UNIVERSITÉ PARIS I - PANTHÉON-SORBONNE
SCIENCES ÉCONOMIQUES
UFR 02 - CES

PhD Thesis

Submitted for the Degree of Doctor of Philosophy in Economics

Defended on
January 25th, 2019 by

Ignacio FLORES

On the Empirical Measurement of Inequality

Thesis Advisor:
Pr. Jérôme GAUTIÉ

Jury

Facundo ALVAREDO:	Professor at PSE
François BOURGUIGNON:	Emeritus Professor at PSE
Emmanuel FLACHAIRE (R):	Professor at Aix-Marseille Univ.
Jérôme GAUTIE	Professor at UP1
Elvire GUILLAUD	Associate Professor at UP1
Julian MESSINA:	Lead Economist at IDB
Nora LUSTIG (R):	Professor at Tulane Univ.

UNIVERSITÉ PARIS I - PANTHÉON-SORBONNE
SCIENCES ÉCONOMIQUES
UFR 02 - CES

Thèse pour l'obtention du grade de docteur en sciences économiques

Soutenue publiquement
le 25 Janvier 2019 par

Ignacio FLORES

De la Mesure Empirique des Inégalités

Directeur de thèse:
Pr. Jérôme GAUTIÉ

Membres du Jury

Facundo ALVAREDO:	Maître de Conférences, PSE
François BOURGUIGNON:	Professeur Émérite, PSE
Emmanuel FLACHAIRE (R):	Professeur, Aix-Marseille Univ.
Jérôme GAUTIE	Professeur, UP1
Elvire GUILLAUD	Maître de Conférences, UP1
Julian MESSINA:	Lead Economist IDB
Nora LUSTIG (R):	Professeure, Tulane Univ.

L'Université Paris 1 Panthéon-Sorbonne n'entend donner aucune approbation ni improbition aux opinions émises dans cette thèse; elles doivent être considérées comme propres à leur auteur.

Acknowledgements

The work that is presented here was possible thanks to many people that have inspired, accompanied and supported me throughout time.

I warmly thank the guidance of my thesis supervisor, Jérôme Gautié, and co-supervisor Elvire Guillaud. Both were extremely supportive, patient and available throughout the whole process. I also thank my co-authors Jorge Atria, Thomas Blanchet, Marc Morgan, Claudia Sanhueza and Ricardo Mayer, with whom I have created special bonds that resulted from the conception of the babies that are our papers. I am immensely thankful to Facundo Alvaredo for having introduced me to the literature of top incomes and wealth and having invited me to participate in the World Inequality Database project at the Paris School of Economics. I would also like to thank François Bourguignon, Emmanuel Flachaire, Julián Messina and Nora Lustig for accepting to be part of my jury and giving remarkable feedback for my research.

I am delighted to have worked at Paris 1 Panthéon-Sorbonne University. There I met curious and fascinating people with whom I enjoyed many literally endless discussions. I especially refer to Neils Boissonnet, Sophie Dessein, Thomas Delcey, Pablo Cardoso, Jean-Éric Hyafill, Francesco Sergi, Quitterie Labbé-Roquebert, Quentin Couix, Malo Mofakhami and Guillaume Noblet. I am also lucky to have benefited, during the last two years of my thesis, from the friendly and stimulating environment that is the World Inequality Lab. I am thankful to the whole team, including those recently arrived, for their good will and kindness. I am pleased to work with people that I respect and admire greatly. I especially thank those that also contributed with insightful comments and remarks to the research in this volume: Facundo Alvaredo (×2), Lydia Assouad (best title consultant), Luis Estévez Bauluz, Yonathan Berman, Thomas Blanchet (×2), Lucas Chancel, Mauricio De Rosa, Amory Gethin, Clara Martinez-Toledano, Marc Morgan (×2), Thomas Piketty, Daniel Waldenstrom and Li Yang ($\times$1.5). Other noteworthy mentions are Pablo Gutierrez and Damián Vergara. I would also like to thank my students from Paris 1. From them I learned many things, but mostly patience.

I finally thank my family and friends for their unconditional support. Not only they were present during the development of my thesis but also during the whole adventure that was to migrate from Chile to France. I thank those with whom I shared the path, while being young and almost alone, in the search for knowledge and good ideas. I think of Stéphane, the other Stéphane, Christophe, Domi, Manuel, Vicente, Remi, Tano, Francesco, Isa, Clarinha, Julia. During the last nine years, they have been my second family. Or maybe it is that we have all been orphans together. Of course, my appreciation is not only defined by the amount of time spent together, but I have the biggest gratitude for those who where there from day one. Papá, Mamá, they taught me the skills that enable me to be where I am today. They taught me for instance, as a toddler, how to stand still without falling like a stick and landing on my face, which has been a rather useful skill so far. They also taught me the complexity of social interactions and signals, as for instance that it is socially desirable to wear matching socks in public. I have to say, I do not always excel at this particular point. More seriously speaking, I mostly thank them for having raised my heart and my mind with both care and passion. I also thank them for having decided to raise two supplementary kids for my own amusement. Matías and Gabriel are good lads. I love them more than I can put into words. Talking about love, a huge thank you to Camille, without whom this project would have probably never started.

Summary

This PhD dissertation contains three essays on the empirical measurement of economic inequality. The first chapter "Top Incomes in Chile: A Historical Perspective of Income Inequality (1964-2015)", which was written with Jorge Atria, Claudia Sanhueza and Ricardo Mayer, is a country-case study. It presents a historical series of Chilean top income shares over a period of almost half a century, mostly using data from tax statistics and national accounts. We distinguish between *adjusted* (1990-2015) and *unadjusted* (1964-2015) series. The latter only includes personal income, while the former includes the imputation of corporate *undistributed profits*, which results in higher inequality levels. Unadjusted estimates follow a decreasing trend over the course of the 1960s, followed by an inverted U-shape that reaches a peak during the dictatorship (1980s). By contrast, the adjusted series contradicts the evidence based on survey data, according to which inequality has fallen constantly over the past 25 years. Rather, it changes direction, increasing from around the year 2000. Chile ranks as one of the most unequal countries among both OECD and Latin American countries over the whole period of study. This chapter is, of course, of special interest for readers who share my concern about Latin American inequality. Yet maybe the broader audience could find interest on the fact that it shows the struggle that is, from an empirical point of view, to construct inequality estimates when datasources provide contrasting information. It indeed brings the intuition for the following two chapters, which develop more general subjects.

The second chapter "Income Under the Carpet: What Gets Lost Between the Measure of Capital Shares and Inequality" measures the relative underestimation of factor income (i.e., capital and labor) in distributive data, with respect to national accounts' figures. I study a group of countries with available data using surveys (LIS-Database), but also tax and Distributional National Accounts (DINA) estimates for the US (WID). I find that households receive around only half of national gross capital income, as opposed to private and public corporations,

and the trend decreases in most countries over 1995-2015 (panel, 19 countries). Due to heterogeneous non-response and misreporting, household-surveys only capture around 20% of this aggregate, versus 70% of labor income (sub-panel, 13 countries). This structure understates inequality estimates, which become insensitive to changes in the capital share (gross and net estimates) and its distribution. These distortions are weaker in tax data but still present, while DINA estimates are not subject to them by construction. I formalize this system in a novel theoretical framework based on accounting identities. I then use it to compute marginal effects and contributions to changes in fractile shares.

The final chapter "The Weight of the Rich: Improving Surveys Using Tax Data", which was written with Thomas Blanchet and Marc Morgan, presents a novel method to adjust household surveys. Indeed, tax data show that household surveys generally fail to properly capture the top of the income distribution, and therefore need to be adjusted to estimate inequality correctly. To date, there is no consensus on how to approach this problem. We introduce a method to combine both data-sets that has several advantages over previous ones: it is consistent with standard survey calibration methods; it has explicit probabilistic foundations and preserves the continuity of density functions; it introduces the concept of a 'trustable span' in tax data; it provides an option to overcome the limitations of bounded survey-supports; and it preserves the microdata structure of the survey, maintaining the representativeness of socio-demographic variables. Our procedure is illustrated by applications in five countries, covering both developed and less-developed contexts.

