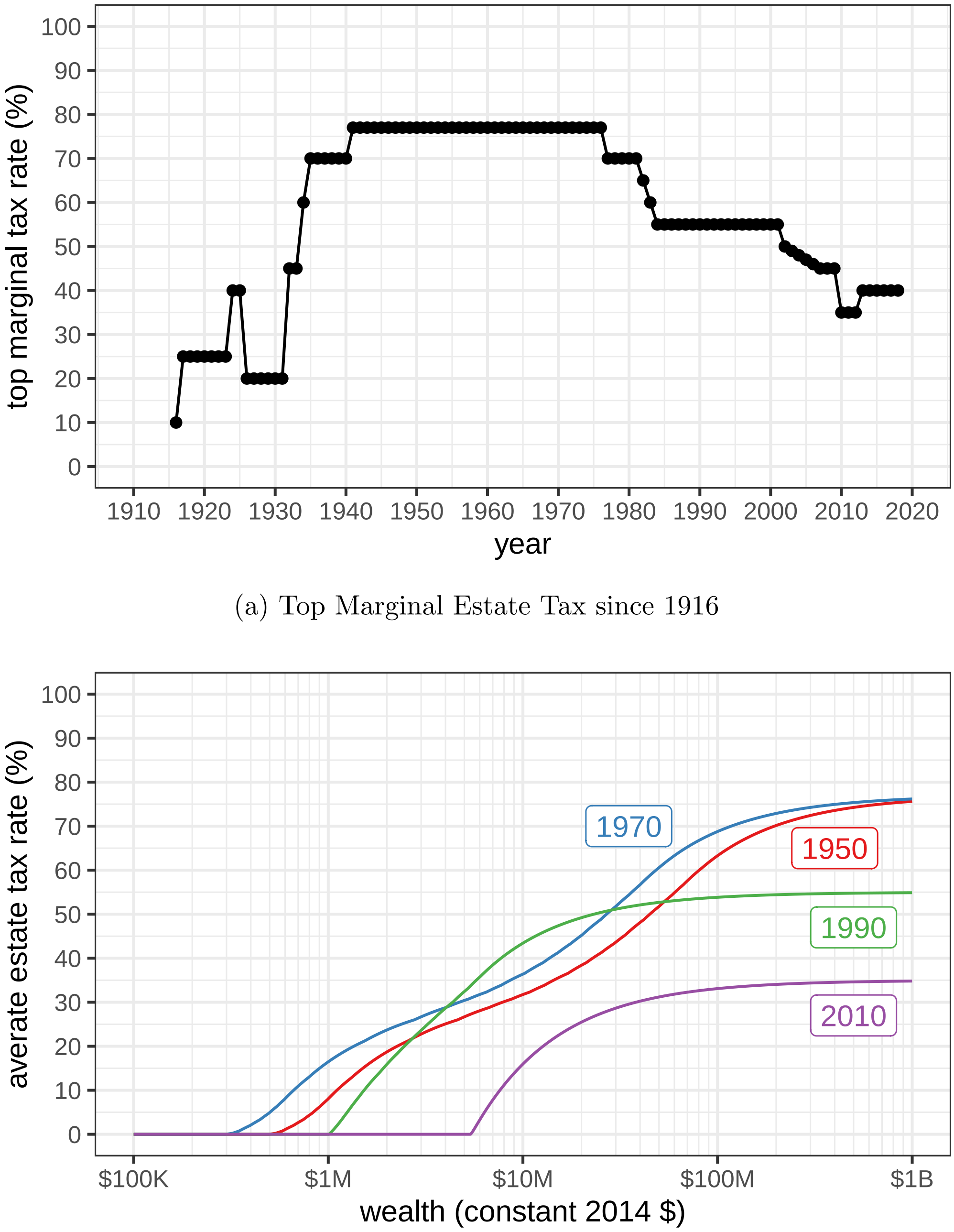
The dependence between the two ranks is weak, but significant (see figure 4.4d): their Kendall's tau is equal to 7.2%. I represent this dependency parametrically using a bivariate copula. I select the most appropriate model out of a large family of 15 single-parameter copulas by finding the best fit according to the Akaike information criterion (AIC), which is the Joe copula.[7][8] I estimate its parameter so as to match the empirical value for Kendall's tau.

Estate Tax I account for the federal estate tax using the complete estate tax schedule and exemption amount for each year. The top marginal estate tax rate has followed a clear inverted U-shaped pattern over the 20th century (figure 4.5a), having been reduced by half since its mid-century peak. However, the changes to the overall progressivity of the estate tax are more ambiguous (figure 4.5b). While the top marginal tax rate was very high in the 1950s, the top bracket did not kick in

[6] It is the direct result of a change of variable $r = F_{A_i=a}(w)$ and using the fact that $\frac{\partial}{\partial w} F_{A_i=a}(w) = f_{A_i=a}(w)$, so that $\int_{-\infty}^{+\infty} \phi(F_{A_i=a}(w)) f_{A_i=a}(w) \, \mathrm{d}w = \int_0^1 \phi(r) \, \mathrm{d}r = 1$.

[7] The list of copulas includes the Gaussian copula, Student's t copula, the Clayton copula, the Gumbel copula, the Frank copula, the Joe copula, and rotated versions of these copulas.

[8] The Joe copula has the parametric form $C_\theta(u, v) = 1 - \left[(1-u)^\theta + (1-v)^\theta - (1-u)^\theta (1-v)^\theta\right]^{1/\theta}$.



(a) Top Marginal Estate Tax since 1916

(b) Average Estate Tax Rate, by Wealth

Source: Author's computation using the tax schedules of the federal estate tax.

Figure 4.5: Estate Tax

until extremely high levels of wealth. The 1980s reforms significantly reduced the top tax rate and increased the exemption amount, so that by 1990, the very top and the upper middle of the wealth distribution were facing lower average tax rates. But individual owning about $10M of wealth were actually facing slightly higher average tax rates. By now, however, the estate has been lowered so much that its profile is unambiguously less progressive than in the 1950s.

4.3.3 Income and Wealth

For the income and wealth data, I primarily rely on the DINA public microdata from Piketty, Saez, and Zucman (2018). These files are annual (except for 1963 and 1965) since 1962. Each observation corresponds to an adult individual (20 or older), and each variable correspond to an item of the national accounts, that is distributed to the whole adult population. These files distribute the entirety of the income and wealth of the United States. The public version regroups observations for anonymity, so it has smaller sample sizes than the one they use internally, and does not exactly reproduce results from their more complete internal files (Saez and Zucman, 2018). The discrepancies, however, are small.

This data has several advantages. It provides distributional estimates that are consistent with macroeconomic aggregates. It has rather large samples (from about 35 000 in the 1960s to about 65 000 today), with oversampling of the richest. And because it is based on tax data, it captures the top tail of the distribution well. It does have some drawbacks, though. First, it has limited socio-demographic information: in particular, age information is only available in the form of very broad age groups. Second, it estimates wealth using the capitalization method: that is, it assumes that everyone gets the same rate of return from the same type of asset. Under the right assumptions (Saez and Zucman, 2016), that method provides accurate estimates of the distribution of wealth, and of average income conditional on wealth. But it almost certainly underestimate the variance of capital income conditional on wealth. Third, the data does not include capital gains, because they are not part of national income as defined by the national accounts. For these reasons, I make some adjustments and imputations to these data, using the SCF and national accounts.

I use post-tax national income as my income concept of reference. It corresponds to income after all taxes and transfers. It also distributes government expenditures and the income of the corporate sector to individuals, so as to sum up to net national income.

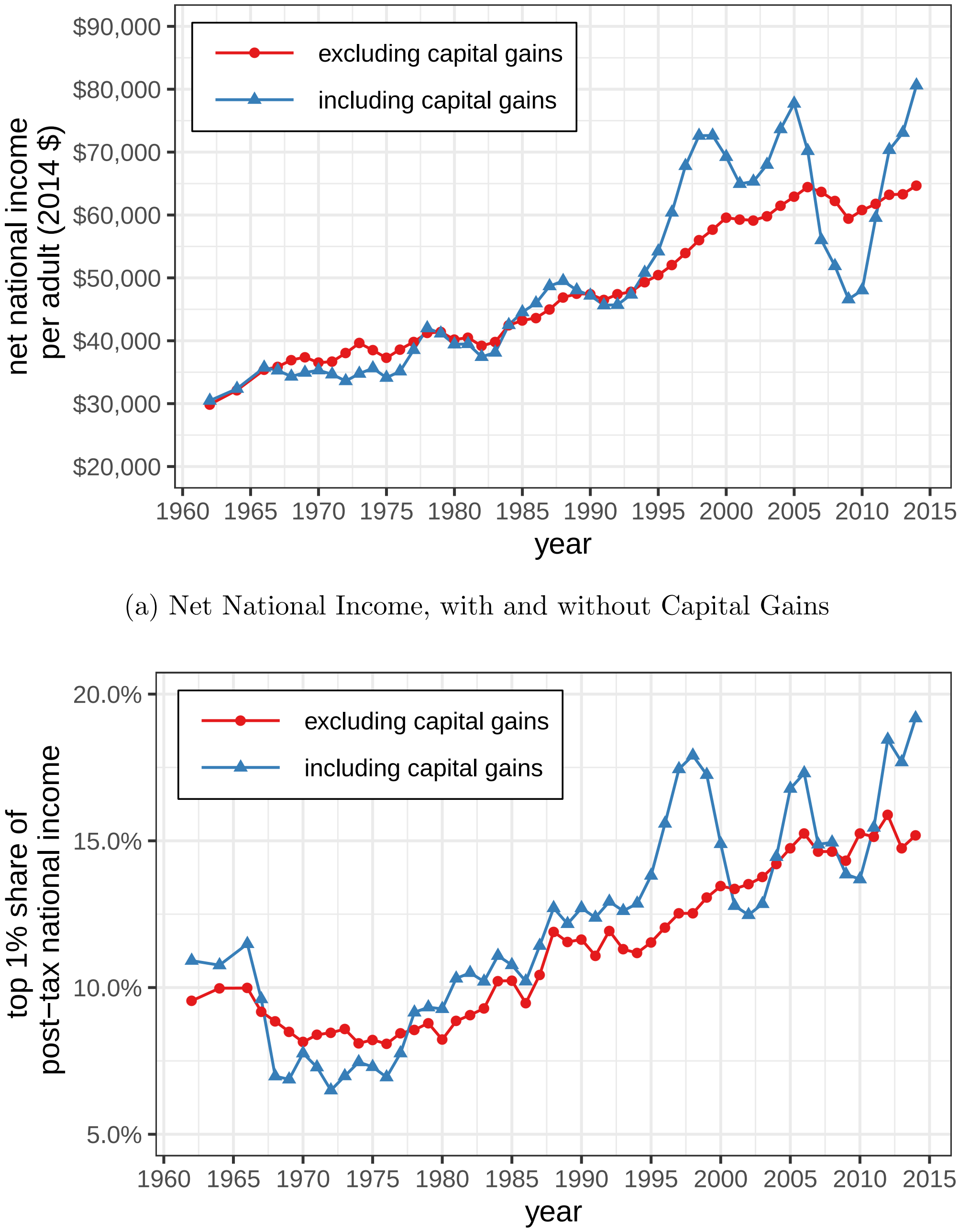
Capital Gains We can measure capital gains when they accrue to individuals, or when they are realized. For our purpose, accrued capital gains are more useful than realized ones, because they are the one that reconcile changes in the value of the balance sheet with national income and savings. Whether a capital gain is realized now or later, on the other hand, is the result of various tax and economic incentives that not relevant here and does not correspond to any meaningful economic aggregate.

The DINA data only records taxable capital gains, which is essentially a measure of realized capital gains. These are a poor proxy for accrued capital gains (Alstadsæter et al., 2017). Instead, I estimate them individually using the capitalization approach of Robbins (2018). I retrieve the rate of capital gains by year and asset type from the national accounts (Piketty, Saez, and Zucman, 2018, table TSD1 in appendix). Then, I assume for a given asset type, everyone gets the same rate of capital gains. By construction, these micro-level estimates of capital gains are consistent with macro totals. Their distribution follows the logic of the Saez and Zucman (2016) capitalization method.[9] Robbins (2018) provides a thorough discussion of why that measure is more appropriate to analyze the role of asset prices changes to inequality and the economy.

National income including capital gains can be quite volatile (figure 4.6a), but on average their inclusion matters on several fronts. Robbins (2018) shows that their inclusion overturns certain stylized facts about the United States economy (such as the long run decline of saving rates) and strengthen others (such as the rising capital share and increase of income inequality). As shown in figure 4.6b, capital gains were dampening the top 1% share of post-tax national income during most of the 1970s, but since then they have consistently increased it.

Wealth by Age The age information in the DINA data is very limited so I cannot use it. Instead, I import it from the SCF and demographic estimates using constrained statistical matching. I calculate the rank in the wealth distribution in both the DINA and the SCF data, and the rank in the age distribution by sex and household type (single or couple) in the SCF data. Then, I match the DINA observations one by one to SCF observations based on their wealth rank to give them

[9] Although the income measure in the DINA data does not include capital gains, it does distribute income from the corporate sector to the owners of capital, with may partly account for changes in asset prices. My measure of capital gains is net of retained earnings, so that there is no double counting.



(a) Net National Income, with and without Capital Gains

(b) Post-tax National Income Share, with and without Capital Gains

Source: Author's computations using the public DINA microdata and table TSD1 (online appendix) from Piketty, Saez, and Zucman (2018). *Note:* The unit of analysis is the adult individual (20 or older). Income is split equally between members of couples. Capital gains are estimated assuming a constant rate of capital gains by asset type. Rates of capital gains by asset types smoothed using a 5-year moving average.

Figure 4.6: The Impact of Capital Gains on National Income and Inequality

a rank in the age distribution.[10] Finally, I use the population structure from the demographic data to attribute an age to every DINA observation. By construction, the method preserve the wealth distribution in DINA, the population by age and sex from demographic sources, and the copula between wealth and age from the SCF.

Variance of Income by Wealth Because the capitalization approach in the DINA data assumes a fixed rate of return by asset type, it is likely to understate the variance of capital income conditional on wealth. Indeed, it will only account for the "between assets" component of total variance, not the "within assets" component. Given that the variance of income conditional on wealth is one of the drivers of the dynamic of wealth, I make an adjustment the DINA estimates using the SCF.

Since 1989 (previous waves provide insufficient data due to lack of oversampling), the standard deviation of the income/wealth ratio at the very top of the distribution is equal to 10.7%, compared to 5.2% in DINA. I use this difference to compute a "within assets" variance component by wealth that I add to the DINA estimates of the income variance conditional on wealth. Note that the survey estimate of this variance is by no means perfect, and is in fact likely to be inflated for two reasons. First, measurement error for either income and wealth might increase the spread of the income/wealth ratio in the survey for spurious reasons. Second, the income in the SCF refers to the previous year, while the wealth refers to the time of the interview: this disconnect introduces additional noise that will have a tendency to also increase the variance of the income/wealth ratio. However, I stress that by construction this adjustment can only affect the interpretation of some parameters of the model, not the overall dynamics of wealth. Indeed, the evolution of wealth ultimately depends on $\sigma_t^2(w) + \tau_t^2(w)$, the sum of the variance of consumption and income. Therefore, as will be explained in section 4.4, in effect the model will directly estimate the overall variance $\sigma_t^2(w) + \tau_t^2(w)$, and then estimate $\sigma_t^2(w)$ by subtracting $\tau_t^2(w)$. To the extent that we overestimate the variance of income, we will underestimate the variance of consumption, and *vice versa*. In any case, the results of section 4.5 will be unaffected.

[10] Note that both datasets are weighted, so that observations end up being duplicated and partially matched to one another. When the samples contain M and N observations respectively, the resulting dataset contains at most $M + N - 1$ observations.

4.4 Identification and Estimation

For concision, define in equation (4.5):

$$\phi_t(w) \equiv \beta_t h(w) - \delta_t g_t(w) - n_t f_t(w) \qquad \text{(the birth/death effect)}$$

$$\psi_t(w) \equiv \int \pi_t(w-\lambda) f_t(w-\lambda) s_t(\lambda|w-\lambda) \; \mathrm{d}\lambda - \pi_t(w) f_t(w) \quad \text{(the inheritance effect)}$$

So that the Fokker-Planck equation (4.5) becomes:

$$\begin{aligned}\frac{\partial}{\partial t} f_t(w) = &- \frac{\partial}{\partial w} \big[(\nu_t(w) - \mu_t(w)) f_t(w)\big] \\ &+ \frac{1}{2} \frac{\partial^2}{\partial w^2} \big[(\tau^2(w) + \sigma^2(w)) f_t(w)\big] + \phi_t(w) + \psi_t(w)\end{aligned}$$

I will use uppercase letters to denote integrated quantities, in particular:

$$F_t(w) = \int_{-\infty}^{w} f_t(s) \; \mathrm{d}s \qquad \Phi_t(w) = \int_{-\infty}^{w} \phi_t(s) \; \mathrm{d}s \qquad \Psi_t(w) = \int_{-\infty}^{w} \psi_t(s) \; \mathrm{d}s$$

4.4.1 Identification

General Result I integrate the Fokker-Planck equation with respect to w, borrowing a suggestion from Lund, Hubbard, and Halter (2014) in the context of physical chemistry.[11] After reordering terms, I get:

$$\begin{aligned}\frac{\frac{\partial}{\partial t} F_t(w)}{f_t(w)} - \frac{\Phi_t(w)}{f_t(w)} - \frac{\Psi_t(w)}{f_t(w)} + \nu_t(w) - \frac{1}{2} \frac{\partial}{\partial w} \tau_t^2(w) - \frac{1}{2} \tau_t^2(w) \frac{\frac{\partial}{\partial w} f_t(w)}{f_t(w)} = \\ \mu_t(w) + \frac{1}{2} \frac{\partial}{\partial w} \sigma_t^2(w) + \frac{1}{2} \sigma_t^2(w) \frac{\frac{\partial}{\partial w} f_t(w)}{f_t(w)} \quad (4.7)\end{aligned}$$

The left-hand side of the equation only contains estimable quantities, while the right-hand side is a linear function of $\frac{\partial}{\partial w} f_t(w)/f_t(w)$ whose slope and intercept relate to the unknown parameters $\mu_t(w)$ and $\sigma_t^2(w)$.

Therefore, if these quantities are stable over time, then for a level of wealth w, we should expect $\frac{\partial}{\partial w} f_t(w)/f_t(w)$ and the left side of the equation to fall alongside a straight line. Assuming that there is some variability of both sides of the equation, we are able to estimate the parameters of interest simply by fitting a line. This leads to the following result.

[11] To integrate the equation, we must be able to invert the time derivative with the integral sign, which is allowed either it we assume that the support of wealth is bounded from below, or if the density of wealth is Lipschitz-continuous (i.e. has a bounded derivative).

Theorem 8 (Identifiability of the Model)**.** *Assume that there is at least two dates, t_1 and t_2, for which:*

(i) For all w, we observe all the quantities in (4.7), except $\mu_t(w)$, $\sigma_t^2(w)$ and $\frac{\partial}{\partial w}\sigma_t^2(w)$.

(ii) The parameters $\mu_t(w)$ and $\sigma_t^2(w)$ are the same in t_1 and t_2: for all w, $\mu_{t_1}(w) = \mu_{t_2}(w) = \mu(w)$ and $\sigma_{t_1}^2(w) = \sigma_{t_2}^2(w) = \sigma^2(w)$.

(iii) The distribution is different between t_1 and t_2, such that $\frac{\partial}{\partial w} f_{t_1}(w)/f_{t_1}(w) \neq \frac{\partial}{\partial w} f_{t_2}(w)/f_{t_2}(w)$ almost surely.

Then the functions $\mu(w)$ and $\sigma^2(w)$ that satisfy (4.7) are unique, i.e. the model is identified.

The assumptions required to estimate the model are relatively innocuous. Assumption (i) states that we can observe, or at least separately estimate, all the relevant quantities except consumption, which we seek to identify. Assumption (ii) states that we need some stability in the consumption process over time to be able to estimate it. And assumption (iii) states that we need some variability in the distribution of wealth, so that we cannot already be at the steady state. This is clearly the for the United States since the 1960s. In theory only two observations are needed to estimate the model. In practice it is better to have many more. First, because we need to estimate the time derivative of the CDF of wealth in (4.7), which requires several data points. Second, because there will always be some measurement error for the different quantities, which can be averaged out when using many data points.

Interpretation in a Simplified Case To better understand the dynamics implied by the estimating equation, consider the following simplified case, which nonetheless capture all the main intuitions of the more complete setting. Ignore the role of demographics ($\Phi_t(w) = 0$), inheritance ($\Psi_t(w) = 0$) and the conditional variance of income ($\tau_t^2(w) = 0$). Consider a high level of wealth w, and assume that at these levels the mean and the standard deviation of consumption are proportional to wealth ($\mu_t(w) \equiv \mu w$ and $\sigma_t(w) \equiv \sigma w$). Define the conditional income-to-wealth ratio $\gamma(w) \equiv \nu_t(w)/w + g$ (note the apparition of the economy's growth rate that was previously included in $\nu_t(w)$ because wealth was normalized by average income). After dividing both sides by w, the estimating equation (4.7) simplifies to:

$$\frac{\frac{\partial}{\partial t}F_t(w)}{wf_t(w)} + \gamma(w) - g = \mu - \sigma^2\left(-\frac{1}{2}\frac{w\frac{\partial}{\partial w}f_t(w)}{f_t(w)} - 1\right) \tag{4.8}$$

Under these circumstances, the top tail converges towards a power law (Gabaix, 2009). Thus, assume that wealth is Pareto-distributed with Pareto coefficient $\alpha > 1$, i.e. $f_t(w) \propto x^{-\alpha-1}$. Then $-\frac{1}{2}w\frac{\partial}{\partial w}f_t(w)/f_t(w) - 1 = (\alpha - 1)/2 > 0$ can serve as a proxy for inequality: the higher it is, the lower inequality.[12] On the left-hand side, inequality increases when $\frac{\partial}{\partial t}F_t(w)/(wf_t(w))$ is negative, and decreases otherwise. We can write equation (4.8) as $Y_t(w) = \mu - \sigma^2 X_t(w)$.

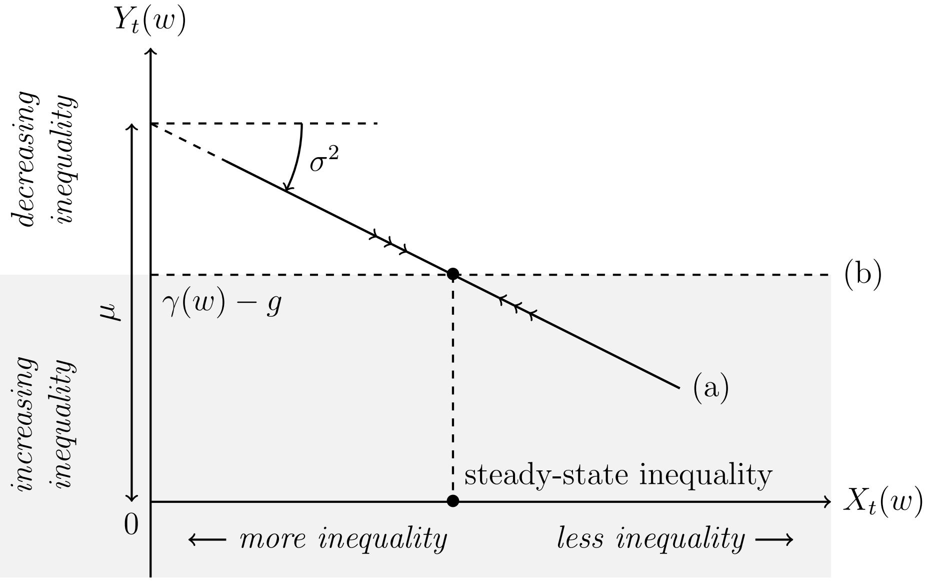


Figure 4.7: Simplified Dynamics of Wealth Inequality in the Top Tail

Figure 4.7 describes the situation. The Pareto coefficient is on the x-axis: the right side of the figure corresponds to low inequality, and the left side to high inequality. The y-axis relates to changes in inequality. We can separate the plane into two regions: the gray one where $\frac{\partial}{\partial t}F_t(w)/(wf_t(w)) < 0$ and therefore inequality increases, and the white one where it decreases. The system moves alongside the (a) line, either up or down depending on whether we are in the gray area or the white area. We keep moving up or down until we meet the (b) line that delimits both these areas: thus, the intersection between (a) and (b) indicates the steady-state level of inequality. The slope of (a) is determined by the diffusion coefficient σ^2, which captures mobility, while the intercept μ corresponds to the average consumption/wealth ratio.

This diagram helps perform some comparative statics. An increase in mean consumption at the top implies that the line (a) shifts upwards, leading to a steady state with lower inequality. A higher mobility (that is, an increase in σ^2) increases the slope of (a) while keeping its intercept constant: so it meets the line (b) at a lower value of $X_t(w)$, which implies higher steady-state inequality. If labor income

[12] We can assume in general that $\alpha > 1$, otherwise mean is infinite.

is negligible at the top, then $\gamma(w) \approx r$, so that the line (b) is positioned at $r - g$. Therefore, inequality is an increasing function of $r - g$ (see Piketty, 2014; Piketty and Zucman, 2015).

We can also use the figure to explain what makes the model identifiable. If we solely focus on the steady state, there is an infinity of values of (μ, σ^2) that can reach a given level of inequality, making the model impossible to estimate. Yet these different parameter values would yield very different dynamics of inequality. Having both low consumption and low mobility means that (a) is very flat, therefore $\frac{\partial}{\partial t} F_t(w)/(wf_t(w))$ is very small, and we converge very slowly to the steady state. Reaching the same steady state by having both high consumption and high mobility happens a lot faster. That line of reasoning breaks down if we are already at the steady state, however, which explains why assumption (ii) is required to identify the model.

4.4.2 Estimation

In essence, the estimation of the model involves fitting the line (a) from figure 4.7. This section covers how to do so in practice.

Transformation of Wealth Because of its fat tail, it can be difficult to estimate the density of wealth. To overcome the problem, I will be working with wealth transformed using the inverse hyperbolic sine function: $x \mapsto \operatorname{asinh}(x)$. This practice is common is the literature on wealth inequality (e.g. Thompson and Suarez, 2015; Kakar, Daniels, and Petrovska, 2019; Steinbaum, 2019). The transformation is bijective, strictly increasing, behaves linearly from low values and logarithmically for high values. Hence, it acts as a logarithmic transform for the top tail without creating problems for zero or negative wealth. I use Itō's lemma to move from the dynamics of wealth to that of its transform:

$$\mathrm{d}\operatorname{asinh}(w_t) = \left[\frac{\nu_t(w_t) - \mu_t(w_t)}{\sqrt{1+w_t^2}} - \frac{1}{2}\frac{\tau_t^2(w_t) + \sigma_t^2(w_t)}{1+w_t^2}\frac{w_t}{\sqrt{1+w_t^2}}\right]\mathrm{d}t \\ + \frac{(\tau_t^2(w_t) + \sigma_t^2(w_t))^{1/2}}{\sqrt{1+w_t^2}}\,\mathrm{d}B_t$$

There are two changes compared to the dynamics of untransformed wealth. First, all quantities are divided by $\sqrt{1+w_t^2}$, meaning that we use ratio quantities for high values of wealth, and absolute quantities for low values. Second, the drift term is adjusted by a factor that depends on the diffusion. I use tildes to designate to quantities that pertain to transformed wealth: that is, I will write $\tilde{\nu}_t$, $\tilde{\mu}_t$, etc. to denote the variables ν_t, μ_t, etc. divided by $\sqrt{1+w_t^2}$, and use $\tilde{F}_t$ and $\tilde{f}_t$ for the CDF and density of transformed wealth.

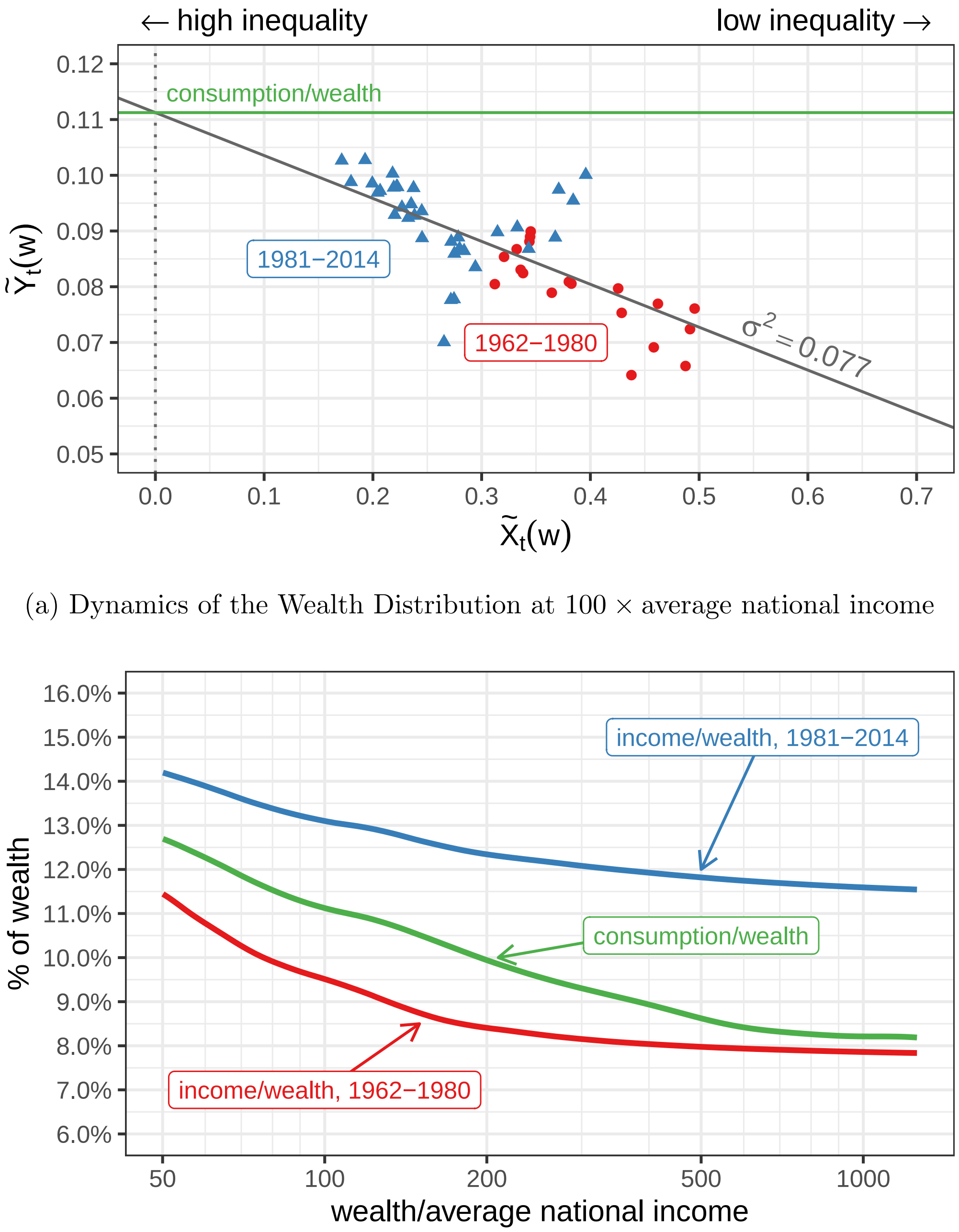
Estimating Equation Assume that μ_t and σ_t^2 are the same for all t. To simplify analysis and limit the number of parameters, assume that $\sigma(w) = \tilde{\sigma}\sqrt{1+w^2}$. This assumption meets the usual requirements of the literature for models of the wealth distribution. Because the standard deviation of consumption is scale invariant at the top, it can produce Pareto-shaped tails (Gabaix, 2009). At the same time, by breaking the scale invariance at the bottom, it makes it possible to get a stationary process. This is similar in spirit to what was done by Gabaix (1999) with a strictly positive reflecting barrier, but smoother (see Saichev, Malevergne, and Sornette (2010, p. 16), for more details on that approach).

With that assumption, the estimating equation (4.7) for transformed wealth becomes:

$$\begin{aligned}&\frac{\frac{\partial}{\partial t}\tilde{F}_t(w)}{\tilde{f}_t(w)} - \frac{\tilde{\Phi}_t(w)}{\tilde{f}_t(w)} - \frac{\tilde{\Psi}_t(w)}{\tilde{f}_t(w)} + \tilde{\nu}_t(w) - \frac{1}{2}\frac{\partial}{\partial w}\tilde{\tau}_t^2(w) \\ &+ \frac{1}{2}\tilde{\tau}_t^2(w)\left[-\frac{\frac{\partial}{\partial w}\tilde{f}_t(w)}{\tilde{f}_t(w)} - \frac{w}{\sqrt{1+w^2}}\right] = \tilde{\mu}_t(w) - \frac{1}{2}\tilde{\sigma}_t^2\left[-\frac{\frac{\partial}{\partial w}\tilde{f}_t(w)}{\tilde{f}_t(w)} - \frac{w}{\sqrt{1+w^2}}\right]\end{aligned} \tag{4.9}$$

For concision, write this equation as $\tilde{Y}_t(w) = \tilde{\mu}(w) - \tilde{\sigma}^2\tilde{X}_t(w)$. To see the link with the simplified estimating equation (4.8), note that the logarithm of a Pareto distributed variable follows an exponential distribution. Consider that for the top of the distribution, $f(w) \propto w^{-\alpha-1}$. Then, transformed wealth approximately follows an exponential distribution with coefficient α, so that $-\frac{\partial}{\partial w}\tilde{f}_t(w)/\tilde{f}_t(w) \approx \alpha$. Furthermore, $w/\sqrt{1+w^2} \approx 1$. Therefore, we have $\tilde{X}_t(w) \approx (\alpha-1)/2$, as in equation (4.8). Matters are somewhat more complicated for the left-hand side, though the intuition is similar. The time derivative $\frac{\partial}{\partial t}\tilde{F}_t(w)/\tilde{f}_t(w)$ is equal to zero when the rest of the left-hand side equals the right hand side, which determines the steady state. However, it is not possible anymore to separate the plane neatly into two fixed regions because the effects of demographics ($\tilde{\Phi}_t(w)$) and inheritance ($\tilde{\Psi}_t(w)$) are endogenous to the distribution of wealth, so the steady-state can only be determined through simulations.

Fitting the Model Assume that we observe the system over at a series of dates $(t_1, \dots, t_k)$. Define a grid of wealth values $(w_1, \dots, w_n)$. Using equation (4.9), the estimation of the model reduces to the estimation of a fixed-effect regression with



(a) Dynamics of the Wealth Distribution at 100 × average national income

(b) Estimated Consumption/Wealth and Income/Wealth Profiles

Source: Author's computations. *Note:* The unit is the adult (20 or older) individual, and wealth is split equally between members of couples.

Figure 4.8: Estimation of the Main Model

the specification:

$$\forall t \in \{t_1, \dots, t_k\} \quad \forall i \in \{1, \dots, n\} \qquad \tilde{Y}_t(w_i) = \tilde{\mu}(w_i) - \tilde{\sigma}^2 \tilde{X}_t(w_i) + \varepsilon_{it}$$

where $\tilde{\mu}(w_i)$ is a wealth-specific fixed effect, $\tilde{\sigma}^2$ is the opposite of the slope, and ε_{it} captures measurement error.[13] I estimate the CDF and density for the distribution of transformed wealth using kernel density estimators. I simulate the demographic and inheritance effects (see section 4.3.2 and 4.3.1), and also estimate resulting distributions using kernel density. For derivatives, both with respect to time and wealth, I run a local polynomial estimator of degree one.[14] For income conditional of wealth, I separate the sample into two periods (1962–1980, and 1981–2014) that constitute the two branches of the U-shaped pattern of wealth. I average income over these two periods so that the model focuses on the long-run mechanisms, rather than short-run dynamics that would introduce noise. I perform the estimation for levels of wealth greater than 50 times average national income, which roughly correspond to the top 1% wealth threshold today. This follows from the fact that the evolution of the top 1% has determined the trajectory of wealth inequality in the United States (see section 4.1.1) and that the model requires meaningful variation in the distribution to properly identify the effects at hand (cf. assumption (ii)).

Figure 4.8 graphically shows the results from the estimation. Panel 4.8a shows a diagram somewhat similar to figure 4.7, but with actual data, for a level of wealth corresponding to 100 times average national income. As we see, the data points for the periods 1962–1980 and 1981–2014 are more or less spread alongside the same line, despite the sharp change in the wealth/income ratio between both periods that impacts the left-hand side of the equation. This suggests that there hasn't been any strong structural changes since 1962 in terms of consumption/wealth profiles. There is a handful of points that stand out: these all correspond to the 1981–1989 period. That can be attributed to the reversal of many dynamics, so that several derivatives change sign at the same, making it harder to estimate them properly. In practice, removing or including these points does not change the results.

The slope is the same for all levels of wealth, and correspond to the variance of consumption/wealth. It is equal to $\sigma^2 = 0.077$. The fixed effects capture the

[13] In fact, measurement error should affect both the dependent and the independent variable, so the standard within estimator for fixed effect regression may give biased results. That being said, I have tested alternative estimators such as orthogonal least squares that account for error on both terms of the equation, and results were virtually identical.

[14] For the term $\frac{\partial}{\partial w}\tilde{f}_t(w)/\tilde{f}_t(w)$, note that it is equal to $\frac{\partial}{\partial w}\log\tilde{f}_t(w)$, so that I directly estimate the derivative of the logarithm of density. Because the logarithm of the density of an exponential distribution is linear, this yields more robust results.

average consumption/wealth ratio: they correspond to the intercept in panel 4.8a. In panel 4.8b, I plot that consumption/wealth ratio for all levels of wealth. That profile is a decreasing function of wealth. During the 1962–1980 period, that ratio was consistently higher than income, so that the decrease in wealth inequality was driven by dissavings at the top. Since 1981, however, income has increased, allowing people at the top to maintain the same relative levels of consumption while still accumulating wealth.

Bottom of the Distribution To fit the model, I restricted myself to wealth below 50 times average national income. To account for the distribution of wealth below that level in simulations (see section 4.5), I assume that the diffusion parameter $\tilde{\sigma}$ is the same for the whole distribution. Under that assumption, we can directly estimate $\tilde{\mu}(w)$ for all levels of wealth by taking the average of $Y_t(w_i) + \tilde{\sigma}^2 \tilde{X}_t(w_i)$ over all time periods.

That approach is an approximation, because it assumes that we can infer wealth mobility throughout the entire distribution based on mobility in the top tail. But in practice it is the simplest and most robust way to match the actual dynamics of wealth at all wealth levels. In particular, it is preferable to the inclusion of low and medium wealth when fitting the model, or to the fitting of a separate complete model for these levels of wealth. Indeed, some quantities for the bottom and middle are harder to estimate (especially the derivative of the density), and the amount of meaningful variation is lower (because the largest changes have happened to the top tail, see section 4.1.1). In addition, there may be some other, time-varying phenomenons that we doe not properly capture. All of this makes the signal-to-noise ratio less favorable. As a result, if we tried to include that part of the data with the top when fitting the model, we would lower the quality of the fit for the top — which has been driving most of the increase in inequality — and therefore diminish our ability to reproduce the main facts about wealth inequality. If we tried to estimate both the diffusion and the drift by fitting a complete model separately, we would get unstable and problematic results (including negative variances in some cases).

However, this approximation has a very limited impact on the overall results. First, because what matters to the shape of the wealth distribution is not the value of diffusion itself, but the joint effect of both drift and diffusion. By taking the average of $Y_t(w_i) + \tilde{\sigma}_t^2 \tilde{X}_t(w_i)$ to estimate average consumption, I ensure that, taken together, the estimate of drift and diffusion reproduce observed patterns. Second, because what matters the most is to faithfully reproduce the top of the distribution, which has driving the dynamics of inequality (see section 4.1.1). In practice, the assumptions for the bottom and the middle are here to ensure a roughly stable wealth distribution the bottom.

4.5 Results

Having estimated the model of wealth accumulation for the United Sates economy, I can now reproduce the observed trajectory of wealth inequality. More importantly, I can also change certain parameters and observe how these changes would affect wealth inequality. I can also make projections of how high wealth inequality is likely to go under current circumstances, and see what could affect that level. I start by studying the past evolution of the wealth distribution in section 4.5.1, and then the future in section 4.5.2.

4.5.1 Past Evolution (1962–2014)

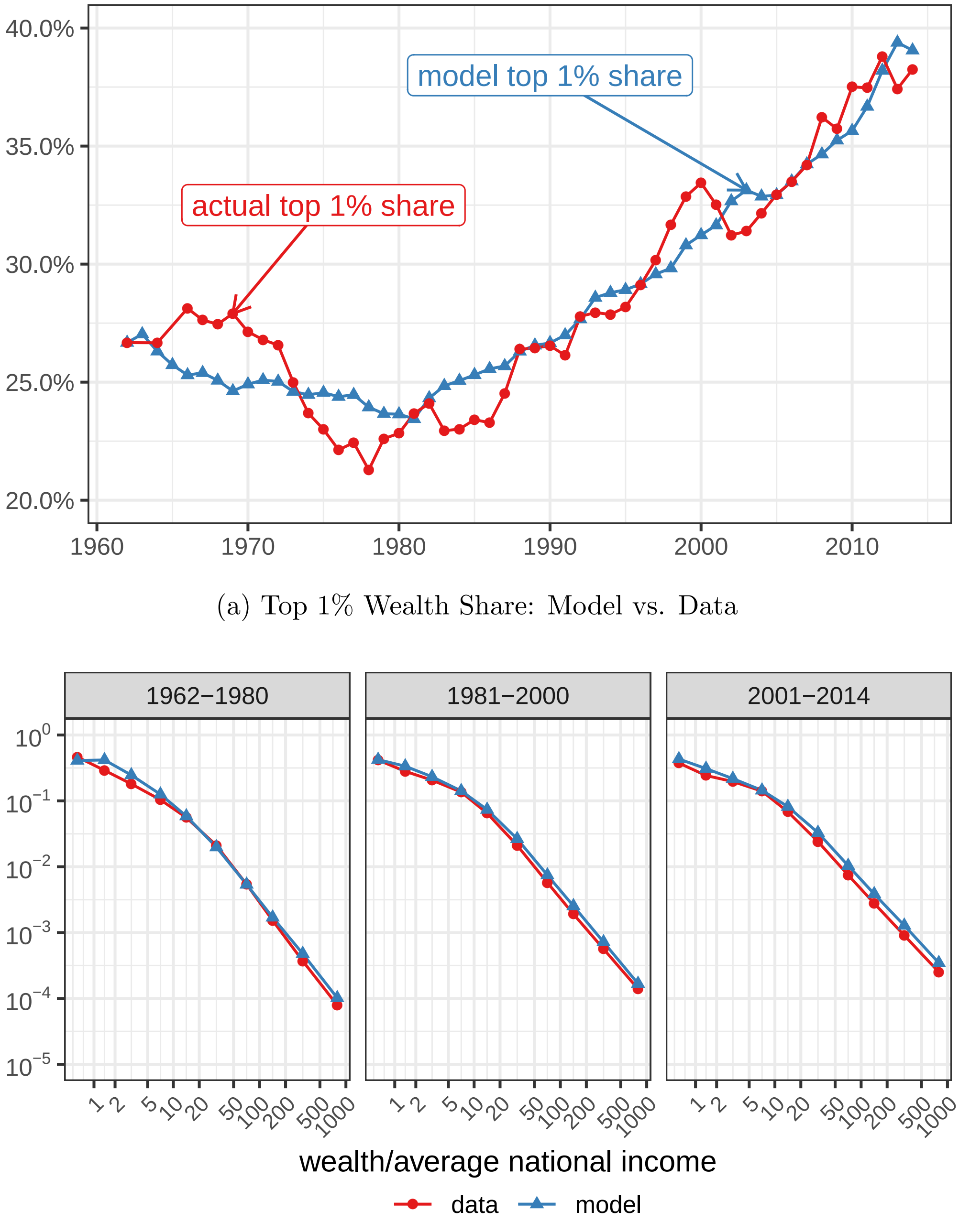
For the past evolution of wealth, we may first check that if we feed the model the actual parameters of the economy, we can reproduce the observed evolution of the wealth distribution. That is, I assume that simulated observations get the same labor income as people with the same rank in the true wealth distribution, and that they the rate of return on their capital income. I also use observed demographic parameters, and the actual estate tax schedule.