Contents

0 Summary in French - *Résumé en Français* **1**

1 Top Incomes in Chile: A Historical Perspective of Income Inequality (1964-2015) **17**

1.1 Literature 20

1.1.1 General Trends in Household Surveys 20

1.1.2 Top Incomes and Tax Data 22

1.2 Tax Data, Definitions and Methodological Issues 23

1.2.1 Income Definitions and Data 23

1.2.2 Tax Incentives and Undistributed Profits 30

1.2.3 Total Population and Interpolation Method 31

1.3 Results 32

1.3.1 Unadjusted Series (1964-2015) 32

1.3.2 Adjusted Series Including Undistributed Profits (1990-2015) 37

1.3.3 The Distribution of Income Growth 40

1.4 Comparison with Other Estimates 43

1.4.1 International Comparisons 43

1.4.2 Local Surveys 44

1.5 Trend Robustness 47

2 Income Under the Carpet: What Gets Lost Between the Measure of Capital Shares and Inequality **52**

2.1 Stylized Facts 57

2.1.1 Data and Concepts 57

2.1.2 The Capital Share of National Income 60

2.1.3 The Household Share of Capital Income 61

2.1.4 Measurement Error in Distributive Data 66

2.2 Theoretical Framework 68

2.2.1 Setup 69
2.2.2 Identities 69
2.2.3 Marginal effects 71
2.3 Applications 72
2.3.1 General Trends 72
2.3.2 Sensitivity Analysis 75
2.3.3 Estimated Contributions 79

3 The Weight of the Rich: Improving Surveys with Tax Data 84
3.1 Related Literature on Correcting Surveys 88
3.1.1 Reweighting Observations 88
3.1.2 Rescaling Incomes 89
3.1.3 Combined Reweighting and Rescaling 90
3.2 Theory and Methodology 91
3.2.1 Reweighting 91
3.2.2 Calibration 98
3.2.3 Replacing and Matching 106
3.2.4 Standard Errors and Confidence Intervals 108
3.3 Applications 108
3.3.1 Definitions and Data 109
3.3.2 Empirical Bias and Corrected Population 110
3.3.3 Income Distribution 114

A Top Incomes in Chile, A Historical Perspective of Income Inequality (1964-2015) 121
A.1 Changes in tax legislation and income definition. 121
A.2 Additional tables and figures 125

B Income Under the Carpet: What Gets Lost Between the Measure of Capital Shares and Inequality 137
B.1 Excluding the Public Sector 149
B.2 Figures using Net Capital Income 151
B.3 Additional Figures 160

C The Weight of the Rich: Improving Surveys with Tax Data 162
C.1 Country Specific Income Concepts and Observational Units 162
C.1.1 Brazil 162
C.1.2 Chile 164

C.1.3 European Countries 166
C.2 Shape of the Bias 167
C.3 Structure of the Corrected Population 167
C.4 Gini Coefficients 178
C.5 The Impact of Misreporting 180

D List of Figures and Tables **183**

E References **189**

Chapter 0

Summary in French - *Résumé en Français*

Histoire des hauts revenus au Chili, nouvelles données sur un pays en développement

Suite aux articles fondateurs de Piketty (2001) et de Piketty et Saez (2003), le développement de la littérature sur les hauts revenus, au cours des deux dernières décennies, a signifié un véritable progrès pour l'étude des inégalités économiques. Basées principalement sur l'utilisation de données fiscales et de comptes nationaux, des études sur plus de 40 pays ont été menées pour explorer la concentration du revenu dans les 10%, 1%, 0,1% et 0,01% les plus riches de la population.[1] Ces travaux ont démontré que, pourvu que le précautions nécessaires soient prises, les données fiscales peuvent révéler une partie de la distribution qui était invisible auparavant. Celles-ci ont ainsi permis d'examiner une plus grande partie de cette distribution et de remonter plus loin dans le temps que ce que les données d'enquêtes ne le permettent. En effet, la vraie valeur des statistiques fiscales est de se concentrer sur des petits groupes de personnes qui accumulent une partie significative du revenu total et dont l'évolution est susceptible d'influencer les tendances globales en matière d'inégalité (Alvaredo, Atkinson, Piketty et al., 2013).

Dans les pays en développement il existe à présent encore peu d'informations sur les niveaux et les tendances des plus hauts revenus. En particulier, les études basées sur des données fiscales sont encore relativement rares. Le premier chapitre de cette thèse, "*Top Incomes in Chile : A Historical Perspective of Income Inequality (1964-2015)*", qui a été co-écrit avec Jorge Atria, Claudia Sanhueza et Ricardo Mayer, contribue à combler cette lacune en ajoutant le Chili à la littérature des hauts revenus. Celui-ci permet d'étudier l'évolution des inégalités dans ce pays sur le long terme, ce qui n'est d'ailleurs pas possible au niveau national en utilisant des données d'enquêtes. Le Chili est un cas intéressant pour diverses raisons. Bien que classé parmi les pays les plus inégalitaires de l'OCDE (OECD, 2015), le Chili est considéré comme l'un des États les plus forts d'Amérique latine en termes de capacité, de niveau de corruption et d'efficacité de la politique fiscale. Néanmoins, le pays a encore un faible niveau de redistribution et la politique budgétaire a une capacité limitée pour réduire des inégalités de marché extrêmement élevées (OECD, 2015).

Les estimations produites dans ce premier chapitre sont comparables à celles des quelques autres pays d'Amérique latine disposant de données similaires,

[1] Voir la base de données des inégalités mondiales : `http://www.WID.world`

tels que l'Argentine, le Brésil, la Colombie et l'Uruguay, mais également aux autres pays inclus dans la base de données sur les inégalités mondiales. Quelques tentatives préalables ont été faites pour introduire des statistiques fiscales dans le cas particulier du Chili. Celles-ci nous permettent d'avoir une idée assez précise sur le niveau de concentration du revenu pour des années récentes, mais elles ne permettent pas de tirer des conclusions sur son l'évolution (LÓPEZ, FIGUEROA et GUTIÉRREZ, 2013/2016, FAIRFIELD et JORRATT DE LUIS, 2016). Ces études mettent l'accent sur l'importance, au niveau local, de la prise en compte des bénéfices non distribués des firmes. Les auteurs expliquent que ceux-ci ont probablement un impact via des incitations structurelles à la rétention des profits au sein des firmes. Pour répondre à cette inquiétude, les estimations produites dans ce chapitre font la distinction entre les séries fiscales *non ajustées* pour la période 1964-2015, qui ne considèrent que le revenu personnel, et les séries *ajustées*, qui incluent l'imputation des bénéfices non distribués aux individus pour la période 1990-2015.

Les principaux résultats indiquent que la concentration des revenus reste relativement élevée dans les deux séries tout au long de la période observée. La tendance des estimations non-ajustées est à la baisse au cours des années précédant le coup d'état (1964-1973). Celle-ci est ensuite inversée au cours des premières années de dictature, où le niveau de concentration a nettement augmenté (1973-1981). À partir de 1990, avec le retour de la démocratie, la série non-ajustée reprend une tendance générale à la baisse jusqu'en 2015. Contrairement, la série ajustée, qui ne couvre que la période démocratique récente (1990-2015) et qui attribue les revenus non-distribués des entreprises à leurs propriétaires, montre non seulement des niveaux d'inégalité bien plus élevés, mais également un changement de tendance. En effet, la part des 1% les plus riches est supérieure de 4 à 10 points de pourcentage, selon les années, et la tendance à la baisse observée après 1990 s'inverse autour de l'année 2000. Cette dernière conclusion est particulièrement pertinente car elle contredit le consensus existant, basé sur des données d'enquête, selon lesquelles les inégalités auraient diminué au cours des deux dernières décennies (Annexe 1.1.1). En outre, quand on compare la part des revenus accumulée par le percentile le plus riche, le Chili est parmi les pays les plus inégaux d'Amérique latine et aussi d'autres pays membres de l'OCDE au cours de la majeure partie de la période. De plus, nous montrons que les niveaux de concentration mesurés avec des données d'enquête sont généralement plus bas et plus volatiles que les mesures basées sur déclarations fiscales et comptes nationaux.