As shown in panel 4.9a, I match the U-shaped pattern of wealth inequality since 1962. Because the model focuses on long-run dynamics, it does not reproduce the small variations that are primarily driven by short-run changes in asset prices. In the long run, however, the model matches the data well.

Panel 4.9b further compares the simulated data with the whole distribution for all levels of wealth by looking at the density. The tail is getting increasingly fatter, as expected given the rise in inequality: the model matches that rise, but also reproduces the overall shape of the distribution for lower levels of wealth.

4.5.1.1 Labor and Capital Income

In figure 4.10, I estimate what the distribution of wealth would look like today if the distribution of labor income or returns on capital had stayed the same after 1980 as it was over the 1962–1980 period. That is, in panel 4.10a, I give people with a given rank in the wealth distribution after 1980 the average mean and variance of labor income from people with the same rank over 1962–1980. By construction, this



(a) Top 1% Wealth Share: Model vs. Data

(b) Density of Wealth: Model vs. Data

Source: Author's simulations and DINA data from Piketty, Saez, and Zucman (2018). *Note:* The unit is the adult individual (20 or older), and wealth is split equally between members of couples.

Figure 4.9: Past Evolution of Wealth: Data vs. Model

implies that the distribution of labor income is held fixed after 1980. In panel 4.10b, I do the same for the rates of return on capital (including capital gains).

Both labor income and capital returns have been significant drivers of wealth inequality: taken together, they account for most of the 15 pp. increase in the top 1% wealth share observed since 1980. Most of it can be attributed to increases in mean income conditional on wealth, as the conditional variance of income has not changed much. The role of labor income inequality is somewhat larger, but both factors are major contributors.

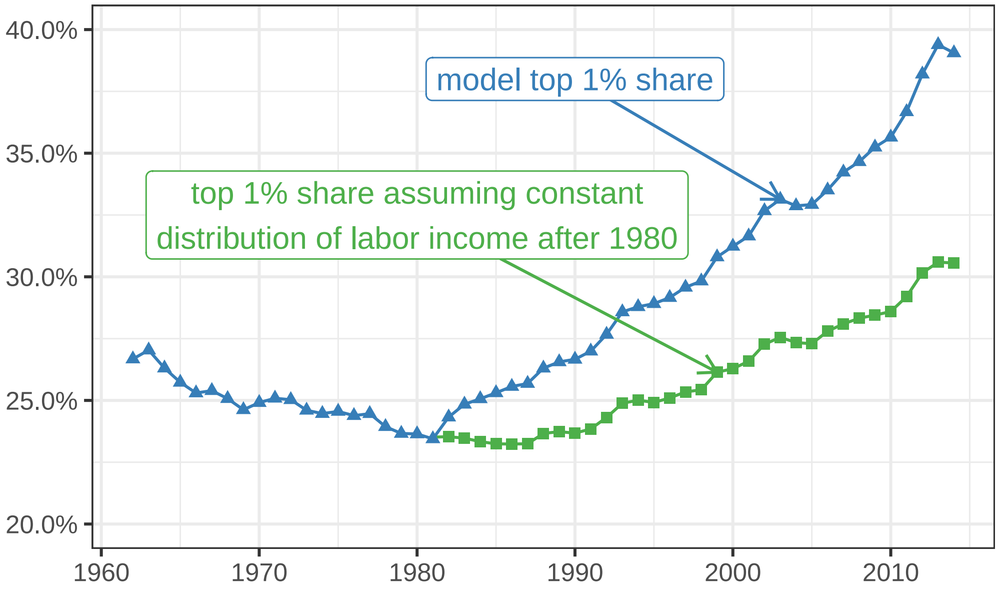
4.5.1.2 Deciphering the Role of $r - g$: Capital Gains and the Growth Rate

The role played by capital rates if return in figure 4.10b is directly connected to the impact of the spread between the rate of return on capital and the growth rate ($r - g$) that was popularized by Piketty (2014) (see also Piketty and Zucman, 2015).

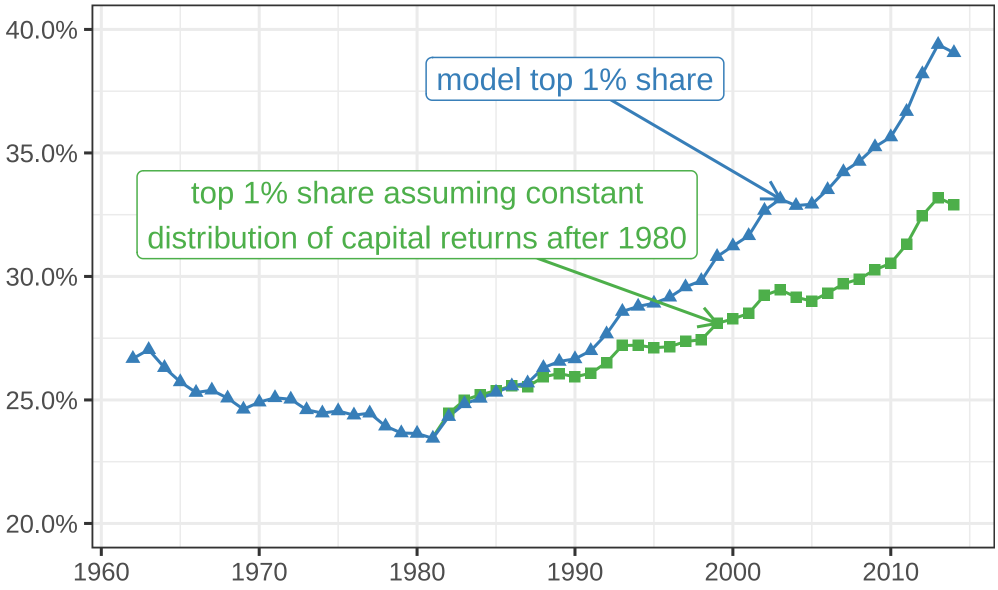
In figure 4.10b, I fixed r but not g. In figure 4.11a, I do the opposite exercise and fix g but not r. We can see that lower economic growth since 1980 has played some role in increasing wealth inequality, but that this role remains more limited than that of capital returns. Still it implies that the overall impact of increasing $r - g$ has been an important contributor to wealth inequality, on par with the rise of labor income inequality.

Though there is a twist. The usual story behind $r - g$ emphasizes normal capital returns (i.e. excluding capital gains), but these cannot explain rising wealth inequality. In fact, according to the DINA data, the average rate of return at the top has been somewhat *lower* since 1980 than before 1962. It is capital gains that explain most of the increase: as figure 4.11b shows, the rise of wealth inequality assuming no capital gains after 1980 is essentially the same as that assuming the overall rate of return as the 1962–1980 period.

The crucial role of capital gains is somewhat at odds with many models of wealth accumulation in the long run that tend to focus on normal capital returns: capital gains tend to be treated as short-run phenomenons that can be ignored when it comes to long-term trends. One of the reasons behind this view is that capital gains represent a change in relative prices. And, almost by definition, relative prices should not be changing when the economy is at its steady state, so there cannot be capital gains in the long run. That view is challenged by the fact that capital gains have been a persistent and economically meaningful phenomenon for the United



(a) Top 1% Wealth Share: Impact of Labor Income



(b) Top 1% Wealth Share: Impact of Capital Rates of Return

Source: Author's simulations. *Note:* The unit is the adult individual (20 or older), and wealth is split equally between members of couples.

Figure 4.10: Past Evolution of Wealth: The Role of Labor and Capital Income

States economy, especially since the 1980. It remains possible that the role of capital gains will eventually disappear, and the period from 1980 to today will become a historical anomaly. But there is an alternative view put forward by Robbins (2018), who shows that in a neoclassical model with imperfect competition, it is possible for capital gains to represent a meaningful fraction of national income at the steady state. Which of these view holds true would have a significant impact on the future evolution of the wealth distribution.

Panel 4.11a further shows how lower growth can increase wealth inequality. Assuming, at the extreme, that the growth rate fell to zero after 1980, top 1% wealth share would be about 5 pp. higher than it is today.

4.5.2 Future Evolution (2015–2100)

We can run the model to get projections of future inequality levels under various scenarios, and determine the steady state of wealth inequality, if any. I stress that these forecasts are always conditional on various parameters (regarding consumption, income, etc.) I do not attempt to endogenize these parameters: the point is to get some idea of how the wealth distribution reacts to them in the long run, and how high inequality can go under their current value.

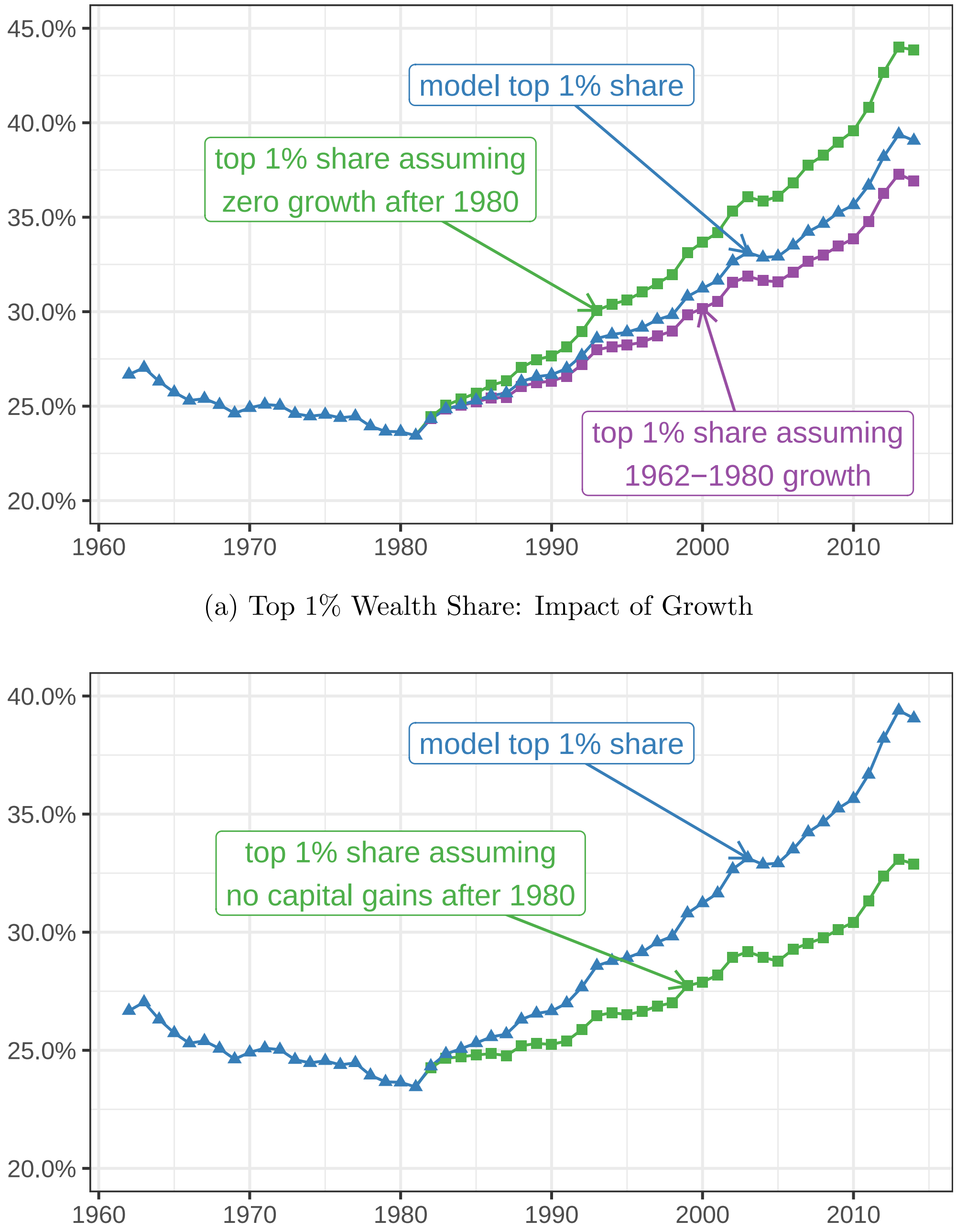
The first result is presented in figure 4.12a. Under current parameters, the wealth distribution in the United States would reach its steady state by the 2040s, with a top 1% share around 45%. This would put it at a level similar to that of the early 20th century — or even slightly higher.

The steady state would correspond to a much lower level of inequality, had the distribution of labor or the distribution of capital rates of return stayed at its 1962–1980 level. The level of inequality in the long run would correspond to a top 1% wealth share of 33% and 35%, respectively.

4.5.3 The Taxation of Wealth

In this section, I use the model of this paper to assess the long run effect of wealth taxes at the top of the distribution. The literature on the topic has grown significantly over the past few years. Recent theoretical contributions have stressed that the long-run elasticity of wealth with respect to the net-of-tax rate is a sufficient statistic for optimal capital taxation (Saez and Stantcheva, 2018; Piketty and Saez, 2013).[15]

[15] The famous result of Chamley (1986) and Judd (1985) — tax the optimal tax rate on capital is zero in the long-run — can be interpreted as the result of an implicit assumption that wealth

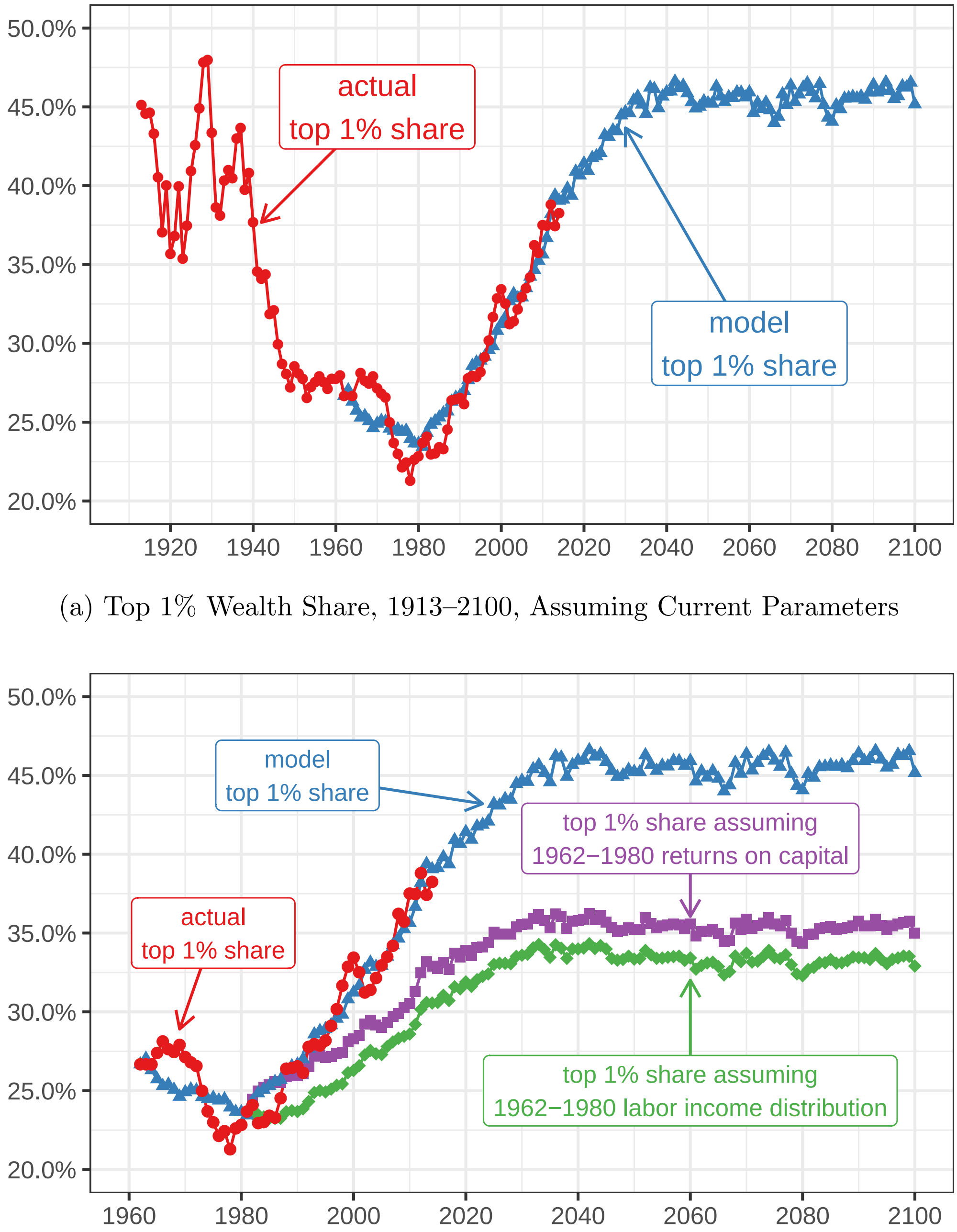


(a) Top 1% Wealth Share: Impact of Growth

(b) Top 1% Wealth Share: Impact of Capital Gains

Source: Author's simulations. *Note:* The unit is the adult individual (20 or older), and wealth is split equally between members of couples.

Figure 4.11: Past Evolution of Wealth: The Role of Growth and Capital Gains



(a) Top 1% Wealth Share, 1913–2100, Assuming Current Parameters

(b) Top 1% Wealth Share: Long-Term Impact of Labor and Capital Income

Source: Author's simulations and DINA data from Piketty, Saez, and Zucman (2018). *Note:* The unit is the adult individual (20 or older), and wealth is split equally between members of couples.

Figure 4.12: Future Evolution of Wealth

Unfortunately little is known about the value of that elasticity.

Several empirical papers have used quasi-experimental settings to estimate the *short-run* elasticity of wealth with respect to the net-of-tax rate: Seim (2017) in Sweden, Londoño-Vélez and Àcila-Mahecha (2018) in Colombia, Brülhart et al. (2016) in Switzerland, and Jakobsen et al. (2019) in Denmark. With the exception of Switzerland, these elasticities tend to be small. This is consistent with the view that a government trying to raise revenue with a one-off, unexpected wealth tax can indeed choose a very high marginal rate.

But the ability to raise revenue sustainably from a wealth tax depends on the *long-run* elasticity. That elasticity is likely to be larger than in the short run. The short-run elasticity only captures tax avoidance or short-run saving responses. But over time, wealth taxes also keep people from accumulating wealth, either through mechanical (lower post-tax rates of return) or behavioral effects (lower savings). This leads to a slow erosion of the tax base. Because it takes a long time to materialize, it is hard to get a clean identification of this effect in the data. As a result, we lack a clear understanding of how the stock of capital would react to wealth tax in the long run.

Recently, two papers have tackled that question. Jakobsen et al. (2019) use their short-run elasticity estimates to calibrate a structural model of savings at the top. They indeed find a higher elasticity in the long-run. Saez and Zucman (2019) consider the problem of taxing the very top of the wealth distribution (billionaires) using data from the Forbes rankings. These two papers provide models that shed different lights on the problem. Jakobsen et al. (2019) model wealth accumulation using an deterministic model of intertemporal choice. This model features standard preferences over a consumption path and a taste for end-of-life wealth (i.e. bequests). They use it to derive analytical expression linking the long-run elasticity of wealth to the short-run elasticity and preference parameters. This model emphasizes the role of behavioral responses on consumption, but it is deterministic so it does not account for the role that mobility plays in shaping the distribution of wealth. This stands in contrast to the model of Saez and Zucman (2019). They focus on billionaire wealth, and therefore assume that the role of consumption is negligible. They consider a simple model in which billionaires are subjected to a given tax rate on their total wealth (not just above a threshold), while everything else remains the same. In that model, the sole determinant of the elasticity of wealth in the long-run has to do with mobility. If wealth mobility is low, then a wealth tax ends up taxing the

is infinitely elastic. Various contributions have overturned the result by introducing, for example, uncertainty (Aiyagari, 1994), incomplete markets (Farhi, 2010) or heterogenous altruism (Farhi and Werning, 2013).

same people again and again: as their wealth mechanically goes down, so does the tax base. Therefore, the elasticity of taxable wealth is high, and the ability to tax wealth is the long-run is limited. However, if mobility is high, the tax base often gets renewed. Individual people are subjected to the tax during shorter periods of time, with new, previously untaxed wealth entering the tax base on a regular basis: as a result, the elasticity is lower.

I contribute to that literature by providing a simple, practical and transparent method to determine how the tax base reacts to a wealth tax in the steady-state. It connects short-run elasticities with the dynamics of wealth using the dynamic model of this paper to estimate a long-run elasticity. This allows me to incorporate insights from Jakobsen et al. (2019) and Saez and Zucman (2019) into single formula. In the short run, I account for behavioral response on savings and tax avoidance using reduced-form elasticities. Then, I use the model of this paper to compute how these effects accumulate over time to produce long-run responses of the tax base to the wealth tax.

I show that, under very general conditions, the steady-state density of wealth *with* a wealth tax is equal to the steady-state density of wealth *without* a wealth tax, multiplied by an additional term that only depends on the tax schedule, wealth mobility, and some behavioral elasticities. This makes it easy to simulate how the tax base would eventually react to any given wealth tax.

I start by considering the pure mechanical effect of a wealth tax to present the key result of this section. Then I show how we can account for various behavioral response by using the same result with a modified "effective" tax schedule that is slightly different from the statutory one.

Dynamic Mechanical Effect Absent a wealth tax, assume that the dynamic of wealth follows the SDE:

$$\mathrm{d}w_{it} = a(w_{it})\,\mathrm{d}t + b(w_{it})\,\mathrm{d}B_{it} \tag{4.10}$$

where $a(w_{it}) \equiv (\nu_t(w_{it}) - \mu_t(w_{it}))$ correspond the average saving by wealth, and $b(w_{it}) \equiv (\tau_t^2(w_{it}) + \sigma_t^2(w_{it}))^{1/2}$ is the standard deviation of savings by wealth. For the rest of this section, I neglect the impact of demographics and inheritance for the sake of tractability. Note that these processes have a limited impact on the long-run dynamics of wealth, so this should not significantly affect the conclusions. We can assess their impact using simulations.

I then introduce a wealth tax with rate α for wealth above the threshold w_0. The dynamic of wealth becomes:

$$\mathrm{d}w_{it} = (a(w_{it}) - \alpha(w_{it} - w_0)_+)\,\mathrm{d}t + b(w_{it})\,\mathrm{d}B_{it} \tag{4.11}$$

where $(x)_+ = \max\{x, 0\}$. I will further assume that, for $w \geq w_0$, the standard deviation of shocks is proportional to wealth, i.e. $b(w) = bw$. That last assumption is not very restrictive, since it is required at the top for Pareto-shaped tails to arise, in line with the literature and the findings of this paper.

At this stage, I do not assume any behavioral response: yet, in the long-run, the distribution of wealth changes in response to that wealth tax, because it lowers post-tax returns on capital. The following result gives the steady-state distribution of wealth with the tax as a function of the steady-state distribution of wealth without the tax (see appendix D.2.2 for proof).

Theorem 9 (Steady-State Distribution With a Wealth Tax)**.** *Assume that, without a wealth tax, the dynamic of wealth follows the equation (4.10). Introduce a wealth tax with rate α on wealth above w_0, so that wealth now evolves according to (4.11). Let f_α be the steady-state density of wealth with the tax, and f_0 the steady-state density without the tax. Define:*

$$\zeta(w) \equiv \begin{cases} \exp\left\{-\frac{2\alpha}{b^2}\left(\frac{w_0}{w} - 1\right)\right\}\left(\frac{w}{w_0}\right)^{-2\alpha/b^2} & \text{if } w \geq w_0 \\ 1 & \text{if } w < w_0 \end{cases}$$

and $K^{-1} \equiv \int_{-\infty}^{+\infty} \zeta(w) f_0(w)\,\mathrm{d}w$. We have $f_\alpha(w) = K f_0(w)\zeta(w)$.

That result makes it possible to estimate how the tax base would react to a wealth tax in the long-run, effectively by reweighting the steady-state distribution of untaxed wealth using the function ζ. I have considered the effect of a linear tax above an exemption threshold, but the result could be extended to an arbitrary number of brackets with different rates without difficulties. The setting mentions the introduction of a new wealth tax where there previously was none, but we could apply the same result to an increase or a decrease of an existing wealth tax by redefining α as a change in the rate of the wealth tax.

The result emphasizes the role of mobility, as explained by Saez and Zucman (2019). As we can see, the impact on the tax base depends on α/b^2 and not just α. Therefore, doubling the parameter b quadruples the parameter b^2, which implies the same change in the tax base despite a tax rate four times as high. The intuition is the same as in Saez and Zucman (2019): high mobility means that people only gets taxed for a short period of time and that new, previously untaxed wealth keeps entering the tax base. As a result, the tax base does not react too much to wealth taxation. When mobility goes to zero, however, the same wealth from the same people is taxed repeatedly, so that the tax base eventually goes to zero.

Let $B = \int_{w_0}^{+\infty} (w - w_0) f_\alpha(w) \; \mathrm{d}w$ be the steady-state tax base. We can calculate it as follows:

$$B = \frac{\mathbb{E}[(w - w_0)_+ \zeta(w)]}{\mathbb{E}[\zeta(w)]}$$

where the expectations should be taken according to the steady-state distribution without tax, i.e. f_0. If we know f_0, then this quantity is directly estimable. Finding that true steady-state density does require some assumptions and additional modelling, as was done in previously paper. The steady-state tax revenue is equal to αB.

Behavioral Response Through Tax Reporting People can react to a wealth tax by hiding some of their wealth, either through tax evasion or tax avoidance. Assume that, in response to a tax α, people only report a fraction $(1 - \alpha)^\varepsilon$ of their wealth. The parameter ε is the elasticity of declared wealth to the net-of-tax rate $1 - \alpha$. For a small rate $\alpha \ll 1$, people react by approximately hiding a fraction $\alpha\varepsilon$ of their wealth. When $\varepsilon = 0$, people truthfully report all of their wealth. As ε goes to infinity, people start hiding all of their wealth to avoid paying the tax. With tax avoidance, people that own w in wealth pay:

$$\alpha[(1 - \alpha)^\varepsilon w - w_0]_+ = \alpha(1 - \alpha)^\varepsilon [w - w_0(1 - \alpha)^{-\varepsilon}]_+$$

instead of $\alpha(w - w_0)_+$. In effect, this is equivalent to having a wealth tax with a lower rate $\alpha(1 - \alpha)^\varepsilon$ and a higher exemption threshold $w_0(1 - \alpha)^{-\varepsilon}$. Therefore, the results for the purely mechanical model hold with minimal modifications. It suffices to replace the true tax parameters α and w_0 by their effective counterparts $\alpha(1 - \alpha)^\varepsilon$ and $w_0(1 - \alpha)^{-\varepsilon}$.

Tax evasion has two effects on the dynamic of wealth. Most importantly, it directly lowers the tax base since people under-report their assets. But as a secondary effect, it increases the post-tax rate of return, allowing people to accumulate more, which grows the tax base in the long-run.

Behavioral Response Through Savings People may also react to a wealth by actually accumulating less wealth. Changes to savings have different implications than tax evasion. Indeed, tax evasion affects both the dynamic of wealth and the tax base. Savings, on the other hand, affect the dynamic of wealth but do not directly reduce the tax base.

Theory provide little constraints regarding how a wealth tax ought to affect saving rates, given the vast number of settings and mechanisms that we could consider. The following reduced-form specification can nonetheless account for the overall effect in a direct and intuitive way. Assume that, in response to a tax rate α on wealth above w_0, people reduce their savings by an amount $(1-(1-\alpha)^\eta)(w-w_0)_+$. The parameter η captures the elasticity of savings with respect to the net-of-tax rate $1-\alpha$. If $\eta = 0$, savings do not respond to the wealth tax. If $\eta > 0$, people start to consume some of their wealth in excess of w_0 rather than pay taxes. At the limit, when α approaches one or η approaches infinity, people immediately consume all wealth above w_0 to avoid paying the wealth tax.[16]

Under those circumstances (and ignoring tax evasion for now), the drift term in the dynamic of wealth become:

$$a(w_{it}) - (\alpha + 1 - (1-\alpha)^\eta)(w_{it} - w_0)_+$$

so the results from the pure mechanical model still hold, except that we need to replace the tax rate α by $\alpha + 1 - (1-\alpha)^\eta$. The behavioral response on savings amplifies the impact of the wealth tax.

Complete Model When combining the behavioral response through savings and tax evasion, it makes sense to assume that savings respond to the effective tax schedule (which accounts for tax evasion) rather than the statutory one. That is, people increase their consumption by an amount $(1-(1-\alpha(1-\alpha)^\varepsilon)^\eta)[w-w_0(1-\alpha)^{-\varepsilon}]$. Therefore, the drift term for the dynamic of wealth is:

$$a(w_{it}) - [\alpha(1-\alpha)^\varepsilon + 1 - (1-\alpha(1-\alpha)^\varepsilon)^\eta][w_{it} - (1-\alpha)^{-\varepsilon} w_0]_+$$

[16] I will ignore the cases where $\eta < 0$, even though they are a theoretical possibility, because it is problematic to assume in a taxation context that the tax base respond positively to the tax. Moreover, the elasticity has to change sign at some point, otherwise a 100% wealth tax would correspond to infinite savings. However, if true, it would imply that wealth tax rates could be higher.

So the results from the mechanical model still apply if we replace the statutory exemption threshold w_0 by $w_0(1-\alpha)^{-\varepsilon}$, and if we replace the statutory tax rate α by $\alpha(1-\alpha)^{\varepsilon}+1-(1-\alpha(1-\alpha)^{\varepsilon})^{\eta}$.

Estimates for Behavioral Elasticities To calibrate ε and η, I rely on the recent empirical literature that exploit various quasi-experimental settings to assess behavioral reactions to a wealth tax.

Several of these papers present bunching evidence (Seim, 2017; Londoño-Vélez and Àcila-Mahecha, 2018; Jakobsen et al., 2019). Bunching provides the cleanest estimates of pure tax avoidance elasticity. Indeed, the true value of wealth in the short run tend to follow unpredictable asset movements, so that it would be very hard for a household to precisely bunch at kink points. Seim (2017) finds an elasticity of 0.5 in Sweden, and Jakobsen et al. (2019) find elasticities that are even lower in Denmark. Londoño-Vélez and Àcila-Mahecha (2018) find a higher estimate (2–3) in Colombia.

As their main identification strategy, Jakobsen et al. (2019) pursue a difference-in-difference approach that exploit various tax reforms. This allows them to compute elasticities that incorporate dynamic and saving responses over larger time spans. Over an 8-year time frame, they find a sizeable elasticity at the top of about 18 with respect to the net-of-tax rate. The authors argue that most (90%) of it can be attributed to a behavioral effect (as opposed to a mechanical effect). Assuming that the elasticities cumulates multiplicatively over time, this would correspond to a yearly behavioral elasticity of 1.4 for both the saving and tax avoidance response. Seim (2017) also analyze saving responses to the wealth tax, but does not find any.

Brülhart et al. (2016) find a much higher elasticity (23–34) in Switzerland using both between canton variations of the tax rate and within variation in the Bern canton. They also look at bunching evidence, but find much lower effects there.

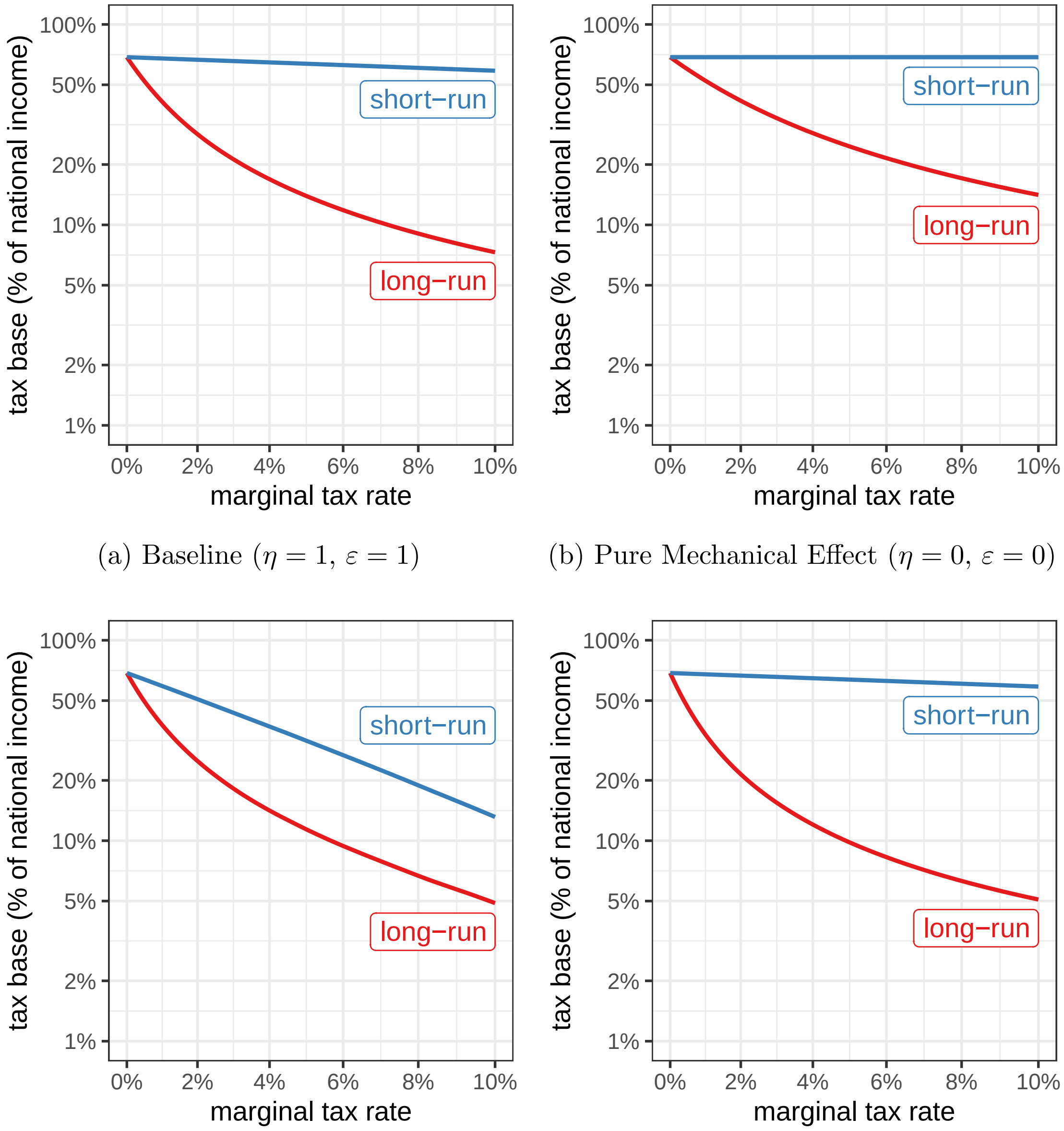
Note that the tax avoidance elasticity is not a pure structural parameter, but also results from how strongly a wealth tax is enforced. For the baseline calibration, I will consider a limited tax avoidance response ($\varepsilon = 1$), which is around the values found by Seim (2017), Londoño-Vélez and Àcila-Mahecha (2018) and Jakobsen et al. (2019). I will also consider a medium savings response ($\eta = 1$), in line with Jakobsen et al. (2019), but higher than zero as opposed to Seim (2017). Then I consider alternative scenarios with a higher saving response ($\eta = 2$) and a higher tax avoidance response ($\varepsilon = 10$). I could consider even higher tax avoidance responses ($\varepsilon = 20$ or $\varepsilon = 30$), as found by Brülhart et al. (2016), but the interest would be limited. Indeed, with such a severe tax avoidance, the dynamic effects under study become negligible compared to static tax avoidance.

Implications for a Wealth Tax For the different values of ε and η, I simulate how the tax base would react to various marginal tax rates. I consider a linear tax with rate α on wealth above \$50m (in 2014 dollars). The tax applies to equal-split wealth (meaning that the threshold for couples is actually \$100m). I assume that, in the long run, the threshold would rise in line with average income (so that there is a stationary solution), and look at the value of the tax base as a fraction of national income for different values of α. I use the steady-state wealth distribution under current parameters as estimated in section 4.5.2. Note, however, that results are very similar if we use the last year of data available (the levels would change, but the elasticities would be the same). Therefore, the results of this section do not depend too much on the outcome of the long-run simulations, and one can use the method with the current distribution of wealth as a short-cut. Figure 4.13 shows the results.

The long-run response of the tax base is naturally stronger than the short-run, which only accounts for tax avoidance. In the baseline specification, a 1% wealth tax decreases the tax base by 50% in the long, which would imply a long-run elasticity around 50. That number may seem high, yet is not out of line with the findings of Jakobsen et al. (2019). At the top of the distribution, they find an elasticity of 18 when looking 8 years after the reform. Using their structural model, this translates into an elasticity of 33 after 30 years in a low return environment, and even higher in a high return environment (which would be closer to the present setting).

However, that elasticity is not constant: it is indeed high for low wealth taxes, but quickly tempers off. As a result, if we were to compute it based on a high 10% wealth tax, the elasticity would be lower (around 21). This nonlinearity results from the dynamic mechanical effects. It implies that one should be careful when extrapolating empirical estimates of the long-run response of the capital stock that are based on relatively small tax changes.

The nonconstant elasticity does impact the tax rate that would maximize revenue in the long run. Indeed, for small marginal tax rates, because we start from a baseline of zero, the wealth tax does raise revenue even though the tax base diminishes quickly. By the time adverse revenue effects arise, the tax base has become less elastic. As consequence, the revenue-maximizing tax rate may in theory be quite high, in several cases north of 10%. In these cases, we nonetheless tend to quickly reach a relatively flat revenue plateau after 10%, so that revenue gains from a wealth tax above 10%



(a) Baseline ($\eta = 1$, $\varepsilon = 1$)

(b) Pure Mechanical Effect ($\eta = 0$, $\varepsilon = 0$)

(c) High Tax Avoidance ($\eta = 1$, $\varepsilon = 10$)

(d) High Saving Response ($\eta = 2$, $\varepsilon = 1$)

Source: Author's computation. *Note:* Results correspond to the steady-state tax base for a linear wealth tax above \$50m in 2014 US dollars. I assume that the tax threshold rises in line with average national income. I assume that the tax applies to equal-split wealth (each members of a couple pay tax on half the wealth of the couple). The parameter ε is the elasticity of tax reporting, and η is the elasticity of the saving response. The short-run response only accounts for tax reporting, while the long-run response also incoporates dynamic mechanical effects et saving effects.

Figure 4.13: Impact on the Tax Base of a Linear Wealth Tax over \$50m

are limited.

The pure dynamic mechanical model (figure 4.13b) is closest to the model Saez and Zucman (2019). Using data from the Forbes 400 rankings and a simplified model, they find an elasticity of 16 (their elasticity is constant by construction). For a 1% wealth tax, I find a somewhat higher elasticity of 27, but for higher wealth tax of 10%, I get a similar value of 15. Note that their estimates are concerned with the extreme top of the distribution (billionaires), whose dynamic of wealth may arguably differ from the rest of the very rich.

I stress that is still significant uncertainty regarding these values. In particular, behavioral elasticities are extrapolated from quasi-experiments based on rather small tax rates, and it is difficult to know what the true reactions would be with more radical policies such as a 10% wealth tax. Note also that I focused on the steady-state. In practice that steady-state may take a very long time to materialize, so it is not necessarily the relevant time horizon. The model nonetheless carries useful insights on the underlying dynamics of the wealth distribution and wealth taxes. Finally, note that I only consider partial equilibrium effects. I do not consider how rates of return or the labor market may react to changes in the capital stock. The estimated elasticities are still useful to calibrate more complex general equilibrium models.

Conclusion

In this paper, I have presented a simple model of the wealth distribution that can decompose the impact of labor income, rates of return, growth, savings, demography and inheritance on the wealth distribution. In spite its simplicity, that model can incorporate a realistic modeling of these various factors. I show that this model can be estimated solely using repeated cross-sectional data, and I estimate it using DINA data on income and wealth for the United States since 1962.

I find that, out of the 15 pp. increase in the top 1% wealth share observed since 1980, about 7 pp. can be attributed to rising labor income inequality, 6 pp. to rising returns on wealth, and 2 pp. to lower growth. Importantly, the role played by rising rates of return on wealth can entirely be attributed to capital gains. In the future, and holding constant the present parameters of the economy, the United States economy would reach a steady state with a top 1% wealth share about 45%. I use the model to investigate the impact of progressive wealth taxes on the capital stock and the wealth distribution. I develop a simple a simple formula to characterize how the tax base would react to a wealth tax in the long-run in terms of observable quantities and key behavioral elasticities. I find that the elasticity of wealth with respect to the net-of-tax rate is sizeable, but also nonconstant, so that revenue-maximizing tax rates may be quite high.

These findings are very general in the sense that they apply to any model of the wealth distribution that generates Pareto-shaped fat tails using an accumulation of multiplicative random shocks (to income, consumption, etc.), as is usually the case. Note that all of the counterfactuals coming out of the model assume that everything else remains equal in the economy, which in reality would not always be the case. This is where more tightly specified, fully microfounded models of the wealth distribution can be useful. Such model can endogenously account for the way in which, say, savings might react to a change in the labor income distribution, and include that is its predictions. Yet, it remains true that the findings of this paper have to apply to the more microfounded model. In that sense, both approaches are complementary, and the methodology of this paper is useful to discipline more complex models.

The key insight of this paper — that the Fokker-Planck equation can be used as an empirical tool to identify certain parameters — may be applied to a wide set of problems. For wealth inequality, it could be used to analyze the dynamics of various phenomenons, such as the racial wealth gap. But in theory it could also be applied to any economic situation that involves stochastic growth, such as the income distribution, or the distribution of firms and city sizes.

Bibliography

Acemoglu, Daron and James Robinson (2015). "The Rise and Decline of General Laws of Capitalism". In: *Journal of Economic Perspectives* 29.1, pp. 3–28. URL: https://www.jstor.org/stable/43194693.

Aiyagari, S. R. (1994). "Uninsured Idiosyncratic Risk and Aggregate Saving". In: *The Quarterly Journal of Economics* 109.3, pp. 659–684. URL: https://academic.oup.com/qje/article-lookup/doi/10.2307/2118417.

Alstadsæter, Annette et al. (2017). "Accounting for Business Income in Measuring Top Income Shares: Integrated Accrual Approach Using Individual and Firm Data from Norway". URL: http://www.nber.org/papers/w22888.

Atkinson, A. B. (1971). "The Distribution of Wealth and the Individual Life-Cycle". In: *Oxford Economic Papers* 23.2, pp. 239–254. URL: https://www.jstor.org/stable/2662236?seq=1#page_scan_tab_contents.

Batty, Michael et al. (2019). "Introducing the Distributional Financial Accounts of the United States". URL: https://doi.org/10.17016/FEDS.2019.017.

Benhabib, Jess, Alberto Bisin, and Shenghao Zhu (2011). "The Distribution of Wealth and Fiscal Policy in Economies With Finitely Lived Agents". In: *Econometrica* 79.1, pp. 123–157. URL: http://dx.doi.org/10.3982/ECTA8416.

Berman, Yonatan, Eshel Ben-Jacob, and Yoash Shapira (2016). "The dynamics of wealth inequality and the effect of income distribution". In: *PLoS ONE* 11.4, pp. 8–10.

Böhl, Gregor and Thomas Fischer (2017). "Can taxation predict US top-wealth share dynamics?"

Bricker, Jesse, Alice Henriques, and Peter Hansen (2018). "How much has wealth concentration grown in the United States? A re-examination of data from 2001-2013". URL: https://doi.org/10.17016/FEDS.2018.024..

Brülhart, Marius et al. (2016). "Taxing Wealth: Evidence from Switzerland". In: *NBER Working Paper Series*, p. 37.

Carroll, Christopher (1998). "Why Do the Rich Save So Much?" URL: http://www.nber.org/papers/w6549.pdf.

Chamley, Christophe (1986). "Optimal Taxation of Capital Income in General Equilibrium with Infinite Lives". In: *Econometrica* 54.3, p. 607.

Cowell, Frank A (1998). "Inheritance and the Distribution of Wealth".

Farhi, Emmanuel (2010). "Capital taxation and ownership when markets are incomplete". In: *Journal of Political Economy* 118.5, pp. 908–948.

Farhi, Emmanuel and Iván Werning (2013). "Estate taxation with altruism heterogeneity". In: *American Economic Review* 103.3, pp. 489–495.

Favilukis, Jack (2013). "Inequality, stock market participation, and the equity premium". In: *Journal of Financial Economics* 107.3, pp. 740–759. URL: http://dx.doi.org/10.1016/j.jfineco.2012.10.008.

Fritsch, F. N. and R. E. Carlson (1980). "Monotone Piecewise Cubic Interpolation". In: *SIAM Journal on Numerical Analysis* 17.2, pp. 238–246. URL: https://doi.org/10.1137/0717021.

Gabaix, Xavier (1999). "Zipf's Law for Cities: an Explanation". In: *Quarterly Journal of Economics* 114.August, pp. 739–767.

– (2009). "Power Laws in Economics and Finance". In: *Annual Review of Economics* 1.1, pp. 255–293.

Gabaix, Xavier et al. (2016). "The Dynamics of Inequality". In: *Econometrica* 84.6, pp. 2071–2111.

Gapminder (2019). *Children per women since 1800.* URL: https://www.gapminder.org/news/children-per-women-since-1800-in-gapminder-world/.

Garbinti, Bertrand, Jonathan Goupille-Lebret, and Thomas Piketty (2016). "Income Inequality in France, 1900-2014: Evidence from Distributional National Accounts (DINA)". URL: http://piketty.pse.ens.fr/filles/GGP2016DINA.pdf.

Góes, Carlos (2016). "Testing Piketty's Hypothesis on the Drivers of Income Inequality: Evidence from Panel VARs with Heterogeneous Dynamics". In: *IMF Working Papers* 16.160, p. 1.

Gomez, Matthieu (2016). "Asset Prices and Wealth Inequality". In: pp. 1–60.

Gyöngy, I (1986). "Mimicking the one-dimensional marginal distributions of processes having an Ito differential". In: *Probability Theory and Related Fields* 71.4, pp. 501–516. URL: https://link.springer.com/article/10.1007/BF00699039.

Haines, Michael R. (1998). "Estimated life tables for the united states, 1850–1910". In: *Historical Methods* 31.4, pp. 149–169.

Hubmer, Joachim and Anthony A Smith (2018). "A Comprehensive Quantitative Theory of the U.S. Wealth Distribution".

Human Fertility Collection (2019). Max Planck Institute for Demographic Research in Rostock, Germany and Vienna Institute of Demography in Vienna, Austria. URL: https://www.fertilitydata.org/.

Human Fertility Database (2019). Max Planck Institute for Demographic Research in Rostock, Germany and Vienna Institute of Demography in Vienna, Austria. URL: https://www.humanfertility.org/.

Human Life Table Database (2019). Max Planck Institute for Demographic Research in Rostock, Germany, Department of Demography at University of California at

Berkeley, USA and Institut d'études démographiques (INED) in Paris, France. URL: https://www.lifetable.de/.

Human Mortality Database (2019). Max Planck Institute for Demographic Research in Rostock, Germany, Department of Demography at University of California at Berkeley, USA. URL: https://www.mortality.org.

Jakobsen, Katrine et al. (2019). "Wealth Taxation and Wealth Accumulation: Theory and Evidence from Denmark". In: *Quarterly Journal of Economics*. URL: http://gabriel-zucman.eu/files/JJKZ2019.pdf.

Judd, Kenneth L. (1985). "Redistributive taxation in a simple perfect foresight model". In: *Journal of Public Economics* 28.1, pp. 59–83.

Kakar, Venoo, Gerald Eric Jr. Daniels, and Olga Petrovska (2019). "Does Student Loan Debt Contribute to Racial Wealth Gaps? A Decomposition Analysis". In: *The Journal of Consumer Affairs*, pp. 1–28. URL: https://onlinelibrary.wiley.com/doi/pdf/10.1111/joca.12271.

Kesten, Harry (1973). "Random Difference Equations and Renewal Theory for Products of Random Matrices". In: *Acta Mathematica* 131.1, pp. 207–248. arXiv: -.

Koenker, Roger (2019). *quantreg: Quantile Regression*. R package version 5.40. URL: https://CRAN.R-project.org/package=quantreg.

Kopczuk, Wojciech (2015). "What Do We Know about the Evolution of Top Wealth Shares in the United States?" In: *Journal of Economic Perspectives* 29.1, pp. 47–66.

Kopczuk, Wojciech and Emmanuel Saez (2004). "Top Wealth Shares in the United States, 1916-2000: Evidence from Estate Tax Returns". In: *National Tax Journal* 57.2, Part 2, pp. 445–487.

Laitner, John (1979). "Household Bequest Behaviour and the National Distribution of Wealth". In: *The Review of Economic Studies* 46.3, pp. 467–483.

Londoño-Vélez, Juliana and Javier Àcila-Mahecha (2018). "Can Wealth Taxation Work in Developing Countries? Quasi-Experimental Evidence from Colombia".

Lund, Steven P, Joseph B Hubbard, and Michael Halter (2014). "Nonparametric Estimates of Drift and Diffusion Profiles via Fokker–Planck Algebra". In: *The Journal of Physical Chemistry B* 118, pp. 12743–12749.

Menchik, Paul L (1980). "Primogeniture, Equal Sharing, and the U.S. Distribution of Wealth". In: *The Quarterly Journal of Economics* 94.2, pp. 299–316. URL: https://www.jstor.org/stable/1884542.

Piketty, Thomas (2014). *Capital in the Twenty First Century*. Cambridge, Massachusetts: The Belknap Press of Harvard University Press.

Piketty, Thomas and Emmanuel Saez (2013). “A Theory of Optimal Inheritance Taxation”. In: *Econometrica* 81.5, pp. 1851–1886. URL: http://doi.wiley.com/10.3982/ECTA10712.

Piketty, Thomas, Emmanuel Saez, and Gabriel Zucman (2018). “Distributional National Accounts: Methods and Estimates for the United States”. In: *Quarterly Journal of Economics* 133.May, pp. 553–609.

Piketty, Thomas and Gabriel Zucman (2014). “Capital is Back: Wealth-Income Rations in Rich Countries 1700–2010”. In: *Quarterly Journal of Economics* 129.3, pp. 1255–1310.

– (2015). “Wealth and Inheritance in the Long Run”. In: *Handbook of Income Distribution* 2, pp. 1303–1368.

Robbins, Jacob A (2018). “Capital Gains and the Distribution of Income on the United States”.

Ruggles, Steven et al. (2019). *IPUMS USA: Version 9.0 [dataset]*. Minneapolis.

Saez, Emmanuel and Stefanie Stantcheva (2018). “A simpler theory of optimal capital taxation”. In: *Journal of Public Economics* 162.September 2017, pp. 120–142. URL: https://doi.org/10.1016/j.jpubeco.2017.10.004.

Saez, Emmanuel and Gabriel Zucman (2016). “Wealth Inequality in the United States since 1913: Evidence from Capitalized Income Tax Data”. In: *Quarterly Journal of Economics* 131.May, pp. 519–578.

– (2018). “Creating Homogeneous Synthetic Individual Tax Files for Distributional Analysis”. URL: https://www.irs.gov/pub/irs-soi/18rpsynthindtaxfiles.pdf.

– (2019). “Progressive Wealth Taxation”. URL: https://www.brookings.edu/wp-content/uploads/2019/09/Saez-Zucman_conference-draft.pdf.

Saichev, Alexander, Yannick Malevergne, and Didier Sornette (2010). *Theory of Zipf's Law and Beyond*. Berlin, Heidelberg: Springer Berlin Heidelberg. URL: http://dx.doi.org/10.1007/978-3-642-02946-2.

Schepsmeier, Ulf et al. (2018). *VineCopula: Statistical Inference of Vine Copulas*. R package version 2.1.8. URL: https://CRAN.R-project.org/package=VineCopula.

Schularick, Moritz, Moritz Kuhn, and Ulrike I. Steins (2018). “Income and Wealth Inequality in America, 1949-2016”. Minneapolis. URL: https://doi.org/10.21034/iwp.9.

Seim, David (2017). “Behavioral responses to wealth taxes: Evidence from Sweden”. In: *American Economic Journal: Economic Policy* 9.4, pp. 395–421.

Steinbaum, Marshall (2019). “Student Debt and Racial Wealth Inequality”.

Stiglitz, Joseph E (1969). “Distribution of Income and Wealth Among Individuals”. In: *Econometrica* 37.3, pp. 382–397. URL: https://www.jstor.org/stable/1912788.

Thompson, Jeffrey P. and Gustavo A. Suarez (2015). “Updating the Racial Wealth Gap”. Washington. URL: https://doi.org/10.17016/FEDS.2015.076r1.

United Nations (2017). *World Population Prospects*. Tech. rep. URL: https://population.un.org/wpp/Publications/.

United States Census Bureau (n.d.). *National Intercensal Tables: 1900-1990*. URL: https://www.census.gov/content/census/en/data/tables/time-series/demo/popest/pre-1980-national.html.

Vaughan, R. N. (1979). “Class Behaviour and the Distribution of Wealth”. In: *The Review of Economic Studies* 46.3, pp. 447–465.

Wold, H. O. A. and P. Whittle (1957). “A Model Explaining the Pareto Distribution of Wealth”. In: *Econometrica* 25.4, pp. 591–595.

Wolff, Edward N (2010). “Recent Trends in Household Wealth in the United States: Rising Debt and the Middle-Class Squeeze”. URL: https://ssrn.com/abstract=1585409%20or%20http://dx.doi.org/10.2139/ssrn.1585409.

General Conclusion

This PhD thesis contributes to the literature on the distribution of income and wealth in several ways. The first two chapter tackle methodological issues in measurement. The first one develop a new method to estimate complete distributions of income and wealth based tabulate income data, such as the one published by tax authorities. The second chapter shows how to properly combine surveys with tax data to correct for the underrepresentation of the rich, while preserving the basic structure of the microdata. The third chapter combines surveys, tax data and national accounts to produce distributional national accounts in Europe since the 1980s. The fourth chapter uses distributional national accounts data in the United States to decompose the long-run drivers of wealth inequality using a simple dynamic stochastic model of the wealth distribution.

A lot of the work presented here can be viewed in realtion to a much larger research project, one that involves many other researchers: the production of consistent, harmonized statistics on the distribution of income and wealth on the largest possible scale, and their exploitation to inform policy debates. This is the goal of the distributional national accounts (DINA) project being undertaken at the World Inequality Lab (Alvaredo et al., 2017). The methodologies presented in this thesis are meant to be used and extended by other researchers, as exhibited by the creation of a R package (`https://github.com/thomasblanchet/gpinter`) and a Stata command (`https://github.com/thomasblanchet/bfmcorr`) to implement the methods of the first two chapters.

There is a need for such an enterprise. Despite a certain amount of progress, we still lack an official, well-established data source for inequality statistics that uses consistent concepts and methodologies for the entire world. This issue is well-recognized by official institutions, as exhibited by various initiatives such as the expert group on disparities in national accounts (EG-DNA) at the OECD. Ideally, the goal would be for official institutions to one day start publishing their own distributional national accounts estimates as part of the standard framework of national accounts. This, after all, is what happened with existing national accounts, which were developed by academics before being taken over by official statistical agencies.

The systematic confrontation of macro and micro sources would also force us to make progress in improving both. This is already true of aggregate national accounts. They combine various, disparate sources into a single consistent accounting framework. This consistency is what gives national accounts their potency. It is also a key driver of better data quality, as national accountants cannot afford to provide widely disparate estimates for the same concepts.

A lot has been said about the oversized importance of GDP in economic discourse. This importance has sometimes been exaggerated: "does it increase GDP?" has never been the only question on policymakers' mind. But there is some truth to it. GDP has shaped economic discourse in remarkable ways. We talk about how "the economy" is doing while implicitly referring to it. Policymaker track its quarterly variations with great attention. Yet if most of that growth accrues to small share of the population, it can creates wide discrepancies between economic statistics and the reality perceived by economic actors. In the future, the introduction of distributional estimates would be a powerful and palatable way to move "beyond GDP."

Appendix A

Appendix to "Generalized Pareto Curves: Theory and Applications"

A.1 Generalized Pareto curves: Omitted Proofs

A.1.1 Proof of proposition 1

That $b(p) > 1$ follows directly from the definition. For the rest of the proposition, we have for $p \geq \bar{p}$:

$$(1-p)Q(p)b(p) = \int_p^1 Q(u)\,\mathrm{d}u$$

We differentiate that equality with respect to p:

$$(1-p)Q(p)b'(p) + (1-p)b(p)Q'(p) - b(p)Q(p) = -Q(p)$$

We assumed that $Q(p) > 0$ for $p \geq \bar{p}$, so we can divide both sides by $Q(p)$:

$$(1-p)b'(p) + (1-p)b(p)\frac{Q'(p)}{Q(p)} - b(p) = -1$$

Hence:

$$(1-p)b(p)\frac{Q'(p)}{Q(p)} = b(p) - 1 - (1-p)b'(p)$$

Because the quantile function is increasing, the left hand side is nonnegative, which concludes the proof. □

A.1.2 Proof of proposition 2

From the proof of proposition 1, we have:

$$\frac{Q'(p)}{Q(p)} = \frac{1}{1-p} - \frac{1}{(1-p)b(p)} - \frac{b'(p)}{b(p)}$$

After integration:

$$\begin{aligned}
Q(p) &= Q(\bar{p}) \exp\left(\int_{\bar{p}}^{p} \frac{1}{1-u} \,\mathrm{d}u - \int_{\bar{p}}^{p} \frac{1}{(1-u)b(u)} \,\mathrm{d}u - \int_{\bar{p}}^{p} \frac{b'(u)}{b(u)} \,\mathrm{d}u \right) \\
&= Q(\bar{p}) \frac{(1-\bar{p})b(\bar{p})}{(1-p)b(p)} \exp\left(- \int_{\bar{p}}^{p} \frac{1}{(1-u)b(u)} \,\mathrm{d}u \right)
\end{aligned}$$

with $Q(\bar{p}) = \bar{x}$ by definition. □

A.1.3 Proof of proposition 3

The following representation of $b^*(x)$ will useful throughout the proofs.

Lemma 1.

$$b^*(x) = 1 + \frac{1}{x(1-F(x))} \int_x^{+\infty} 1 - F(z) \,\mathrm{d}z$$

Proof. Using integration by parts:

$$\begin{aligned}
\int_x^{+\infty} z f(z) \,\mathrm{d}z &= \int_x^{+\infty} (-z)(-f(z)) \,\mathrm{d}z \\
&= [-z(1-F(z))]_{z=x}^{+\infty} + \int_x^{+\infty} 1 - F(z) \,\mathrm{d}z
\end{aligned}$$

Because $\mathbb{E}[|X|] < +\infty$, Markov's inequality implies $1 - F(x) = o(1/x)$, so the bracketed term vanishes for $x \to +\infty$. Hence:

$$\int_x^{+\infty} z f(z) \,\mathrm{d}z = x(1-F(x)) + \int_x^{+\infty} 1 - F(z) \,\mathrm{d}z$$

replacing in the expression of $b^*(x)$ yields the result. □

First, note that since $\lim_{p\to 1} Q(p) = +\infty$, $\lim_{p\to 1} b(p) = \lim_{x\to +\infty} b^*(x)$. The assumption that L is slowly varying is equivalent to the assumption that $1 - F$ is regularly varying of index $-\alpha < -1$.

Direct half Applying the direct half of Karamata's theorem (Bingham, Goldie, and Teugels, 1989, 1.5.11, p. 28) to the representation of lemma 1, we have:

$$\lim_{x\to+\infty} \frac{1}{b^*(x)} = 1 + \frac{1}{\alpha-1} = \frac{\alpha}{\alpha-1}$$

Converse half We assume that $\lim_{p\to 1} b(p) = \alpha/(\alpha-1)$. Hence:

$$\lim_{x\to+\infty} \frac{1}{b^*(x)-1} = \alpha - 1$$

Then, we apply the converse half of Karamata's theorem (Bingham, Goldie, and Teugels, 1989, 1.6.1, p. 30) (with $\sigma = 0$) to the representation of lemma 1, proving that $1 - F$ is regularly varying of index $-\alpha$. □

A.1.4 Proof of proposition 4

Direct half According to lemma 1, we have:

$$b^*(x) = 1 + \frac{1}{x(1-F(x))} \int_x^{+\infty} 1 - F(z)\,\mathrm{d}z$$

After a change of variable $z = tx$:

$$\begin{aligned} b^*(x) &= 1 + \int_1^{+\infty} \frac{1-F(tx)}{1-F(x)}\,\mathrm{d}t \\ &= 1 + \int_1^{K} \frac{1-F(tx)}{1-F(x)}\,\mathrm{d}t + \int_K^{+\infty} \frac{1-F(tx)}{1-F(x)}\,\mathrm{d}t \end{aligned}$$

for some $K > 1$. The function $t \mapsto (1 - F(xt))/(1 - F(x))$ is continuous over the compact interval $[1, K]$, so it is bounded. Therefore, Lebesgue's dominated convergence theorem implies:

$$\lim_{x\to+\infty} \int_1^K \left[\frac{1-F(tx)}{1-F(x)}\right] \mathrm{d}t = \int_1^K \left[\lim_{x\to+\infty} \frac{1-F(tx)}{1-F(x)}\right] \mathrm{d}t = 0$$

Moreover, we assumed that $1-F$ is regularly varying. Therefore, using corollary 2.4.2 in Bingham, Goldie, and Teugels (1989, p. 85), the limit:

$$\lim_{x\to+\infty} \frac{1-F(xt)}{1-F(t)} = 0$$

holds uniformly for t over $[K, +\infty[$. The uniform convergence theorem implies:

$$\lim_{x\to+\infty} \int_K^{+\infty} \left[\frac{1-F(tx)}{1-F(x)}\right] \mathrm{d}t = \int_K^{+\infty} \left[\lim_{x\to+\infty} \frac{1-F(tx)}{1-F(x)}\right] \mathrm{d}t = 0$$

Therefore, we have $\lim_{x\to+\infty} b^*(x) = 1$. □

Converse half We assume that $\lim_{x\to+\infty} b^*(x) = 1$. Therefore:

$$\lim_{x\to+\infty} \int_1^{+\infty} \frac{1-F(tx)}{1-F(x)} \mathrm{d}t = 0$$

Let $\lambda > 1$ and $x \geq \bar{x}$. Because $t \mapsto (1-F(xt))/(1-F(x))$ is decreasing, we have for all $t < \lambda$:

$$\frac{1-F(\lambda x)}{1-F(x)} < \frac{1-F(tx)}{1-F(x)}$$

After integration with respect to t between 1 and λ:

$$\begin{aligned}\frac{1-F(\lambda x)}{1-F(x)} &< \frac{1}{\lambda-1}\int_1^{\lambda} \frac{1-F(tx)}{1-F(x)} \mathrm{d}t \\ &< \frac{1}{\lambda-1}\int_1^{+\infty} \frac{1-F(tx)}{1-F(x)} \mathrm{d}t\end{aligned}$$

because $(1-F(tx))/(1-F(x)) \geq 0$ for all t. Since the inequality holds for all $x > \bar{x}$, and the left hand side is nonnegative, we have for all $\lambda > 1$:

$$\lim_{x\to+\infty} \frac{1-F(\lambda x)}{1-F(x)} = 0$$

Therefore, $1-F$ is rapidly varying. □

A.2 Other Concepts of Local Pareto Coefficients

The inverted Pareto coefficient $b(p)$ is not the only local concept of Pareto coefficient that can be used to nonparametrically describe power law behavior. Using a simple principle, we can in fact define an infinite number of such coefficients, some of which have already been introduced in the literature (eg. Gabaix, 1999). First, notice that if $G(x) = 1 - F(x) = Cx^{-\alpha}$ is a strict power law, then for $n > 0$:

$$-\frac{xG^{(n)}(x)}{G^{(n-1)}(x)} - n + 1 = \alpha \tag{A.1}$$

which does not depend on x or n. But when the distribution isn't strictly Paretian, we can always define $\alpha_n(x)$ equal to the left-hand side of (A.1), which may now depend on x and n. For example $\alpha_2(x)$ correspond to the "local Zipf coefficient" as defined by Gabaix (1999).[1] For $n = 1$, we get $\alpha_1(x) = xf(x)/(1 - F(x))$. As long as $\alpha > -n + 1$, we can also extend formula (A.1) to zero or negative n, substituting integrals for negative orders of differentiation. More precisely, we set:

$$\forall n < 0 \qquad G^{(n)}(x) = (-1)^n \underbrace{\int_x^{+\infty} \cdots \int_{t_2}^{+\infty}}_{|n| \text{ times}} G(t_1)\,\mathrm{d}t_1 \ldots \mathrm{d}t_{|n|}$$

The definition above ensures that $G^{(n_1)(n_2)} = G^{(n_1+n_2)}$ for all $n_1, n_2 \in \mathbb{Z}$. We call $\alpha_n(x)$, $n \in \mathbb{Z}$, the *local Pareto coefficient* of order n. We have for $n = 0$:

$$\alpha_0(x) = 1 + \frac{x(1 - F(x))}{\int_x^{+\infty} 1 - F(t)\,\mathrm{d}t}$$

which implies:[2]

$$b(p) = \frac{\alpha_0(x)}{\alpha_0(x) - 1}$$

That formula corresponds to the inverted Pareto coefficient for a strict Pareto distribution $b = \alpha/(\alpha - 1)$. In fact, $b(p)$ is an alternative way of writing $\alpha_0(x)$, with a clearer interpretation in terms of economic inequality. We could similarly define inverted Pareto coefficients $b_n(p) = \alpha_n(x)/(\alpha_n(x) - 1)$ for any order n, and $b(p) = b_0(p)$. But $b_0(p)$ has the advantage of being the most simple to estimate, because it only involves quantiles and averages. Other estimators require estimating the density or one of its successive derivatives, which is much harder, especially when we have limited access to data.[3]

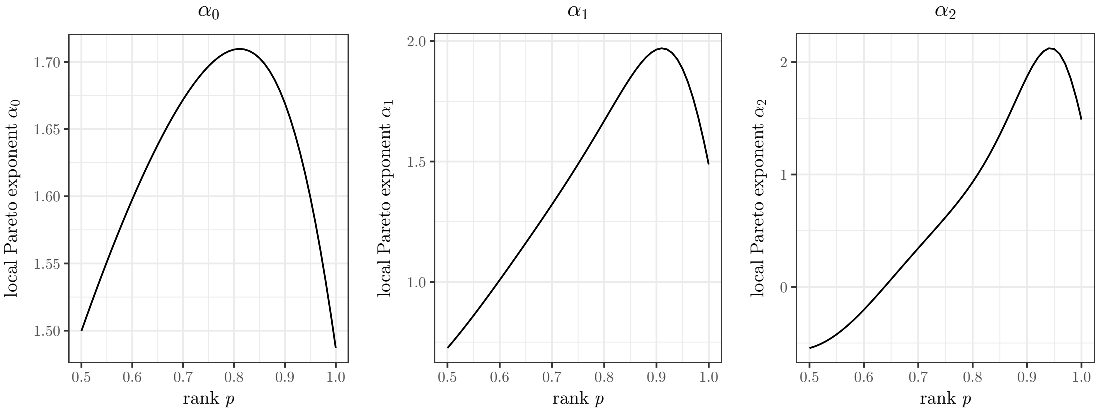
The most natural members of the family of local Pareto coefficients are α_0, α_1 and α_2 (other involve many orders of differentiation or integration). Figure A.1 shows how these different coefficients compare for the 2010 distribution of pre-tax national income in the United States. There are some differences regarding the range of values taken by the different coefficients. The inverted U-shape is less pronounced for α_1 and α_2 than α_0. But we reach similar conclusions regardless of the one we pick: the

[1] Which we can write more simply as $\alpha_2(x) = -xf'(x)/f(x) - 1$.

[2] See lemma 1 in appendix A.3.

[3] We can move from one coefficient to the next using the following recurrence relation:

$$\alpha_{n+1}(x) = \alpha_n(x) - \frac{x\alpha_n'(x)}{\alpha_n(x) + n - 1}$$



Distribution of pre-tax national income in the United States, 2010. α_0 estimated fitting a polynomial of degree 5 on empirical data. Source: authors' calculations using Piketty, Saez, and Zucman (2016)

Figure A.1: Different concepts of local Pareto exponent

coefficient is not constant (including at the very top) and there is an inversion of the slope near the top of the distribution. All these coefficients have fairly regular shapes, and looking at them conveys more information about the tail than solely looking at the quantile or the Lorenz curve. This is why it is better to work directly with them rather than with quantiles or shares.

A.3 Dynamic Model of Income Growth an Wealth Accumulation

A.3.1 Proofs omitted from the main text

We prove simultaneously the theorems 1 and 2. We assume that $\mu(x) \to \mu$, $\sigma^2(x) \to \sigma^2$. Recall that the process X_t follows the stochastic differential equation:

$$\frac{\mathrm{d}X_t}{X_t} = \mu(X_t)\,\mathrm{d}t + \sigma(X_t)\,\mathrm{d}W_t$$

which means that the evolution of its density $f(x,t)$ is described by the Fokker-Planck equation:

$$\frac{\partial}{\partial t} f(x,t) = -\frac{\partial}{\partial x}[x\mu(x)f(x,t)] + \frac{1}{2}\frac{\partial^2}{\partial x^2}[x^2\sigma^2(x)f(x,t)] \tag{A.2}$$

We also write:

$$\zeta(x) = 1 - \frac{2\mu(x)}{\sigma^2(x)}$$

and $\zeta = \lim_{x\to+\infty} \zeta(x) = 1 - 2\mu/\sigma^2$. For the stationary distribution $f(x)$, we have $\frac{\partial}{\partial t} f(x) = 0$, so equation (A.2) implies:

$$0 = -\frac{\partial}{\partial x}[x\mu(x)f(x)] + \frac{1}{2}\frac{\partial^2}{\partial x^2}[x^2\sigma^2(x)f(x)]$$

We can integrate that equation into:

$$\begin{aligned} x\mu(x)f(x) &= \frac{1}{2}\frac{\partial}{\partial x}[x^2\sigma^2(x)f(x)] \\ &= [x\sigma^2(x) + x^2\sigma(x)\sigma'(x)]f(x) + \frac{1}{2}x^2\sigma^2(x)f'(x) \end{aligned}$$

Reordering terms, we get:

$$\frac{f'(x)}{f(x)} = -\frac{\zeta+1}{x} - \frac{2\sigma'(x)}{\sigma(x)} - \frac{\zeta(x)-\zeta}{x}$$

And after integration:

$$f(x) \propto x^{-\zeta-1} \exp\left(-\log(\sigma^2(x)) - \int_1^x \frac{\zeta(t)-\zeta}{t}\,\mathrm{d}t\right)$$

Rewrite that expression as $f(x) = L(x)x^{-\zeta-1}$. Because $x \mapsto \zeta(x) - \zeta$ converges to zero and $x \mapsto \sigma^2(x)$ converges to a positive constant, Karamata's (1930) representation theorem (Bingham, Goldie, and Teugels, 1989, p. 12) implies that L is slowly varying. Then, we can use the following property of slowly varying functions:

$$1 - F(x) = \int_x^{+\infty} L(t)t^{-\zeta-1}\,\mathrm{d}t \sim L(x)\frac{x^{-\zeta}}{-\zeta}$$

to see that the stationary distribution is in fact an asymptotic power law. □

A.3.2 Alternative Calibrations

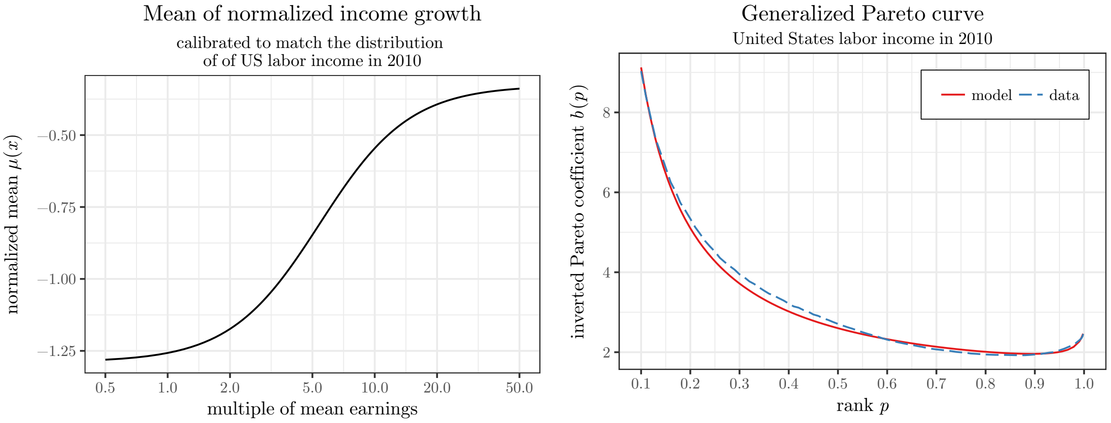
Income Distribution: Calibration of the mean Here we match the increase of $b(p)$ at the top by adjusting the mean of income shocks. We set:

$$\mu(x) = -c_1 + \frac{c_2x^2}{1 + c_3x^2} \tag{A.3}$$

with $c_1, c_2, c_3 > 0$. The baseline income growth of non-reflected units is $-c_1$, which is negative because we have normalized income by the overall income growth: since reflected units have positive growth, non-reflected units must have negative growth to compensate (Gabaix, 2009). The other part of the formula, $c_2x^2/(1 + c_3x^2)$, is here to make $\mu(x)$ increase at the top of the distribution. It also ensures that μ converges to a constant, so that we get a power law in the end. For the variance, we set:

$$\sigma(x) = \sqrt{\frac{c_4 + x^2}{x^2}}$$

which ensures a stationary process, and normalizes $\sigma(x)$ to 1 at infinity. (This normalization is necessary because μ and σ can only be identified up to a scaling constant.)



Model calibrated to match the US distribution of labor income in 2010 ($c_1 = 1.289, c_2 = 0.033, c_3 = 0.034, c_4 = 2.574$). $\mu(x)$ corresponds to the difference between the growth of non-reflected units and average income growth, expressed as a multiple of $\sigma(x)$ at infinity.

Figure A.2: Calibration of $\mu(x)$ on the US distribution of labor income

Figure A.2 shows the increase in the average of income shocks that is necessary to match the increase of $b(p)$ at the top.[4] We can see that the final rise in $b(p)$ is consistent with an increase of $\mu(x)$ of about one standard deviation between the middle and the very top of the distribution. The model with varying mean income growth is also able to precisely match almost the entire income distribution.

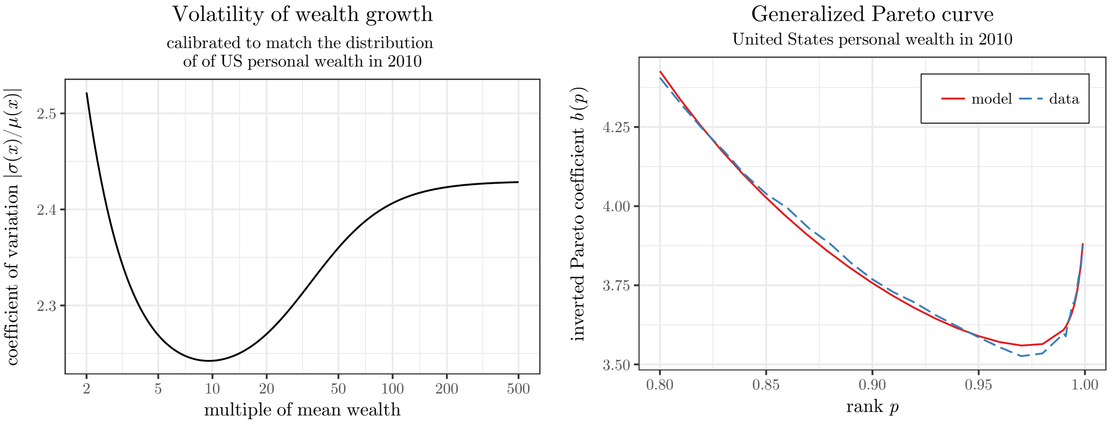
Gabaix et al. (2016) suggested that scale dependence in $\mu(x)$ is necessary to account for the speed of the increase in inequality in the United States. Our finding corroborates theirs, showing that scale dependence is can explain the shape of the distribution in a static framework, not just in a dynamic one.

Wealth Distribution: Calibration of the variance We can do a similar exercise for the distribution of wealth. The generalized Pareto curves for wealth and income are similar (U-shaped with a smaller increase at the top end). However, for

[4] The value of $\mu(x)$ only concerns the non-reflected units, but units at the top are unlikely to hit the reflecting barrier, so $\mu(x)$ constitute a good indicator of effective average income growth at the top.

wealth, $b(p)$ is higher overall, and the final increase happens later. What does this mean for the underlying process generating wealth? To answer that question, we consider a wealth generating process similar to that of income. We drop the reflective barrier because wealth can go below zero, but focus on the top 20%.[5]

We calibrate the profile of variance using the same formula as for income (1.3). We see in figure A.3 that we also get a U-shaped profile: wealth is more volatile at the very top of the distribution than at the middle. However, the increase starts much later, around ten times average wealth, which correspond roughly to the top 1%. For income, the increase started to happen around the top 10%. The difference between the lowest point and the top of the distribution is also more modest, at about 8% instead of 30%.

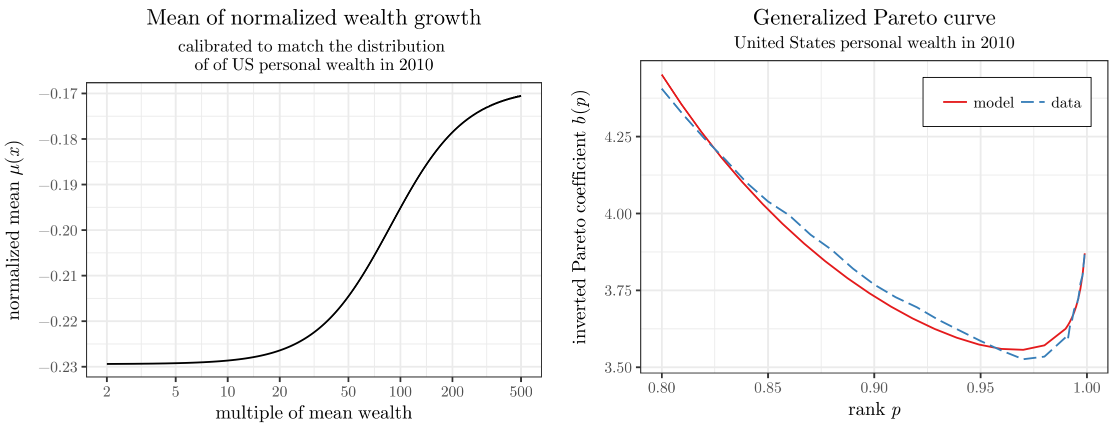


Model calibrated to match the US distribution of personal wealth in 2010 ($c_1 = 5.84, c_2 = 4.90, c_3 = 0.000809, c_4 = 0.000804$). The coefficient of variation correspond to the standard deviation divided by the absolute value of the mean growth.

Figure A.3: Calibration of $\sigma(x)$ on the US distribution of personal wealth

Wealth Distribution: Calibration of the mean We use again the formula (A.3) to model mean wealth growth. Figure A.4 shows the result. Again, we do observe an increasing mean of wealth growth (wealthier people experience higher returns, saving rates and/or higher incomes as a fraction of their wealth). But the increase is much more modest than for income (around 6% of a standard deviation). It also happens much later, starting at 10 times the average wealth (which roughly correspond to the top 1%).

[5] We focus on the top because we view these processes primarily as a model of the top of the distribution, even though it can sometimes fit well the bottom of the distribution too, as we saw for income. Wealth goes to zero fast once we leave top of the distribution, so providing a good fit for the bottom presents more difficulties. We do not explicitly model the negative part of the distribution because it is not necessary for our calibration.



Model calibrated to match the US distribution of labor income in 2010 ($c_1 = 0.229, c_2 = 0.00000788, c_3 = 0.000130, c_4 = 1.43$). The coefficient of variation correspond to the standard deviation divided by the absolute value of the mean growth.

Figure A.4: Calibration of $\mu(x)$ on the US distribution of personal wealth

This type of scale dependence is consistent with available microdata. Fagereng et al. (2016) document using administrative Norwegian data that returns are positively correlated with wealth. Because these higher returns partly reflect the fact that wealthier people hold riskier assets, it also implies higher variance at the top. Therefore, we have scale dependence for both the variance and the mean. This is also consistent with the work of Bach, Calvet, and Sodini (2017) on Swedish administrative data. The model of Kacperczyk, Nosal, and Stevens (2014), in which investors have different levels of sophistication, can account for these findings. Scale dependence can also arise if the very wealthy have higher saving rates (Saez and Zucman, 2016). Benhabib, Bisin, and Luo (2015) study a model where saving rates increase in wealth because the bequest function is more elastic than the utility of consumption.

A.4 Detailed interpolation method

Recall that $\hat{\varphi}_k$ correspond to the quintic spline over the interval $[x_k, x_{k+1}]$ ($1 \leq k < K$). We parametrized the spline (ie. the polynomial of degree 5) with $(y_k, y_{k+1}, s_k, s_{k+1}, a_k, a_{k+1})$ so that:

$$\begin{aligned} \hat{\varphi}_k(x_k) &= y_k & \hat{\varphi}'_k(x_k) &= s_k & \hat{\varphi}''_k(x_k) &= a_k \\ \hat{\varphi}_k(x_{k+1}) &= y_{k+1} & \hat{\varphi}'_{k+1}(x_{k+1}) &= s_{k+1} & \hat{\varphi}''_{k+1}(x_{k+1}) &= a_{k+1} \end{aligned}$$

The parameters $y_1, \dots, y_K$ and $s_1, \dots, s_K$ are directly given by the interpolation problem. But we still need to determine a_k, a_{k+1}. We first have $K-2$ equations to ensure $\mathcal{C}^3$ continuity at the junctures:

$$\forall k \in \{2, \dots, K-1\} \qquad \hat{\varphi}'''_{k-1}(x_k) = \hat{\varphi}'''_k(x_k)$$

Then, we impose the natural spline constraint at the first knot:

$$\hat{\varphi}'''_1(x_1) = 0$$

And we use a two points finite difference for the value of $\hat{\varphi}''_{K-1}(x_K)$:

$$\hat{\varphi}''_{K-1}(x_K) = \frac{s_K - s_{K-1}}{x_K - x_{K-1}}$$

That leads to a linear system of K equations for the K unknowns $a_1, \dots, a_K$. We can put that system in matrix form to solve it numerically using standard methods. Define $\Delta_k = x_{k+1} - x_k$. Then $\boldsymbol{a} = [a_1 \quad \cdots \quad a_K]'$ is given by $\boldsymbol{a} = \boldsymbol{A}^{-1}\boldsymbol{B}\boldsymbol{z}$, where:

$$\boldsymbol{z} = [y_2 - y_1 \quad \cdots \quad y_K - y_{K-1} \quad s_1 \quad \cdots \quad s_K]'$$

$$\boldsymbol{B} = [\boldsymbol{B}_1 | \boldsymbol{B}_2]$$

$$\boldsymbol{B}_1 = \begin{bmatrix} 60/\Delta_1^3 & 0 & 0 & \cdots & 0 & 0 \\ -60/\Delta_1^3 & 60/\Delta_2^3 & 0 & \cdots & 0 & 0 \\ 0 & -60/\Delta_2^3 & 60/\Delta_3^3 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 60/\Delta_{K-1}^3 & 0 \\ 0 & 0 & 0 & \cdots & -60/\Delta_{K-1}^3 & 60/\Delta_{K-1}^3 \\ 0 & 0 & 0 & \cdots & 0 & 0 \end{bmatrix}$$

$$\boldsymbol{B}_2 = \begin{bmatrix} -36/\Delta_1^2 & -24/\Delta_1^2 & 0 & \cdots & 0 & 0 \\ 24/\Delta_1^2 & 36/\Delta_1^2 - 36/\Delta_2^2 & -24/\Delta_2^2 & \cdots & 0 & 0 \\ 0 & 24/\Delta_2^2 & 36/\Delta_2^2 - 36/\Delta_3^2 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 36/\Delta_{K-1}^2 - 36/\Delta_{K-1}^2 & -24/\Delta_{K-1}^2 \\ 0 & 0 & 0 & \cdots & -1/\Delta_{K-1} & 1/\Delta_{K-1} \end{bmatrix}$$

$$\boldsymbol{A} = \begin{bmatrix} 9/\Delta_1 & -3/\Delta_1 & 0 & \cdots & 0 & 0 \\ -3/\Delta_1 & 9/\Delta_1 + 9/\Delta_2 & -3/\Delta_2 & \cdots & 0 & 0 \\ 0 & -3/\Delta_2 & 9/\Delta_2 + 9/\Delta_3 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 9/\Delta_{K-2} + 9/\Delta_{K-1} & -3/\Delta_{K-1} \\ 0 & 0 & 0 & \cdots & 0 & 1 \end{bmatrix}$$

A.5 Other comparisons of interpolation methods

We present here extended tables for the comparison of our new interpolation method with others. Those tables include a fourth interpolation method, described below, which was suggested by Cowell (2000, p. 158), yet virtually unused in the empirical literature. This method has a good pointwise performance, in many cases comparable to the generalized Pareto interpolation. However, it does not lead to a smooth quantile function or a continuous density.