Il convient de noter que les tableaux de déclarations pour l'impôt sur le revenu, qui sont la principale source employée pour cette étude, présentent une limitation majeure en ce qu'elles ne contiennent de l'information que sur le revenu total et pas sur sa composition (salaires, pensions, intérêts, dividendes, etc.). FAIRFIELD et JORRATT DE LUIS (2016) suggèrent que l'évasion fiscale est principalement liée aux dividendes et au revenu des travailleurs indépendants. Mais comme nous ne pouvons pas distinguer les différents types de revenus, nous ne pouvons pas effectuer d'ajustement pour évasion. Ces limitations biaisent probablement le niveau de nos estimations à la baisse. Nous considérons donc que nos estimations n'établissent qu'une borne inférieure sur les niveaux de concentration. Celles-ci sont alors principalement utiles pour l'interprétation de tendances.

Cette recherche confirme l'inquiétude exprimée dans la littérature, selon laquelle la structure institutionnelle spécifique au Chili inciterait les entreprises à retenir leurs bénéfices, tout en permettant à leurs propriétaires d'y accéder de manière moins détectable et donc moins taxable. Nous allons plus loin en constatant que non seulement le niveau, mais aussi la tendance de la concentration du revenu peuvent être biaisés par ce phénomène. Nous remettons donc en question la tendance à la baisse de la concentration des revenus apparaissant à la fois dans les estimations des enquêtes et des données fiscales strictement personnelles, du moins depuis le début des années 2000. L'évolution des bénéfices non distribués a très probablement contribué à faire baisser ces tendances. Il est donc crucial d'étudier l'évolution conjointe du revenu des entreprises et des revenus personnels afin d'analyser la situation dans son ensemble et d'identifier les tendances des inégalités plus solides dans le scénario chilien. Naturellement, des recherches supplémentaires sont nécessaires pour déterminer si ce changement de tendance pourrait aussi se traduire dans des mesures d'inégalité globale.

Le premier chapitre de cette thèse peut susciter un intérêt particulier chez les lecteurs qui partagent une préoccupation spéciale vis-à-vis des inégalités dans la région latino-américaine. Mais peut-être, un public plus large pourrait trouver de l'intérêt sur le fait qu'il montre la difficulté, d'un point de vue empirique, de la construction des données distributives lorsque les sources de données fournissent des informations divergentes. Ce chapitre apporte en effet l'intuition pour les deux chapitres suivants, qui développent des sujets plus généraux.

Des revenus cachés sous le tapis, une analyse globale sur les données distributives

Au cours des 50 à 60 dernières années, la plupart des pays développés ont enregistré une croissance substantielle dans la part du capital dans leur revenu national (IMF, 2007 ; ARPAIA, PÉREZ et PICHELMANN, 2009 ; PIKETTY et ZUCMAN, 2014 ; KARABARBOUNIS et NEIMAN, 2014). Autrement dit, la part du revenu macro-économique qui rémunère le capital, par opposition au travail, augmente depuis des décennies. Ce phénomène se produit parallèlement à l'augmentation de la concentration des revenus personnels enregistrée par ATKINSON, PIKETTY et SAEZ (2011) et ALVAREDO, CHANCEL et al. (2018), ce qui a amené plusieurs chercheurs à explorer la relation entre la division factorielle du revenu national (c'est-à-dire les parts du capital et du travail) et les inégalités. Ce domaine de recherche a pour objectif fondamental d'établir un meilleur lien entre les agrégats macro-économiques, qui sont généralement utilisés pour mesurer le progrès économique, et la répartition des revenus et des richesses, souvent utilisées pour étudier le bien-être.

Lorsque les revenus du capital sont plus concentrés que les revenus du travail, intuitivement, on pourrait s'attendre à ce qu'une augmentation de la part du capital provoque nécessairement une augmentation de l'inégalité totale. Cependant, la concentration relative des revenus factoriels ne fournit pas suffisamment d'informations pour définir une telle relation. Comme le dit MILANOVIC (2017) :

> La concentration élevée d'une source de revenu donnée ne garantit pas que celle-ci contribue positivement à l'inégalité totale. L'indice de Gini calculé sur les allocations de chômage est généralement supérieur à 90 (puisque la plupart des personnes ne reçoivent aucune allocation de chômage au cours d'une année donnée), mais puisque les bénéficiaires des allocations chômage se situent généralement au bas de la distribution, une augmentation de la part des allocations de chômage dans le revenu total se traduit dans une réduction des inégalités.

Étant donné que les individus perçoivent des revenus de différentes sources au même temps et que l'on trouve les bénéficiaires de chaque type de revenu tout au long de la distribution, il convient toujours de prendre en compte la distribution conjointe des revenus factoriels avant de faire des hypothèses. ATKINSON et BOURGUIGNON (2000) ATKINSON (2009) et MILANOVIC (2017) contribuent avec des démonstrations formelles basées sur des identités comptables, appuyant cette idée. Les deux premiers articles analysent le coefficient de variation d'une distribution théorique. Ils l'utilisent comme mesure de l'inégalité, celle-ci est alors définie par la part du capital dans les revenus totaux, le coefficient de variation de chaque revenu factoriel et la corrélation entre les revenus du travail et ceux du capital. Atkinson (2009) définit la valeur critique pour laquelle la part de capital commence à avoir un impact positif sur l'inégalité.[2] Bien que ce chiffre puisse être positif, ce qui signifie qu'à certains niveaux, une augmentation de la part du capital pourrait théoriquement entraîner une réduction des inégalités, le point critique devrait être plutôt faible dans les scénarios convexes plausibles. Milanovic (2017) décrit un cadre similaire utilisant le coefficient de Gini. L'auteur définit trois contraintes majeures pour établir une relation positive entre la part du capital et les inégalités : premièrement, l'existence d'une forte épargne relative sur les revenus du capital ; deuxièmement, une forte concentration des actifs ; troisièmement, une forte corrélation entre le classement du revenu du capital et le classement du revenu total. Dans les cas réels, toutes ces exigences sont facilement remplies.

Avec une approche plutôt empirique, Bengtsson et Waldenström (2017) utilisent un panel de 21 pays pour évaluer la relation statistique entre nos variables d'intérêt. Ils estiment la part du capital dans le revenu national de ces pays en utilisant des données provenant des comptes nationaux historiques, puis ils appliquent des régressions pour tenter d'expliquer les variations dans la concentration du revenu. Ils utilisent deux types de mesures d'inégalité : les parts des hauts revenus, qu'ils obtiennent grâce à la base de données sur les inégalités mondiale (WID)[3], et les coefficients de Gini, qu'ils tirent de Atkinson et Morelli (2012).[4] Leurs estimations, avec effets fixes par pays, confirment leurs attentes. Les auteurs constatent un fort effet marginal positif de la part du capital sur les deux mesures d'inégalité. Lorsqu'ils introduisent un ensemble de variables de contrôle dans leurs régressions, l'effet estimé diminue mais reste

[2] La part de capital (π) doit satisfaire l'inégalité suivante : $\pi > (1-\lambda\rho)/(1-\lambda^2-2\lambda\rho)$, où λ est le rapport entre le coefficient de variation carré du capital et celui du revenu du travail et ρ est la corrélation entre le capital et le revenu du travail.

[3] Les auteurs citent la base de données en utilisant son ancien nom (celui qu'elle avait au moment de l'écriture de leur article : la *World Top Incomes Database* (WTID).

[4] Atkinson et Morelli (2012) comptent deux types d'estimations des coefficients de Gini, soit ils les estiment directement à partir d'enquêtes aux ménages, soit ils les tirent de centres de données internationaux bien connus. Ces estimations sont disponibles pour un sous-ensemble de pays et une période plus courte comparée à celle où les données sur les hauts revenus sont disponibles

significatif.