We also include fiscal income in addition to pre-tax national income, as in the main text. Fiscal income tend to include a large fraction of individual with zero income, hence an important singularity near zero. To avoid that problem, we use a different tabulation in input, namely $p = 40\%, 70\%, 90\%, 99\%$.

Finally, we provide in table 1.2 the extrapolation results when the tabulation includes the top 10% and top 1%, and we seek the top 0.1%.

Method 4: piecewise Pareto distribution The method uses the Pareto distribution with information on both the thresholds and the means. It works by adjusting both the constant μ and the Pareto coefficient α of a Pareto distribution inside each bracket. The density over $[q_k, q_{k+1}]$ is:

$$f(x) = c_k x^{-\alpha_k - 1}$$

so that we get a nonlinear system of two equation with two unknowns (α_k and c_k), and two knowns $\xi_k = \int_{q_k}^{q_{k+1}} f(x) \, \mathrm{d}x$ and $\zeta_k = \int_{q_k}^{q_{k+1}} x f(x) \, \mathrm{d}x$. α_k is the solution of:

$$\frac{\alpha_k}{\alpha_k - 1} \frac{q_{k+1}^{1-\alpha_k} - q_k^{1-\alpha_k}}{q_{k+1}^{-\alpha_k} - q_k^{-\alpha_k}} = \zeta_k$$

which has no explicit solution but can be solved numerically. Then:

$$c_k = \frac{\alpha_k \xi_k}{q_{k+1}^{-\alpha_k} - q_k^{-\alpha_k}}$$

For $k = K$, so that $p_{K+1} = 1$ and $q_{K+1} = +\infty$, it becomes equivalent to method 1.

Table A.1: Mean relative error for different interpolation methods (fiscal income)

		mean percentage gap between estimated and observed values				
		M0	M1	M2	M3	M4
United States (1962–2014)	Top 50% share	0.042% (ref.)	0.59% (×14)	5.4% (×129)	0.035% (×0.83)	0.019% (×0.46)
	Top 20% share	0.037% (ref.)	0.34% (×9.3)	5.7% (×156)	0.021% (×0.56)	0.072% (×2)
	Top 5% share	0.11% (ref.)	1.3% (×11)	11% (×96)	0.54% (×4.8)	0.11% (×1)
	P50/average	0.57% (ref.)	14% (×25)	7.7% (×14)	0.39% (×0.68)	0.34% (×0.6)
	P80/average	0.13% (ref.)	2% (×16)	2.8% (×21)	1.2% (×9.2)	0.19% (×1.5)
	P95/average	0.42% (ref.)	6.9% (×16)	4.2% (×9.9)	1.6% (×3.7)	0.6% (×1.4)
France (1994–2012)	Top 50% share	0.055% (ref.)	0.42% (×7.6)	1.8% (×32)	0.019% (×0.34)	0.043% (×0.78)
	Top 20% share	0.032% (ref.)	0.35% (×11)	1.4% (×42)	0.02% (×0.63)	0.056% (×1.7)
	Top 5% share	0.05% (ref.)	0.35% (×6.8)	2.5% (×49)	0.43% (×8.5)	0.039% (×0.78)
	P50/average	0.48% (ref.)	7.3% (×15)	4.1% (×8.5)	0.31% (×0.65)	0.41% (×0.86)
	P80/average	0.058% (ref.)	2% (×35)	1.6% (×27)	1.1% (×18)	0.12% (×2)
	P95/average	0.11% (ref.)	1.4% (×13)	0.74% (×6.9)	2.1% (×20)	0.12% (×1.1)

Pre-tax national income. Sources: author's calculation from Piketty, Saez, and Zucman (2016) (United States) and Garbinti, Goupille-Lebret, and Piketty (2016) (France). The different interpolation methods are labeled as follows. M0: generalized Pareto interpolation. M1: constant Pareto coefficient. M2: log-linear interpolation. M3: mean-split histogram. M4: piecewise Pareto distribution. We applied them to a tabulation which includes the percentiles $p = 40\%$, $p = 70\%$, $p = 90\%$, and $p = 99\%$. We included the relative increase in the error compared to generalized Pareto interpolation in parentheses. We report the mean relative error, namely:

$$\frac{1}{\text{number of years}} \sum_{t=\text{first year}}^{\text{last year}} \left| \frac{\hat{y}_t - y_t}{y_t} \right|$$

where y is the quantity of interest (income threshold or top share), and $\hat{y}$ is its estimate using one of the interpolation methods. We calculated the results over the years 1962, 1964 and 1966–2014 in the United States, and years 1994–2012 in France.

Table A.2: Mean relative error for different interpolation methods (pre-tax national income)

		mean percentage gap between estimated and observed values				
		M0	M1	M2	M3	M4
United States (1962–2014)	Top 70% share	0.059% (ref.)	2.3% (×38)	6.4% (×109)	0.054% (×0.92)	0.055% (×0.94)
	Top 25% share	0.093% (ref.)	3% (×32)	3.8% (×41)	0.54% (×5.8)	0.55% (×5.9)
	Top 5% share	0.058% (ref.)	0.84% (×14)	4.4% (×76)	0.83% (×14)	0.22% (×3.8)
	P30/average	0.43% (ref.)	55% (×125)	29% (×67)	1.4% (×3.3)	0.48% (×1.1)
	P75/average	0.32% (ref.)	11% (×35)	9.9% (×31)	5.8% (×18)	0.31% (×0.99)
	P95/average	0.3% (ref.)	4.4% (×15)	3.6% (×12)	1.3% (×4.5)	0.88% (×3)
France (1994–2012)	Top 70% share	0.55% (ref.)	4.2% (×7.7)	7.3% (×13)	0.14% (×0.25)	0.082% (×0.15)
	Top 25% share	0.75% (ref.)	1.8% (×2.4)	4.9% (×6.5)	0.37% (×0.49)	0.34% (×0.46)
	Top 5% share	0.29% (ref.)	1.1% (×3.9)	8.9% (×31)	0.49% (×1.7)	0.095% (×0.33)
	P30/average	1.5% (ref.)	59% (×40)	38% (×26)	2.6% (×1.8)	0.26% (×0.18)
	P75/average	1% (ref.)	5.2% (×5.1)	5.4% (×5.3)	4.7% (×4.6)	0.28% (×0.27)
	P95/average	0.58% (ref.)	5.6% (×9.6)	3.2% (×5.5)	1.8% (×3.2)	0.48% (×0.82)

Pre-tax national income. Sources: author's calculation from Piketty, Saez, and Zucman (2016) (United States) and Garbinti, Goupille-Lebret, and Piketty (2016) (France). The different interpolation methods are labeled as follows. M0: generalized Pareto interpolation. M1: constant Pareto coefficient. M2: log-linear interpolation. M3: mean-split histogram. M4: piecewise Pareto distribution. We applied them to a tabulation which includes the percentiles $p = 10\%$, $p = 50\%$, $p = 90\%$, and $p = 99\%$. We included the relative increase in the error compared to generalized Pareto interpolation in parentheses. We report the mean relative error, namely:

$$\frac{1}{\text{number of years}} \sum_{t=\text{first year}}^{\text{last year}} \left| \frac{\hat{y}_t - y_t}{y_t} \right|$$

where y is the quantity of interest (income threshold or top share), and $\hat{y}$ is its estimate using one of the interpolation methods. We calculated the results over the years 1962, 1964 and 1966–2014 in the United States, and years 1994–2012 in France.

Table A.3: Mean relative error on the top 0.1% for different extrapolation methods, knowing the top 10% and the top 1%

		mean percentage gap between estimated and observed values		
		M0	M1	M2
United States (1962–2014)	Top 0.1% share	3.4% (ref.)	4.2% (×1.2)	46% (×13)
	P99.9/average	5.5% (ref.)	4% (×0.72)	23% (×4.2)
France (1994–2012)	Top 0.1% share	1.4% (ref.)	4.5% (×3.2)	20% (×15)
	P99.9/average	1% (ref.)	2.1% (×2)	8.2% (×7.9)

Fiscal income. Sources: author's calculation from Piketty, Saez, and Zucman (2016) (United States) and Garbinti, Goupille-Lebret, and Piketty (2016) (France). The different extrapolation methods are labeled as follows. M0: generalized Pareto distribution. M1: constant Pareto coefficient. M2: log-linear interpolation. We applied them to a tabulation which includes the percentiles $p = 90\%$, and $p = 99\%$. We included the relative increase in the error compared to generalized Pareto interpolation in parentheses. We report the mean relative error, namely:

$$\frac{1}{\text{number of years}} \sum_{t=\text{first year}}^{\text{last year}} \left| \frac{\hat{y}_t - y_t}{y_t} \right|$$

where y is the quantity of interest (income threshold or top share), and $\hat{y}$ is its estimate using one of the interpolation methods. We calculated the results over the years 1962, 1964 and 1966–2014 in the United States, and years 1994–2012 in France.

A.6 Error estimation

A.6.1 Decomposition of the error

Recall that the tabulation is based on $K \geq 3$ fractiles of the population $p_1, \dots, p_K$ such that $0 \leq p_1 < \cdots < p_K < 1$. Let $k \in \{1, \dots, K-1\}$ and $p \in [0,1]$ a fractile such that $p_k \leq p \leq p_{k+1}$. We also define $x = -\log(1-p)$.

Let n be the size of the population covered by the tabulation. Income or wealth are represented as a iid. copies $(X_1, \dots, X_n)$ of the random variable X. The empirical quantile is $\hat{Q}_n(p) = X_{(\lfloor np \rfloor)}$ where $X_{(r)}$ is the r-th order statistic (i.e. the r-th largest value of the sample). We note $\overline{X}_n$ empirical average, and the empirical Lorenz curve is:

$$\hat{L}_n(p) = \frac{\sum_{i=1}^{\lfloor np \rfloor} X_{(i)}}{n\overline{X}_n}$$

Formally, we define the tabulation as the $(2K+1)$-tuple:

$$\boldsymbol{T}_n = [\overline{X}_n \quad \overline{X}_n \hat{L}_n(p_1) \quad \cdots \quad \overline{X}_n \hat{L}_n(p_K) \quad \hat{Q}_n(p_1) \quad \cdots \quad \hat{Q}_n(p_K)]$$

And its theoretical counterpart is:

$$\boldsymbol{T}_\infty = [\mathbb{E}[X] \quad \mathbb{E}[X]L(p_1) \quad \cdots \quad \mathbb{E}[X]L(p_K) \quad Q(p_1) \quad \cdots \quad Q(p_K)]$$

We define $\hat{\varphi}_n$ the function that we obtain through the procedure of section 3.1 on the tabulation $\boldsymbol{T}_n$. We also define $\hat{\varphi}_\infty$ the function that would be obtained with the same method on the tabulation $\boldsymbol{T}_\infty$. Then, we define $\varphi_n(x) = -\log((1-\hat{L}_n(p))\overline{X}_n)$ the plug-in estimator of φ (hence, we may write $\varphi_\infty = \varphi$). We use analogous notations for φ'. The error at point x is:

$$\begin{aligned} e_n(x) &= \hat{\varphi}_n(x) - \varphi_n(x) \\ &= \underbrace{\hat{\varphi}_\infty(x) - \varphi_\infty(x)}_{\text{misspecification error}} + \underbrace{\hat{\varphi}_n(x) - \hat{\varphi}_\infty(x) + \varphi_\infty(x) - \varphi_n(x)}_{\text{sampling error}} \end{aligned}$$

We can set $u(x) = \hat{\varphi}_\infty(x) - \varphi_\infty(x)$ and $v_n(x) = \hat{\varphi}_n(x) - \hat{\varphi}_\infty(x) + \varphi_\infty(x) - \varphi_n(x)$, which proves the first part of theorem 7.

A.6.2 Misspecification error

The magnitude of the misspecification error depends on two features. First, the tightness of the tabulation in input: we can better estimate the true shape of the distribution if we have access to many percentiles of the data. Second, the "regularity" of the function we seek to approximate: in loose terms, for the interpolation to work well, the function φ should not stray too far away from a polynomial of sufficiently low degree.

It is possible to express the misspecification error in a way that disentangle both effects. To that end, define an operator $\mathcal{E}$ which, to a function g over $[x_1, x_K]$, associates the interpolation error $\hat{g} - g$. It satisfies the three following properties:[6]

- **linearity:** $\mathcal{E}(f + \lambda g) = \mathcal{E}(f) + \lambda\mathcal{E}(g)$
- **inversion with integral sign:** $\mathcal{E}\left\{\int_{x_1}^{x_K} f(\,\cdot\,, t)\,\mathrm{d}t\right\} = \int_{x_1}^{x_K} \mathcal{E}\{f(\,\cdot\,, t)\}\,\mathrm{d}t$
- **exact for polynomials of degree up to 2:** if $f \in \mathbb{R}[X]$, $\deg(f) \leq 2$, then $\mathcal{E}(f) = 0$

Under those conditions, the Peano kernel theorem gives a simple formula for the error term. Consider the Taylor expansion of the true function φ with integral remainder:

$$\varphi(x) = \varphi(x_1) + (x - x_1)\varphi'(x_1) + \frac{1}{2}(x - x_1)^2\varphi''(x_1) + \frac{1}{2}\int_{x_1}^{x_K} (x - t)_+^2 \varphi'''(t)\,\mathrm{d}t$$

where $(x - t)_+ = \max\{x - t, 0\}$. Using the properties of $\mathcal{E}$, we have:

$$\hat{\varphi}_\infty(x) - \varphi_\infty(x) = \frac{1}{2}\int_{x_1}^{x_K} \mathcal{E}(K_t)(x)\varphi'''(t)\,\mathrm{d}t \tag{A.4}$$

where $K_t : x \mapsto (x - t)_+^2$, so that $\mathcal{E}(K_t)(x)$ is independent from φ. That last expression corresponds to the Peano kernel theorem. We get a similar expression for the first derivative $\varphi'(x)$. Therefore, setting $\varepsilon(x, t) = \mathcal{E}(K_t)(x)/2$ proves theorem 8.

The interpolation error at t can therefore be written as a scalar product between two functions. The first one, $t \mapsto \mathcal{E}(K_t)(x)$, depends only on the position of the brackets in terms the percentiles of the distribution. If the fractiles $p_1, \dots, p_K$ included in the tabulation get more numerous and closer to each other, its value will get closer to zero. The other term, $t \mapsto \varphi'''(t)$, characterizes the regularity of the distribution. When $\varphi''' = 0$, the interpolated function is a polynomial of degree 2, so the interpolation error is zero. That is the case, in particular, of strict Pareto distributions, which the method can interpolate exactly. φ''' is best viewed as a "residual": it summarizes all the properties of the underlying distribution that are not properly captured by the

[6] Many other interpolation methods would satisfy those three properties (possibly with different degrees for the polynomial), so that the results of this section could be extended to them with minimal changes.

functional form used in the interpolation.

We can obtain a first inequality on the absolute value of the error using the triangular inequality:

$$|\hat{\varphi}_\infty(x) - \varphi_\infty(x)| \leq \frac{||\varphi'''||_\infty}{2} \int_{x_1}^{x_K} |\mathcal{E}(K_t)(x)| \, \mathrm{d}t$$

where $||\varphi'''||_\infty = \sup\{|\varphi'''(t)| : x_1 \leq t \leq x_K\}$. That last formula is a conservative bound on the error, which would only be attained in the worst-case scenario where φ''' would frequently switch signs so as to systematically amplify the value of $\mathcal{E}(K_t)(x)$. Still, it remains interesting because we can evaluate it (using numerical integration), independently of φ, up to a multiplicative constant, and it gives insights on the shape of the error that will remain valid even after refinements. Figure A.5 show this bound for a tabulation with fractiles 10%, 50%, 90% and 99%.

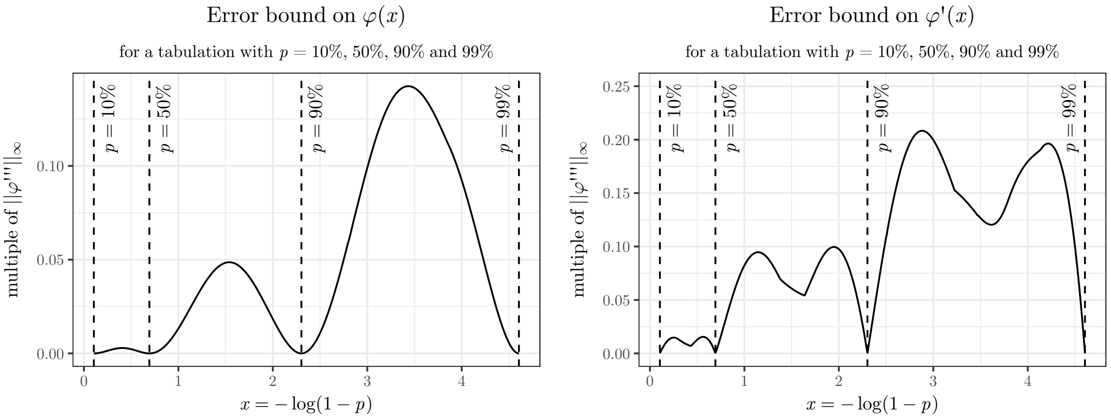


Figure A.5: Bounds on the misspecification error term for φ and φ'

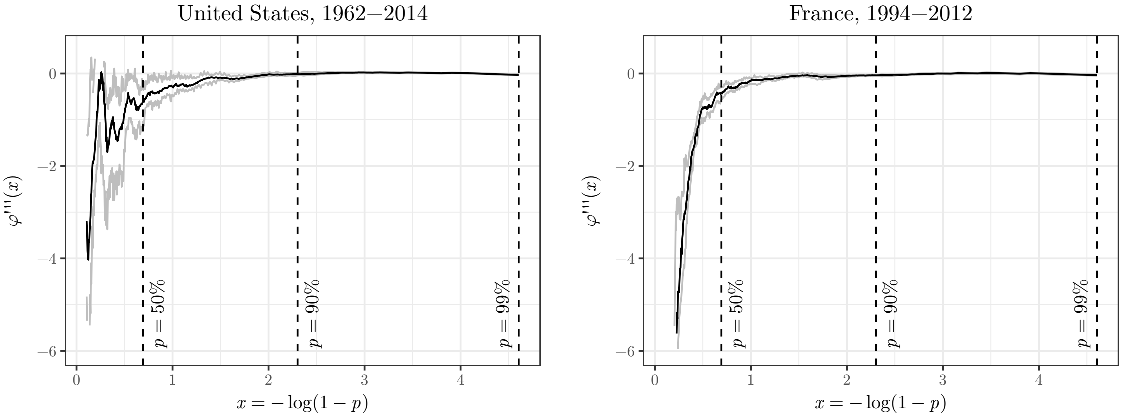
As expected, the error term is equal to zero for both φ and φ' at all the fractiles included in the tabulation. The error is also larger when the log-transformed bracket $[-\log(1-p_k), -\log(1-p_{k+1})]$ is wider. The overall shape of the error is quite different for φ and φ'. For φ, the error bound is bell-shaped within brackets, and its maximal value is attained near the middle of it. The error bound on φ' admits two peaks within each bracket, with the maximal error occuring somewhere near the 1/4th and the 3/4th of it. Estimates at the middle of each bracket are actually more precise than at those two values. That somewhat atypical profile is explained by the fact that the integral of φ' over $[x_k, x_{k+1}]$ is known and equal to $\varphi(x_{k+1}) - \varphi(x_k)$. Therefore, if we overestimate φ' in the first half of the bracket, we will have to underestimate it in the second half to compensate. By continuity, the error will have to equal to zero at some point, and that will happen somewhere near the middle of the interval.

Going back to the quantities that are ultimately of interest: the error on top shares follows the shape of the error for φ, while the error on quantiles follows the shape of the error for φ'. In fact, for top shares, the error can be written as:

$$\begin{aligned}\left|\frac{\mathrm{e}^{-\hat{\varphi}(x)} - \mathrm{e}^{-\varphi(x)}}{\mathbb{E}[X]}\right| &= \left|\frac{\mathrm{e}^{-\varphi(x)-\varepsilon(x)} - \mathrm{e}^{-\varphi(x)}}{\mathbb{E}[X]}\right| \\ &= \left|\frac{\mathrm{e}^{-\varphi(x)}}{\mathbb{E}[X]}(\mathrm{e}^{-\varepsilon(x)} - 1)\right| \\ &\approx \frac{\mathrm{e}^{-\varphi(x)}}{\mathbb{E}[X]}\left|\varepsilon(x)\right|\end{aligned}$$

where $\varepsilon(x)$ is the interpolation on φ. If it is small, then at the first order, the absolute error on φ corresponds to the relative error on top shares.

$\mathcal{E}(K_t)(x)$ only depends on known parameters, but we still need φ''' to use (A.4) in practice comes from . With sufficiently detailed tabulations, we can estimate it nonparametrically via local polynomial fitting on the empirical values of ϕ and ϕ'. Figure A.6 shows the results, performed separately on the United States and France over all available years.



Pre-tax national income. Sources: author's computation from Piketty, Saez, and Zucman (2016) (for the United States) and Garbinti, Goupille-Lebret, and Piketty (2016) (for France). Median value over all years in black, first and tenth deciles in gray. Estimation by local polynomial fitting of degree 3 on both φ and φ' with Gaussian kernel and adaptive bandwidth so that 5% of observations are within one standard deviation of the Gaussian kernel.

Figure A.6: Estimations of $\varphi'''(x)$

There is some similarity between both countries. The function φ''' can take relatively high values in the bottom half of the distribution, but then quickly converges to zero. Although it takes fairly different shapes in the bottom half, the shapes are actually very similar above $p = 50\%$. Within each country, there is also a certain stability over time (especially for France), as exhibited by the gray lines showing the first and the tenth decile of estimated values over all years.

A.6.3 Sampling error

We first prove that the sampling error converges to zero as the sample size increases (second part of theorem 4). This result is a natural consequence of the fact that the tabulation of the data used in input eventually converges to the theoretical values implied by the underlying distribution, so that the sampling error eventually vanishes.

Proof. Consider the function Θ:

$$\begin{array}{cccc} & \mathbb{R}^{2K+3} & \to & \mathbb{R}^2 \\ \Theta : & \begin{bmatrix} \boldsymbol{T}_n \\ \overline{X}_n \hat{L}_n(p^*) \\ \hat{Q}_n(p^*) \end{bmatrix} & \mapsto & \begin{bmatrix} \hat{\varphi}_n(x^*) - \varphi_n(x^*) \\ \hat{\varphi}_n'(x^*) - \varphi_n'(x^*) \end{bmatrix} \end{array}$$

which is continuously differentiable since it is a combination of continuously differentiable functions. This function takes as input the tabulation $\boldsymbol{T}_n$, and the value of the Lorenz curve and the quantile at p^*. It returns the difference the estimated value of $\varphi_n(x^*)$ and its actual value. The sampling error, as we defined it, correspond to:

$$\Theta[\boldsymbol{T}_n, \overline{X}_n \hat{L}_n(p^*), \hat{Q}_n(p^*)] - \Theta[\boldsymbol{T}_\infty, \mathbb{E}[X] L(p^*), Q(p^*)]$$

The strong law of large number — alongside analogous results for the sample quantile — implies:

$$\begin{aligned} \boldsymbol{T}_n &\xrightarrow{\text{a.s.}} \boldsymbol{T}_\infty \\ \overline{X}_n \hat{L}_n(p^*) &\xrightarrow{\text{a.s.}} \mathbb{E}[X] L(p^*) \\ \hat{Q}_n(p^*) &\xrightarrow{\text{a.s.}} Q(p^*) \end{aligned}$$

Therefore, the continuous mapping theorem implies:

$$\Theta[\boldsymbol{T}_n, \overline{X}_n \hat{L}_n(p^*), \hat{Q}_n(p^*)] - \Theta[\boldsymbol{T}_\infty, \mathbb{E}[X] L(p^*), Q(p^*)] \xrightarrow{\text{a.s.}} 0$$

□

To prove theorem 6 on the speed of convergence of the error, we need to make additional regularity assumptions on the distribution.

Assumption 1. *$f > 0$ and f' is bounded in a neighborhood of* $[Q(p_1), Q(p_K)]$.

Assumption 1 covers most relevant cases.[7] It allows the Bahadur (1966) representation of the sample quantile to hold, so that it has a regular asymptotic behavior. It also implies asymptotic normality of the trimmed mean (Stigler, 1973), and, by extension, of the Lorenz curve.

Next, we distinguish two cases, depending on the tail behavior of the distribution. If it has a sufficiently thin upper tail, then the distribution will have finite variance. But it is common for the distribution of income (and *a fortiori* wealth) to have much fatter upper tails, leading to infinite variance. This distinction has important consequences for the asymptotic behavior of the sample mean, and by extension of our estimator.

A.6.3.1 The finite variance case

For a strict power law, finite variance corresponds to $b(p) > 2$. More generally, we can state the finite variance assumption using the second-order moment.

Assumption 2. *X has a finite second-order moment, i.e.* $\mathbb{E}[X^2] < +\infty$.

When variance is finite, the central limit theorem applies. Hence, we get the standard result of asymptotic normality and convergence rate $n^{-1/2}$ using the delta method.

Proof. We start by deriving the asymptotic joint distribution of all the quantiles and the means in the tabulation, which is multivariate normal. Then, theorem 9 for finite variance follows from the delta method applied to Θ.

For $k, k_1, k_2 \in \{1, \dots, K+1\}$, define:

$$\begin{aligned} U_i^k &= X_i \mathbb{1} Q(p_k) < X_i \leq Q(p_{k+1}) \\ V_k = \mathbb{X} \leq \mathbb{Q}(_{\mathbb{k}}) \mu_k &= \mathbb{E}[U_i^k] \\ \sigma_k^2 &= \mathrm{Var}(U_i^k) \\ \sigma_{k_1,k_2} &= \mathrm{Cov}(U_i^{k_1}, U_i^{k_2}) \end{aligned}$$

[7] There is one situation where that assumption may seem problematic, namely if $p_1 = 0$ and the distribution has a finite lower bound. However, in such cases, the value of $Q(p_1)$ is generally known *a priori* (typically, because we assumed income is nonnegative) and is therefore not subject to sampling variability. Hence, it will not affect the results.

We will also use the following matrix notations:

$$\begin{aligned}
\boldsymbol{U}_i &= [U_i^1 \quad U_i^2 \quad \cdots \quad U_i^{K+1}]' \\
\boldsymbol{V}_i &= [V_i^1 \quad V_i^2 \quad \cdots \quad V_i^{K+1}]' \\
\boldsymbol{\mu} &= [\mu_1 \quad \mu_2 \quad \cdots \quad \mu_{K+1}]'
\end{aligned}$$

We start with a lemma that gives the joint asymptotic distribution of $\overline{\boldsymbol{U}}_n$ and $\overline{\boldsymbol{V}}_n$.

Lemma 2.

$$\sqrt{n}\begin{bmatrix} \overline{\boldsymbol{U}}_n - \boldsymbol{\mu} \\ \overline{\boldsymbol{V}}_n - \boldsymbol{p} \end{bmatrix} \overset{\mathcal{D}}{\to} \mathcal{N}\left(\boldsymbol{0}, \begin{bmatrix} \boldsymbol{A} & \boldsymbol{C} \\ \boldsymbol{C}' & \boldsymbol{B} \end{bmatrix}\right)$$

where:

$$\boldsymbol{A} = \begin{bmatrix} \sigma_1^2 & \sigma_{1,2} & \cdots & \sigma_{1,K+1} \\ \sigma_{2,1} & \sigma_2^2 & \cdots & \sigma_{2,K+1} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{K+1,1} & \sigma_{K+1,2} & \cdots & \sigma_{K+1}^2 \end{bmatrix}$$

$$\boldsymbol{B} = \begin{bmatrix} \tilde{p}_1(1-\tilde{p}_1) & \tilde{p}_1(1-\tilde{p}_2) & \cdots & \tilde{p}_1(1-\tilde{p}_{K+1}) \\ \tilde{p}_1(1-\tilde{p}_2) & \tilde{p}_2(1-\tilde{p}_2) & \cdots & \tilde{p}_2(1-\tilde{p}_{K+1}) \\ \vdots & \vdots & \ddots & \vdots \\ \tilde{p}_1(1-\tilde{p}_{K+1}) & \tilde{p}_2(1-\tilde{p}_{K+1}) & \cdots & \tilde{p}_{K+1}(1-\tilde{p}_{K+1}) \end{bmatrix}$$

$$\boldsymbol{C} = \begin{bmatrix} -\tilde{p}_1\mu_1 & \mu_1(1-\tilde{p}_2) & \cdots & \mu_1(1-\tilde{p}_K) & \mu_1(1-\tilde{p}_{K+1}) \\ -\tilde{p}_1\mu_2 & -\tilde{p}_2\mu_2 & \cdots & \mu_2(1-\tilde{p}_K) & \mu_2(1-\tilde{p}_{K+1}) \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ -\tilde{p}_1\mu_K & -\tilde{p}_2\mu_K & \cdots & -\tilde{p}_K\mu_K & \mu_K(1-\tilde{p}_{K+1}) \\ -\tilde{p}_1\mu_{K+1} & -\tilde{p}_2\mu_{K+1} & \cdots & -\tilde{p}_{K+1}\mu_{K+1} & -\tilde{p}_{K+1}\mu_{K+1} \end{bmatrix}$$

Proof. We have, for $i \in \{1, \dots, n\}$ and $k_1, k_2 \in \{1, \dots, K+1\}$ with $k_1 < k_2$:

$$\begin{aligned}
\mathbb{E}[V_i^{k_1}] &= p_{k_1} \\
\operatorname{Var}(V_i^{k_1}) &= p_{k_1}(1-p_{k_1}) \\
\operatorname{Cov}(V_i^{k_1}, V_i^{k_2}) &= p_{k_1}(1-p_{k_2})
\end{aligned}$$

and for $k_1, k_2 \in \{1, \ldots, K+1\}$:

$$\mathrm{Cov}(U_i^{k_1}, V_i^{k_2}) = \begin{cases} -p_{k_2}\mu_{k_1} & \text{if} \quad k_2 \le k_1 \\ \mu_{k_1}(1 - p_{k_2}) & \text{if} \quad k_2 > k_1 \end{cases}$$

Therefore, by the central limit theorem:

$$\sqrt{n}\begin{bmatrix} \overline{\boldsymbol{U}}_n - \boldsymbol{\mu} \\ \overline{\boldsymbol{V}}_n - \boldsymbol{p} \end{bmatrix} \overset{\mathcal{D}}{\to} \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \boldsymbol{A} & \boldsymbol{C} \\ \boldsymbol{C}' & \boldsymbol{B} \end{bmatrix}\right)$$

with $\boldsymbol{A}$, $\boldsymbol{B}$ and $\boldsymbol{C}$ defined as in lemma 2. □

We know define for all $i \in \{1, \ldots, n\}$ and $k \in \{1, \ldots, K\}$:

$$W_i^k = X_i \mathbb{1}\{\hat{Q}_n(p_k) < X_i \le \hat{Q}_n(p_{k+1})\}$$

and for $k = K+1$:

$$W_i^{K+1} = X_i \mathbb{1}\{\hat{Q}_n(p_{K+1}) < X_i\}$$

and in matrix form:

$$\boldsymbol{W}_i = [W_i^1 \quad \cdots \quad W_i^{K+1}]$$

The definition of $\boldsymbol{W}_i$ is similar to that of $\boldsymbol{U}_i$, except that the quantile function was replaced by its empirical counterpart. We may now prove a second lemma, which correspond to the joint asymptotic distribution of $\overline{\boldsymbol{W}}_n$ and $\hat{\boldsymbol{Q}}_n$.

Lemma 3.

$$\sqrt{n}\begin{bmatrix} \overline{\boldsymbol{W}}_n - \boldsymbol{\mu} \\ \hat{\boldsymbol{Q}}_n - \boldsymbol{q} \end{bmatrix} \overset{\mathcal{D}}{\to} \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \boldsymbol{A} + \boldsymbol{MC}' + \boldsymbol{CM}' + \boldsymbol{MBM}' & -\boldsymbol{CN} - \boldsymbol{MBN} \\ -\boldsymbol{NC}' - \boldsymbol{NBM}' & \boldsymbol{NBN} \end{bmatrix}\right)$$

where $\boldsymbol{A}$, $\boldsymbol{B}$ and $\boldsymbol{C}$ are defined as in lemma 2, $\boldsymbol{N} = \mathrm{diag}(1/f(\tilde{q}_1), \ldots, 1/f(\tilde{q}_{K+1}))$, and:

$$\boldsymbol{M} = \begin{bmatrix} \tilde{q}_1 & -\tilde{q}_2 & 0 & \cdots & 0 & 0 \\ 0 & \tilde{q}_2 & -\tilde{q}_3 & \cdots & 0 & 0 \\ 0 & 0 & \tilde{q}_3 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \tilde{q}_K & -\tilde{q}_{K+1} \\ 0 & 0 & 0 & \cdots & 0 & \tilde{q}_{K+1} \end{bmatrix}$$

Proof. Note that for $k \in \{1, \dots, K\}$:

$$\begin{aligned}
\overline{W}_n^k &= \frac{1}{n} \sum_{i=\lfloor np_k \rfloor + 1}^{\lfloor np_{k+1} \rfloor} X_{(i)} \\
&= \frac{1}{n} \left[\sum_{i=\lfloor np_k \rfloor + 1}^{n\overline{V}_n^k} X_{(i)} + \sum_{i=n\overline{V}_n^k + 1}^{n\overline{V}_n^{k+1}} X_{(i)} + \sum_{i=n\overline{V}_n^{k+1} + 1}^{\lfloor np_{k+1} \rfloor} X_{(i)} \right] \\
&= \overline{U}_n^k + \frac{1}{n} \left[\sum_{i=\lfloor np_k \rfloor + 1}^{n\overline{V}_n^k} X_{(i)} + \sum_{i=n\overline{V}_n^{k+1} + 1}^{\lfloor np_{k+1} \rfloor} X_{(i)} \right]
\end{aligned}$$

Where $\sum_{i=a}^{b} x$ should be understood as $-\sum_{i=b}^{a} x$ if $a > b$. Therefore:

$$\sqrt{n}(\overline{W}_n^k - \mu_k) = \sqrt{n}(\overline{U}_n^k - \mu_k) + \frac{1}{\sqrt{n}} \left[\sum_{i=\lfloor np_k \rfloor + 1}^{n\overline{V}_n^k} X_{(i)} + \sum_{i=n\overline{V}_n^{k+1} + 1}^{\lfloor np_{k+1} \rfloor} X_{(i)} \right]$$

We have:

$$\begin{aligned}
\frac{1}{\sqrt{n}} \sum_{i=\lfloor np_k \rfloor + 1}^{n\overline{V}_n^k} X_{(i)} &= \frac{1}{\sqrt{n}} \sum_{i=\lfloor np_k \rfloor + 1}^{n\overline{V}_n^k} (X_{(i)} - q_k + q_k) \\
&= \frac{1}{\sqrt{n}} \sum_{i=\lfloor np_k \rfloor + 1}^{n\overline{V}_n^k} (X_{(i)} - q_k) + q_k \sqrt{n} \left(\overline{V}_n^k - \frac{\lfloor np_k \rfloor}{n} \right)
\end{aligned}$$

The first term converges in probability to zero because:

$$\left| \frac{1}{\sqrt{n}} \sum_{i=\lfloor np_k \rfloor + 1}^{n\overline{V}_n^k} (X_{(i)} - q_k) \right| \le \left| \frac{n\overline{V}_n^k - \lfloor np_k \rfloor}{\sqrt{n}} \right| \max\{|X_{(\lfloor np_k \rfloor + 1)} - q_k|, |X_{(n\overline{V}_n^k)} - q_k|\}$$

where the first term is bounded in probability, $|X_{(\lfloor np_k \rfloor + 1)} - q_k| \xrightarrow{\mathbb{P}} 0$ and $|X_{(n\overline{V}_n^k)} - q_k| \xrightarrow{\mathbb{P}} 0$. Hence:

$$\begin{aligned}
\frac{1}{\sqrt{n}} \sum_{i=\lfloor np_k \rfloor + 1}^{n\overline{V}_n^k} X_{(i)} &= q_k \sqrt{n}(\overline{V}_n^k - p_k) + \sqrt{n} \left(p_k - \frac{\lfloor np_k \rfloor}{n} \right) + o(1) \\
&= q_k \sqrt{n}(\overline{V}_n^k - p_k) + o(1)
\end{aligned}$$

Similarly:

$$\frac{1}{\sqrt{n}} \sum_{i=n\overline{V}_n^{k+1}+1}^{\lfloor np_{k+1} \rfloor} X_{(i)} = -q_{k+1}\sqrt{n}(\overline{V}_n^{k+1} - p_{k+1}) + o(1)$$

Therefore:

$$\sqrt{n}(\overline{W}_n^k - \mu_k) = \sqrt{n}(\overline{U}_n^k - \mu_k) + q_k\sqrt{n}(\overline{V}_n^k - p_k) - q_{k+1}\sqrt{n}(\overline{V}_n^{k+1} - p_{k+1}) + o(1)$$

By similar arguments:

$$\sqrt{n}(\overline{W}_n^{K+1} - \mu_{K+1}) = \sqrt{n}(\overline{U}_n^{K+1} - \mu_{K+1}) + q_{K+1}\sqrt{n}(\overline{V}_n^{K+1} - p_{K+1}) + o(1)$$

Hence, in matrix notation:

$$\sqrt{n}(\boldsymbol{\overline{W}}_n - \boldsymbol{\mu}) = \sqrt{n}(\boldsymbol{\overline{U}}_n - \boldsymbol{\mu}) + \boldsymbol{M}\sqrt{n}(\boldsymbol{\overline{V}}_n - \boldsymbol{p}) + o(1)$$

The Bahadur (1966) representation of the quantile implies:

$$\boldsymbol{\hat{Q}}_n - \boldsymbol{q} = -\boldsymbol{N}(\boldsymbol{\overline{V}}_n - \boldsymbol{p}) + o(n^{-1/2})$$

Therefore, we have:

$$\sqrt{n}\begin{bmatrix} \boldsymbol{\overline{W}}_n - \boldsymbol{\mu} \\ \boldsymbol{\hat{Q}}_n - \boldsymbol{q} \end{bmatrix} = \sqrt{n}\begin{bmatrix} \boldsymbol{I} & \boldsymbol{M} \\ \boldsymbol{0} & -\boldsymbol{N} \end{bmatrix}\begin{bmatrix} \boldsymbol{\overline{U}}_n - \boldsymbol{\mu} \\ \boldsymbol{\overline{V}}_n - \boldsymbol{p} \end{bmatrix} + o(1)$$

Using lemma 2, we get:

$$\sqrt{n}\begin{bmatrix} \boldsymbol{\overline{W}}_n - \boldsymbol{\mu} \\ \boldsymbol{\hat{Q}}_n - \boldsymbol{q} \end{bmatrix} \xrightarrow{\mathcal{D}} \mathcal{N}\left(\boldsymbol{0}, \begin{bmatrix} \boldsymbol{A} + \boldsymbol{MC'} + \boldsymbol{CM'} + \boldsymbol{MBM'} & -\boldsymbol{CN} - \boldsymbol{MBN} \\ -\boldsymbol{NC'} - \boldsymbol{NBM'} & \boldsymbol{NBN} \end{bmatrix}\right)$$

with $\boldsymbol{A}$, $\boldsymbol{B}$ and $\boldsymbol{C}$ defined as in lemma 2. □

Notice that:

$$\forall k \in \{1, \ldots, K+1\} \qquad \overline{X}_n\hat{L}_n(\tilde{p}_k) = \sum_{\ell=k}^{K+1} \overline{W}_n^{\ell} \qquad \text{and} \qquad \mathbb{E}[X]L(\tilde{p}_k) = \sum_{\ell=k}^{K+1} \mu_\ell$$

Therefore, we can write in matrix form that $\boldsymbol{\hat{L}}_n = \boldsymbol{P}\boldsymbol{\overline{W}}_n$ where:

$$\boldsymbol{\hat{L}}_n = [\overline{X}_n\hat{L}_n(\tilde{p}_1) \quad \cdots \quad \overline{X}_n\hat{L}_n(\tilde{p}_K)]$$

and $\boldsymbol{P}$ is the upper triangular matrix with only ones. Define $\nabla\Theta$ the gradient of Θ expressed at $[\boldsymbol{T}_\infty, \mathbb{E}[X]L(p^*), Q(p^*)]$, and:

$$\boldsymbol{R} = \begin{bmatrix} \boldsymbol{P} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{I} \end{bmatrix}$$

Denote $\boldsymbol{S}$ the covariance matrix of lemma 3, and $\boldsymbol{\Sigma} = (\nabla\Theta)'\boldsymbol{R}'\boldsymbol{S}\boldsymbol{R}(\nabla\Theta)$. The delta method (van der Vaart, 2000, p. 25) then implies:

$$\sqrt{n}(\Theta[\boldsymbol{T}_n, \overline{X}_n\hat{L}_n(p^*), \hat{Q}_n(p^*)] - \Theta[\boldsymbol{T}_\infty, \mathbb{E}[X]L(p^*), Q(p^*)]) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \boldsymbol{\Sigma})$$

□

A.6.3.2 The infinite variance case

When $\mathbb{E}[X^2] = +\infty$, the standard central limit theorem does not apply anymore. There is, however, a generalization due to Gnedenko and Kolmogorov (1968) that works with infinite variance. There are two main differences with the standard central limit theorem. The first one is that convergence operates more slowly than $n^{-1/2}$, the speed being determined by the asymptotic power law behavior of the distribution (the fatter the tail, the slower the convergence). The second one is that the limiting distribution belongs to a larger family than just the Gaussian, called *stable distributions*. With the exception of the Gaussian — which is an atypical member of the family of stable distributions — stable distributions exhibit fat tails and power law behavior. In most cases, their probability density function cannot be expressed analytically: only their characteristic function can. Although we designed our interpolation method with power laws in mind, we did not actually restrict the asymptotic behavior of the distribution, until now. But to apply the generalized central limit theorem, we need to make such an assumption explicitly (Uchaikin and Zolotarev, 1999, p. 62).

Assumption 3. $1 - F(x) \sim Cx^{-\alpha}$ *as* $x \to +\infty$ *for* $1 < \alpha \leq 2$ *and* $C > 0$.

Assumption 3 implies that X is an asymptotic power law, but it is a little more restrictive than definition 2: instead of assuming that $x \mapsto L(x)$ in definition 2 is *slowly varying*, we make the stronger assumption that it *converges to a constant*. It still covers a vast majority of cases. We limit ourselves to situations where $1 < \alpha \leq 2$, since when $\alpha > 2$ we are back to the finite variance case, and when $\alpha \leq 1$ the mean is infinite.

Proof. We use the generalized central limit theorem of Gnedenko and Kolmogorov (1968), which gives the asymptotic distribution of the sample mean when $\mathbb{E}[X^2] = +\infty$. Once we apply this theorem, the rest of the proof becomes simpler than with finite variance. Indeed, with infinite variance, the sample mean in the last bracket converges slower than $1/\sqrt{n}$, while quantiles and means in other brackets still converge at the same speed. Therefore, asymptotically, there is only one source of statistical variability that eventually dominates all the other. Hence, we need not be concerned by, say, the joint distribution of the quantiles, because at the first order that distribution will be identically zero. This insight leads to the following lemma.

Lemma 4.

$$r_n \begin{bmatrix} \overline{\boldsymbol{U}}_n - \boldsymbol{\mu} \\ \overline{\boldsymbol{V}}_n - \boldsymbol{p} \end{bmatrix} \xrightarrow{\mathcal{D}} \boldsymbol{S}$$

where:

$$r_n = \begin{cases} n^{1-1/\alpha} & \textit{if} \quad 1 < \alpha < 2 \\ (n/\log n)^{1/2} & \textit{if} \quad \alpha = 2 \end{cases}$$

$$S_i = \begin{cases} \gamma Y & \textit{if} \quad i = K+1 \\ 0 & \textit{otherwise} \end{cases}$$

and Y is a stable distribution with the characteristic function:

$$g(t) = \exp(-|t|^\alpha [1 - i\tan(\alpha\pi/2)\mathrm{sign}(t)])$$

and:

$$\gamma = \begin{cases} \left(\frac{\pi C}{2\Gamma(\alpha)\sin(\alpha\pi/2)}\right)^{1/\alpha} & \textit{if} \quad 1 < \alpha < 2 \\ \sqrt{C} & \textit{if} \quad \alpha = 2 \end{cases}$$

Proof. Standard results on quantiles (David and Nagaraja, 2005) and the trimmed mean (Stigler, 1973) imply that quantiles and means in middle bracket converge in distribution at speed $1/\sqrt{n}$. Because $r_n = o(\sqrt{n})$, they converge to zero in probability when multiplied by r_n. Hence, the only nonzero term in $\boldsymbol{S}$ correspond to $\overline{U}_n^{K+1}$, which converges to γY according to the generalized central limit theorem (Uchaikin and Zolotarev, 1999, p. 62). □

We now move from the asymptotic distribution of $\overline{\boldsymbol{U}}_n$ and $\overline{\boldsymbol{V}}_n$ to the asymptotic distribution of $\overline{\boldsymbol{W}}_n$ and $\hat{\boldsymbol{Q}}_n$, as we did in the previous section. Except that now both distributions are the same, because the disturbances introduced by quantiles and middle bracket averages are asymptotically negligible.

Lemma 5.

$$r_n \begin{bmatrix} \overline{\boldsymbol{W}}_n - \boldsymbol{\mu} \\ \hat{\boldsymbol{Q}}_n - \boldsymbol{q} \end{bmatrix} \overset{\mathcal{D}}{\to} \boldsymbol{S}$$

with the same notations as in lemma 4.

Proof. Using the same method as in the proof of lemma 3, we get:

$$r_n(\overline{\boldsymbol{W}}_n - \boldsymbol{\mu}) = r_n(\overline{\boldsymbol{U}}_n - \boldsymbol{\mu}) + o(1)$$

Moreover, the Bahadur (1966) representation of the quantile implies:

$$r_n(\hat{\boldsymbol{Q}}_n - \boldsymbol{q}) = o(1)$$

Using lemma 4 give the result. □

We may now apply the delta method as we did in the previous section. We are in a somewhat non standard case because the convergence operates more slowly than $\sqrt{n}$, and the asymptotic distribution is not Gaussian, but the basic idea of the delta method applies nonetheless. We get:

$$r_n(\Theta[\boldsymbol{T}_n, \overline{X}_n \hat{L}_n(p^*), \hat{Q}_n(p^*)] - \Theta[\boldsymbol{T}_\infty, \mathbb{E}[X]L(p^*), Q(p^*)]) \overset{\mathcal{D}}{\to} (\nabla\Theta)\boldsymbol{RS}$$

which proves the result of theorem 6 for infinite variance. □

The precise parameters of the stable distribution and the constants γ_1, γ_2 are given in appendix alongside the proof. For practical purposes, that theorem requires in particular the estimation of α and C. Using the generalized Pareto distribution model with parameters ξ, σ, μ as in section 3.3, we have:

$$\alpha = 1/\xi$$
$$C = (1-p)(\xi/\sigma)^{-1/\xi}$$

for $\xi \geq 1/2$ (if $\xi < 1/2$, variance is finite). If we have access to individual data, at the very top of the distribution, it is possible to use better estimates of α and C, using the large literature on the subject in extreme value theory (Haan and Ferreira, 2006, pp. 65–126).[8]

The infinite variance approximation is a rougher than the finite variance approximation for two reasons. First, because it relies on parameters, such as the asymptotic Pareto coefficient, which are harder to estimate than variances or covariances. Second, because it makes a first-order approximation which is less precise. That is, it assumes that all of the error comes from mean of the top bracket (which converges at speed $1/r_n$), and none from the quantiles or the mean of the lower brackets (which converge at speed $n^{-1/2}$). Although that is asymptotically true, because r_n grows more slowly than $n^{1/2}$, it is possible that the second-order term is not entirely negligible for finite n. Still, it gives a good idea of the magnitude of the error.

A.6.3.3 Comparison

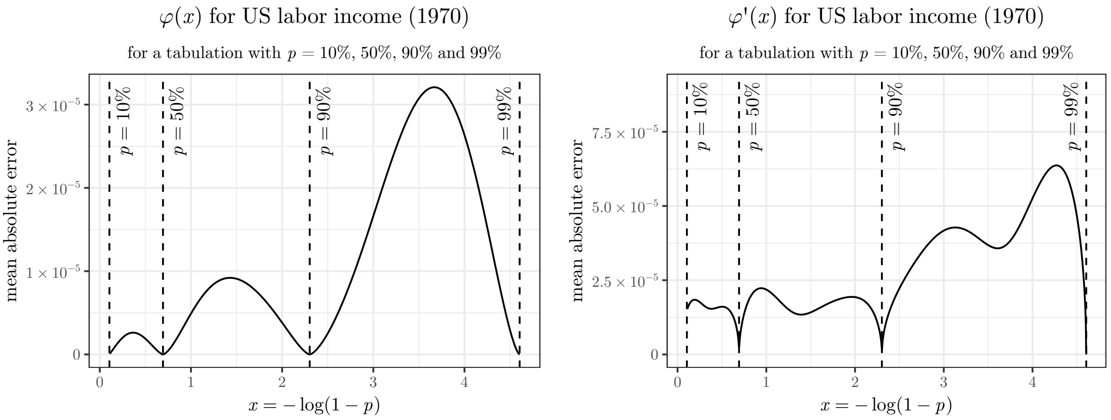


Figure A.7: Asymptotic mean absolute value of the sampling error with finite variance

To observe theorem 6 in practice, we turn to the distribution of labor and capital income in the United States in 1970. Back then, labor income inequality was low enough so that the asymptotic inverted Pareto coefficient was comfortably below 2 (somewhere between 1.4 and 1.6), which means that the distribution has finite variance. Capital income, on the other hand, was as always more unequally distributed, so that its asymptotic inverted Pareto coefficient appeared to be above 2.3, which implies infinite variance.

Figures A.7 and A.8 apply theorem 6 to the distribution of labor and capital income in the United States. The patterns are reminiscent of what we observed for the misspecification error: a bell-shaped, single-peaked error in each bracket for φ, and

[8] For example, wealth rankings such as the *Forbes 400* can give the wealth of a country's richest individuals. See Blanchet (2016).

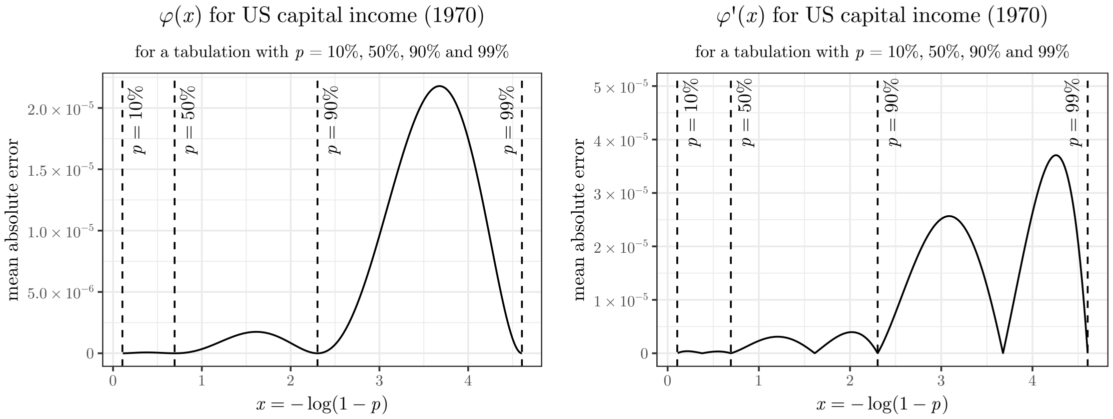
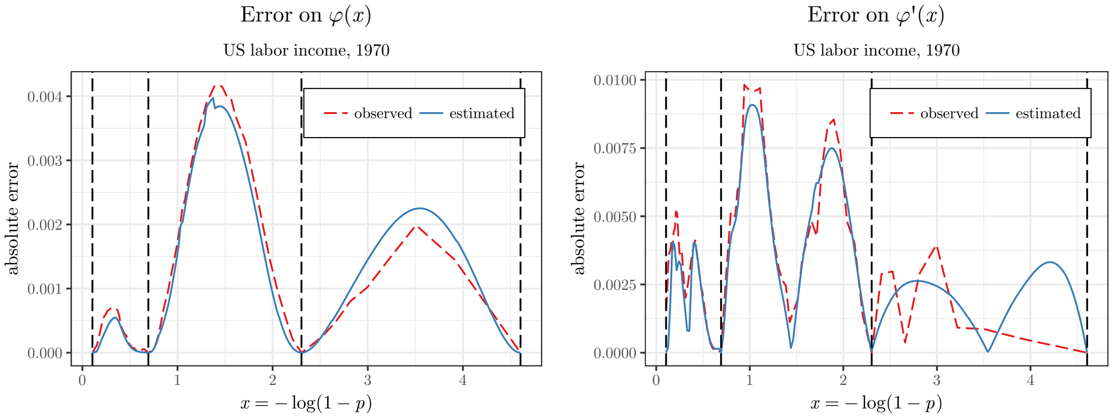


Figure A.8: Asymptotic mean absolute value of the sampling error with infinite variance

a double-peaked error for φ'.

A.6.4 Comparing Misspecification with Sampling Error



Pre-tax national income. Sources: author’s computations from Piketty, Saez, and Zucman (2016). The solid blue line correspond to the misspecification error estimated with formula (A.4) and a nonparametric estimate of φ'''. The dashed red line correspond to the actual, observed error. We smoothed the observed error for φ' using the Nadaraya-Watson kernel estimator to remove excessive variability due to rounding.

Figure A.9: Actual error and estimated misspecification error

The misspecification error largely dominates the sampling error given the sample sizes that are typical of tax tabulations. To see this, we may go back to the previous example of the US distribution of labor income in 1970. Figure A.9 shows the misspecification error is this case estimated using formula (A.4) and a nonparametric estimate of φ''', alongside the actual, observed error. There is some discrepancy between both figures, largely due to the fact that φ''' cannot be estimated perfectly.

Yet the estimated misspecification error appear to be a fairly good estimate of the actual error overall.

We may then look at figure A.7 to see how the sampling error compares. At its highest, it reaches to 3.5×10^{-5} for φ and 7.5×10^{-5} for φ'. The misspecification error is several orders of magnitude higher, around 10^{-3}. Even if the population was 100 times smaller, the magnitude of the mean absolute deviation of the error would be multiplied by $\sqrt{100} = 10$, so it would remain an order of magnitude lower. The figures would be similar for other years, countries or income concept. In practice, we can confidently neglect the sampling error.

Bibliography

Bach, Laurent, Laurent Calvet, and Paolo Sodini (2017). "Rich Pickings? Risk, Return, and Skill in the Portfolios of the Wealthy". URL: https://ssrn.com/abstract=2706207.

Bahadur, R. R. (June 1966). "A Note on Quantiles in Large Samples". In: *The Annals of Mathematical Statistics* 37.3, pp. 577–580. URL: http://dx.doi.org/10.1214/aoms/1177699450.

Benhabib, Jess, Alberto Bisin, and Mi Luo (2015). "Wealth distribution and social mobility in the US: A quantitative approach". In: *NBER Working Paper Series* 21721, pp. 1–32.

Bingham, N. H., C. M. Goldie, and J. L. Teugels (1989). *Regular Variation.* Encyclopedia of Mathematics and its Applications. Cambridge University Press.

Blanchet, Thomas (2016). "Wealth inequality in Europe and the United States: estimates from surveys, national accounts and wealth rankings". MA thesis. Paris School of Economics.

Cowell, Frank A. (2000). *Measuring Inequality.* LSE Economic Series. Oxford University Press.

David, H. A. and H. N. Nagaraja (2005). *Order Statistics.* John Wiley & Sons, Inc.

Fagereng, Andreas et al. (2016). "Heterogeneity and Persistence in Returns to Wealth". URL: http://www.nber.org/papers/w22822.

Gabaix, Xavier (1999). "Zipf's Law for Cities: An Explanation". In: *The Quarterly Journal of Economics* 114.3, p. 739.

– (2009). "Power Laws in Economics and Finance". In: *Annual Review of Economics* 1.1, pp. 255–294.

Gabaix, Xavier et al. (2016). "The Dynamics of Inequality". In: *Econometrica* 84.6, pp. 2071–2111. URL: http://dx.doi.org/10.3982/ECTA13569.

Garbinti, Bertrand, Jonathan Goupille-Lebret, and Thomas Piketty (2016). "Income Inequality in France, 1900-2014: Evidence from Distributional National Accounts (DINA)". In: WID.world Working Paper. URL: http://wid.world/document/b-garbinti-j-goupille-and-t-piketty-inequality-dynamics-in-france-1900-2014-evidence-from-distributional-national-accounts-2016/.

Gnedenko, B.V. and A.N. Kolmogorov (1968). *Limit distributions for sums of independent random variables.* Addison-Wesley series in statistics. Addison-Wesley.

Haan, L. de and A. Ferreira (2006). *Extreme Value Theory: An Introduction.* Springer Series in Operations Research and Financial Engineering. Springer New York. URL: http://www.springer.com/us/book/9780387239460.

Kacperczyk, Marcin, Jaromir B Nosal, and Luminita Stevens (2014). "Investor Sophistication and Capital Income Inequality". URL: http://www.nber.org/papers/w20246.pdf.

Karamata, J. (1930). "Sur un mode de croissance regulière des fonctions". In: *Mathematica* 4.

Piketty, Thomas, Emmanuel Saez, and Gabriel Zucman (Dec. 2016). "Distributional National Accounts: Methods and Estimates for the United States". In: Working Paper Series 22945. URL: http://www.nber.org/papers/w22945.

Saez, Emmanuel and Gabriel Zucman (2016). "Wealth Inequality in the United States since 1913: Evidence from Capitalized Income Tax Data". In: *Quarterly Journal of Economics* 131.May, pp. 519–578.

Stigler, Stephen M. (May 1973). "The Asymptotic Distribution of the Trimmed Mean". In: *The Annals of Statistics* 1.3, pp. 472–477. URL: http://dx.doi.org/10.1214/aos/1176342412.

Uchaikin, V.V. and V.M. Zolotarev (1999). *Chance and Stability: Stable Distributions and their Applications*. Modern Probability and Statistics. De Gruyter. URL: http://staff.ulsu.ru/uchaikin/uchzol.pdf.

van der Vaart, A.W. (2000). *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Appendix B

Appendix to "The Weight of the Rich: Improving Surveys with Tax Data"

B.1 Formal Biases and Adjustments

Our adjustment procedure is based on the interpretation of the whole difference between tax and survey densities as being due solely to nonresponse. However, the misreporting of income by survey respondents may also produce discrepancies. Misreporting tends to be negatively correlated with income.[1] That is, on average, the poor are more likely to overreport their true income while the rich tend to underreport. It is thus fair to ask: what would be the consequences of such behavior in our analytical framework? And how do replacing methods — which aim to adjust for underreporting at the top — compare to reweighting?

B.1.1 Double-Biased Density Functions

To define the misreporting bias, let $f_Y(y)$ be the true income distribution, $f_M(y)$ the distribution of misreported income, $p(y)$ the probability of misreporting for a given level of income, conditional on response, and $\bar{p}$ its average. Then we define f_Z as the income distribution of a sample that is drawn from $f_Y(y)$, including both the

[1] See Bound and Krueger (1991), Bollinger (1998), Pedace and Bates (2000), Cristia and Schwabish (2009), and Abowd and Stinson (2013) for studies on the United States and Angel, Heuberger, and Lamei (2017) and Paulus (2015) for studies on Austria and Estonia respectively.

nonresponse and misreporting biases (the former is defined in equation (2.1)) :

$$f_Z(y) = f_Y(y)\theta(y)(1 - p(y)) + f_M(y)\bar{p} \tag{B.1}$$

The left side of the sum stands for those who report income correctly with a given (relative) probability of response that is defined in equation (2.1), i.e. $\theta(y)$. The right side of the sum accounts for those declaring misreported income equal to y, given that they respond. In this situation, the over- or under-estimation of f_Z with respect to the true distribution f_Y can be formulated as the ratio of the two distributions:

$$\frac{f_Z(y)}{f_Y(y)} = \theta(y)(1 - p(y)) + \frac{f_M(y)}{f_Y(y)}\bar{p} \tag{B.2}$$

If the ratio is higher than 1, the density is overestimated. If it is lower than 1, it is underestimated. Naturally, the shape of such bias depends on the characteristics of each of the variables at play. Following the empirical literature, it is reasonable to define the probability of misreporting as being higher in both ends of the distribution and relatively stable in the middle. However, explicit information on the shape of the misreported-income distribution is rare, since it relies on having individually-linked survey and register micro-data. In order to better understand the potential impact of assumptions on its shape, it can be useful to analyze a simplified situation where misreporting operates in isolation. In that case we have:

$$\frac{f_Z(y)}{f_Y(y)} = 1 - p(y) + \frac{f_M(y)}{f_Y(y)}\bar{p} \tag{B.3}$$

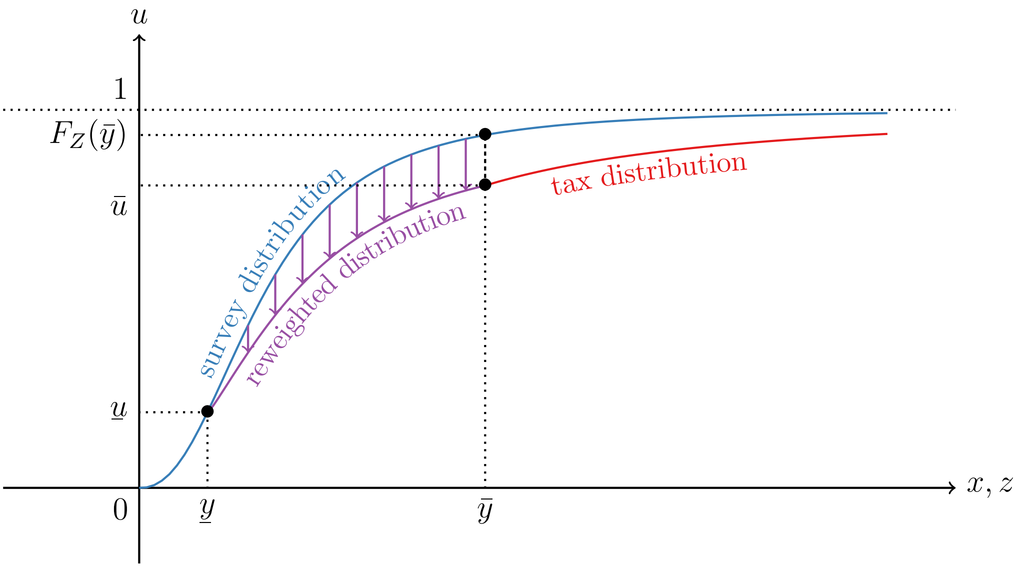
If misreported income follows the same distribution as true income, that is $f_M(y) = f_Y(y)$, then densities are underestimated where the probability of misreporting is higher than its average ($p(y) > \bar{p}$). Symetrically, densities are overestimated where the same probability is lower than its average ($p(y) < \bar{p}$). Of course, it may seem odd to assume that misreported income is distributed exactly as true income. However, we consider this to be a useful simplification which helps to convey that both the nonresponse and misreporting biases can have a similar impact and that we are unable to tell them apart *ex-post*. Indeed, both biases, either working alone or together, can perfectly describe a profile as the one in figure 2.3. If $f_M \neq f_Y$, we can still get a similar result under some circumstances. For instance, if both densities are of the same type but defined by different parameters (e.g. if both are log-normal with a different mean and standard error) — which does not seem to be a strong assumption — the ratio of the sample to true distribution would likely have a form similar to Figure 2.3 but with strong or slight perturbations near the mode of each distribution (assuming the true and misreported income-densities are unimodal).

When we study the ratio of income distributions from actual tax and survey data — in section 2.3.2.2 — the empirical estimate of the θ coefficient should be capturing the effect of both these biases. Figure 2.6 shows that estimates for countries with comprehensive tax coverage (e.g. Norway, France and the UK) depict rather flat shapes through most of the distribution and only fall closer to the right tail. Such a shape implies that, if the misreporting bias is present in the survey, the differences between f_M and f_Y are not big enough to cause perturbations that are easily distiguishable from noise in the θ coefficient. In any case, to our knowledge, it is not possible to measure the relative size of both the nonresponse and misreporting biases without access to individually-matched micro datasets.

B.1.2 Adjustment Methods: Reweighting vs. Replacing

In practice, researchers face the following problem while combining survey and tax data: on one side, survey data supposedly covers the whole population but fails to properly capture the top tail of the income distribution. On the other side, they have a tax data distribution which is assumed to be accurate, at least at the top.[2]

Figure B.1: Correcting for Nonresponse by Reweighting



[2] The issues of tax avoidance and evasion are issues of underreporting, but are more difficult to remedy without access to third-party/offshore data. Therefore it is useful to think of tax data, at least above a certain top threshold, as being an accurate lower bound for incomes.

Reweighting The reweighting solution in this scenario can be represented as in figure B.1, which displays the Cumulative Distribution Functions (CDF) of the survey, tax data and "reweighted" distributions. The tax data start at the value $\bar{y}$, which correspond to the population fractile $\bar{u}$. If nonresponse is higher at the top, the corresponding fractile in the survey ($F_Z(\bar{y})$) will be higher as shown. We can also define a low income level $\underline{y}$ with corresponding fractile $\underline{u}$ below which we do not want to alter the survey (e.g. the national poverty line). If there is no such concern, then we can set $\underline{u} = 0$ and $\underline{y} = -\infty$.

Replacing While the reweighting method adjusts the weight of survey observations, replacing methods adjust their value. The usual rationale behind replacing methods is different. It accounts for the discrepancy between the survey and the tax data by assuming that people misreport their income, rather than by assuming that people refuse to answer the survey or its income-related questions.

Either problem may happen in reality, and mathematically it is not possible to disentangle them without linking tax and survey data directly (see Appendix B.1.1). But the case for reweighting relies in part on the fact that even if misreporting is the problem, it is unclear that pure replacing does a better job of solving it than reweighting. To convey this, let us start with a formulation of the misreporting problem. We have the following relationship between two random variables Y and Z, which represent true income and misreported income respectively:

$$Z = Y\Lambda$$

where Λ is a random variable that may depend on Y. We call $1 - \Lambda$ the rate of underreporting. In this setting, the PDF of Z will depend on the joint PDF of Y and Λ:

$$f_Z(z) = \int_{-\infty}^{+\infty} \frac{1}{|\lambda|} f_{Y\Lambda}[z/\lambda, \lambda] \, \mathrm{d}\lambda$$

The expression above raises some major tractability issues. In particular, it is not possible to recover $f_{Y\Lambda}$ from the knowledge of f_Y and f_Z separately, so Λ may only be estimated when we can link misreported income and its covariates (i.e. Z and X) at the individual level, which is not common in practice. Otherwise, there will infinitely many Λ that satisfy the problem. For these reasons, previous researchers working with replacing methods have made some very strong (even if implicit) assumptions, which we make explicit below.

Assumption 1. The rate of underreporting is a deterministic function of the rank in Y: $\Lambda = \lambda(F_Y(Y))$.

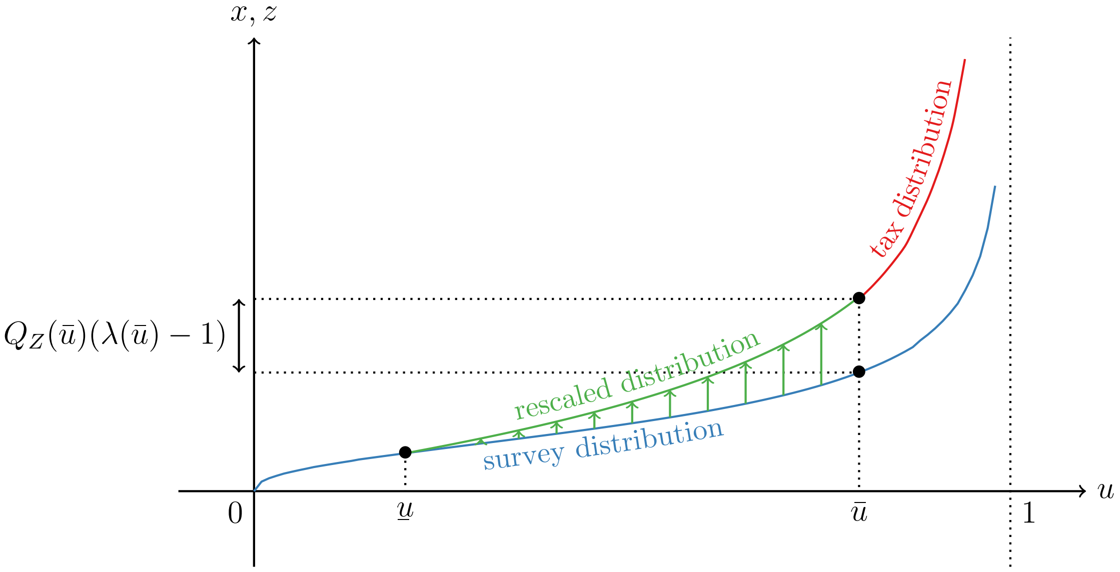
Assumption 2. The rank is the same in the true income distribution and in the survey income distribution: $F_Y(Y) = F_Z(Z) = U$.

These assumptions on λ are very strong and unavoidable. Otherwise, it is not possible to interpret $\lambda(u)$ as an average underreporting given rank u (i.e. $\lambda(u) = \mathbb{E}[\Lambda|F_Y(Y) = u]$)), because f_Z depends on the entire joint distribution of (Y, Λ). Using these assumptions, estimating the underreporting function λ is very simple. Indeed, since misreporting leaves the rank unchanged, we have:

$$\lambda(u) = \frac{Q_Z(u)}{Q_Y(u)} \tag{B.4}$$

where Q_Y and Q_Z are the quantile functions of Y and Z. The replacing approach to correcting survey data proceeds as follows.[3] We assume a rank $\underline{u}$ below which we do not alter the survey data, assuming it is already accurate, so $\lambda(\underline{u}) = 1$. The tax data start at the rank $\bar{u}$, at which the rate of underreporting is observed directly: $\lambda(\bar{u}) = Q_Z(\bar{u})/Q_Y(\bar{u})$. The situation is pictured in figure B.2. Between $\underline{u}$ and $\bar{u}$, we must assume a certain shape of the function λ. A simple and common choice is the linear rescaling profile $\lambda(u) = 1 + (\lambda(\bar{u}) - 1)\frac{u-\underline{u}}{\bar{u}-\underline{u}}$.

Figure B.2: Correcting for Misreporting by Replacing



This procedure may make sense if we view it as a manipulation of the distribution in

[3] Here we present the most extreme class of replacing methods, which we label ‘rescaling’. In this case there is a part of the survey distribution that is adjusted (rescaled) for which there are no tax data values to replace it. See section 2.1.2.

itself. But given the extremely strong and unrealistic assumptions stated above, any interpretation in terms of individual behaviour is slippery. And if we only understand the replacing approach as a manipulation of distribution at the aggregate level, then we should expect reweighting to perform similarly well. Indeed, reweighting simply involves adjusting the survey distribution in figure B.2 horizontally rather than vertically. Therefore, we have the following equivalence between reweighting and replacing coefficients for income y and rank u, so that reweighting may be interpreted as a specific case of replacing with:

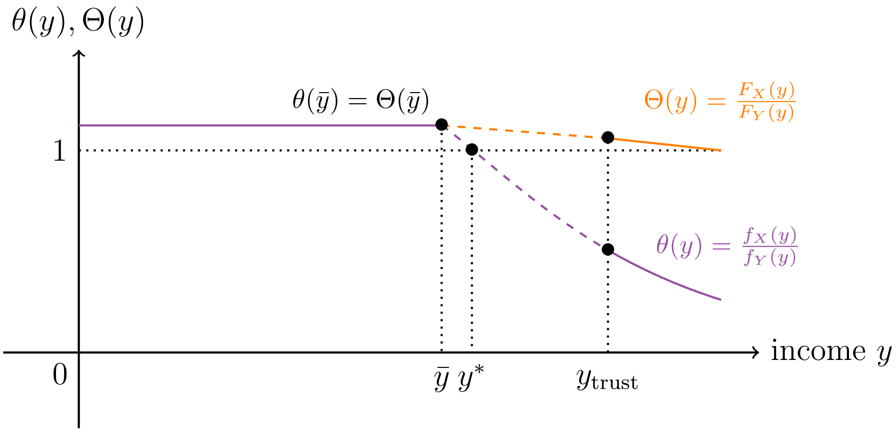
$$\Theta(y) = \frac{F_Z[Q_Z(u)/\lambda(u)]}{u}$$

In the end, unlike reweighting, it is unclear what problem exactly replacing methods end up solving. In any case, reweighting does, at least, an equally good job at solving it. Furthermore, reweighting has a clear interpretation, it is consistent with widely accepted calibration methods, it is easier to generalize to more complex settings and it always preserves the continuity of density functions, which is highly desirable and not the case in the replacing procedures, especially those adjusting arbitrary portions of the distribution (e.g. the top 1%).

B.2 Merging Point Below the Trustable Span

Sometimes the part of the distribution covered by the tax data is too limited to observe a merging point such that $\Theta(y) = \theta(y)$. This situation is represented in figure B.3. Below y_{trust}, the value of $\theta(y)$ and $\Theta(y)$ have to be extrapolated until both curves cross, which is where we define the merging point.

Figure B.3: Choice of Merging Point when $\bar{y} < y_{\text{trust}}$


$\theta(y), \Theta(y)$
$\theta(\bar{y}) = \Theta(\bar{y})$
$\Theta(y) = \frac{F_X(y)}{F_Y(y)}$
1
$\theta(y) = \frac{f_X(y)}{f_Y(y)}$
0
$\bar{y}$
y^*
y_{trust}
income y

We need to define a functional form for $\theta(y)$ in order perform the extrapolation (the value of $\Theta(y)$ follows from that of $\theta(y)$). We will assume the following:

$$\log\theta(y) = \gamma_0 - \gamma_1 \log y \tag{B.5}$$

which may also be written $\theta(y) = \mathrm{e}^{\gamma_0} y^{-\gamma_1}$. In addition to fitting the shape of the bias observed in practice, this form has the property of preserving Pareto distributions. Indeed, if $f_Y(y) \propto x^{-\alpha-1}$, then $f_X(y) = \theta(y) f_Y(y) \propto x^{-\gamma_1-\alpha-1}$, which is also a Pareto density. The parameter γ_1 may be interpreted as an elasticity of nonresponse: when the income of people increases by 1%, how much less likely are they to be represented in the survey.

While the equation (B.5) can be estimated by OLS, we need to take into account situations where tax data covers such a small share of the distribution that the number of data points is insufficient to estimate the regression reliably. Since the frontier between having and not having enough data is blurry, our preferred approach is to deal with the two cases at once using a ridge regression. The idea is that we can know from experience a typical value for γ_1 called γ_1^*. In the absence of data, it represents our baseline estimate.[4] As we observe new data, we may be willing to deviate from that value, but only to the extent that there is enough evidence for doing so. The ridge regression formalizes this problem as:

$$\min_{\gamma_0,\gamma_1} \sum_{i=1}^{m} (\log\tilde{\theta}_k - \gamma_0 - \gamma_1 \log y_k)^2 + \lambda(\gamma_1 - \gamma_1^*)^2$$

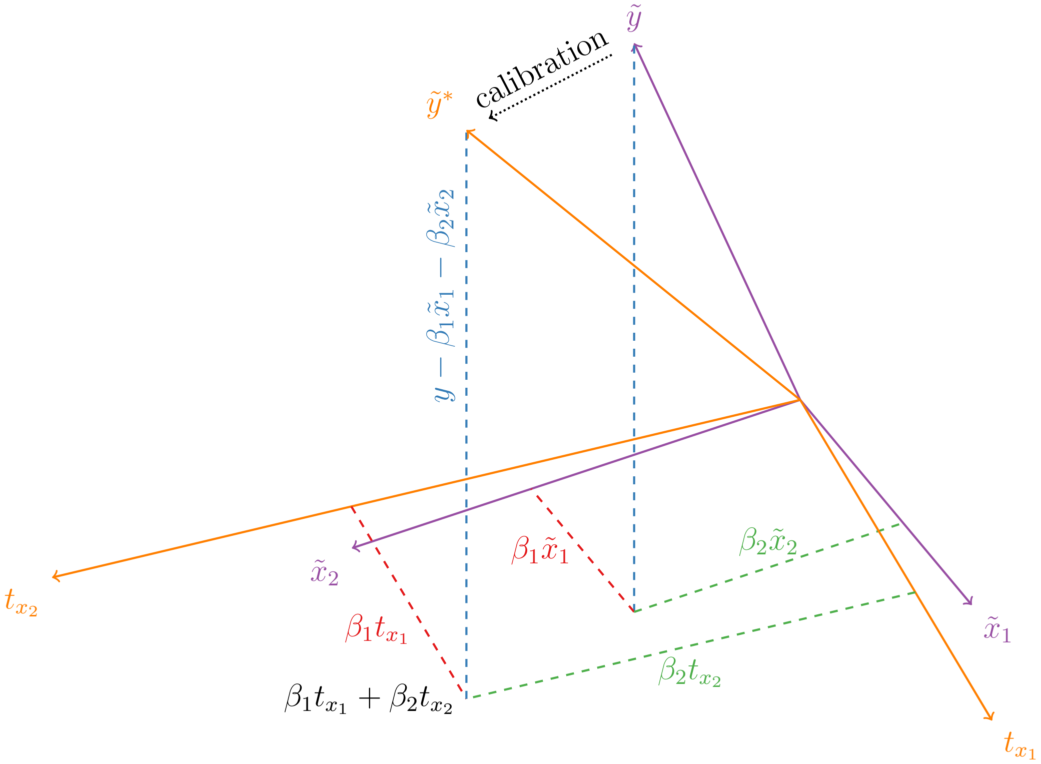
The first term is the same sum of squares as the one minimized by standard OLS. The second term is a Tikhonov regularization parameter that penalizes deviations from γ_1^*. If $m = 1$, then $\gamma_1 = \gamma_1^*$ and the sum of squares only determines the intercept. As we get more data points, the sum of squares gets more weight and results get closer to OLS. The parameter λ determines the strength of the penalization. The problem has an explicit solution expressible in matrix form (e.g. Hoerl and Kennard, 2000). We can have a Bayesian interpretation of the method where our prior for γ_1 is a normal distribution centered around γ_1^* and λ determines its variance. The solution of the ridge regression gives the mean value of the posterior. Once we have the estimation of γ_0, γ_1 we can simulate a tax data distribution by reweighting the survey data: the point at which $\theta(y)$ crosses $\Theta(y)$ becomes the merging point $\bar{y}$, and the reweighted survey from $\bar{y}$ to y_{trust} can be used to complete the tax data.

[4] In practice, γ_1^* can be drawn from other “similar countries” that have sufficient data. For example, in our applications, we use the Brazilian γ_1^* to extrapolate the Chilean merging point (see section 2.3.2.2).

B.3 Geometrical Interpretation of Calibration

A further interpretation of the linear calibration presented in section 2.2.2 is geometrical. It comes from the relationship between (2.4) and the generalized regression estimator (GREG). Assume that we seek to estimate the total of a survey variable y. We can directly use the survey total, which we will write $\tilde{y}$. But if we wish to exploit the information on the true population totals of the auxiliary variables $x_1, \ldots, x_k$, we can use the GREG estimator, whose logic is represented in figure B.4. The idea is to first use the survey to project the variable of interest y onto the auxiliary variables $x_1, \ldots, x_k$ using an ordinary least squares regression. Hence we get a linear prediction $\hat{y}_i = \boldsymbol{\beta x}_i$ of y_i, which corresponds to the part of y that can be explained by the auxiliary variables $x_1, \ldots, x_k$. We can then substitute the survey totals by their true population counterpart in the linear prediction to get a new, corrected prediction of y. Adding back the unexplained part of y leads to the GREG estimator $\tilde{y}^* = \tilde{y} + \boldsymbol{\beta}(\boldsymbol{t} - \tilde{\boldsymbol{x}})$.

Figure B.4: Geometrical Interpretation of Linear Calibration



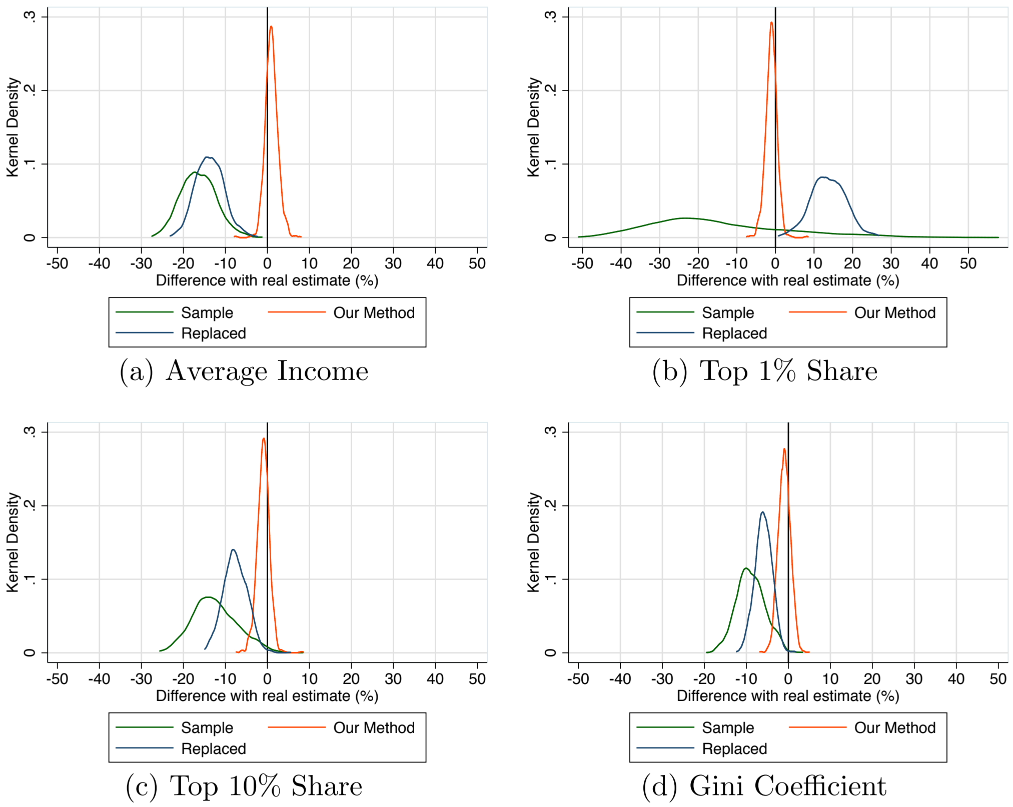
The survey totals $\tilde{y}$, $\tilde{x}_1$ and $\tilde{x}_2$ are shown in purple. The GREG estimator, which is equivalent to linear calibration, first projects $\tilde{y}$ onto $\tilde{x}_1$ and $\tilde{x}_2$ (dashed blue line). This projection is equal to $\beta_1\tilde{x}_1 + \beta_2\tilde{x}_2$. The true population totals t_{x_1} and t_{x_2} are in orange. We substitute them for $\tilde{x}_1$ and $\tilde{x}_2$ in the projection, which gives the value $\beta_1 t_{x_1} + \beta_2 t_{x_2}$. We add back the unexplained part of $\tilde{y}$ (dashed blue line) to get the calibrated total $\tilde{y}^*$.

It can be shown algebraically that linear calibration is identical to the GREG procedure (Deville and Särndal, 1992). By using the calibrated weights, we systematically project the variable of interest on the calibration variables and perform the correction described above, without having to explicitly calculate the GREG estimator every time.

B.4 Further Monte Carlo Simulations

This section presents three supplementary experiments to those presented in section 2.3.1. Each one of them includes punctual changes in the parameters underlying the benchmark experiment, which is a useful way to isolate possible effects and thus to anticipate the method’s performance in different scenarios.