Francese et Mulas-Granados (2015) est une autre contribution au volet empirique de cette littérature, avec des résultats quelque peu contrastés. Les auteurs utilisent les données d'enquête auprès des ménages harmonisées de la base de données luxembourgeoise des revenus (`http://lisdatacenter.org`) pour effectuer une analyse de décomposition du coefficient de Gini dans 43 pays au cours de la période 1978-2010. Ils décomposent le coefficient de Gini en ses composantes comptables et appliquent ensuite une régression similaire à celle de Bengtsson et Waldenström (2017), mais en utilisant uniquement le coefficient de Gini de l'enquête en tant que variable dépendante. Après analyse, ils concluent que la part du capital joue un rôle négligeable dans l'évolution des inégalités mesurées, en particulier comparé au rôle de l'inégalité des revenus du travail, qu'ils considèrent comme le principal facteur explicatif de l'inégalité totale. Bien qu'en théorie la relation entre part du capital et inégalité soit clairement fondée sur des identités comptables, elle peut sembler empiriquement plus opaque et imprévisible.

Le problème avec les modèles existants est qu'ils ne permettent une relation négative (ou nulle) que dans des circonstances très restrictives. Par exemple, dans Milanovic (2017), le seul moyen d'y parvenir est d'avoir un revenu du capital plus (ou également) concentré que le revenu du travail, une corrélation négative (ou nulle) entre le revenu total et le revenu du capital et/ou ayant un taux d'épargne sur les revenus du travail qui soit plus élevé (ou égal) à celui sur les revenus du capital. Ces configurations sont bien sûr plutôt irréalistes. Par conséquent, lorsque des corrélations négatives apparaissent empiriquement, ces modèles ne semblent pas fournir une description convaincante des mécanismes en jeu. En effet, lorsque Bengtsson et Waldenström (2017) observent une corrélation négative pour l'Argentine et le Canada, ils traitent ce résultat comme une anomalie. Sur une plus grande échelle, en Amérique latine, Abeles, Amarante et Vega (2014) enregistrent une augmentation globale de la part du capital qui est parallèle et contradictoire avec la baisse généralisée de l'inégalité observée par López-Calva et Lustig (2010), ce qui paraît étrange et difficile à interpréter. Afin de mieux comprendre ce genre de scénario complexe, nous devons ajouter quelques variables à l'équation.

L'intuition principale du deuxième chapitre de cette thèse, intitulé "*Income Under the Carpet : What Gets Lost Between the Measure of Capital Shares and Inequality*", est assez simple : les divergences entre les définitions du revenu et la qualité des différentes sources de données peuvent filtrer l'impact des parts factorielles sur les estimations des inégalités. Les agrégats de revenus issus des comptes nationaux, en particulier ceux des revenus du capital, sont souvent nettement plus élevés que ceux présentés dans la plupart des données distributives (enquêtes ou statistiques fiscales). Au moins une partie de ce phénomène est due au fait que les comptes nationaux, qui sont souvent utilisés pour estimer les parts factorielles, ont une définition plus large du revenu. Cependant, on retrouve le même résultat lorsque l'on compare des définitions harmonisées. Étant donné que les comptes nationaux sont généralement utilisés comme point de repère, on peut donc considérer qu'au moins une partie du «revenu global agrégé» est absente des données distributives.[5] Même si la littérature a accordé une certaine attention à ce sujet, elle ne l'a pas explicitement inclus dans les modèles empiriques ni théoriques. Dans ce chapitre, nous examinerons deux voies principales par lesquelles les enquêtes et les données fiscales pourraient ignorer du revenu (les revenus cachés sous le tapis). Premièrement, tous les revenus du capital (dividendes, intérêts, profits, par exemple) ne sont pas perçus par des personnes physiques. Au moins une partie atterrit dans les caisses des entreprises privées ou publiques. Il est donc logique de mesurer systématiquement la part des revenus du capital qui est effectivement reçue par les ménages. Deuxièmement, les bases de données distributives sont souvent sujettes à des erreurs de mesure, principalement en raison de l'hétérogénéité des taux de réponse et de l'existence de déclarations fausses ou erronées. L'erreur avec laquelle les statistiques distributives mesurent les revenus factoriels agrégés peut, elle aussi, être tracée et analysée.

La première contribution de ce deuxième chapitre consiste à établir des faits stylisés à la fois sur la part du revenu du capital des ménages et sur l'erreur de mesure dans les enquêtes aux ménages. Les séries sur ce premier élément sont construites principalement à l'aide des comptes nationaux des 43 pays dont les statistiques sont suffisamment détaillées auprès de la division de statistiques officielles des Nations Unies. La principale constatation sur cette question est une diminution généralisée et forte de la part des revenus du capital reçue par le secteur des ménages, par opposition aux entreprises privées et publiques. Cette constatation est valable lorsque l'on étudie le revenu agrégé de 19 pays dans un panel équilibré au cours de la période 1995-2015. Mais aussi au niveau individuel, au cours de la même période, dans un panel déséquilibré comprenant 43 pays.

[5]Bien entendu, la reconnaissance des différences conceptuelles entre les différentes sources de données n'est pas nouvelle. En fait, Atkinson (2009 : Section II) énumère de façon détaillée les éléments pertinents à cette question.

Les séries les plus longues disponibles montrent que les tendances commencent à baisser vers 1990 dans la plupart des cas. Un tel phénomène peut potentiellement impliquer une multiplicité d'effets sur le plan économique, mais notre étude porte exclusivement sur l'impact qu'il aura sur la mesure de l'inégalité. En outre, les estimations sur l'erreur de mesure dans les enquêtes sont calculées à l'aide à la fois des comptes nationaux et des micro-données harmonisées d'enquêtes (base de données luxembourgeoise sur le revenu), sur un panel équilibré de 13 pays pour la période 1995-2013. Les revenus du travail et du capital semblent être tous les deux sous-évalués dans tous les pays. Les revenus du capital sont, dans tous les cas, plus fortement sous-estimés, avec seulement 20% environ de d'agrégat brut de ceux-ci étant enregistrés dans les enquêtes, contre 70% chez les revenus du travail. Cette dernière relation reste généralement stable sur la période. Pour les États-Unis, nous comparons le niveau de sous-estimation à celui des données fiscales, qui est nettement inférieur pour les revenus du capital. Les estimations des comptes nationaux distributifs (DINA) ne sont pas, par construction, soumises à ce type de sous-estimation.[6]

La deuxième contribution de ce chapitre est d'introduire un cadre théorique simple basé sur des identités comptables qui retracent le chemin parcouru entre les revenus du capital et du travail au niveau national jusqu'à la distribution des revenus entre les ménages, telles quelles sont enregistrées dans des données d'enquêtes ou des données fiscales. On trouve que le produit entre la part des ménages dans les revenus du capital et la sous-estimation relative des revenus du capital agit comme facteur de distorsion. Il filtre l'effet des revenus du capital. Le résultat est généralement une sous-estimation des niveaux et des tendances des inégalités qui affecte la sensibilité des estimations des inégalités à la part des revenus du capital dans le revenu national et à la distribution de ceux-ci. Cette représentation simple et directe est ensuite utilisée dans le deuxième chapitre afin d'explorer la sensibilité empirique des estimations de la concentration du revenu à chacune des variables du modèle selon les pays et à travers le temps. Les estimations issues d'enquêtes sous-estiment en grande partie l'influence du revenu du capital et semblent donc suivre presque exclusivement la distribution du revenu du travail. Les données fiscales sont relativement plus sensibles à la

[6]Les comptes nationaux distributifs désignent la méthodologie développée par Alvaredo, Atkinson, Chancel et al. (2016) dans le cadre d'un projet global visant à combiner des données d'enquêtes, des données de taxes et des comptes nationaux pour mieux étudier la répartition de l'ensemble du revenu national. Grâce à diverses procédures d'imputation, les estimations résultantes sont cohérentes avec les valeurs du revenu agrégé figurant dans les comptes nationaux.

part du capital et à sa distribution, du moins pendant la période 1975-2015. Les estimations DINA ne sont, là encore, par construction, pas sujettes à ces distorsions.

Devrait-on viser à corriger les données distributives ?