Figure B.5: Experiment with more Misreporting



(a) Average Income

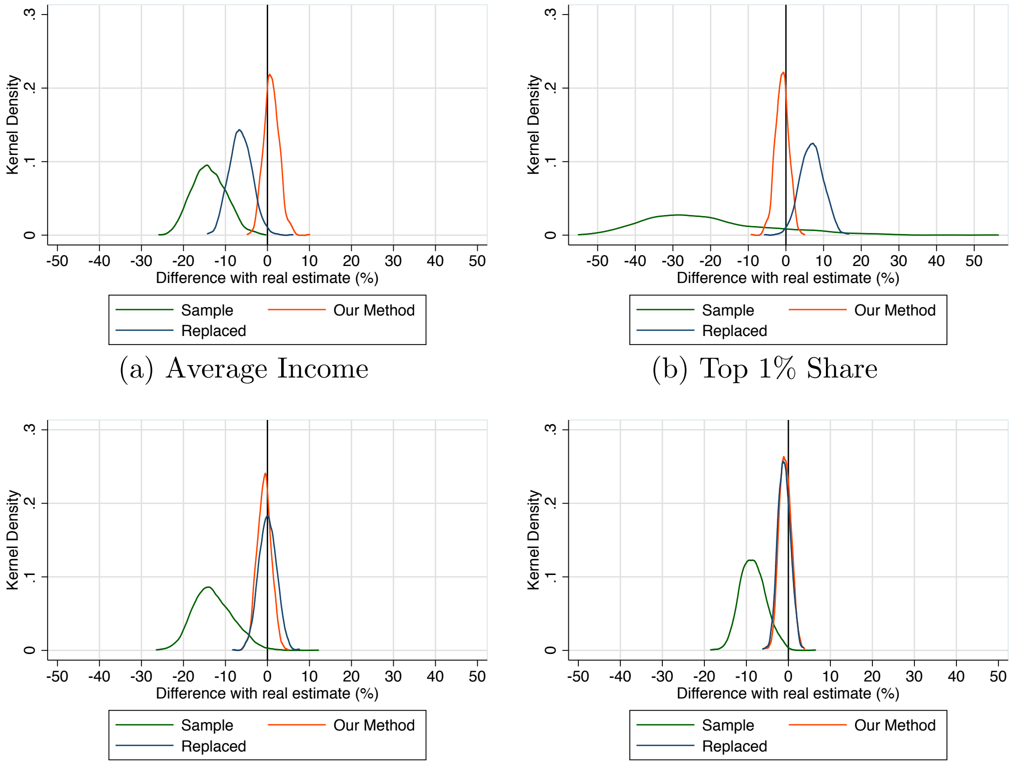
(b) Top 1% Share

(c) Top 10% Share

(d) Gini Coefficient

Figure B.5 displays results of an experiment which only differs from the benchmark in that the misreporting bias is stronger. That is, the probability of misreporting starts increasing from percentile 85 (P85) instead of 95 (P95). This mechanically affects the accuracy of estimates produced using the raw survey, which is expected given that more people are actually misreporting their income. Indeed, the variance of each of the estimates from the raw sample increases substantially, which is visible by comparing the width of their kernel densities to the corresponding ones in the benchmark setting. Although replaced surveys still appear to get somewhat closer than raw estimates to true values after correction, they are also substantially affected by the increased variability. However, this setting only affects the performance of our method marginally, as the resulting densities of estimates are almost indistinguishable from those displayed in figure 2.5, which is a good proof of adaptability. Other experiments were conducted, where we assumed stronger non-response biases. But we do not display results because they are almost identical to those presented in figure B.5. This can be explained by the fact that both biases have a similar effect — only in distributive terms, as opposed to individual representativeness — on resulting distributions (see Appendix B.1.1).

Figure B.6: Experiment with Replacing the Top 5%



(a) Average Income

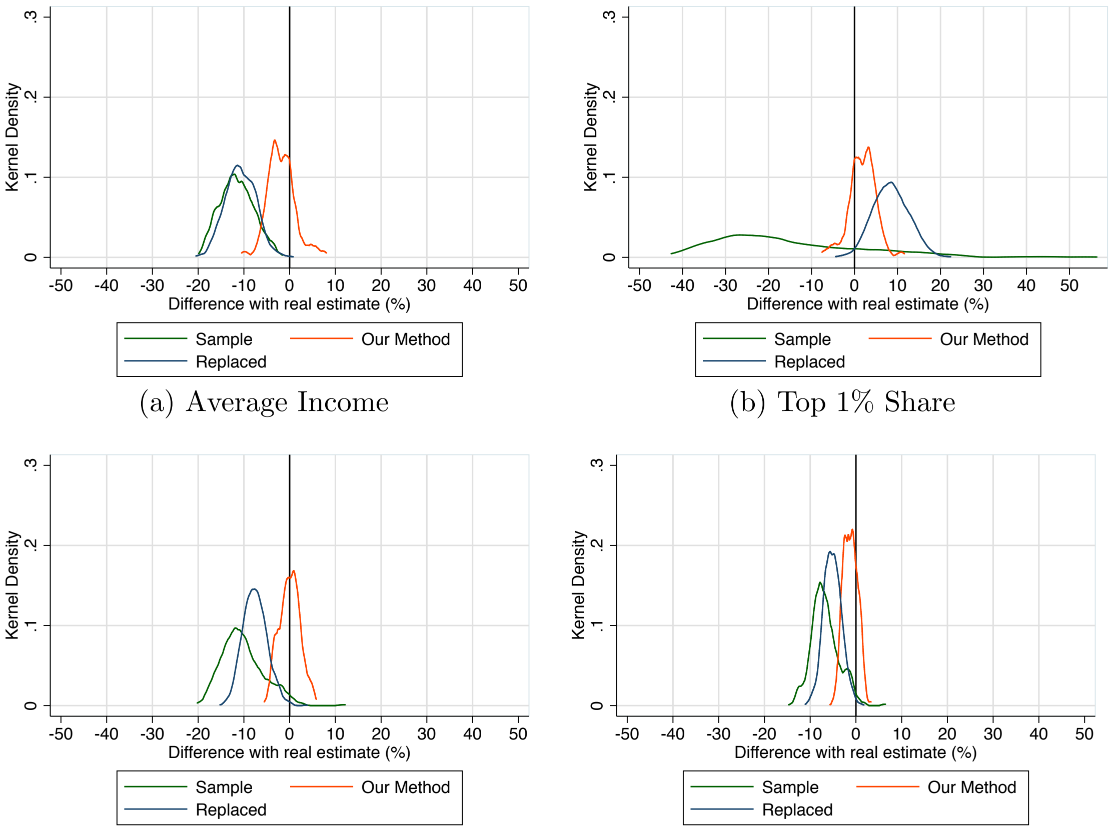
(b) Top 1% Share

(c) Top 10% Share

(d) Gini Coefficient

Figure B.6 depicts another setting, where the only difference with the benchmark experiment is that the replacing procedure uses the top 5% instead of only the top 1%. The estimates produced by our adjustment method are virtually the same than the benchmark. Since, by definition, biases are active in the top decile, the increase in the replaced population results in estimates that are more accurate, especially in the case of the top 10% share and the Gini coefficient (see figures B.6c and B.6d), which appear to be substantially closer to our estimates and thus to true values. However, the same is not true for both the estimated average income and the top 1% share, which still tend to be substantially underestimated and overestimated, respectively (figures B.6a and B.6b). Although we could go further and try to find the exact portion of the population that has to be replaced to get a similar result to that obtained with our method, we judge this to be an unnecessary exercise. As we argue in section B.1.2, the equivalence between our method and replacing can be found in some cases, yet it would only would be valid in a purely distributional perspective because replacing implies extremely unrealistic assumptions at the individual level and, thus, does not preserve the consistency of the resulting observations.

Figure B.7: Experiment with Poor Tax Data



(a) Average Income

(b) Top 1% Share

(c) Top 10% Share

(d) Gini Coefficient

Figure B.7 represents a somewhat extreme case where we limit access to tax declarations to only the top 5% of respondents, thus forcing our method to extrapolate adjustment factors in a large majority of cases. Our estimates appear to be less precise than in the benchmark, where a larger part of the information was used, yet they still perform better than both the raw survey and the replacing alternative. The resulting distribution of estimates is to our judgement rather satisfactory, since estimates remain closely centered around true values. This experiment shows that under extreme circumstances, where tax data covers a very small part of the population, estimates resulting from our correction method are still accurate, yet reasonably less precise.

B.5 Data Details and Supplementary Results

B.5.1 Country Specific Income Concepts and Observational Units

B.5.1.1 Brazil

To reconcile incomes in surveys with those in tax data, we use the latter as the benchmark for the top of the distribution. We thus require that the survey definition of income, from the micro-data, be consistent with the definition of income in the tax tabulations in order for the comparison to make sense. The total income assessed in tax data is pre-tax-and-transfer income, but including pensions and unemployment insurance. It is the sum of three broad fiscal categories: taxable income, exclusively taxed income and tax-exempt income (reported in table 9 of the tax report *Grandes Números DIRPF*). We describe each of these in turn before describing how we construct the survey definition of income.

Taxable income comprises of wages, salaries, pensions and property rent. These are incomes that are subject to assessment for the personal income tax. Exclusively-taxed income is income that has been already been taxed at source according to a separate tax schedule. It also contains capital income and labour income components. The labour component is the sum of the 13th monthly salary received by the contributor and their dependents, wages received cumulatively by contributors or dependents, and worker participation in company profits. The capital component comprises of the sum of fixed income investment income, interests on own capital ("juros sobre capital próprio"), variable income investment income, capital gains and other capital income. Non-taxable incomes are the last fiscal category, whose decomposition is presented in table 20 of the tax reports. These are incomes that are declared but which are not subject to any personal taxation when received. Close to one-fifth of these exempt incomes can be classified as labour income. These comprise of compensation for laid-off workers, the exempt portion of pension income for over 65s, withdrawals from employment security fund, scholarships, and other labour incomes.

The remaining items can be classified as capital income (distributed company profits, dividends, interests from savings accounts/mortgage notes) or mixed income (the exempt portion of agricultural income).

We construct survey income to be as close to the tax definition as possible. The total income we analyse from the PNAD surveys is the sum of labour income, mixed income and capital income. Labour income is the sum of all reported income from primary, secondary or all other jobs (variables V9532, V9982, V1022) for all employed individuals who do not classify themselves as own-account (self-employed) workers or employers. For employers, we assume that labour income is the portion of their work income that is below the annual exemption limit for the DIRPF, as set by the Receita Federal. Thus, values above the first tax paying threshold are taken to be capital withdrawals. Also in labour income are pensions (V1252, V1255, V1258, V1261), work allowances (V1264), *abono salarial* and unemployment insurance. Of the latter two, the first is imputed as one minimum wage for eligible formal private sector employees, while the second is imputed for respondents who claimed to have received unemployment benefits at some point in the 12 months before the PNAD interview. Benefit levels were imputed as yearly averages of shares of the minimum wage from current legislation. Values of V1273 equal to or below 1 monthly minimum wage are interpreted as social benefits, which are excluded from the analysis.

Mixed income is the reported income of own-account workers. Capital income is estimated as the sum of rent (V1267), financial income, and the capital portion of employer work income (i.e. reported amounts exceeding the annual exemption limit for DIRPF). Financial income (interests and dividends) is taken from other income sources declared (V1273) and estimated as any income from this source that exceeds 1 monthly minimum wage. Finally, we add a 13th monthly salary to the annual calculation of the incomes of formal employees and retirees. In total, the income we calculate from the surveys represents close to 80% of the equivalent (fiscal income) total from the household sector in the national accounts, on average between 2007 and 2015. The total income we use from tax statistics accounts for about 63% of the same fiscal income total from the national accounts over the same period.

Given that the unit of assessment in the tax data can either be the individual or the couple, in cases where the latter opt to declare jointly, we cannot strictly restrict ourselves to the analysis of individual income as it is received by each person. Therefore, we decide follow the tax legislation by identifying the number of married couples appearing jointly on the declaration and splitting their total declared income equally between them when carrying out the generalized Pareto interpolation

(Blanchet, Fournier, and Piketty, 2017) from the tabulation. This allows us to bring the analysis to the individual level by assuming that all spouses equally share their income. We use the information available in the tax statistics to estimate the share of joint declarations, which overall represent about 30% of all filed declarations (see (Morgan, 2018)). To be consistent in the comparison, we also use individual income in the surveys, with the income of married couples being split equally between the composite adults. We consider all adults aged 20 or over in our analysis.

B.5.1.2 Chile

Following the same logic as that applied to the Brazilian case, we construct from the Chilean survey an income definition that is as close as possible to the one used in tax data. The resulting definition is the one we use when merging datasets. However, in Chile, unlike Brazil, the survey reports post-tax incomes. In broad terms, we estimate pre-tax income retrospectively from declared post-tax income. In order to do so, we make a priori assumptions on whether certain types of income pay income taxes or not. Additionally, some self-reported characteristics are used to determine if the income of certain individuals should be treated as taxable or not. For instance, dependent workers that do not have a contract (and will not sign any soon) are considered to be informal, thus they are assumed to not pay the income tax. A similar mechanism is used for independent workers – depending on if they emit invoices (both commercial or for services) we define them as formal or informal. Table B.1 gives a comprehensive view on what types of income are assumed to pay taxes or not. For further comments on the definition of income corresponding to tax data, please refer to Flores et al. (n.d.).

B.5.1.3 European Countries

Tax Data For the three European countries we use tabulated tax data from official sources. In the case of Norway and the United Kingdom, the data come directly from institutional sources: "Tax Statistics for Personal Taxpayers" from Statistics Norway (`https://www.ssb.no/en/statbank/list/selvangivelse`) for the former, and the "Survey of Personal Incomes" (SPI) from HM Revenue & Customs (`https://www.gov.uk/government/statistics/income-tax-liabilities-by-income-range`), for the latter. The tax unit for both countries is the individual. As explained in Section 2.3.2.2, we interpolate the tabulations using a generalized Parteo interpolation Blanchet, Fournier, and Piketty (2017). For France, we use detailed tabulations produced by Garbinti, Goupille-Lebret, and Piketty (2016) from the micro-files of French taxpayers. These are available in the Appendix C Tables

Table B.1: From Post-Tax to Pre-Tax Income in Chilean Surveys

Type of income	**Taxable Income**		**Tax Exempt Income**	
	Variable name	**Code**	**Variable name**	**Code**
Labor Income	Wage (1ry occup.).	y1a	Occasional work.	y16a
	Wage (2ry occup.).	y6, y10	Unemp. insurance.	y14c
	Inc. from previous months (if dependent).	y14b	Tips, travel expenses.	y3c, y3e
	Extra hours, commissions & allowances.	y3a, y3b, y3d y3f	Christmas bonus.	y4a
	Rewards & additional salary.	y4b, y4c, y4d	Inc. of the inactive.	y11a
			Wage of informals.	o17, o14
Pensions	Old age pension.	y27am		
	Disability pension.	y27bm		
	Widow’s pension.	y27cm		
	Orphan’s pension.	y27dm		
Mixed Income	Inc. of indep. (1ry occup.)	y7a	Inc. of indep. (2ry occup.).	y6,y10
	Inc. from previous months (if indep.).	y14b	Inc. of non-qualified, informal, small minery & craftsmen.	oficio1, oficio4, o14
Capital Income	Rent (agricultural).	y12b	Rent (urban).	y12a
	Interest.	y15a	Rent (seasonal).	y16b
	Dividends.	y15b		
	Withdrawals.	y15c		
	Rent (equipment).	y16a		

Notes: Codes correspond to those of CASEN 2011-2013. Formality is defined as conditional to having a contract and/or emitting ”*boletas de honorarios*” (invoices by independents). Information on formality is only available for primary occupation. Formality is assumed to be the same for 1ry and 2ry occupations. In the survey, income is post-tax. Pre-tax formal income of contract-workers is calculated using tables of IUSC (*Impuesto Único de Segunda Categoría*) retrospectively. Pre-tax income of formals emitting invoices is added of mandatory provisional deductions (e.g. 10%) and standard presumptive expenses (e.g. 30%). Pre-tax capital income is calculated using the IPC (*Impuesto de Primera Categoría*) single tax-rate (e.g. 20%). Rent of urban properties is assumed to be untaxed because of law D.F.L.2 (1959) of their Data (see `http://piketty.pse.ens.fr/en/publications`). We use the individual-level tabulations that present the distribution of gross total fiscal income for 127 percentiles.

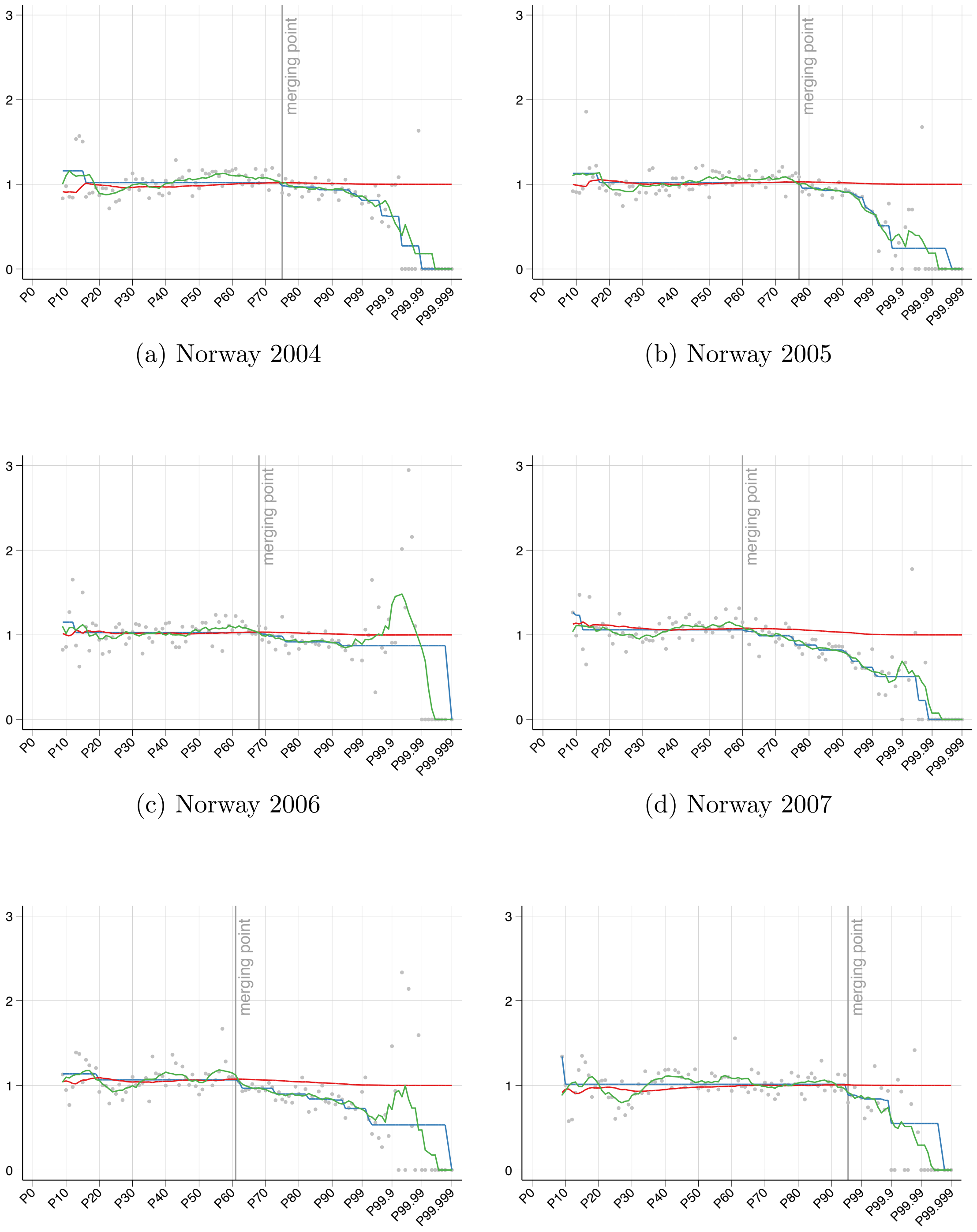
EU-SILC Data The advantage of using EU-SILC data is that it is a harmonized household survey dataset for European countries. However, given that we anchor our estimation method to the tax data, the definition of income used from surveys must match that accounted for in tax statistics. To do so we take the sum for each observation of employee cash or near cash income (variable PY010), self-employment cash income (PY050), Pensions received from private plans (PY080), a host of benefits related to unemployment, old-age, suvivors, sickness and disability (PY090, PY100, PY110, PY120, PY130), and capital income components (rent from property or land (HY040) and interests, dividends, profit from capital investments (HY090)). These capital incomes are reported at the household level. We individualise them by equally splitting the income among spouses and civil partners. For Norway and the UK, consistent with the fiscal income in tax data, we take gross incomes (before income taxes and individual social contributions levied at source). Since fiscal income in the French tax data is before income tax but after social contributions levied at source, we take net income values from the French SILC dataset. Income taxes are not levied at source in France for the period we analyse so the definition of net income in SILC is apt to be used for this case. We also select the reference population to be kept in accordance with the tax statistics. In Norway, the tax tabulations refer to individuals aged 17 and over, so we discard individuals under the age of 17 in the survey. For the UK, the tax data does not provide comparable information, so we follow the practice by Atkinson (2007) in taking a reference population of individuals aged 15 and over. In France, consistent with the use of the population aged 20 and over in Garbinti, Goupille-Lebret, and Piketty (2016), we keep persons aged 20 and over in the survey.

B.5.2 Further Tables and Figures

B.5.2.1 Shape of the Bias

Figures B.8–B.12 show the shape of the bias we estimate for the other years among our sampled countries. Each coverage of the data points are determined by the trustable span of the tax data in each country, which is defined as the portion of the population that are subject to positive income tax payments.

Figure B.8: Merging Points in Norway, 2004–2013



(a) Norway 2004

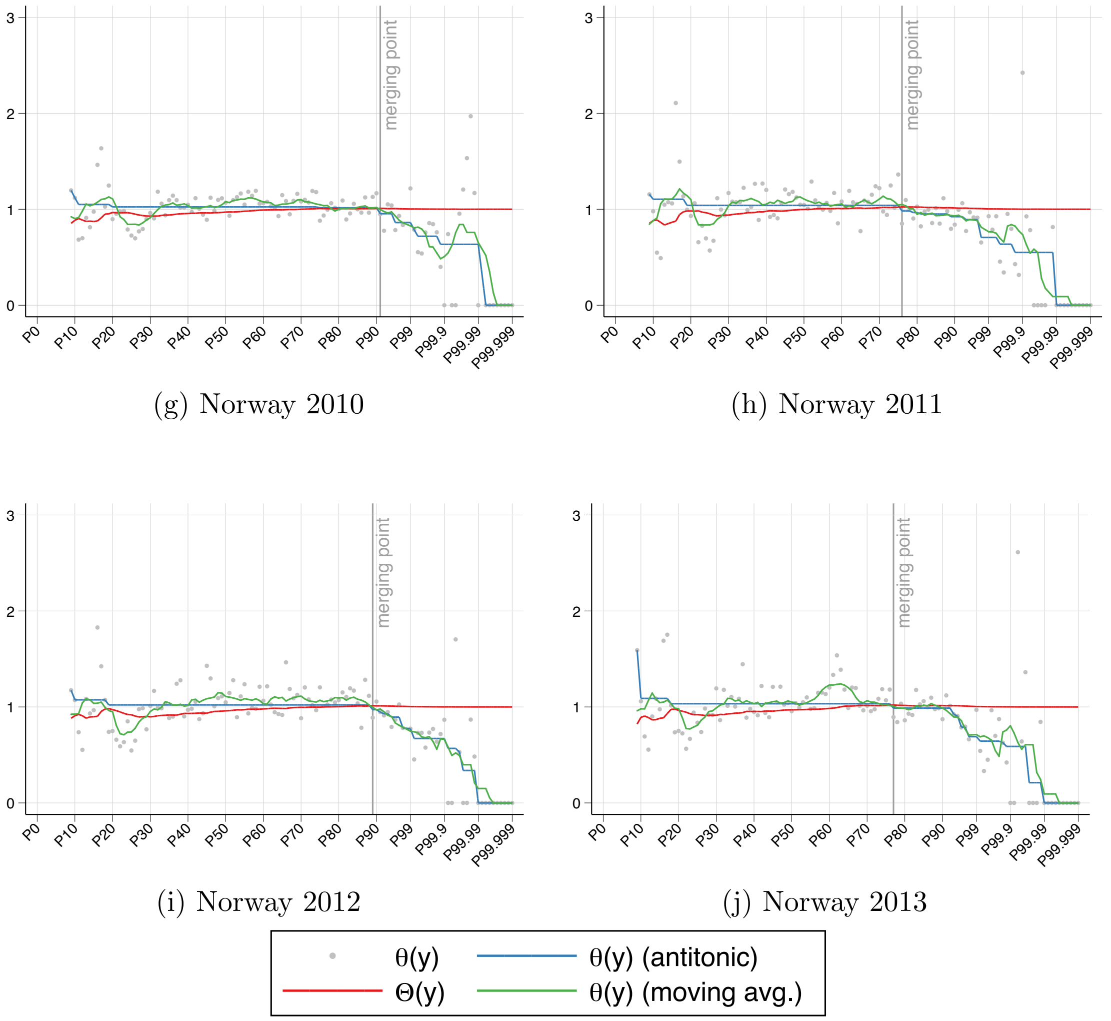
(b) Norway 2005

(c) Norway 2006

(d) Norway 2007

(e) Norway 2008

(f) Norway 2009



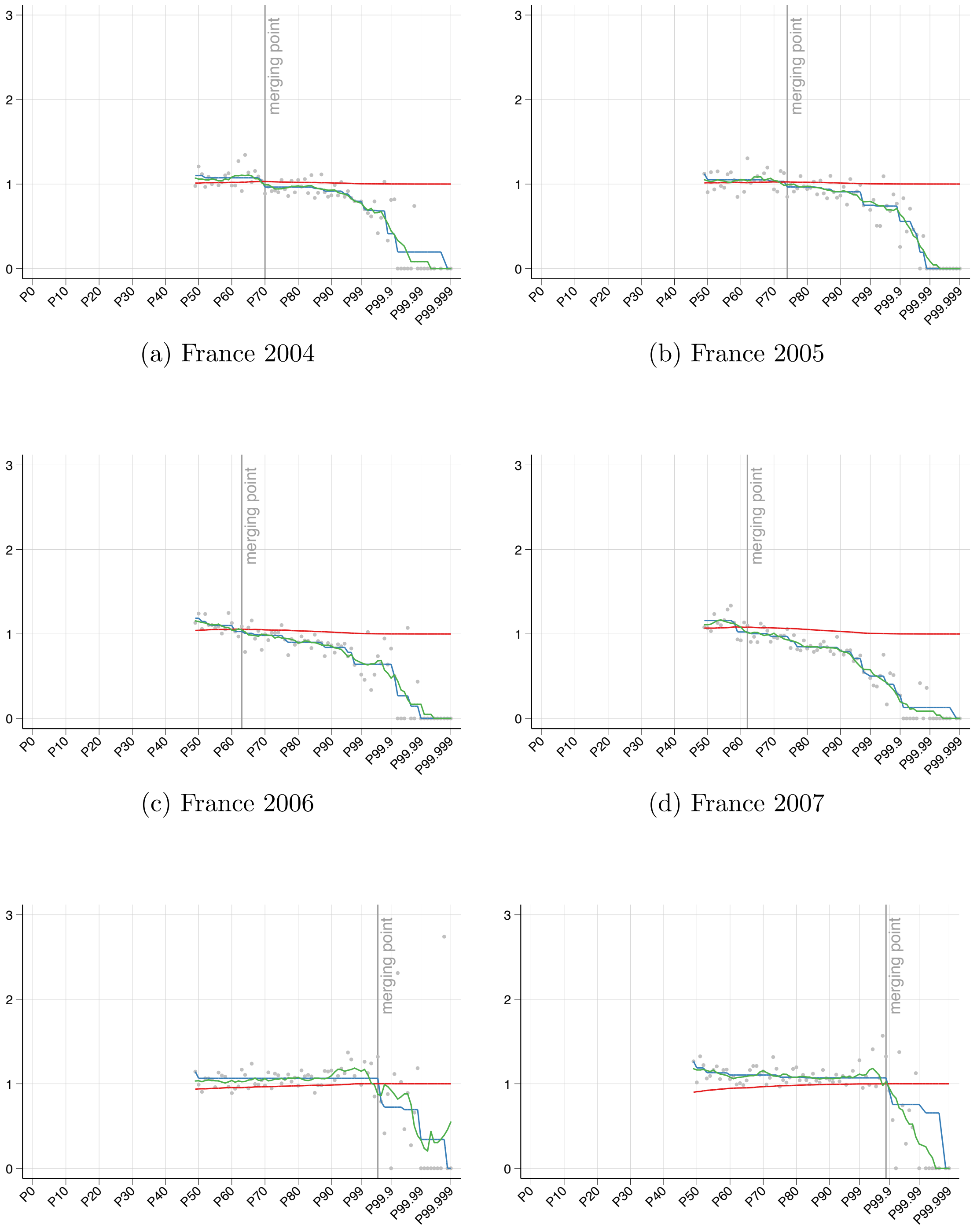
(g) Norway 2010

(h) Norway 2011

(i) Norway 2012

(j) Norway 2013

Figure B.9: Merging Points in France, 2004–2013



(a) France 2004

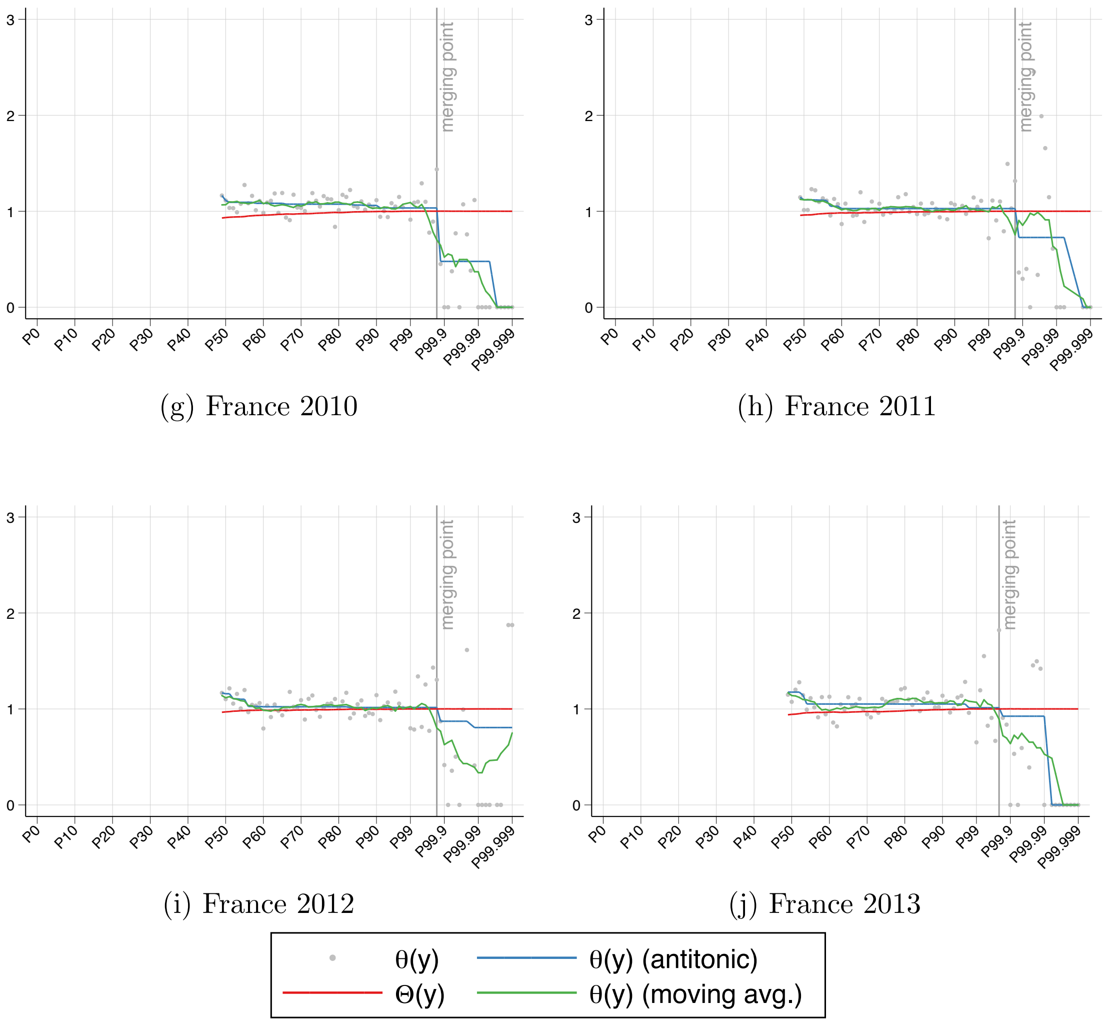
(b) France 2005

(c) France 2006

(d) France 2007

(e) France 2008

(f) France 2009



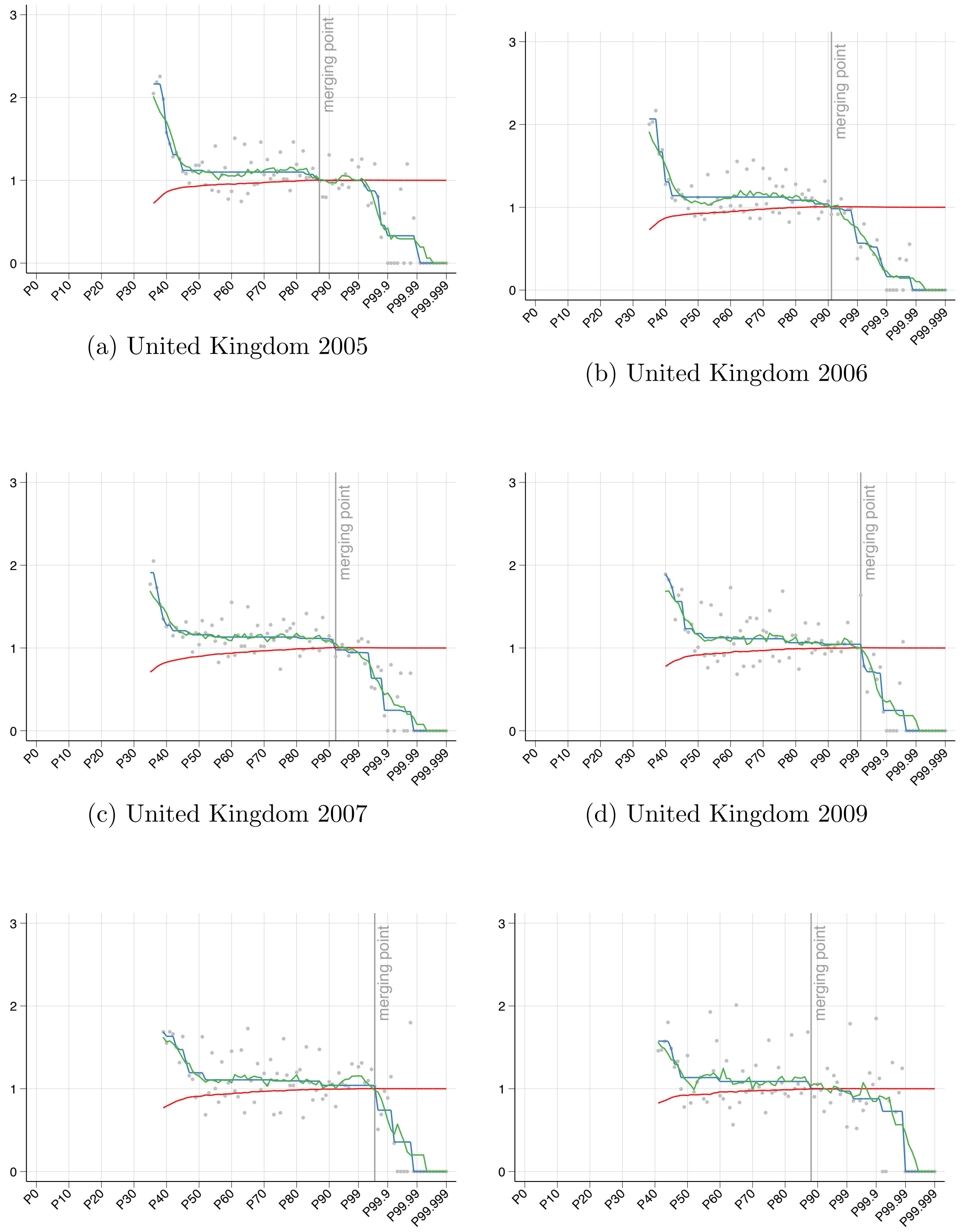
(g) France 2010

(h) France 2011

(i) France 2012

(j) France 2013

Figure B.10: Merging Points in United Kingdom, 2005–2013



(a) United Kingdom 2005

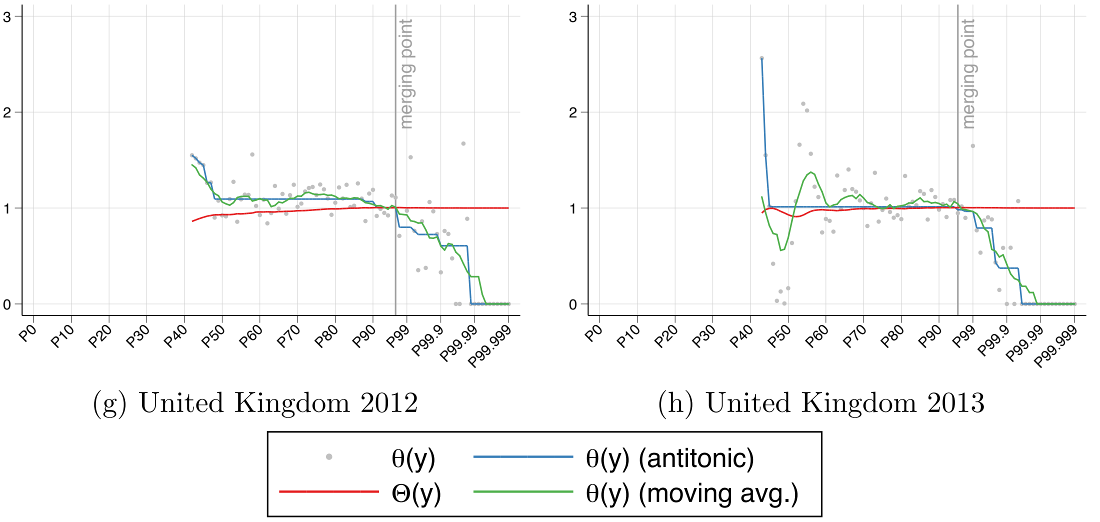
(b) United Kingdom 2006

(c) United Kingdom 2007

(d) United Kingdom 2009

(e) United Kingdom 2010

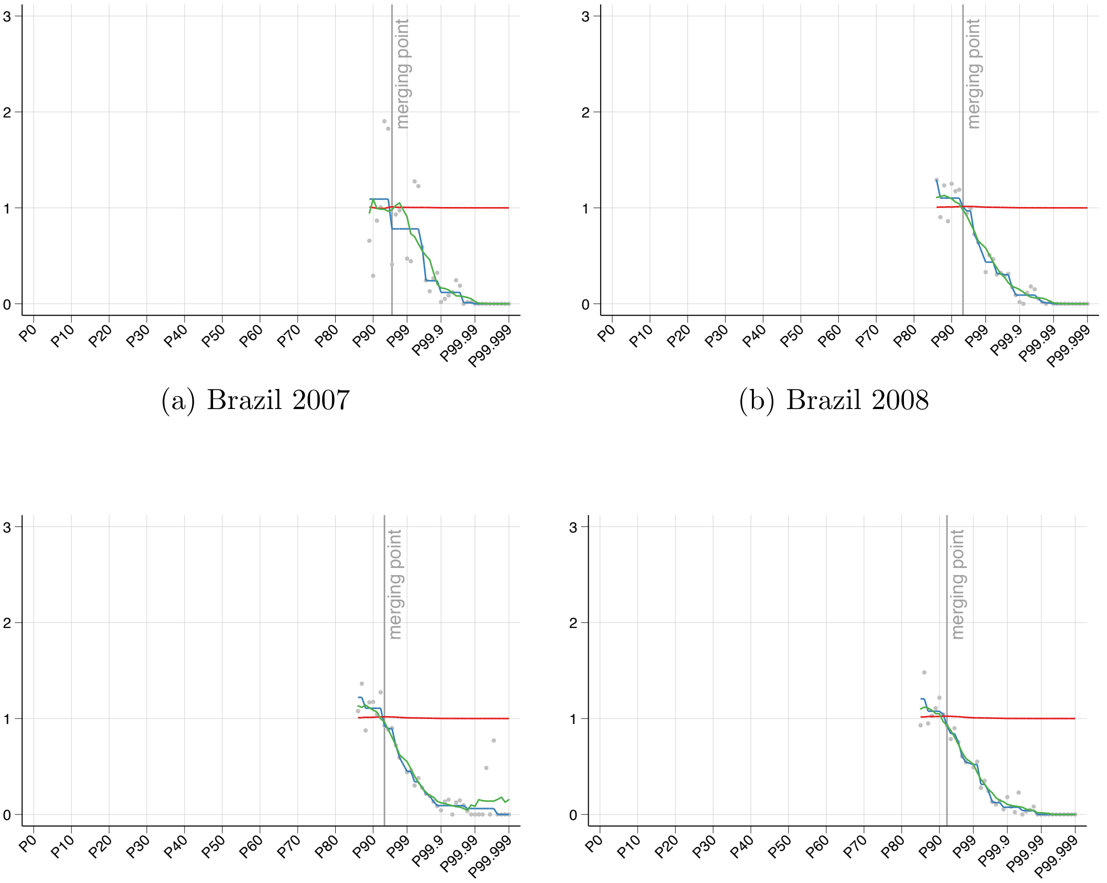
(f) United Kingdom 2011



(g) United Kingdom 2012

(h) United Kingdom 2013

Figure B.11: Merging Points in Brazil, 2007–2014

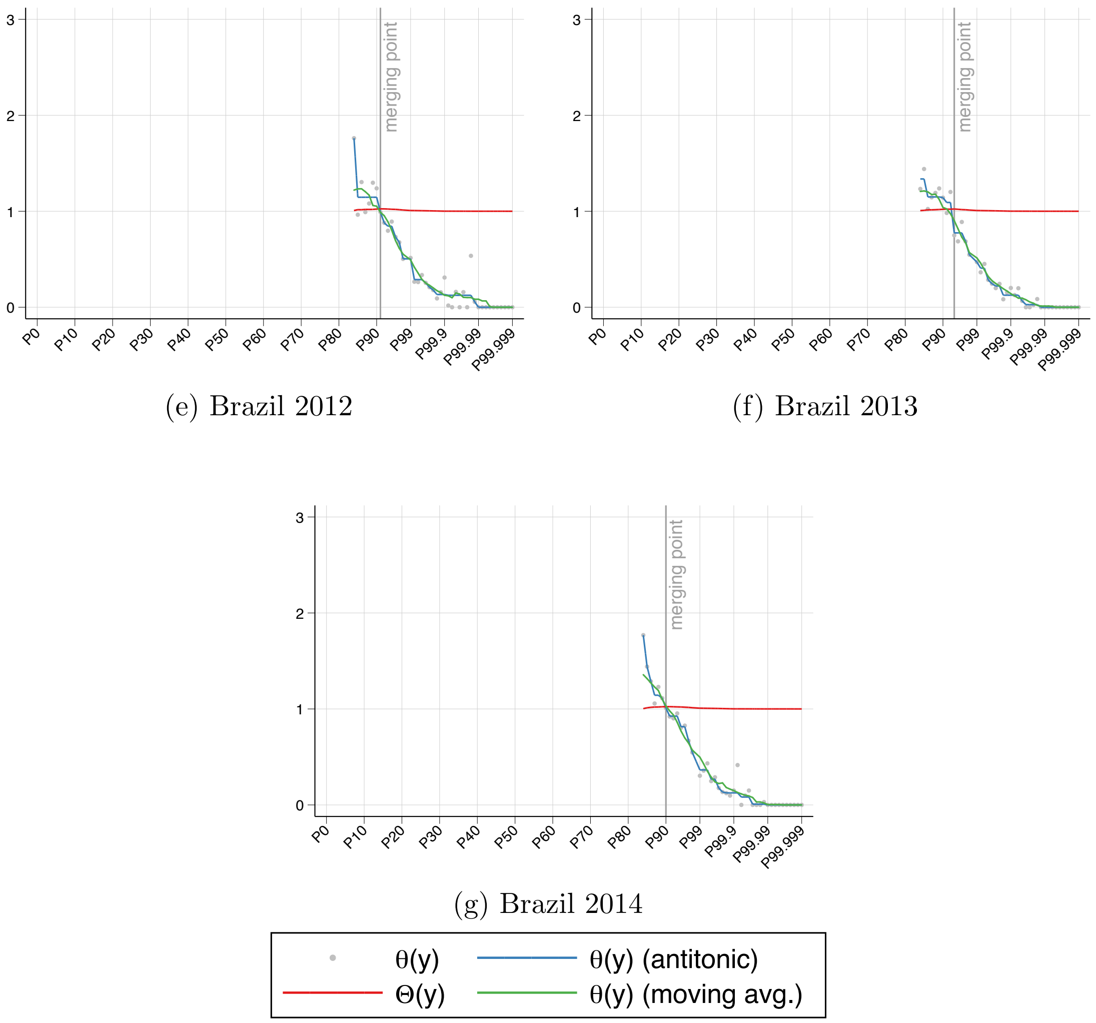


(a) Brazil 2007

(b) Brazil 2008

(c) Brazil 2009

(d) Brazil 2011

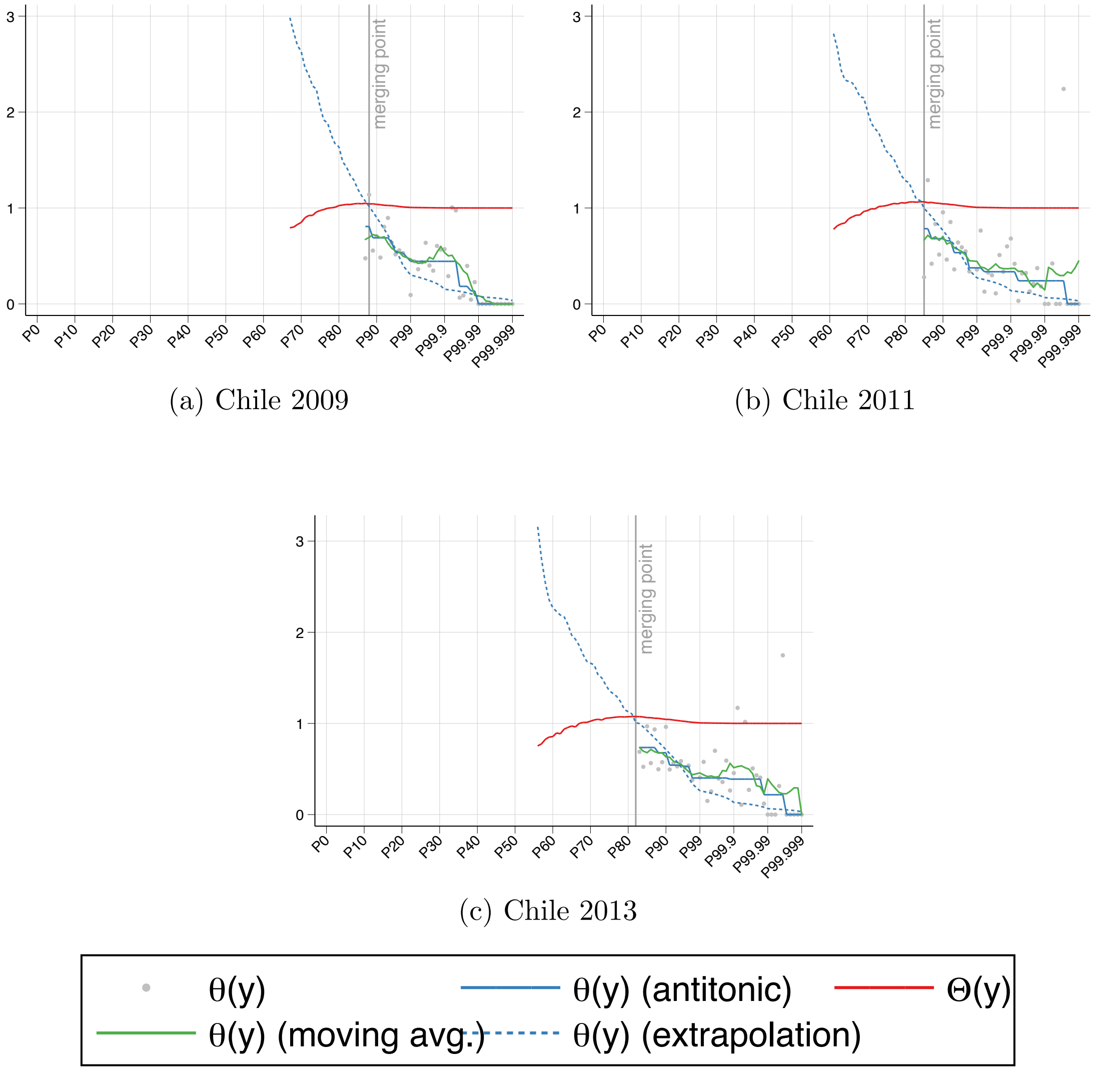


(e) Brazil 2012

(f) Brazil 2013

(g) Brazil 2014

Figure B.12: Merging Points in Chile, 2009–2013



(a) Chile 2009

(b) Chile 2011

(c) Chile 2013

B.5.2.2 Structure of the Corrected Population

Tables B.2-B.6 show the structure of the corrected population for all years in all sampled countries.

Table B.2: Structure of Corrected Population in Brazil, 2007-2015

Year	Population over Merging Point (% total population)		Corrected population		
	Tax data	Survey	Total	Share inside survey support	Share outside survey support
	[2]	[3]	[4] = [2] − [3]	[5]	[6]
2007	1.0%	0.7%	0.33%	98.2%	1.8%
2008	1.0%	0.6%	0.44%	97.2%	2.8%
2009	1.0%	0.5%	0.51%	99.3%	0.7%
2011	2.0%	1.4%	0.57%	95.9%	4.1%
2012	3.0%	2.3%	0.70%	98.3%	1.7%
2013	2.0%	1.4%	0.62%	97.1%	2.9%
2014	2.0%	1.2%	0.76%	98.8%	1.2%
2015	2.0%	1.3%	0.70%	97.2%	2.8%

Notes: Column [2] shows the proportion of the population that is above this merging point in the tax data. Column [3] shows the proportion that is above the merging point in survey data. The difference between the two is the proportion of the survey population that is corrected (Column [4]). As explained in the text, we adjust survey weights below the merging point by the same proportion. The corrected proportion above the merging point can be decomposed into the share of the corrected population that is inside the survey support (up to the survey's maximum income) and the share that is outside the support (observations with income above the survey's maximum).

Table B.3: Structure of Corrected Population in Chile, 2009-2015

Year	Population over Merging Point (% total population)		Corrected population		
	Tax data	Survey	Total	Share inside survey support	Share outside survey support
	[2]	[3]	[4] = [2] − [3]	[5]	[6]
2009	11.0%	7.2%	3.8%	99.6%	0.4%
2011	14.0%	8.5%	5.5%	99.9%	0.1%
2013	17.0%	10.6%	6.4%	99.9%	0.1%
2015	17.0%	11.1%	5.7%	99.99%	0.01%

Notes: Column [2] shows the proportion of the population that is above this merging point in the tax data. Column [3] shows the proportion that is above the merging point in survey data. The difference between the two is the proportion of the survey population that is corrected (Column [4]). As explained in the text, we adjust survey weights below the merging point by the same proportion. The corrected proportion above the merging point can be decomposed into the share of the corrected population that is inside the survey support (up to the survey's maximum income) and the share that is outside the support (observations with income above the survey's maximum).

Table B.4: Structure of Corrected Population in France, 2004-2014

Year	Population over Merging Point (% total population)		Corrected population		
	Tax data	Survey	Total	Share inside survey support	Share outside survey support
	[2]	[3]	[4] = [2] − [3]	[5]	[6]
2004	29.0%	26.8%	2.17%	99.9%	0.1%
2005	25.0%	23.1%	1.95%	98.5%	1.5%
2006	36.0%	32.5%	3.50%	99.5%	0.5%
2007	37.0%	32.0%	4.99%	99.96%	0.04%
2008	0.4%	0.3%	0.11%	97.6%	2.4%
2009	0.1%	0.1%	0.02%	89.8%	10.2%
2010	0.2%	0.1%	0.11%	94.5%	5.5%
2011	0.2%	0.1%	0.06%	94.3%	5.7%
2012	0.2%	0.2%	0.03%	96.5%	3.5%
2013	0.3%	0.3%	0.03%	72.3%	27.7%
2014	0.1%	0.0%	0.05%	99.0%	1.0%

Notes: From 2008, the French survey was supplemented with register data for increased precision in the responses. Column [2] shows the proportion of the population that is above this merging point in the tax data. Column [3] shows the proportion that is above the merging point in survey data. The difference between the two is the proportion of the survey population that is corrected (Column [4]). As explained in the text, we adjust survey weights below the merging point by the same proportion. The corrected proportion above the merging point can be decomposed into the share of the corrected population that is inside the survey support (up to the survey's maximum income) and the share that is outside the support (observations with income above the survey's maximum).

Table B.5: Structure of Corrected Population in Norway, 2004-2014

Year	Population over Merging Point (% total population)		Corrected population		
	Tax data [2]	Survey [3]	Total [4] = [2] − [3]	Share inside survey support [5]	Share outside survey support [6]
2004	24.0%	22.5%	1.49%	99.3%	0.7%
2005	22.0%	19.7%	2.27%	99.8%	0.2%
2006	31.0%	28.8%	2.16%	99.9%	0.1%
2007	39.0%	34.2%	4.75%	99.5%	0.5%
2008	38.0%	33.4%	4.59%	99.95%	0.05%
2009	4.0%	3.5%	0.54%	99.4%	0.6%
2010	8.0%	7.1%	0.88%	99.0%	1.0%
2011	23.0%	21.1%	1.93%	99.0%	1.0%
2012	10.0%	8.9%	1.13%	98.6%	1.4%
2013	22.0%	20.5%	1.49%	99.1%	0.9%
2014	5.0%	4.6%	0.39%	96.0%	4.0%

Notes: Column [2] shows the proportion of the population that is above this merging point in the tax data. Column [3] shows the proportion that is above the merging point in survey data. The difference between the two is the proportion of the survey population that is corrected (Column [4]). As explained in the text, we adjust survey weights below the merging point by the same proportion. The corrected proportion above the merging point can be decomposed into the share of the corrected population that is inside the survey support (up to the survey's maximum income) and the share that is outside the support (observations with income above the survey's maximum).

Table B.6: Structure of Corrected Population in United Kingdom, 2005-2014

Year	Population over Merging Point (% total population)		Corrected population		
	Tax data [2]	Survey [3]	Total [4] = [2] − [3]	Share inside survey support [5]	Share outside survey support [6]
2005	12.0%	11.7%	0.26%	99.5%	0.5%
2006	8.0%	7.3%	0.72%	96.9%	3.1%
2007	7.0%	6.5%	0.53%	95.5%	4.5%
2009	0.8%	0.5%	0.33%	85.5%	14.5%
2010	0.4%	0.3%	0.14%	84.9%	15.1%
2011	11.0%	10.8%	0.18%	93.0%	7.0%
2012	3.0%	2.6%	0.37%	92.2%	7.8%
2013	4.0%	3.6%	0.45%	86.1%	13.9%
2014	3.0%	2.5%	0.54%	93.6%	6.4%

Notes: Column [2] shows the proportion of the population that is above this merging point in the tax data. Column [3] shows the proportion that is above the merging point in survey data. The difference between the two is the proportion of the survey population that is corrected (Column [4]). As explained in the text, we adjust survey weights below the merging point by the same proportion. The corrected proportion above the merging point can be decomposed into the share of the corrected population that is inside the survey support (up to the survey's maximum income) and the share that is outside the support (observations with income above the survey's maximum).

B.5.3 Detailed Distribution

Table B.7 depicts a more detailed picture of the impact of our adjustment method on the income distribution of our 5 countries, compared to the raw survey results and those from the replacing alternative. We take the last available year as an illustration. With respect to income shares across the distribution, the main conclusions drawn from the analysis of top shares in Section 2.3 can be generally extended, more or less, to other top shares, from the top 10% to the top 0.001% shares. As is to be expected, both the middle 40% and Bottom 50% shares are reduced in all countries after our adjustment. This is consistent with the mechanics of our method, where higher aggregate weight for top fractile incomes must be compensated by a lowering of the amount of middle and lower incomes observed in the population. Replacing produces results in the same direction, except that, by not decreasing the weight of lower incomes, it results in higher shares for the Bottom 50% than those from our method in all countries. The same is true for the Middle 40% for Brazil and Chile, but not for the three European countries. Overall, replacing produces inconsistent results across the distribution, which are difficult to explain.

Figure B.13 presents in more detail the impact of our method on total income. For our two country case studies with the largest corrections to total income, we are able to show that the total income in the corrected surveys is closer to the reference total of “fiscal income” from national accounts. For the cases of Chile and Brazil respectively, our correction bridges about 80% and 60% of the gap between survey income and the reference total from national accounts.

Table B.7: Income Shares: Raw Survey and Corrected Survey

Raw Survey

Income groups	**Brazil**	**Chile**	**France**	**Norway**	**UK**
Bottom 50%	16.9%	8.0%	23.4%	25.2%	14.8%
Middle 40%	45.3%	45.2%	47.0%	48.6%	49.6%
Top 10%	37.7%	46.9%	29.6%	26.2%	35.5%
Incl. Top 1%	*10.2%*	*14.3%*	*7.2%*	*5.8%*	*9.4%*
Incl. Top 0.1%	*2.2%*	*3.4%*	*1.5%*	*1.4%*	*2.5%*
Incl. Top 0.01%	*0.5%*	*0.7%*	*0.4%*	*0.3%*	*0.4%*
Incl. Top 0.001%	*0.09%*	*0.2%*	*0.1%*	*0.03%*	*0.04%*
Average income	€8,691	€8,101	€23,367	€37,431	€22,389
Gini	0.505	0.64	0.40	0.37	0.52

Corrected Survey (Our Method)

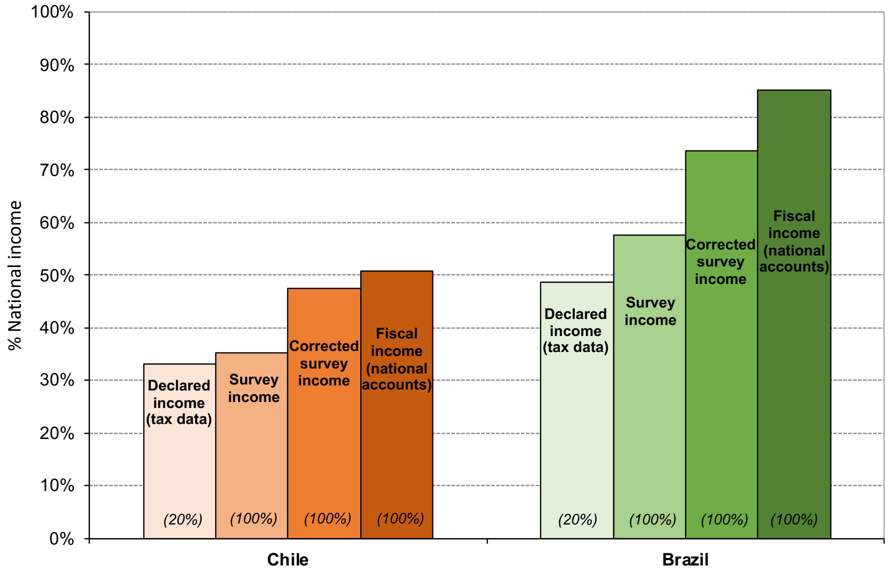
Income groups	**Brazil**	**Chile**	**France**	**Norway**	**UK**
Bottom 50%	12.7%	6.7%	23.2%	24.6%	13.9%
Middle 40%	35.1%	40.1%	46.5%	47.7%	46.6%
Top 10%	52.3%	53.2%	30.3%	27.6%	39.6%
Incl. Top 1%	*23.7%*	*16.7%*	*8.2%*	*7.1%*	*13.7%*
Incl. Top 0.1%	*11.2%*	*4.5%*	*2.2%*	*2.2%*	*5.4%*
Incl. Top 0.01%	*5.6%*	*1.3%*	*0.6%*	*0.7%*	*2.1%*
Incl. Top 0.001%	*2.8%*	*0.4%*	*0.2%*	*0.26%*	*0.89%*
Average income	€11,935	€11,097	€23,621	€38,320	€24,081
Gini	0.619	0.69	0.41	0.38	0.55

Corrected Survey (Replacing)

Income groups	**Brazil**	**Chile**	**France**	**Norway**	**UK**
Bottom 50%	14.4%	7.9%	24.0%	25.7%	14.8%
Middle 40%	36.4%	41.0%	45.9%	47.1%	46.4%
Top 10%	49.2%	51.2%	30.0%	27.2%	38.8%
Incl. Top 1%	*26.7%*	*21.1%*	*7.9%*	*7.1%*	*14.0%*
Incl. Top 0.1%	*12.6%*	*5.7%*	*2.2%*	*2.2%*	*5.5%*
Incl. Top 0.01%	*6.3%*	*1.6%*	*0.6%*	*0.7%*	*2.1%*
Incl. Top 0.001%	*3.1%*	*0.5%*	*0.2%*	*0.26%*	*0.90%*
Average income	€10,647	€8,792	€23,439	€37,956	€23,578
Gini	0.624	0.70	0.44	0.40	0.57

Notes: The table presents the distribution of pre-tax fiscal income per adult, in the survey before the correction and after the correction using our method and the replacing alternative used in Section 2.3. Average incomes are expressed in French Euros PPP. Brazil and Chile refer to 2015, while all the European countries refer to 2014.

Figure B.13: Discrepancy of income across datasets in Chile and Brazil: 2015



Reading: in 2015 the total income declared in tax data in Brazil, which covers 20% of the population represents 49% of national income. The total income in the raw survey represents 58% of national income and 74% in the corrected survey, which are both representative of the entire population. The equivalent income calculated from national accounts represents 85% of national income. Authors' calculations using data from surveys, income tax declarations and national accounts.

Bibliography

Abowd, John M. and Martha H. Stinson (2013). "Estimating measurement error in annual job earnings: A comparison of survey and administrative data". In: *Review of Economics and Statistics* 95.5, pp. 1451–1467.

Angel, Stefan, Richard Heuberger, and Nadja Lamei (2017). "Differences Between Household Income from Surveys and Registers and How These Affect the Poverty Headcount: Evidence from the Austrian SILC". In: *Social Indicators Research* 138.2, pp. 1–29.

Atkinson, A. B. (2007). "The distribution of top incomes in the United Kingdom 1908–2000". In: *Top Incomes over the Twentieth Century: A Contrast between Continental European and English-Speaking Countries*. Ed. by A. B. Atkinson and Thomas Piketty. Vol. 1. Oxford University Press, pp. 82–140.

Blanchet, Thomas, Juliette Fournier, and Thomas Piketty (2017). "Generalized Pareto Curves: Theory and Applications".

Bollinger, Christopher R (1998). "Measurement error in the Current Population Survey: a nonparametric look." In: *Journal of labor economics* 16.3, pp. 576–594.

Bound, John and Alan B. Krueger (1991). "The Extent of Measurement Error in Longitudinal Earnings Data: Do Two Wrongs Make a Right?" In: *Journal of Labor Economics* 9.1, p. 1.

Cristia, Julian and Jonathan A. Schwabish (2009). "Measurement error in the SIPP: Evidence from administrative matched records". In: *Journal of Economic and Social Measurement* 34.1, pp. 1–17.

Deville, Jean-Claude and Carl-Erik Särndal (1992). "Calibration Estimators in Survey Sampling". In: *Journal of the American Statistical Association* 87.418, pp. 376–382.

Flores, Ignacio et al. (n.d.). "Top Incomes in Chile: A Historical Perspective on Income Inequality, 1964–2017". In: *Review of Income and Wealth* 0.0 (). eprint: `https://onlinelibrary.wiley.com/doi/pdf/10.1111/roiw.12441`. URL: `https://onlinelibrary.wiley.com/doi/abs/10.1111/roiw.12441`. Forthcoming.

Garbinti, Bertrand, Jonathan Goupille-Lebret, and Thomas Piketty (2016). "Income Inequality in France, 1900-2014: Evidence from Distributional National Accounts (DINA)". URL: `http://piketty.pse.ens.fr/files/GGP2016DINA.pdf`.

Hoerl, Arthur E and Robert W Kennard (2000). "Ridge Regression: Biased Estimation for Problems Nonorthogonal". In: *Technometrics* 42.1, pp. 80–86.

Morgan, Marc (2018). "Essays on Income Distribution: Methodological, Historical and Institutional Perspectives with Applications to the Case of Brazil (1926–2016)". PhD Dissertation in Economics. Paris: Paris School of Economics & EHESS.

Paulus, Alari (2015). “Income underreporting based on income expenditure gaps: Survey vs tax records”. URL: `http://hdl.handle.net/10419/126467`.

Pedace, Roberto and Nancy Bates (2000). “Using administrative records to assess earnings reporting error in the survey of income and program participation”. In: *Journal of Economic and Social Measurement* 26, pp. 173–192.

Appendix C

Appendix to “How Unequal is Europe? Evidence from Distributional National Accounts”

This is the main appendix of our paper. For the detailed data appendix, including country-specific discussions, see `https://wid.world/europe2019`.

C.1 Detailed Methodology

The issues that affect the validity and the comparability of existing income inequality estimates may be divided into three categories: conceptual discrepancies, nonsampling error, and sampling error.

Conceptual discrepancies are not errors in themselves but refer to differences as to what, precisely, is being measured. Existing estimates of income inequality may be concerned with different types of income and different populations units. While there may be a case for measuring inequality using any of these concepts and units, the existence of such a wide range of definitions makes it hard to compare inequality estimates both over time and between countries. As we have seen, both survey tabulations and fiscal data suffer from important conceptual discrepancies, as they are measured on different groups of individuals and using different income concepts. One of the contributions of this paper is to provide a new method to harmonize these different distributions.[1]

[1] Previous studies on European or global income distributions typically relied on a combination of non-harmonized income and consumption sources, see for instance Lakner and Milanovic (2016).

Sampling and non-sampling errors apply to surveys. Sampling error refers to problems that arise purely out of the limited sample size of survey data. Low sample sizes affect the variance of estimates, which means they may vary a lot around their expected value. But low sample sizes may also create biases, especially when measuring inequality at the top of the distribution (Taleb and Douady, 2015). Estimates based on raw survey data do not account for any of these biases and therefore tend to underestimate incomes at the top end. Non-sampling error refers to the systematic biases that affect survey estimates in a way that is not directly affected by the sample size. These mostly include people refusing to answer surveys and misreporting their income in ways that are not observed, and therefore not corrected, by the survey producers.

The general methodology we introduce in this paper aims at correcting all three biases. We correct conceptual discrepancies by training a machine learning algorithm (Chen and Guestrin, 2016) that systematically analyzes how they affect estimates of the income distribution. We correct for non-sampling error in survey data by combining them with harmonized top income shares using a nonlinear survey calibration method (Deville and Särndal, 1992; Lesage, 2009). And we correct for sampling error by modeling the top tail of the income distribution based on extreme value theory (Ferreira and Haan, 2006). We view this methodology as a consistent and straightforward framework to exploit all published survey and tax information, while correcting for the weaknesses of these different sources. We feed to our methodology virtually all the data available and obtain estimates of inequality in Europe that reflect latest data and methodological developments.

C.1.1 Machine Learning Algorithm to Harmonize the Survey Data

The first step of our methodology consists in harmonizing surveys for which we are unable to recover directly the distribution of pre-tax and post-tax incomes among equal-split adults. This is the case of all survey tabulations, as well as some surveys for which we have microdata but for which pre-tax income or post-tax income was not measured. For these data sources, we have to develop a strategy to transform the distribution of the observed ”source concept” (such as consumption per capita or pre-tax income among households, for instance) into an imputed distribution measured in a ”target concept” (pre-tax or post-tax income per adult).

The distributions for the different income concepts across country-years are correlated: therefore, we can use the distribution for one income concept to impute the distribution for another whenever the former is observed but not the latter. To do so, we use all the cases where the income distribution is simultaneously observed for two different concepts to learn how one tends to relate to another. In practice, we use survey microdata (EU-SILC, LIS and ECHP) to compute distributions for all equivalence scales and all income concepts available in a given country-year. We then use these estimates — as well as survey tabulations observed in similar country-years but measured using different concepts — to model how different income concepts and population units relate to one another at different points of the distribution.

To clarify this idea, we can first consider a straightforward, but naive approach. We can observe the p-th quantile of both the source and the target distributions for a variety of countries i and a variety of years t: write them $Q_{it}^{\text{target}}(p)$ and $Q_{it}^{\text{source}}(p)$. Therefore, we can estimate the average ratio between the two distributions for each percentile as $\alpha(p) = \mathbb{E}[Q_{it}^{\text{target}}(p)/Q_{it}^{\text{source}}(p)]$. Say that for a country j in year s, we only observe the source concept $Q_{js}^{\text{source}}(p)$. Then we can approximate the target concept as $Q_{js}^{\text{target}}(p) = \alpha(p)Q_{js}^{\text{source}}(p)$. While this remains an approximation, it at least corrects for some systematic discrepancies that we can observe in the data.

That approach has the merit of simplicity. When we tried it with our data, it gave passable results. But there are several problems with it, both in theory and in practice. The main issue is that it makes a very restrictive assumption about the way different income concepts may relate to one another: it considers that the sole predictor of, say, the 25th percentile of income for equal-split adults is the 25th percentile of income for households. Furthermore, it assumes that the relationship is entirely linear. There is no good theoretical reason for any of that to be true: a better, more general model would allow that 25th percentile of the target distribution to depend on any percentile of the source distribution, including but not limited to the 25th. It would also allow these relationships to be nonlinear and potentially with interactions. That relationship could also depend on auxiliary variables Z_{it} capturing demographic, political and institutional factors. The simple approach also cannot ensure that the estimated distribution for the target concept will be increasing, which creates problems that have to be dealt with in an *ad hoc* way (e.g. by re-ranking percentiles) and imply inefficient use of information. This in particularly true for the bottom of the distribution for which incomes can be close to zero and the ratios may therefore be very unstable.

Therefore, to construct the best mappings between the different concepts, we consider a much more general model. In that model, each percentile of the target distribution is an arbitrary function of every percentile of the source distribution, and of additional covariates. We write:

$$\mathbb{E}[Q_{it}^{\text{target}}(p)] = \varphi(Q_{it}^{\text{source}}(p_1), \dots, Q_{it}^{\text{source}}(p_m), p, t, Z_{it})$$

for a grid $0 \leq p_1 < \cdots < p_m < 1$ of fractiles. Estimating such a model raises some challenges. Linear regression will not be flexible enough due to its parametric assumptions and will tend to overfit the data if m is large due to the number of covariates.

To estimate this model, we therefore rely on more recent advances in high-dimensional, nonparametric regression, also known as *machine learning* methods. The algorithm we use is known as *boosted regression trees*, a powerful and commonly used method introduced by Friedman (2001). We rely on an implementation known as XGBoost (Chen and Guestrin, 2016), which has enjoyed great success due to its speed and performance, to the point that is has earned a reputation for "winning every machine learning competition" (Nielsen, 2016). On top of their performance, boosted regression makes it easy to deal with missing values, or to impose certain constraints, such as the fact that the quantile function $Q(p)$ must be increasing with p.

The algorithm starts from regression trees, a fast and simple nonlinear prediction method that successively cuts the space of predictors into two subspaces in which the predicted variable has lower variance. This leads to a "tree" of simple decision rules based on the value of the predictors. Following these rules the algorithm places any observation into a subspace where the predictor should have a relatively low variance, and the predicted value for that observation is the average of the predictor within that subspace.

Regression trees provide predictions that are simple, but rough. "Boosting" is a method that combines many of these simple but low accuracy prediction methods into a high-accuracy one. It starts by estimating a regression tree. It then runs a second regression tree to predict the residual from the previous regression: this is called a "boosting round." The process is repeated several times: each round of boosting forces the algorithm to concentrate on the part of the data where the previous predictions failed. In the end, all the regression trees are combined into a single prediction.

The appropriate number of boosting rounds is determined by cross-validation: the sample is divided into K subsamples. For each subsample, we train the algorithm on the data excluding the subsample, and we test the prediction on the excluded subsample: we use the number of boosting rounds for which the cross-validation prediction error is lowest. By excluding the sample on which we perform the prediction, we make sure to avoid overfitting to the data on which we estimate the model.

Since our dataset is made up of countries that we follow over the years, it has a panel dimension, which we take into account as follows. We assume that the country-specific prediction error is independent conditional on all observed variables (i.e. that it is a *random* rather than a *fixed* effect.) Under that assumption, the imputation method remains valid because the error term remains exogenous. However, there is a risk of over-fitting if we do not make sure that the different subsamples used in the cross-validation are not independent, because then we would force the algorithm to try to predict the country random effect. To avoid that problem, we perform the cross-validation by making sure that all the observations for one country are in the same cross-validation subsample, which is known as leave-one-cluster-out cross validation (Fang, 2011). When possible, we also estimate and include the country random effect into our imputation. The random effect is estimated as a function of the percentile using the mean prediction error by country and percentile.

In the end, for any target concept of interest, we get as many predictions as there are sources available. Let $\boldsymbol{y} = (\hat{Q}_{it}^{\text{target},1}, \dots, \hat{Q}_{it}^{\text{target,n}})'$ the n different predictions. Using the cross-validation estimation of the prediction error, we can estimate the variance-covariance matrix $\boldsymbol{\Sigma}$ between the different predictions. Following the logic of generalized least squares, the optimal way of combining the n predictions into one is to average them, weighted by the row or column sums of the symmetric matrix $\boldsymbol{\Sigma}$. This yields our harmonized estimate of the distribution, taking into account observed regularities across concepts and percentile groups.

As table C.1 shows, the mean (cross-validation) prediction error for the value of the average of a percentile is between 2% and 11% depending on the concept that was used for the prediction.[2] Adjusting for the statistical unit while keeping the income concept identical creates the least difficulties. Consumption, on the other hand, is a rather poor predictor of income. Moving from post-tax to pre-tax income is a somewhat intermediary situation. The auxiliary variables that we use to improve the performance of the prediction are the average national income, the share of

[2] Before training the model, we transform the data using the transform $y \mapsto \operatorname{asinh}(y)$ for the value of the quantiles and $x \mapsto -\log(1-x)$ for the corresponding rank. This stabilizes the mode without changing the nature of the data. The use of asinh rather than the logarithm avoid issues with having zero or near-zero incomes at the bottom of the distribution. All distributions are normalized by their average since we are only concerned with the distribution of income. When we report prediction errors, these are computed for distributions that have been properly transformed back to their original value.

Table C.1: Mean relative error on the average by percentile when imputing pre-tax and post-tax income by adults from a different concept using the machine learning algorithm

		mean relative prediction error	
income/consumption concept	statistical unit	pre-tax income	post-tax income
consumption	equal-split per adults	10.1%	10.6%
consumption	equal-split per capita	10.6%	10.7%
consumption	households	11.0%	9.4%
consumption	OECD equivalence scale	9.8%	10.4%
consumption	square root equivalence scale	9.8%	9.8%
pre-tax income	equal-split per adults	n/a	5.7%
pre-tax income	equal-split per capita	4.2%	6.1%
pre-tax income	households	3.9%	6.8%
pre-tax income	OECD equivalence scale	2.6%	6.1%
pre-tax income	square root equivalence scale	2.8%	6.2%
post-tax	equal-split per adults	5.8%	n/a
post-tax	equal-split per capita	7.0%	3.8%
post-tax	households	7.1%	4.1%
post-tax	OECD equivalence scale	6.1%	2.2%
post-tax	square root equivalence scale	6.0%	2.7%

Source: authors' computations. *Note*: Error calculated only for the top 90% of the distributions to avoid problems of denominator equal to zero. *Interpretation*: When trying to impute pre-tax income per equal-split adult from consumption per household, the mean relative error for the average income of a given percentile is 11%.

households with different sizes, the population structure by age and gender, the top tax rates and social expenditures. While the inclusion of these variables has only second-order effects on our harmonized series, they do improve the prediction error by about 15–20%.

C.1.2 Calibration on Top Income Shares to Correct for Non-sampling Error

We correct survey data for non-sampling error using known top income shares estimated from administrative tax data. We do so by adjusting the survey weights using survey calibration methods (Deville and Särndal, 1992). Statistical institutes already routinely use these methods to ensure that their surveys are representative, typically in terms of age and gender. Our approach is a natural extension of theirs, in the sense that we enforce representativity in terms of taxable income in addition to age and gender.

Let $d_1, \dots, d_n$ be the original survey weights, and let $w_1, \dots, w_n$ be the corrected survey weights. The objective of survey calibration is to minimize the distortion between the original survey weights and the corrected survey weights:

$$\min_{w_1,\dots,w_n} \sum_{i=1}^{n} \frac{(d_i - w_i)^2}{d_i}$$

under the constraint that the top shares in the corrected survey are equal to their value in the tax data. However, traditional survey calibration methods only work with constraints that can be written as a linear function of the data (such as a mean or a frequency), which is not the case with top shares.

Lesage (2009) suggested two methods to solve such problems. The first one involves linearizing the top shares using their *influence function.* Informally, the influence measures the marginal contribution of the weight of each observation to the overall statistic. For the case of the top $(1 - \alpha) \times 100\%$ share, we show in the technical appendix that it is equal to:

$$z_i = y_i H\left(\frac{\alpha N - W_{i-1}}{w_i}\right) + (\alpha - \mathbb{1}_{y_i < \hat{Q}_\alpha})\hat{Q}_\alpha$$

where y_i is the income of observation i, w_i is the weight of observation i, $\hat{Q}_\alpha$ is the α-quantile of income in the survey, and H is a function such that $H(x) = 0$ if $x < 0$, $H(x) = x$ if $0 \leq x < 1$ and $H(x) = 1$ if $x \geq 1$. As Lesage (2009) explains, it then suffices to impose the linear constraint $\sum_{i=1}^{n} w_i z_i = 0$ in standard calibration methods to approximately enforce, up to a first order approximation, the value of the top income share. Intuitively, the survey calibration performs a trade-off between spreading the adjustment of the weights over as many observations as possible (hence minimizing overall distortion) and concentrating the adjustment on the observations with the largest impact on the top share (hence satisfying the constraint with fewer distortions). The optimum is attained when the marginal penalty of adjusting each observation is equal to their marginal contribution to the constraint, which is given by the influence function. The first-order approximation comes from the fact that the influence of each observation is assumed to be constant.

The second solution of Lesage (2009) involves the introduction of a *nuisance parameter.* For the top $(1 - \alpha) \times 100\%$ share, the nuisance parameter is the true value of the α-quantile of income. Given that value, one can apply standard calibration methods to impose the proper number of people and their proper amount of income on both sides of the quantile. The advantage is that this leads to the constraint being exactly satisfied. But for that method to give acceptable results, we need a good guess for the value of the nuisance parameter. Lesage (2009) suggests using its value in the original survey.

We obtained the best results by combining both methods. In the first step, we use the influence function method. This performs the majority of the required adjustment, but still leaves a small discrepancy between the survey and the tax data. In the second step, we get rid of the remaining discrepancy by applying the second approach, with the nuisance parameter estimated in the survey corrected through the first step.

Statistically, survey calibration can be interpreted as the estimation of a non-response function, in which non-response depends on the variables introduced in the constraints. In that interpretation, we are assuming that nonresponse has the same shape as the influence function for top shares. This shape is that of a continuous, piecewise linear function with a kink at the threshold corresponding to the top share. It is almost flat below that threshold, meaning that the bottom 90% of the distribution is virtually unchanged. Above the threshold, nonresponse increases linearly with income — though we can capture non-linearity of nonresponse at the top by including several top income groups in the calibration, for example top 10%, 5% and 1%. That shape is what we expect if the richest households refuse to answer surveys at a higher rate, and also corresponds to the share of the nonresponse that we observe with access to richer data (Blanchet, Flores, and Morgan, 2018). Because the nonresponse function is continuous, our correction method preserves the continuity of the density function of income.

The average estimated nonresponse profile over all the survey and tax data is mostly flat for most of the distribution, meaning that survey distribution is mostly preserved. But observations in the top 0.1% are underrepresented by a factor of 3 on average. We may also notice certain regularities: nonresponse is higher at the top when there is more inequality in the survey. This is the result of having more wealthy households that are less likely to answer surveys, a fact partially captured by the level of inequality before correction. Given that high-inequality countries have experienced more nonresponse, surveys have a tendency not just to underestimate inequality, but to compress them in cross-country comparisons.

When we do not directly observe tax data in a country, we still perform a correction based on the profile of nonresponse that we observe in other countries. To capture statistical regularities such as the one describe above, we estimate the nonresponse profile as a function of the distribution of income in the uncorrected survey using the same machine learning algorithm as in section C.1.1. We stress that this remains a rough approximation and that in our view the proper estimation of top income inequality requires access to tax data. Fortunately, our tax data covers a large majority of the European population and an even larger majority of European income, so that the impact of these corrections on our results remain limited.

C.1.3 Extreme Value Theory to Correct for Sampling Error

The sample size of surveys varies a lot and can sometimes be quite low: this, in itself, can seriously affect estimates of inequality at the top and, in general, will underestimate it (Taleb and Douady, 2015). Correcting sampling error requires some sort of statistical modeling. We chose to use methods coming from extreme value theory, which is routinely used in actuarial sciences to estimate the probability of occurrence of very rare events, but can similarly be used to estimate the distribution of income at the very top.

The main tenet of extreme value theory can be understood in analogy to the central limit theorem. According to the central limit theorem, under some regularity assumptions, but regardless of the exact distribution of iid. variables $X_1, \dots, X_n$, the distribution of the sum $\sum_{i=1}^n X_i$ as n goes to infinity will belong to a tightly parametrized family of distributions (a Gaussian one). Similarly, under mild regularity assumptions, the distribution of the largest value of the sample $\max(X_1, \dots, X_n)$ as n goes to infinity will belong to a certain parametric family. The same holds for the second-largest value, the third-largest value, and so on. As a result, the top k largest values will approximately follow a distribution known as the generalized Pareto distribution, which has the cumulative distribution function:

$$F(x) = 1 - \left\{ 1 + \xi \left(\frac{x - \mu}{\sigma} \right) \right\}^{-1/\xi}$$

That result is known as the Pickands–Balkema–de Haan theorem (e.g. Ferreira and Haan, 2006). The generalized Pareto distribution therefore more or less provides a universal approximation of the distribution of the tails of distributions. It includes the Pareto or the exponential distribution as a special case. We use it to model the top 10% of income distributions. Because the likelihood surface of the generalized Pareto distribution is very flat, maximum likelihood estimation often gives poor results unless the sample size is very large. The standard method of moments also fails if the distribution has infinite variance, which can often occur with income distributions. We use a simple and robust alternative known as probability-weighted moments (Hosking and Wallis, 1987). For X following a generalized Pareto distribution, define $a = \mathbb{E}[X]$ and $b = \mathbb{E}[X(1 - F(x))]$. Then we have $\xi = (a - 4b)/(a - 2b)$ and $\sigma = 2ab/(a - 2b)$, while μ is determined a priori from the threshold from which we start to use the model. We obtain the complete distribution by combining the empirical distribution for the bottom 90% with the generalized Pareto model for the top 10%.

C.1.4 Income Distribution by State in the US

To compare the geography of inequality in Europe with that of the United States, we compute estimates of the income distribution by state in the United States since 1980 by combining survey, tax data and national accounts.

Frank et al. (2015) provide estimates of top taxable income shares by US state. These estimates use an income concept and a unit of analysis which is different from DINA studies such as Piketty, Saez, and Zucman (2018). There are in fact more comparable with the older estimates from Piketty and Saez (2003). The data for producing proper DINA estimates by state in the US is still lacking at this point, so we proceed using the following methodology.

We attribute national income to each state based on their share of GDP (the only national account aggregate available at the state level). To that end, we use data on total state domestic products from the Bureau of Economic Analysis, along with state adult populations series from the United States Census Bureau. (State domestic products provided by the Bureau of Economic Analysis go back as far as 1967. For the historical part, we extrapolate these series back to 1929 by using the growth rates in personal income per capita available from Barro and Sala-i-Martin (1992).)

For the distribution of national income, we proceed as follows. For every g-percenile within the top 10% of the distribution (every percentile from 90% to 99%, and narrower brackets above until the top 99.999%), we compute a correction coefficient between average taxable income by tax unit and the average pre-tax national income per equal-split adult at the national level, using the data of Piketty, Saez, and Zucman (2018). We then apply those coeffcients to the distributions of Frank et al. (2015) to get top income shares of pre-tax national income by state. We then combine these top income shares (for the top 10%) with the distribution of pre-tax income by state that we obtain from the CPS. We do so by stitching together the tax data Lorenz curve at the top with the survey Lorenz curve at the bottom.

We stress that this methodology is approximate. The income concept from the CPS that we use is somewhat different from pre-tax income in the DINA sense. The correction that we apply to top fiscal income shares by Frank et al. (2015) is the same even though in reality it would be different from state to state. The production of actual DINA estimates by state is outside the scope of this paper. For the purpose of this paper — a simple decomposition of within vs. between state inequality — we view them as sufficient, at least to get proper orders of magnitude. We aggregating our state-level estimates, we reproduce the national trends well. We exaggerate the national top 1% income share, by 3.5 pp. on average and also exagerate the bottom 50% income share by 3.8 pp.