Le message général du deuxième chapitre est que les statistiques d'enquêtes ne parviennent pas à capturer une part croissante du revenu national qui rémunère le capital. Cela sous-estime probablement les niveaux et les tendances des inégalités. La taille de ce phénomène rend les estimations d'enquête presque totalement insensibles au mouvement macro-économique des parts de capital. À la lumière de l'évidence présentée dans cet ouvrage, nous pouvons mieux comprendre les résultats de Francese et Mulas-Granados (2015). Les auteurs utilisent également des enquêtes LIS. Ils constatent que l'évolution de l'inégalité observée s'explique presque exclusivement avec la répartition des revenus du travail. Cependant, leurs conclusions ne doivent pas être interprétées comme une preuve de la banalité de l'impact des revenus du capital sur l'inégalité. Ceci, car leurs estimations ne prennent tout simplement pas en compte la grande majorité des revenus du capital. Les résultats de ce chapitre peuvent également aider à comprendre que Bengtsson et Waldenström (2017) trouvent un fort effet des parts de capital sur les parts de revenus les plus élevées, telles que mesurées par des données administratives, sachant que celles-ci sont mieux adaptées pour saisir les revenus de capitaux que les enquêtes. En outre, il est quelque peu surprenant qu'ils constatent quand même un impact significatif des parts de capital sur les estimations des coefficients de Gini issus d'enquêtes.

Le fait que les enquêtes représentent mal le revenu du capital dans la plupart des cas n'implique certainement pas qu'il faille s'en débarrasser. Les enquêtes aux ménages sont certainement la source de données la plus riche et la plus facilement disponible pour étudier l'inégalité des revenus dans toutes ses dimensions. Cela est principalement dû au nombre élevé de covariables généralement déclarées par les répondants. En outre, la grande majorité des habitants des pays à revenu élevé et intermédiaire sont rémunérés via le revenu du travail, qui est lui au moins relativement bien capturé par les enquêtes. Il est en effet généralement admis que les enquêtes fournissent des informations précieuses sur ce qui se passe à la fois au milieu et au bas de la répartition du revenu. Par conséquent, devrions-nous essayer d'ajuster les enquêtes pour inclure des informations externes et fiables disponibles à la fois dans les comptes nationaux et dans les données administratives ? Ou devrions-nous reconnaître leurs limites et utiliser chaque base de données pour étudier des différents aspects spécifiques de la répartition du revenu et du patrimoine ? D'une part, la deuxième option évite le risque d'introduire des distorsions indésirables dans les enquêtes et les estimations qui en résultent. En effet, en fonction de la qualité et du niveau de détail initial de chacune des bases de données, on peut être amené à émettre des hypothèses plus ou moins inconfortables lors de l'application de corrections aux enquêtes. Même si les ajustements sont effectués avec soin, des différences de qualité des données entre les pays pourraient potentiellement introduire du bruit dans les comparaisons internationales. D'autre part, une bonne correction permettrait de mieux étudier l'incidence de la croissance des revenus macro-économiques. Sachant que cette croissance est souvent soulignée dans les discours politiques comme étant universellement bénéfique, mais que cette distribution est rarement et très mal mesurée. Cela permettrait également, dans l'idéal, de fonder les études sur l'inégalité économique et ses différentes dimensions sur des données plus solides et plus fiables. En outre, ce type d'ajustement serait particulièrement utile lorsque les bases de données se contredisent en termes de tendances de la concentration.

Quoi qu'il en soit, il ne faut pas oublier que, malgré les efforts considérables déployés pour harmoniser les enquêtes auprès des ménages d'un pays à l'autre, l'objectif est encore lointain. Des différences substantielles sont observées en termes de taux de réponse moyens, de définitions de revenus et de méthodes d'échantillonnage. Celles-ci représentent des sources de biais potentiellement importantes. De plus, on sait que les poids des observations des enquêtes sont aujourd'hui ajustés de façon très courante, principalement à l'aide de techniques de post-stratification et de calibration de poids, qui utilisent toutes les deux des données externes pour effectuer des corrections (souvent des données issues de recensements). Ces techniques d'ajustement, qui sont rarement remises en question par les utilisateurs des enquêtes, visent à corriger la distribution inégale des taux de réponse parmi les personnes présentant des caractéristiques socio-économiques différentes ; elles apportent toutefois généralement des corrections en fonction des totaux de population (par exemple, âge, sexe, région géographique) et non sur la distribution de variables telles que le revenu.

Le poids des riches, ou comment corriger les enquêtes aux ménages avec des données fiscales

Pendant longtemps, l'essentiel de ce que nous savions sur la répartition des revenus provenait d'enquêtes, dans lesquelles il est demandé à des ménages choisis au hasard de remplir un questionnaire. Ces enquêtes ont été un outil précieux pour suivre l'évolution de la société. Mais ces dernières années, la littérature s'est de plus en plus soucié de leurs limites. En particulier, les enquêtes ont du mal à capturer les revenus de la partie haute de la distribution.

Pour cette raison, la recherche s'est tournée vers une source différente : les données fiscales. L'idée n'est pas nouvelle. nous pouvons la retrouver dans le travail précurseur de KUZNETS (1953), ou même de PARETO (1896). Plus récemment, PIKETTY et SAEZ (2003) et PIKETTY (2003) ont appliqué leur méthode aux données les plus récentes pour la France et les États-Unis. Ce travail a été étendu à un plus grand nombre de pays par de nombreux chercheurs dont les contributions ont été rassemblées par ATKINSON (2007, 2010) et ont servi de base à la base de données sur les inégalités mondiales (`http://wid.world`).

Mais les données fiscales ont leurs propres limites. Elles ne couvrent que le sommet de la distribution et incluent au mieux un ensemble limité de covariables. Souvent, elles ne sont pas disponibles sous forme de microdonnées, mais plutôt sous forme de tabulations résumant la distribution, ce qui limite leur utilisation. L'unité statistique qu'ils utilisent (individus ou ménages) dépend de la législation locale et peut ne pas être comparable d'un pays à l'autre. C'est pourquoi de nombreux indicateurs, tels que les taux de pauvreté ou les écarts entre les sexes, doivent encore être calculés à partir d'enquêtes. L'utilisation de différentes sources - parfois contradictoires - pour calculer des statistiques sur la répartition du revenu et de la richesse peut compliquer la tâche de tracer une image cohérente et précise des tendances en matière d'inégalité. Ceci explique les efforts en cours pour combiner les différentes sources de données à notre disposition de manière à exploiter leurs forces et à corriger leurs faiblesses.

Le projet des comptes nationaux distributifs (DINA) est un bon exemple de cet effort. Ses instructions (ALVAREDO, ATKINSON, CHANCEL et al., 2016) insistent sur la nécessité d'examiner la distribution dans son ensemble, d'harmoniser les concepts et, si possible, de les décomposer en fonction de l'âge et du sexe. PIKETTY, SAEZ et ZUCMAN (2018a) aux États-Unis et GARBINTI, GOUPILLE-LEBRET et PIKETTY (2016) en France ont utilisé à la fois des données d'enquêtes et des données fiscales pour créer des statistiques de répartition tenant compte de tous les revenus enregistrés dans les comptes nationaux. Mais ces exemples reposent en grande partie sur l'existence de microdonnées administratives accessibles aux chercheurs, auxquelles des informations provenant d'enquêtes peuvent être ajoutées.

Dans de nombreux pays, développés et moins développés, un accès à des données d'une telle qualité est assez rare. Au lieu, on trouve des tabulations sur le revenu fiscal contenant des informations sur le nombre et le revenu déclaré des contribuables par tranche de revenu. La couverture de la population est souvent inférieure à celle de la population adulte totale et la différence varie selon les pays étudiés. Dans de tels cas, nous devons procéder dans l'ordre inverse : au lieu d'incorporer des informations d'enquête dans les données fiscales, nous devons incorporer des informations fiscales dans les données d'enquête.

Un certain nombre d'approches ont été suggérées pour traiter ce problème, mais la littérature n'a pas réussi à converger vers un consensus. Dans le troisième chapitre de cet ouvrage, intitulé : "*The Weight of the Rich : Improving Surveys Using Tax Data*", co-écrit avec Thomas Blanchet et Marc Morgan, on développe une nouvelle méthodologie qui présente des avantages importants par rapport aux précédentes et qui devrait couvrir la plupart des cas pratiques dans un seul cadre unifié. Notre méthode est basée sur des fondements probabilistes explicites avec des interprétations claires et intuitives. Cela évite également de s'appuyer, dans la mesure du possible, sur des hypothèses paramétriques *ad hoc*. Nous présentons une méthode guidée par les données sous-jacentes pour déterminer où le biais commence dans les données d'enquête et au-delà de quel point nous fusionnons les revenus provenant de données fiscales dans l'enquête. Nous effectuons les ajustements nécessaires de manière à minimiser les distorsions par rapport à l'enquête originale et à préserver les propriétés souhaitables, telles que la continuité de la fonction de densité. Plutôt que de faire directement des hypothèses sur le comportement de statistiques complexes telles que des quantiles ou des moyennes par intervalles, notre méthode émet des hypothèses facilement interprétables au niveau des observations. En conséquence, nous pouvons préserver la richesse des informations dans les enquêtes, à la fois en termes de covariables et de structure des ménages. En examinant simultanément toutes les variables, nous garantissons la représentativité de l'enquête en termes de revenu tout en maintenant sa représentativité en termes d'âge, de sexe ou de toute autre dimension.

Notre méthode se déroule en deux étapes, qui visent à corriger les deux principaux types d'erreur dans les enquêtes : l'erreur non due à l'échantillonnage et l'erreur d'échantillonnage. Les erreurs non dues à l'échantillonnage font référence à des problèmes qui ne peuvent pas être facilement résolus avec un échantillon de plus grande taille et proviennent généralement de taux de réponse hétérogènes non observés. Dans la première étape, nous corrigeons ces problèmes en utilisant une procédure de repondération basée sur la théorie de calibration d'enquêtes (Deville et Särndal, 1992). Ce faisant, nous corrigeons une incohérence de longue date entre la littérature empirique sur les hauts revenus et la pratique établie de la plupart des producteurs d'enquêtes. En effet, depuis que Deming et Stephan (1940) ont introduit leur algorithme *raking*, les instituts de statistique ont régulièrement repondéré leurs enquêtes pour correspondre aux totaux démographiques connus des données de recensement. Cependant, la littérature sur les revenus a principalement consisté à ajuster la valeur attachée à des observations, plutôt que leur poids, pour assurer la cohérence entre les données fiscales et les données d'enquête. Les fondements théoriques de cette approche sont moins explicites et plus difficiles à justifier.

Cette première étape traite des erreurs non dues à l'échantillonnage, mais sa capacité à corriger l'erreur d'échantillonnage est limitée, ce qui signifie un manque de précision dû à la taille limitée de l'échantillon.[7] Un exemple radical est le revenu maximum, qui est presque toujours inférieur dans l'enquête par rapport aux données fiscales, un phénomène qu'aucune repondération ne peut résoudre. La part des très hauts revenus est également fortement biaisée à la baisse dans des petits échantillons (Taleb et Douady, 2015), de sorte que les inégalités seront sous-estimées même si toutes les erreurs non dues à l'échantillonnage ont été corrigées. Pour résoudre ce problème, nous complétons la calibration de l'enquête par une seconde étape, dans laquelle nous remplaçons les valeurs des observations du haut par une distribution générée à partir des données fiscales en y faisant correspondre les covariables de l'enquête. Pour ce faire, l'algorithme préserve la distribution des covariables dans l'enquête initiale, leur relation avec le revenu et la structure des ménages, quelle que soit l'unité statistique dans les données fiscales. Le résultat est un ensemble de données où la variabilité d'échantillonnage en termes de revenu au sommet a été en grande partie éliminée et dont les covariables ont les mêmes propriétés statistiques que l'enquête repondérée. Comme nous

[7]Les méthodes de calibration peuvent, dans une certaine mesure, corriger l'erreur d'échantillonnage. Mais leur capacité à le faire n'est valable que de façon asymptotique (Deville et Särndal, 1992), elle ne s'applique donc pas aux groupes de revenus étroits situés en haut de la distribution.

préservons la nature des microdonnées d'origine, nous pouvons utiliser les résultats pour construire différentes unités statistiques, des échelles d'équivalence, calculer des indicateurs complexes et effectuer des décompositions en fonction de l'âge, du sexe ou de toute autre dimension.

Notre méthode peut être utilisée pour tous les pays avec les données requises, à savoir des microdonnées d'enquête couvrant l'ensemble de la population et des données fiscales en couvrant au moins une fraction de celle-ci.[8] Pour illustrer le fonctionnement de la méthode, nous l'appliquons aux données de cinq pays, trois développées (France, Royaume-Uni, Norvège) et deux pays moins développés (Brésil, Chili). Les études de cas que nous avons choisies montrent la grande applicabilité de la méthode à la fois aux pays développés et aux pays moins développés dont la qualité des données est plus limitée.

Pour une utilisation pratique, nous avons développé une commande Stata complète qui applique la méthodologie décrite dans cet article. Le programme fonctionne avec plusieurs types d'entrées, assurant une flexibilité pour les utilisateurs. Notre méthode peut donc être facilement utilisée par les chercheurs intéressés à analyser les différentes dimensions de l'inégalité, telles que celles concernant le genre, l'éducation, les habitudes de vote, etc. [9]

L'objectif principal de ce chapitre est de fournir un outil méthodologique rigoureux permettant aux chercheurs de combiner des enquêtes sur le revenu ou le patrimoine avec des données administratives de manière simple et cohérente. Nous présentons une nouvelle méthodologie sur la combinaison de ces sources, qui intègre une compréhension formelle plus claire des biais potentiels en jeu et une solution pour y remédier. Nous soutenons que le résultat de notre approche de repondération devrait consister en un ensemble de données plus représentatif pouvant servir de base à l'étude des différentes dimensions de l'inégalité sociale. Notre algorithme est construit de manière à générer automatiquement, à partir d'enquêtes brutes et de données fiscales, un jeu de micro-données ajusté comprenant de nouvelles pondérations modifiées et de nouvelles observations, tout en préservant la cohérence des autres variables sociodémographiques préexistantes, tant au niveau individuel qu'au niveau global.

Étant donné que nous mettons les outils statistiques à la disposition du

[8]Dans le cas où les utilisateurs ne disposent que de données d'enquête tabulées, notre méthode effectuera toujours la correction en utilisant les informations par centile des microfichiers synthétiques produits par le programme gpinter.

[9]Les *packages* à télécharger sont `bfmcorr` pour la méthode de correction, et `postbfm` pour la post-estimation. Les deux commandes sont accompagnées d'instructions complètes pour l'utilisateur.

public, ils pourraient fournir les bases d'une plus grande collaboration entre les instituts nationaux de statistiques et les administrations fiscales afin d'améliorer les ensembles de données représentatifs au niveau national. La combinaison de données d'enquêtes et administratives existe déjà dans certains pays, les premières s'ancrant progressivement aux secondes dans les cas des pays les plus développés. Les statisticiens participant à la production d'enquêtes pourraient utiliser notre méthode de correction s'ils ont directement accès aux données sur les revenus et autres covariables des ministères. Pour de nombreux pays dans lesquels la majorité de la population n'est pas incluse dans les statistiques de l'impôt sur le revenu ou des cotisations de sécurité sociale, notre ajustement pourrait générer des gains importants.

Chapter 1

Top Incomes in Chile: A Historical Perspective of Income Inequality (1964-2015)

This chapter presents a historical series of Chilean top income shares over a period of almost half a century, mostly using data from tax statistics and national accounts. We distinguish between *adjusted* (1990-2015) and *unadjusted* (1964-2015) series. The latter only includes personal income, while the former includes the imputation of corporate *undistributed profits*, which results in higher inequality levels. Unadjusted estimates follow a decreasing trend over the course of the 1960s, followed by an inverted U-shape that reaches a peak during the dictatorship (1980s). By contrast, the adjusted series contradicts the evidence based on survey data, according to which inequality has fallen constantly over the past 25 years. Rather, it changes direction, increasing from around the year 2000. Finally, Chile ranks as one of the most unequal countries among both OECD and Latin American countries over the whole period of study.