C.2 Methodology and Alternative Assumptions

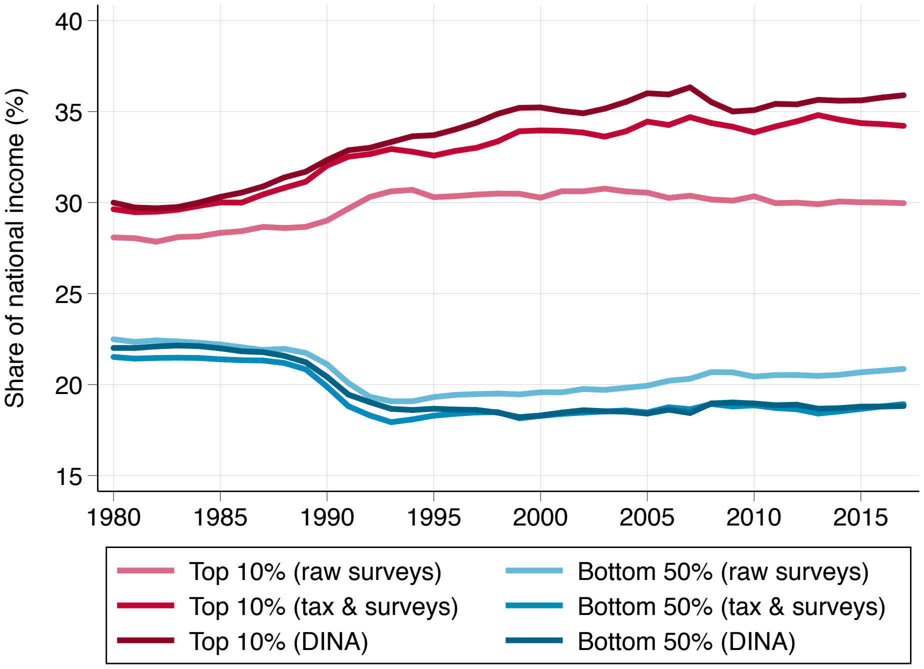


Figure C.1: Role of the Various Steps:
From Raw Surveys to Corrected Surveys to DINA Income

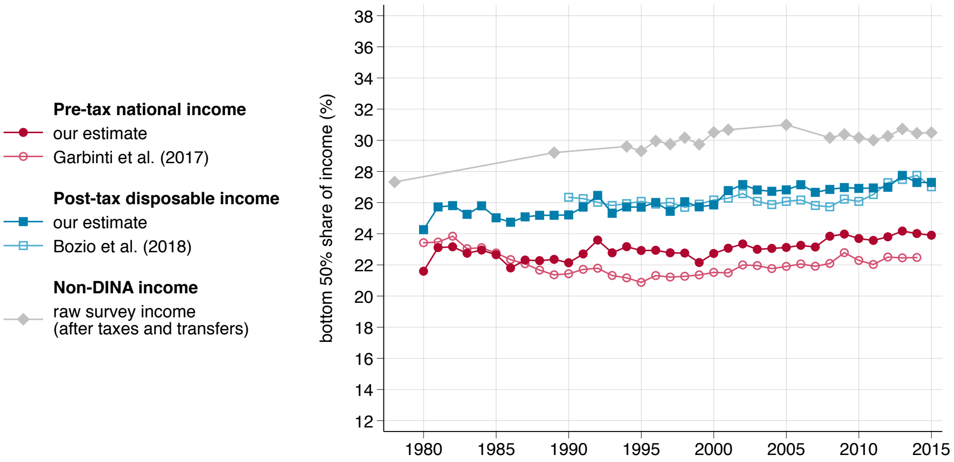


Figure C.2: Bottom 50% Income Shares in France: Comparison

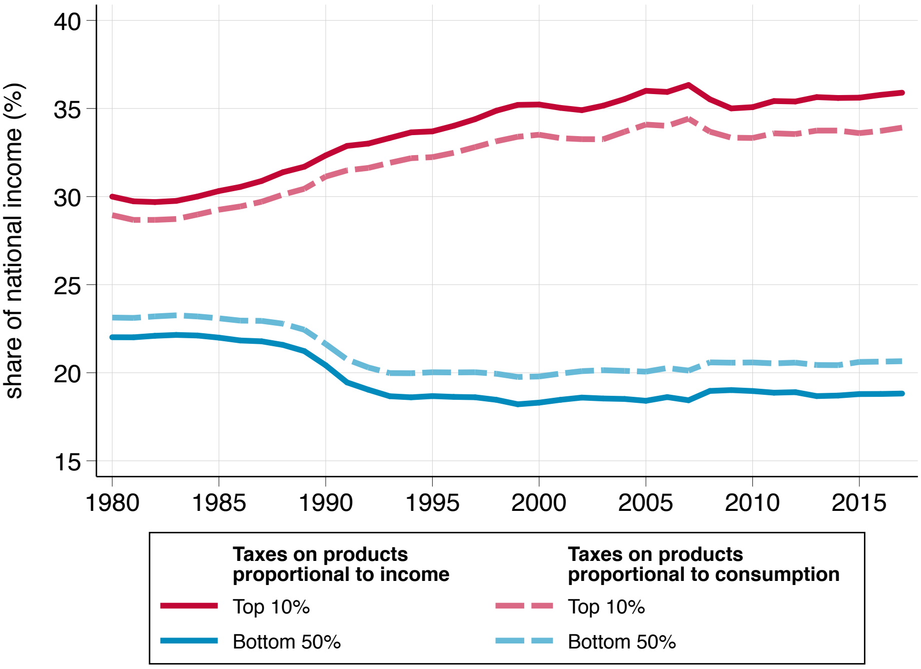
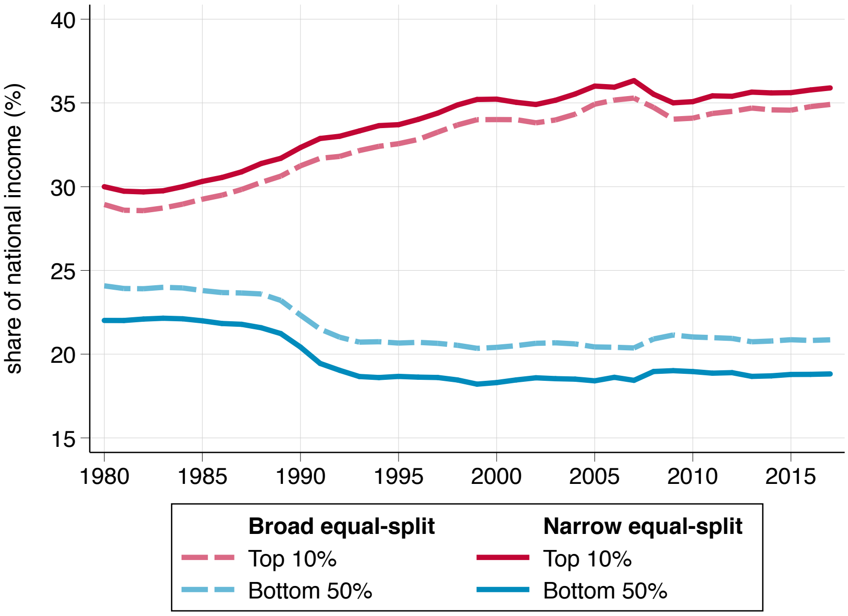
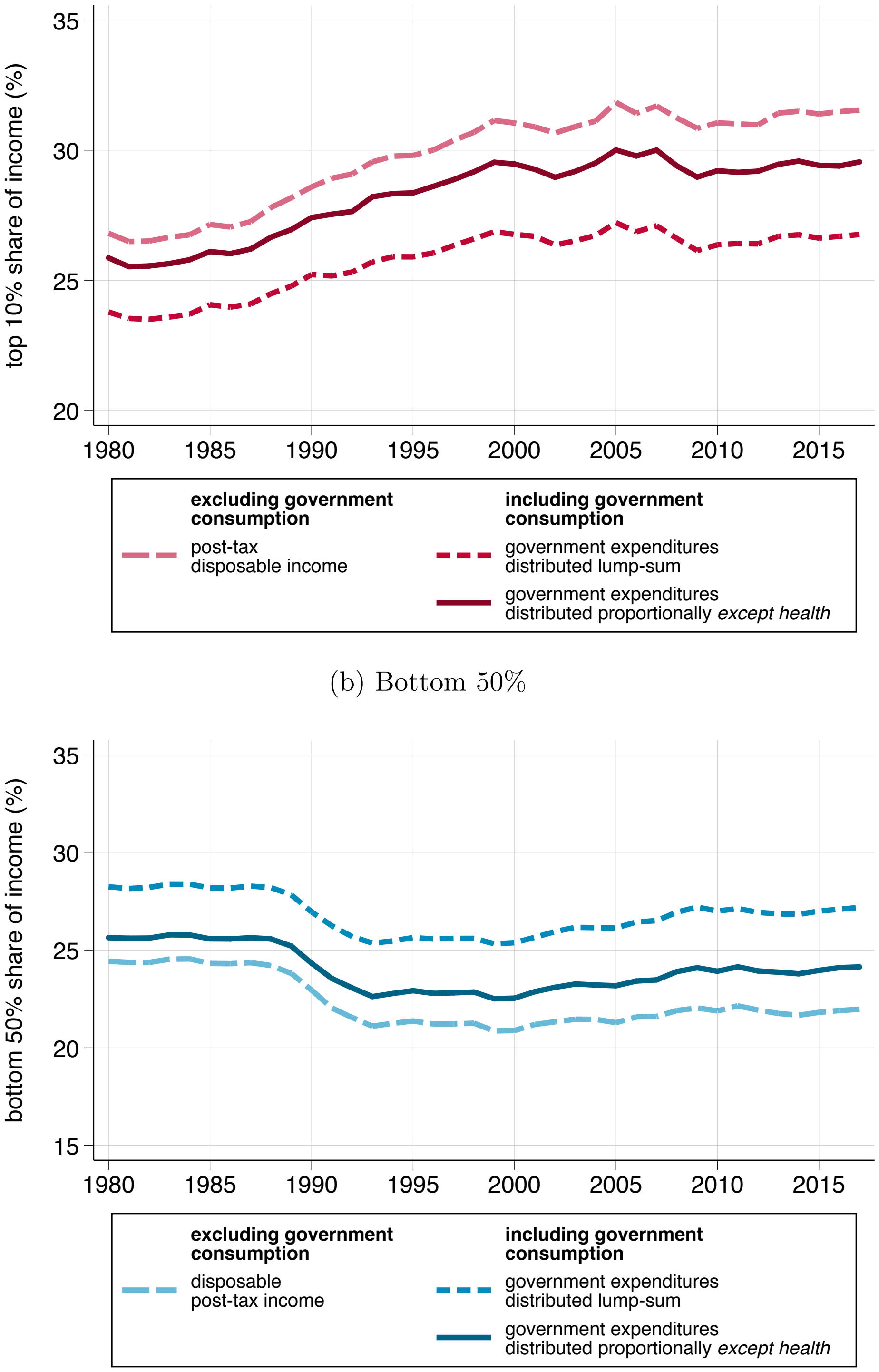


Figure C.3: Distribution of Production Taxes



Narrow equals-split (our benchmark) only split income within couples. Broad equals-split split incomes within entire households.

Figure C.4: Narrow vs. Broad Equal-split



(a) Top 10%

(b) Bottom 50%

Figure C.5: Distribution of Government Expenditures

C.3 Coverage of Data Sources

Table C.2: Coverage of data sources

Country	Surveys	Tax data	Undistrib. prof.	Imp. rents	Tax data source	Quality score
Western Europe						
Austria	1987-2016		1995-2017	1995-2017	Authors	Medium
Belgium	1985-2016	1990-2016	1994-2017	1985-2017	Decoster, Dobbeleer, and Maes (2017)	High
France	1989-2015	1980-2014	1995-2017	1980-2017	Garbinti, Goupille-Lebret, and Piketty (2018)	Very high
Germany	1981-2016	1980-2013	1991-2017	1991-2017	Bartels (2017)	High
Ireland	1980-2016	1980-2015	2001-2017	1995-2017	Jäntti et al. (2007)	High
Luxembourg	1985-2016	2011	1999-2016	1995-2017	Authors	High
Netherlands	1983-2016	1981-2012	1990-2017		Salverda and A. B. Atkinson (2007)	High
Switzerland	1982-2016	1981-2014			Foellmi and Martínez (2017)	High
United Kingdom	1986-2016	1981-2014	1990-2017	1990-2017	A. B. Atkinson (2007)	High
Northern Europe						
Denmark	1981-2016	1980-2010	1994-2017	1990-2017	A. Atkinson and Søgaard (2013)	High
Finland	1981-2016	1980-2009	1995-2017	1980-2017	Jäntti et al. (2010)	High
Iceland	2004-2015	1990-2016		2005-2014	Authors	High
Norway	1986-2016	1981-2011	1995-2017	1980-2017	Aaberge and A. B. Atkinson (2010)	High
Sweden	1981-2016	1980-2013	1995-2017	1980-2017	Roine and Waldenström (2010)	High
Southern Europe						
Cyprus	1990-2016			1995-2017		Medium Low
Greece	1995-2016	2004-2011	1995-2016	1995-2016	Chrissis and Koutentakis (2017)	High
Italy	1981-2016	1980-2009	1995-2017	1980-2017	Alvaredo and Pisano (2010)	High
Malta	2007-2016			2000-2017		Medium Low
Portugal	1980-2016	1980-2005	1995-2017	1995-2017	Alvaredo (2009)	High
Spain	1980-2016	1981-2012	1995-2017	1995-2017	Alvaredo and Saez (2010)	High
Eastern Europe						
Albania	1996-2012					Low
Bosn. & Herz.	1983-2011					Medium Low
Bulgaria	1980-2016					Medium
Croatia	1983-2016	1983-2013		2002-2012	Kump and Novokmet (2018)	High
Czech Republic	1980-2016	1980-2015	1993-2017	1993-2017	Novokmet, Piketty, and Zucman (2018)	High
East Germany		1980-1988			Authors	Medium High
Estonia	1988-2016		1994-2017			High
Hungary	1982-2016	1980-2008	1995-2017	1995-2017	Mavridis and Mosberger (2017)	High
Kosovo	2003-2013					Medium Low
Latvia	1988-2016		1994-2017	1995-2017		Medium
Lithuania	1988-2016		1995-2017	1995-2017		Medium
Macedonia	1983-2014					Medium Low
Moldova	1993-2015					Low
Montenegro	1983-2014					Medium Low
Poland	1983-2016	1983-2015	1995-2016	1995-2016	Bukowski and Novokmet (2017)	High
Romania	1989-2016			2004-2013		Medium
Serbia	1983-2016			1997-2011		Medium
Slovakia	1980-2016		1995-2017	1995-2017		Medium
Slovenia	1987-2016	1991-2012	1995-2017	1995-2017	Kump and Novokmet (2018)	High

C.4 Average Incomes in Europe

Table C.3: Average national incomes in Europe, 1980-2017

	Average national income per adult					% of European average income				
	1980	1990	2000	2007	2017	1980	1990	2000	2007	2017
European regions										
Europe	21160	24120	27600	30910	32130	100	100	100	100	100
EU-15 (West)	24010	27810	31930	34950	35250	113	116	116	113	110
EU-13 (East)	12940	13230	13440	17720	21690	61	55	49	57	68
Other West	35050	40960	48610	51900	51700	165	170	177	168	161
Other East	9100	8110	6460	8700	10080	43	34	23	28	31
Eastern Europe										
Albania	6630	5720	6670	9340	11000	31	24	24	30	34
Bosnia and Herzegovina	2540	2070	7750	9110	10630	12	9	28	30	33
Bulgaria	8450	10440	8760	12500	17080	40	43	32	41	53
Croatia	18370	16190	13680	18910	19070	87	67	50	61	59
Czech Republic	17660	20280	17100	21980	24260	83	84	62	71	76
Estonia	13130	15120	14320	23010	24700	62	63	52	75	77
Hungary	14200	15070	14710	18690	20890	67	63	53	61	65
Latvia	13180	15280	8920	17350	19510	62	64	32	56	61
Lithuania	14760	15560	11380	20390	24620	70	65	41	66	77
Macedonia	11160	9630	8680	9110	11140	53	40	32	30	35
Moldova	6010	6530	2350	3570	4880	28	27	9	12	15
Montenegro	19710	15160	10720	13560	15750	93	63	39	44	49
Poland	11550	10480	14630	17200	22510	55	44	53	56	70
Romania	11400	11920	10120	14830	19370	54	50	37	48	60
Serbia	12690	11520	6520	9950	11290	60	48	24	32	35
Slovakia	14180	15260	13710	20140	23980	67	63	50	65	75
Slovenia	22150	18020	19840	25170	24910	105	75	72	82	78
Western Europe										
Austria	25400	29640	34700	38960	38930	120	123	126	126	121
Belgium	24850	29130	34380	37010	37610	117	121	125	120	117
France	24690	28480	32980	34930	35130	117	118	120	113	110
Germany	26740	29820	32520	35920	39420	126	124	118	116	123
Ireland	15590	20730	37870	42740	43960	74	86	138	139	137
Luxembourg	31040	54900	75660	89090	60010	146	228	275	289	187
Netherlands	32030	31590	39890	43840	43580	151	131	145	142	136
Switzerland	36070	42640	44940	45220	45530	170	177	163	147	142
United Kingdom	21070	25850	32300	37010	37490	99	108	117	120	117
Cyprus	16950	26320	30890	37000	31270	80	110	112	120	98
Greece	21690	22180	24610	30110	20670	102	92	89	98	64
Italy	25280	28660	31820	32950	29450	119	119	116	107	92
Malta	14300	18310	23680	25660	33050	67	76	86	83	103
Portugal	14370	18670	22670	23070	23010	68	78	82	75	72
Spain	18770	23300	27230	29340	30230	89	97	99	95	94
Northern Europe										
Denmark	25740	29010	36040	41430	42410	121	121	131	134	132
Finland	20970	25420	31410	37760	35240	99	106	114	122	110
Iceland	27510	30430	35330	42800	45740	130	127	128	139	143
Norway	33810	38800	55480	63880	62510	160	161	202	207	195
Sweden	23470	27670	33860	41530	45880	111	115	123	135	143

Source: authors' computations. Serbia includes Kosovo. *Interpretation*: in 1980, Albania's average national income per adult was 31% of the European average (69% lower).

Bibliography

Aaberge, R. and A. B. Atkinson (2010). “Top incomes in Norway”. In: *Top incomes: a global perspective.* Ed. by A. B. Atkinson and Thomas Piketty. Oxford University Press. Chap. 9, pp. 448–481.

Alvaredo, Facundo (2009). “Top incomes and earnings in Portugal 1936-2005”. In: *Explorations in Economic History* 46.1, pp. 404–417.

Alvaredo, Facundo and Elena Pisano (2010). “Top incomes in Italy, 1974-2004”. In: *Top incomes: a global perspective.* Ed. by A. B. Atkinson and Thomas Piketty. Oxford University Press. Chap. 12, pp. 625–663.

Alvaredo, Facundo and Emmanuel Saez (2010). “Income and wealth concentration in Spain in a historical and fiscal perspective”. In: *Top incomes: a global perspective.* Ed. by A. B. Atkinson and Thomas Piketty. Oxford University Press. Chap. 10, pp. 482–559.

Atkinson, A. B. (2007). “Measuring Top Incomes: Methodological Issues”. In: *Top Incomes over the Twentieth Century: A Contrast Between Continental European and English-Speaking Countries.*

Atkinson, A.B. and J.E. Søgaard (2013). “The long-run history of income inequality in Denmark: Top incomes from 1870 to 2010”. In: *EPRU Working Paper Series 2013-01.*

Barro, Robert and Xavier Sala-i-Martin (1992). “Convergence”. In: *Journal of Political Economy* 100.2, pp. 223–251.

Bartels, Charlotte (2017). “Top incomes in Germany, 1871-2013”. In: *WID.world Working Paper Series 2017/18.*

Blanchet, Thomas, Ignacio Flores, and Marc Morgan (2018). “The Weight of the Rich: Improving Surveys Using Tax Data”.

Bukowski, Pawel and Filip Novokmet (2017). “Top incomes during wars, communism and capitalism: Poland 1892-2015”. In: *WID.world Working Paper Series 2017/22.*

Chen, Tianqi and Carlos Guestrin (2016). “XGBoost: A Scalable Tree Boosting System”. URL: `http://dx.doi.org/10.1145/2939672.2939785`.

Chrissis, Kostas and Franciscos Koutentakis (2017). “From dictatorship to crisis: the evolution of top income shares in Greece (1967-2013)”. In: *Working Paper.*

Decoster, André, Koen Dobbeleer, and Sebastiaan Maes (2017). “Using fiscal data to estimate the evolution of top income shares in Belgium from 1990 to 2013”. In: *Ku Leuven Discussion Paper Series DPS17.18.*

Deville, Jean-Claude and Carl-Erik Särndal (1992). “Calibration Estimators in Survey Sampling”. In: *Journal of the American Statistical Association* 87.418, pp. 376–382.

Fang, Yixin (2011). “Asymptotic Equivalence between Cross-Validations and Akaike Information Criteria in Mixed-Effects Models”. In: *Journal of Data Science* 9, pp. 15–21.

Ferreira, Ana and Laurens de Haan (2006). *Extreme Value Theory: An Introduction.* Springer Series in Operations Research. Springer.

Foellmi, Reto and Isabel Z. Martínez (2017). “Volatile top income shares in Switzerland? Reassessing the evolution between 1981 and 2010”. In: *Review of Economics and Statistics* 99.5, pp. 793–809.

Frank, Mark et al. (2015). “Frank-Sommeiller-Price Series for Top Income Shares by US States since 1917”. In: *WID.world Technical Note Series 2015/7.*

Friedman, Jerome H. (2001). “Greedy Function Approximation: A Gradient Boosting Machine”. In: *The Annals of Statistics* 29.5, pp. 1189–1232. URL: `http://www.jstor.org/stable/2699986`.

Garbinti, B, J Goupille-Lebret, and Thomas Piketty (2018). “Income inequality in France, 1900–2014: Evidence from Distributional National Accounts (DINA)”. In: *Journal of Public Economics* 162.1, pp. 63–77.

Hosking, J R M and J R Wallis (1987). “Parameter and Quantile Estimation for the Generalized Pareto Distribution”. In: *Technometrics* 29.3, pp. 339–349. URL: `https://www.tandfonline.com/doi/abs/10.1080/00401706.1987.10488243`.

Jäntti, M. et al. (2007). “Long-term trends in top income shares in Ireland”. In: *Top incomes over the 20th century.* Ed. by A. B. Atkinson and Thomas Piketty. Oxford University Press. Chap. 12, pp. 501–530.

– (2010). “Trends in top income shares in Finland”. In: *Top incomes: a global perspective.* Ed. by A.B. Atkinson and Thomas Piketty. Oxford University Press. Chap. 8, pp. 371–447.

Kump, Nataša and Filip Novokmet (2018). “Top incomes in Croatia and Slovenia, from 1960s until today”. In: *WID.world Working Paper Series 2018/8.*

Lakner, Christoph and Branko Milanovic (2016). “Global Income Distribution: From the Fall of the Berlin Wall to the Great Recession”. In: *The World Bank Economic Review* 30.2, pp. 203–232. URL: `https://academic.oup.com/wber/article-lookup/doi/10.1093/wber/lhv039`.

Lesage, Éric (2009). “Calage non linéaire”.

Mavridis, Dimitris and Pálma Mosberger (2017). “Income inequality and incentives: the quasi-natural experiment of Hungary, 1914-2008”. In: *WID.world Working Paper Series 2017/17.*

Nielsen, Didrik (2016). *Tree Boosting With XGBoost Why Does XGBoost Win "Every" Machine Learning Competition?* Tech. rep. December, p. 2016. URL:

https://brage.bibsys.no/xmlui/bitstream/handle/11250/2433761/16128_FULLTEXT.pdf?sequence=1&isAllowed=y.

Novokmet, Filip, Thomas Piketty, and Gabriel Zucman (2018). "From Soviets to oligarchs: inequality and property in Russia 1905-2016". In: *The Journal of Economic Inequality* 16.2, pp. 189–223.

Piketty, Thomas and Emmanuel Saez (2003). "Income Inequality in the United States, 1913–1998". In: *Quarterly Journal of Economics* CXVIII.1.

Piketty, Thomas, Emmanuel Saez, and Gabriel Zucman (2018). "Distributional National Accounts: Methods and Estimates for the United States". In: *Quarterly Journal of Economics* 133.May, pp. 553–609.

Roine, Jesper and Daniel Waldenström (2010). "Top incomes in Sweden over the twentieth century". In: *Top incomes: a global perspective*. Ed. by A. B. Atkinson and Thomas Piketty. Oxford University Press. Chap. 7, pp. 299–370.

Salverda, W. and A. B. Atkinson (2007). "Top incomes in the Netherlands over the twentieth century". In: *Top incomes over the 20th century*. Ed. by A. B. Atkinson and Thomas Piketty. Oxford University Press. Chap. 10, pp. 426–471.

Taleb, Nassim Nicholas and Raphael Douady (2015). "On the super-additivity and estimation biases of quantile contributions". In: *Physica A: Statistical Mechanics and its Applications* 429, pp. 252–260. URL: http://dx.doi.org/10.1016/j.physa.2015.02.038.

Appendix D

Appendix to “Modeling the Dynamics of Wealth Inequality in the United States, 1962–2100”

D.1 Comparison to Synthetic Saving Rates

Saez and Zucman (2016) introduced a different method to decompose the dynamics of inequality. Their approach was extended and used by several authors (Garbinti, Goupille-Lebret, and Piketty, 2017; Berman, Ben-Jacob, and Shapira, 2016) to perform decomposition and simulations similar to section 4.5 of this paper.

The setting is more or less the same, in that they observe cross-sections of the joint distribution of income and wealth at different points in time, and use this information to analyze the dynamics of wealth. The difference of their approach is that they do not seek to estimate structural parameters that relate to individual behavior. Instead, they construct “synthetic saving rates” that relate the amount of income that accrue to the various percentiles of the wealth distribution to their evolution over time.

Assume zero growth ($g_t = 0$) for simplicity. Synthetic saving rates are defined as follows. Take n brackets of the wealth distribution. Describe the evolution of wealth for each bracket as:

$$\begin{cases} W_{t+1}^{(1)} & = & W_t^{(1)} + r_t^{(1)} W_t^{(1)} + Z_t^{(1)} - C_t^{(1)} \\ \vdots & & \\ W_{t+1}^{(\mathrm{n})} & = & W_t^{(\mathrm{n})} + r_t^{(\mathrm{n})} W_t^{(\mathrm{n})} + Z_t^{(\mathrm{n})} - C_t^{(\mathrm{n})} \end{cases}$$

where $Z_t^{(\mathrm{i})}$, $r_t^{(\mathrm{i})}$ and $C_t^{(\mathrm{i})}$ refers to labor income, rate of return and consumption for bracket (i). Define income $Y_t^{(\mathrm{i})} \equiv Z_t^{(\mathrm{i})} + r_t^{(\mathrm{i})} W_t^{(\mathrm{i})}$. All variables but $C_t^{(\mathrm{i})}$ can be observed in the data, so $C_t^{(\mathrm{i})}$ is estimated as a residual. The ratio $s_t^{(\mathrm{i})} = 1 - C_t^{(\mathrm{i})}/Y_t^{(\mathrm{i})}$ is the synthetic saving rate of bracket (i).

Using various assumptions on how $s_t^{(\mathrm{i})}$ relates to the average income or wealth of the bracket (i), we can perform forecasts and counterfactuals on the evolution of the wealth distribution. Note, however, that the synthetic saving rate is not the average saving rate of the corresponding bracket: it can only be interpreted as such under stringent assumptions (no mobility between brackets and homogeneous behavior within brackets). The synthetic saving rate is a reduced-form parameter that captures mobility, the inequality of savings, their correlation with wealth, demography, intergenerational wealth mobility, etc.

The framework of this paper can shed some light on the underlying mechanics of the synthetic saving rates method. We can use it to explicitly show what synthetic savings rates capture, and how they are capturing it. In doing so I will clarify how it compres the work in this paper.

Consider the following continuous time formulation of the synthetic saving rates. For all t, and for all $0 \leq p < 1$, define the p-th wealth fractile $W_t(p)$. Also write $Z_t(p) = \mathbb{E}[Z_t | W_t = W_t(p)]$ and $r_t(p) = \mathbb{E}[r_t | W_t = W_t(p)]$. I will write:

$$\frac{\partial}{\partial t} W_t(p) = r_t(p) W_t(p) + Z_t(p) - C_t(p) = Y_t(p) - C_t(p) \tag{D.1}$$

which defines "synthetic consumption" $C_t(p)$. That specification differs from that of this paper in two respects (if we set aside the role of inheritance and demographics). The first issue is the lack of explicit randomness — and, therefore, mobility. If we differentiate equation (D.1) with respect to p and use the change of variable $p = F_t(w)$, we end up with a special case of the Fokker-Planck equation in which the diffusion term ($\sigma^2(w)$) is equal to zero. Therefore, the synthetic saving rates approach is analogous to a stochastic differential equation with only the drift term: we can view it as a specific, somewhat degenerate case of the more general model used in this paper.

The second issue is the formulation of the saving rate: do we consider savings out of income or out of wealth? Are they a function of the level of wealth, or a function of the rank in the wealth distribution?

Formulation of the Saving Rate There are several justifiable ways of expressing saving rates. Traditionally, the literature writes $C_t(p) = (1 - s(p))Y_t(p)$, so that saving out of income is a function of the rank in the wealth distribution. We could also write $C_t(p) = (1 - s[W_t(p)])Y_t(p)$ to make savings a function of the wealth level rather than the wealth rank. We could also consider savings out of wealth instead of income: $C_t(p) = (1 - s(p))W_t(p)$ or $C_t(p) = (1 - s[W_t(p)])W_t(p)$. That last specification is closest to the one adopted in this paper.

Which formulation to choose depends in part on the intention behind the model. For descriptive purposes, it does not matter. All of them describe the same reality in a different way. Things are different when it comes to running forecasts or counterfactuals. Such an exercise requires setting certain parameters constant, so that the way we parameterize the problem has an impact on the outcome. Utimately, the dynamic of wealth depends entirely on the difference between income and consumption at various points of the wealth distribution: nothing would change if we were to increase everybody's income and consumption by the same amount.

To fix ideas, assume that people at the top of the wealth distribution earn a return of 10% on their wealth, and save 50% of their income. This is identical to saying that they consume 5% of their wealth. Now, increase their income by \$100. Assuming a constant saving rate out of income means that saving increase by \$50. However, assuming a constant saving rate out of wealth means that saving increases by the total amount, i.e. \$100. Therefore, an increase in income has a higher impact with the second specification.

Which specification is better? The first concern is that C_t should accurately describe the behavior of agents. All functional forms for C_t considered here are simplified rule of thumbs that only approximate the true saving behavior, yet it remains important to know which is closest to reality. The macroeconomics and household finance literature would suggest a saving rate out wealth that depends on wealth, i.e. $C_t(p) = s[W_t(p)]W_t(p)$. While there is still considerable disagreement regarding the proper model of household saving (e.g. Browning and Lusardi, 1996), models in which consumption depends on "cash-on-hand" (i.e. wealth plus current income) are commonplace. Such a rule can be microfounded using models of precautionary savings (e.g. Weil, 1993) — especially for the top of the distribution — or models with a preference for wealth (e.g. Piketty and Zucman, 2014). Models in which agents always consume a given fraction of their current income is harder are harder to justify theoretically, and rarely seen in the literature (with the atypical exception of “hand-to-mouth” households, that have in effect a saving rate of zero). It is also much more common to assume that behavior depends on the absolute level of wealth, rather than the rank in the wealth distribution.

The second concern is that the choosen parameters remain constant over time, which makes it more likely that they will remain so in the future. As we’ve seen in this paper, we can reproduce the evolution of the wealth distribution since 1962 by assuming constant parameters for consumption out of wealth. However, the average income/wealth ratio at the top has changed between 1962–1980 and 1981–2014. Therefore, the saving rate out of income has changed, even though the saving rate out of wealth has remained stable. This also renders the latter specification preferable.

Mobility The synthetic saving rate parameter is meant to capture both average savings and mobility. We can use the stochastic model of this paper to clarify what that entails. Assume that the dynamic of wealth at the top follows:

$$\mathrm{d}w_{it} = [z(w_{it}) + r(w_{it})w_{it} - \mu(w)w_{it}]\,\mathrm{d}t + \sigma w_{it}\,\mathrm{d}B_{it} \tag{D.2}$$

using the notations of the paper (in particular, $\mu(w)$ is the average consumption out of wealth, σ^2 is its variance, and we ignore income-induced diffusion $\tau^2(w)$ for simplicity.) Assume that, at time t and for high w, wealth follows a Pareto distribution with coefficient α (i.e. $f_t(w) \propto w^{-\alpha-1}$). The Fokker-Planck equation combine the effect of the drift and the diffusion:

$$\frac{\partial}{\partial t} f_t(w) = \underbrace{-\frac{\partial}{\partial w}\left[(z(w) + w(r(w) - \mu(w)))w^{-\alpha-1}\right]}_{\text{drift}} + \underbrace{\frac{1}{2}\frac{\partial^2}{\partial w^2}\left[\sigma^2 w^2 w^{-\alpha-1}\right]}_{\text{diffusion}}$$

Imagine that, following the synthetic saving rates approach, we estimate a “synthetic” value of consumption $\mu(w)$, noted $\mu^*(w)$, by only taking the drift into account. When observing the evolution of wealth, we still see the role that mobility plays, but we will attribute it to the drift. Thus, rewrite the diffusion term as:

$$\frac{1}{2}\frac{\partial^2}{\partial w^2}\left[\sigma^2 w^2 w^{-\alpha-1}\right] = -\frac{1}{2}\sigma^2(\alpha - 1)\frac{\partial}{\partial w}\left[w w^{-\alpha-1}\right]$$

That way, we can include the diffusion term into the drift term such that:

$$\frac{\partial}{\partial t} f_t(w) = -\frac{\partial}{\partial w}\left[\left(y(w) + w\left(r(w) - \mu(w) + \frac{1}{2}\sigma^2(\alpha - 1)\right)\right) f_t(w)\right]$$

Hence, the synthetic consumption $\mu(w)$ that we estimate is $\mu^*(w) = \mu(w) - \sigma^2(\alpha - 1)/2$. It differs from the true parameter $\mu(w)$ in two respects. First, mobility makes synthetic consumption lower (and therefore makes savings higher) than true average consumption. Second, and perhaps more problematically, this difference depends on the shape of the wealth distribution itself. For example, assume, as found in this paper, that $\sigma^2 \approx 0.08$. Assume that we move from the level of wealth inequality of the 1970s ($(\alpha - 1)/2 \approx 0.5$) to today's level ($(\alpha - 1)/2 \approx 0.2$). In the 1970s, the synthetic consumption out of wealth $\mu^*(w)$ will be $0.08 \times 0.5 = 4\%$ lower than true consumption, while today it would be only $0.08 \times 0.2 = 1.6\%$ lower. That correspond to a 2.4% increase in synthetic consumption out of wealth, despite no change in the underlying wealth accumulation process. Assuming an income/wealth ratio of 10% at the top, that represents a spurious change of 24 pp. to the synthetic saving rate out of income.

Steady-State The lack of an explicit diffusion mechanism (i.e. mobility) in the synthetic saving rates approach also has an impact on whether, why and how a steady-state distribution can emerge.

In the stochastic model used in this paper, a steady-state power-law distribution arises naturally from scale invariance at the top. That mechanism does not apply in the absence of diffusion. To fix ideas, assume:

$$\frac{\partial}{\partial t} W_t(p) = r(p)W_t(p) + Z(p) - (1 - s(p))W_t(p)$$

The steady state, if any, is given by setting $\frac{\partial}{\partial t} W_\infty(p) = 0$, so that $W_\infty(p) = Z(p)/(1 - s(p) - r(p))$. Assuming scale invariance (i.e. $s(p) \equiv s$ and $r(p) \equiv r$) and constant labor income ($Z(p) \equiv Z$), the distribution of wealth can either diverge or collapse onto the single value $Z/(1 - s - r)$. To retrieve a smooth steady-state distribution, it is crucial that at least of one of s, r or Z be a smooth function of the rank in the wealth distribution, and not just the level of wealth. Even then, whether a power-law emerges from this type of model will be a direct consequence of the shape of $s(p)$, $r(p)$ or $Z(p)$, not something that the approach explains on its own.

The existence of a non-degenerate steady state requires $1 - s - r > 0$ and $Z > 0$. So it is not possible in this model to have a stationary state in which people at the top of the wealth distribution have no labor income (i.e. a strict separation between workers and capitalists). Wealth at the top is directly proportional to the labor income earned by the same group. This stands in contrast to the steady-state requirements of the stochastic model, which are more general (especially once we introduce demographics as an additional stabilizing force, see Gabaix (2009)). In particular, with a stochastic model it is possible to sustain a stationary steady-state even if labor income plays no role at the top.

D.2 Omitted proofs

D.2.1 Application of Gyöngy's (1986) Theorem

Recall that wealth at the individual level follows the SDE:

$$\mathrm{d}w_{it} = [y_{it} - c_{it}]\,\mathrm{d}t + [\tau_{it}^2 + \sigma_{it}^2]^{1/2}\,\mathrm{d}B_{it}$$

Consider a small time interval $[t, t+\mathrm{d}t]$. Over that interval, the income process y_{it} has mean $\nu_{it}\,\mathrm{d}t$ and variance $\tau_{it}^2\,\mathrm{d}t$, while the consumption process c_{it} has mean $\mu_{it}\,\mathrm{d}t$ and variance $\sigma_{it}^2\,\mathrm{d}t$. Following Gyöngy's (1986) theorem, we can write:

$$\mathrm{d}w_{it} = [\nu_t(w_{it}) - \mu_t(w_{it})]\,\mathrm{d}t + [\tau_t^2(w_{it}) + \sigma_t^2(w_{it})]^{1/2}\,\mathrm{d}B_{it}$$

where:

$$\begin{aligned} \nu_t(w) &= \mathbb{E}[\nu_{it}|w_{it} = w] \qquad & \tau_t^2(w) &= \mathbb{E}[\tau_{it}^2|w_{it} = w] \\ \mu_t(w) &= \mathbb{E}[\mu_{it}|w_{it} = w] \qquad & \sigma_t^2(w) &= \mathbb{E}[\sigma_{it}^2|w_{it} = w] \end{aligned}$$

To simplify notations, consider all expectations conditional on $w_{it} = w$. We can write, somewhat informally, $c_{it} = \mu_{it}\,\mathrm{d}t + \sigma_{it}\,\mathrm{d}B_{it}$. For the drift term, we have directly $\mathbb{E}[c_{it}] = \mathbb{E}[\mu_{it}]\,\mathrm{d}t = \mu_t(w)\,\mathrm{d}t$. For the diffusion term:

$$\begin{aligned} \mathrm{Var}(c_{it}) &= \mathbb{E}[(c_{it} - \mu_t(w)\,\mathrm{d}t)^2] \\ &= \mathbb{E}[(\mu_{it}\,\mathrm{d}t - \mu_t(w)\,\mathrm{d}t + \sigma_{it}\,\mathrm{d}B_{it})^2] \\ &= \underbrace{\mathbb{E}[(\mu_{it}\,\mathrm{d}t - \mu_t(w)\,\mathrm{d}t)^2]}_{=\,0 \text{ because } (\mathrm{d}t)^2 = 0} + \underbrace{\mathbb{E}[\sigma_{it}^2\,\mathrm{d}B_{it}^2]}_{\substack{=\,\mathbb{E}[\sigma_{it}^2]\,\mathrm{d}t \\ \text{because } \mathrm{d}B_{it}^2 = \mathrm{d}t}} + 2\underbrace{\mathbb{E}[\sigma_{it}(\mu_{it} - \mu_t(w))\,\mathrm{d}B_{it}\,\mathrm{d}t]}_{=\,0 \text{ because } \mathrm{d}B_{it}\,\mathrm{d}t = 0} \end{aligned}$$

Therefore, $\mu_t(w)\,\mathrm{d}t = \mathbb{E}[c_{it}]$ and $\sigma_t^2(w)\,\mathrm{d}t = \mathrm{Var}(c_{it})$. Similarly, $\nu_t(w)\,\mathrm{d}t = \mathbb{E}[y_{it}]$ and $\tau_t^2(w)\,\mathrm{d}t = \mathrm{Var}(y_{it})$.

D.2.2 Steady-State Wealth Distribution With a Wealth Tax

Let f_α be the steady-state density of wealth with a tax rate α (assuming it exists). It has to obey the Fokker-Planck equation with the time derivative terms set to zero, i.e.:

$$0 = -\frac{\partial}{\partial w}\left[(a(w_{it}) - \alpha(w_{it} - w_0)_+)f_\alpha(w)\right] + \frac{1}{2}\frac{\partial^2}{\partial w^2}\left[b^2(w)f_\alpha(w)\right]$$

After solving this differential equation, we can write:

$$f_\alpha(w) = C_\alpha \exp\left\{-2\int_{w_0}^{w}\frac{b(s)b'(s) - a(s)}{b^2(s)}\,\mathrm{d}s\right\}\exp\left\{-\frac{2\alpha}{b^2}\int_{w_0}^{w}\frac{(s-w_0)_+}{s^2}\,\mathrm{d}s\right\}$$

where the constant C_α is defined so that the density integrates to one. Note that:

$$f_\alpha(w) = \frac{C_\alpha}{C_0}f_0(w)\exp\left\{-\frac{2\alpha}{b^2}\int_{w_0}^{w}\frac{(s-w_0)_+}{s^2}\,\mathrm{d}s\right\}$$

where f_0 corresponds to the steady-state density of wealth without any wealth tax. For $w < w_0$, this just amounts to $f_\alpha(w) = (C_\alpha/C_0)f_0(w)$. For $w \geq w_0$, we have:

$$\begin{aligned} f_\alpha(w) &= \frac{C_\alpha}{C_0}f_0(w)\exp\left\{-\frac{2\alpha}{b^2}\int_{w_0}^{w}\frac{(s-w_0)}{s^2}\,\mathrm{d}s\right\} \\ &= \frac{C_\alpha}{C_0}f_0(w)\exp\left\{-\frac{2\alpha}{b^2}\left(\frac{w_0}{w} - 1\right)\right\}\left(\frac{w}{w_0}\right)^{-2\alpha/b^2} \end{aligned}$$

Define the function:

$$\zeta_\alpha(w) \equiv \exp\left\{-\frac{2\alpha}{b^2}\left(\frac{w_0}{w} - 1\right)\right\}\left(\frac{w}{w_0}\right)^{-2\alpha/b^2}$$

The steady-state tax base is $T(\alpha) = \int_{w_0}^{+\infty}(w - w_0)f_\alpha(w)\,\mathrm{d}w$, hence:

$$T(\alpha) = \frac{C_\alpha}{C_0}\int_{w_0}^{+\infty}(w - w_0)\zeta_\alpha(w)f_0(w)\,\mathrm{d}w$$

For the constant term, notice that:

$$\left(\frac{C_\alpha}{C_0}\right)^{-1} = F_0(w_0) + \int_{w_0}^{+\infty}\zeta_\alpha(w)f_0(w)\,\mathrm{d}w$$

Bibliography

Berman, Yonatan, Eshel Ben-Jacob, and Yoash Shapira (2016). "The dynamics of wealth inequality and the effect of income distribution". In: *PLoS ONE* 11.4, pp. 8–10.

Browning, Martin and Annamaria Lusardi (1996). "Household Saving: Micro Theories and Micro Facts". In: *Journal of Economic Literature* 34.4, pp. 1797–1855.

Gabaix, Xavier (2009). "Power Laws in Economics and Finance". In: *Annual Review of Economics* 1.1, pp. 255–293.

Garbinti, Bertrand, Jonathan Goupille-Lebret, and Thomas Piketty (2017). "Accounting for Wealth Inequality Dynamics: Methods, Estimates and Simulations for France (1800-2014)".

Gyöngy, I (1986). "Mimicking the one-dimensional marginal distributions of processes having an Ito differential". In: *Probability Theory and Related Fields* 71.4, pp. 501–516. URL: `https://link.springer.com/article/10.1007/BF00699039`.

Piketty, Thomas and Gabriel Zucman (2014). "Capital is Back: Wealth-Income Rations in Rich Countries 1700–2010". In: *Quarterly Journal of Economics* 129.3, pp. 1255–1310.

Saez, Emmanuel and Gabriel Zucman (2016). "Wealth Inequality in the United States since 1913: Evidence from Capitalized Income Tax Data". In: *Quarterly Journal of Economics* 131.May, pp. 519–578.

Weil, Philippe (1993). "Precautionary Savings and the Permanent Income Hypothesis". In: *Review of Economic Studies* 60.2, pp. 367–383. URL: `https://academic.oup.com/restud/article-abstract/60/2/367/1574342`.