Introduction

Following seminal papers by Piketty (2001) and Piketty and Saez (2003), extensive progress has been made by top incomes literature over the past two decades in the field of economic inequality. Papers addressing more than 40 countries have used tax data to explore the evolution of income concentration within the richest 10%, 1%, 0.1% and 0.01% of the population relative to total personal income.[1]. These works have successfully demonstrated that, provided the necessary precautions are taken, tax data can reveal a previously invisible section of the distribution, allowing the examination of a larger part of that distribution and extending farther back in time than any survey statistic. Indeed, the true value of tax statistics is to focus on small groups of people who concentrate substantial parts of total income, and whose evolution is likely to influence overall inequality trends (Alvaredo, Atkinson, Piketty, et al., 2013).

However, in developing countries there is still scant evidence of top income shares based on tax data. This chapter contributes to filling this gap by adding Chile to the Top Incomes literature, making use of tax statistics to shed light on long-term inequality in the developing world. Chile is an interesting case for various reasons. Although ranked among the most unequal OECD countries (OECD, 2015), Chile has been considered one of the stronger states in Latin America in terms of state capacity, corruption levels and the effectiveness of tax policy. However, the country still has a low level of redistribution, and fiscal policy has limited capacity to reduce extremely high market inequalities (OECD, 2015).

Our estimates are comparable to those of other Latin American countries with similar data, such as Argentina, Brazil, Colombia and Uruguay, but also to other countries included in the World Inequality Database (WID). Although previous attempts have been made to introduce tax statistics into the study of Chile's inequality, these are either not fully comparable with the existing literature, as in López, Figueroa, and Gutiérrez (2013/2016), or use precise data but cover a period too brief to make trend interpretations, as in Fairfield and Jorratt De Luis (2016). Both studies have strongly highlighted the local relevance of *undistributed profits*, which likely have a biasing impact via local incentives to retain corporate profits. In fact, we distinguish between *unadjusted* fiscal series for the period 1964-2015, which only includes personal income, and *adjusted* series,

[1]See the works assembled in the World Inequality Database: `http://www.WID.world`

which includes the imputation of undistributed profits for the shorter period of 1990-2015.

Our findings indicate that income concentration remains relatively high in both series throughout the whole observable period. Unadjusted top shares globally decrease during the early years (1964-1973). They then increase during the dictatorship years for which we have data (1973-1981), and finally decrease from 1990 onwards. The shorter adjusted series only covers the recent democratic period (1990-2015). The key characteristic of the latter is to include the imputation of undistributed profits to individuals based on distributive information from Fairfield and Jorratt De Luis (2016). Compared to unadjusted estimates for the same period, this series not only shows an increase in the level of inequality, but also a change in trend. Indeed, the top 1% share is higher by 4 to 10 percentage points, depending on the year, and the decreasing trend that is observed after 1990 is reversed around the year 2000. This latter finding is especially relevant because it contradicts the prevailing consensus, based on survey data, according to which local inequality has been decreasing over the past two decades (Appendix 1.1.1). Furthermore, when comparing the top 1% share, Chile ranks among the most unequal Latin American and developed countries over most of the period. In addition, we show that survey data estimates of top income concentration are generally lower and more volatile than fiscal income-based measures.

It should be noted that our tabulated income tax data has one major limitation in that it only includes total income, and lacks information on income composition by type (e.g., wages, pensions, interest, dividends). Fairfield and Jorratt De Luis (2016) suggest that tax evasion is mostly driven by dividends and the income of independent (self-employed) workers. However, as we cannot distinguish different kinds of income, we are unable to adjust for the tax evasion that is associated with each. These limitations likely bias the estimates downward, and we therefore consider our results strictly as a conservative indication of the level of income concentration.

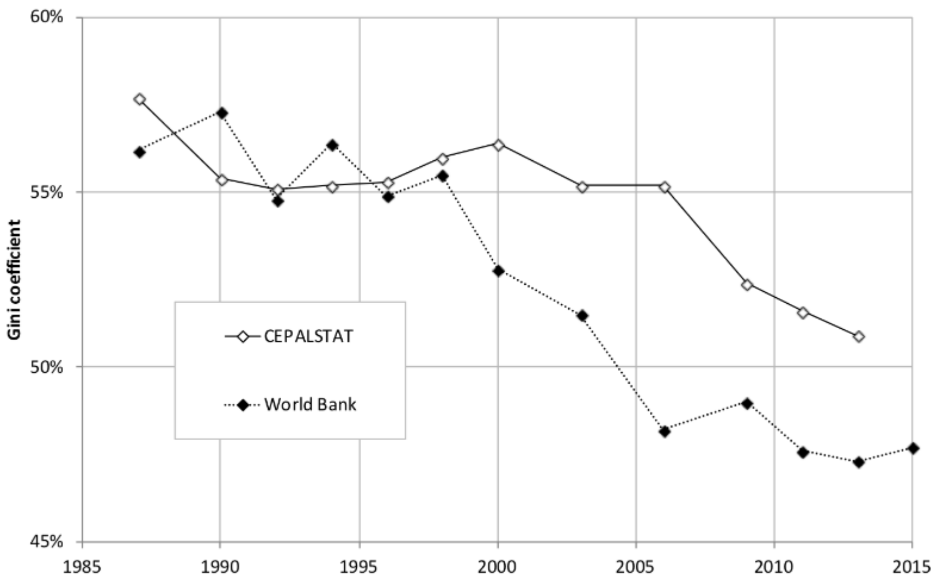
This chapter is organized as follows. Section 1.1 presents a review of previous attempts to calculate top shares in Chile. Section 1.2 discusses the structure of our data along with methodological issues, such as the interpolation method and the construction of totals for both population and income. Section 1.3 presents and analyzes resulting estimates of both adjusted and unadjusted top shares, and offers a dynamic analysis of the distribution of income growth. Section 1.4 compares our results with estimates of top income shares using the CASEN Survey, and presents international comparisons. Section 1.5 discusses trend robustness. Finally, we offer conclusions.

1.1 Literature

1.1.1 General Trends in Household Surveys

In Chile the study of personal income inequality is based predominantly on survey data. The CASEN Survey is considered to be the most precise, mainly because of its large samples which extend nationwide across both urban and rural areas. However, it only started in 1987, and for methodological reasons, some editions are often judged to be incomparable to each other. Despite this, World Bank and CEPALSTAT (ECLAC) – both international data banks – have used the CASEN Survey to build internationally comparable Gini coefficients for Chile since 1987, as shown in Figure 1.1.

Figure 1.1: Gini coefficient by CEPALSTAT and World Bank (1987-2015)



Source: CEPALSTAT and World Bank. Note: the World Bank series was updated in 2017 in response to the release of uncorrected CASEN Survey databases. Previous series were used to counter-adjust official estimates to recover an approximation of the originals (in order to avoid building income estimates that were scaled to fit National Accounts aggregates).

Each of these institutions treats the original data differently, which explains the observed differences in trends and levels. In particular, CEPALSTAT has historically applied a specific adjustment that scales aggregates of different types of income present in the survey to fit values from national accounts. The gap between national accounts and survey aggregates is imputed proportionally throughout the distribution to declared income and separately for each type, with the exception of capital income, which is only imputed proportionally to the top quintile of the distribution. This modification is noteworthy for three reasons. First, it is the main factor behind the substantial differences observed in Figure 1.1 (there are differences in the management of missing values as well, but the impact is marginal). Second, until recently, public CASEN Survey databases included these adjustments as standard, without displaying uncorrected data. Consequently, most research and official estimates on personal income and inequality to date include it. The issue is non-trivial – as Bourguignon (2015) discusses in greater depth – because this kind of adjustment has potentially significant distorting effects on the estimated distribution of income, especially when original data is already biased (as generally happens, at least at the top). Third, one of the reasons for applying such corrections is that most income aggregates from survey data are relatively low compared to both national accounts and fiscal estimates. This results in scaling factors that can multiply some types of income 2 or 3 times (at the individual level), which imposes a rather substantial alteration to the original data.

From a descriptive point of view, it appears that trend interpretation in figure 1.1 is not necessarily clear in the short run, as the gap between the two series is not constant and does not even hold strictly to positive or negative values. In fact, we can identify points in Figure 1.1 where CEPALSTAT and World Bank estimates follow opposite directions from one year to the next. However, at least since the year 2000, World Bank (unadjusted) estimates appear to be around 3 points lower than those of CEPALSTAT (adjusted). However, in the medium term, there is a degree of consensus among scientific and political narratives as to a generally decreasing trend in income inequality between the return to democracy (1990) and recent years.

In order to explain inequality trends more precisely, Contreras and Ffrench-Davis (2012) base their interpretations on a combination of the CASEN Survey and the employment survey conducted by the Universidad de Chile (EOD). The latter is considered to be more consistent over time – hence should be more precise for short-term interpretations – but has some important drawbacks. It only refers to what happens in the capital city, its samples are considerably smaller, and it is designed primarily to study employment, focusing essentially on labor income. Taking previous considerations into account, they conclude that Chile is more unequal nowadays that it was before the dictatorship. They also find that the peak of personal income inequality was somewhere in the mid-1980s, and since the return of democracy, inequality has decreased. In the short term, they find a fast drop in wage dispersion following the return to democracy, and explain this phenomenon by the decline in poverty driven mainly by increasing employment, the minimum wage, and expansion of social security. Nevertheless, this progression started to stagnate around the Asian crisis (1999) as inequalities began rising. Finally, another period of decreasing inequality begins around 2003, with improvements in poverty levels supposedly caused by increasing public spending and counter-cyclical measures during the 2009 crisis.

Clearly, the available personal income datasets present many limitations. However, the main conclusions we can draw from them should be trusted to a certain degree in terms of a fair portrayal of labor income, and their historic analysis and observed trends do help to contextualize our findings.

1.1.2 Top Incomes and Tax Data

The first attempt to study Chilean top incomes was made by Sanhueza and Mayer (2011).[2] Although they used the *Universidad de Chile*'s employment survey (EOD) and not tax data, the authors were able to study the evolution of top incomes over a period of more than fifty years. They show the top 10% of the population with a poorly-defined inverted U-shape over the 1957-2007 period, increasing sharply during the military dictatorship (1973-1990), peaking in 1988, and finally decreasing to 2007. The trend described by the top 1% is considerably more erratic, most likely as a consequence of the low representative power of survey data concerning top earners.

Subsequently, López, Figueroa, and Gutiérrez (2013) study the topic using publicly available tabulations of income declarations provided by the Chilean tax agency for the 2004-2010 period. They focus their attention mainly on the issue of undistributed profits as being a specific concern for Chile. They argue that there are strong institutional incentives for retaining profits artificially, at least during the 2000s. Moreover, the income definition that is used in the tax statistics only

[2]This section refers exclusively to the top incomes literature that is dedicated to the case of Chile. For a review of the findings of international top incomes literature please refer to Atkinson, Piketty, and Saez (2011) and Alvaredo, Atkinson, Piketty, et al. (2013)

includes an insignificant share of capital gains, which is the tool generally used to deal with this matter in the literature. Thus, they cleverly combine information from other papers to impute the whole value of corporate retained profits to the distribution of personal income. Their estimates are magnified by this procedure (nearly 30% of total income for the top 1%).[3]

Fairfield and Jorratt De Luis (2016) access micro data on income tax declarations for two specific years (2004 and 2009). They combine it with corporate tax data to track individual property and impute corporate accrued profits to their owners, following the same logic as López, Figueroa, and Gutiérrez (2013). They are able to accurately impute 80% of firms' accrued profits to their owners, with almost 30% of the latter being foreigners and thus not included in their estimates. The remaining 20% of firms, whose owners are not identified, are then imputed to the distribution. They provide various estimates according to the different assumptions that are made during imputation of the remaining part of accrued profits, and to whether or not they adjust for tax evasion. To implement this latter adjustment, they proportionally scale the revenue of both independent work and distributed profits, using aggregates from national accounts as a benchmark. Their results – all adjustments included – are stable over the period and reach similar levels to those obtained by López, Figueroa, and Gutiérrez (2013).

As we can see, previous research in the area does not provide sufficient estimates for a study of long-run top income trends. Nonetheless, they serve as useful benchmarks. Their work identifying Chilean institutional specificities also contributes with some initial guidance.

1.2 Tax Data, Definitions and Methodological Issues

1.2.1 Income Definitions and Data

Fiscal Income

The definition of income we use as the numerator of top income shares can be broadly described as including all types of revenue that is declared by resident

[3]López, Figueroa, and Gutiérrez (2016) applied more or less the same data treatment to an extended timespan (2004-2013). However, this time they used fundamental accrued capital gains (Gutiérrez, López, and Figueroa, 2015), taking into account the costs that enterprise owners would have to bear if they decided to materialize the amounts that authors are imputing to them.

individuals to tax authorities. In principle, this is a rather broad definition, as both taxed and untaxed incomes should be declared, unless the law suspends it explicitly.[4] More precisely, it corresponds to what is referred to in the Chilean tax system as the *base imponible* of personal income tax, which is the pre-tax revenue that is used to estimate marginal tax rates. Table 1.1 describes the general concepts that are included in this definition. It includes income from both dependent and independent work, both being net of social security contributions.[5] Independent workers and the self-employed report income net of costs incurred to obtain it. All types of pension, public or private, are also included. As is common in the literature, distributed profits (e.g., dividends and withdrawals), interest and rental income are also included.

Table 1.1: General Income Definition in Tax Data

	Included	Deducted
Labor income	Wages, Pensions	Contributions (Mandatory)
Mixed income	Independent Work, Self-Employment	Contributions (Non-Mandatory), Costs
Capital Income	Rents, Distributed Profits, Interest, Capital Gains	Capital Losses

Note: Major deductions and allowances, which are not included, are listed in greater detail in Section 1.3.1 and in Appendix A.1 .

Furthermore, net realized capital gains are theoretically included in the definition presented in Table 1.1. According to Atkinson, Piketty, and Saez (2011) the inclusion of realized capital gains is generally used as a tool to indirectly assess the contribution made by corporate retained profits to top incomes. Since we impute undistributed profits in our adjusted series, which starts in 1990, this could potentially present a problem of double-counting. However, there is evidence that the total amount reported by individuals as capital gains should

[4]In practice, however, the enforcement of declarations for tax-exempt revenue is generally a difficult task for the tax agency, as bank secrecy obstructs access to proper external sources of information in some cases (Fairfield and Jorratt De Luis, 2016).

[5]Although the ideal in literature is to use a definition of pre-tax income *before* deductions, the income we observe is *after* deductions and we are unable to make adjustments in order to impute deductions and allowances back. This is mainly due to data constraints and the characteristics of these contributions. In particular, independent workers are not compelled to contribute, and we cannot differentiate types of income: we only have total income. We are thus unable to make an informed adjustment to our tax tabulations.

be insignificant, at least after 1990 where tax incentives remain globally the same. Indeed, Fairfield and Jorratt De Luis (2016) report that only 3-7% of total dividends are distributed to natural persons, since at least 90% of registered shareholders of publicly traded companies are actually corporations. The vast majority of corporate property, and thus capital gains, is not held by individuals. Thus, we judge the part of realized capital gains that is present in our data to be negligible after 1990, not causing any significant bias on the level or the trend of our estimates.[6]

The structure of our data only allows us to study total income in the long run, as it provides no information in terms of composition. This constraint represents a major drawback that probably provokes an underestimation of the level of inequality in our series. Fairfield and Jorratt De Luis (2016) show that both independent income and dividends are substantially underestimated in tax data compared to National Accounts. The authors thus make a proportional adjustment for these types of income, which results in an increase in the top 1% share of fiscal income by nearly 6 percentage points (from roughly 15% to nearly 21%).[7]

Tax System and Data

The top income share estimates that are presented in this chapter are mainly based on information from tabulated data on income declarations made for tax purposes. Of course, we do not claim that this data is exempt of error as at least some individuals are likely to be subject of an incentive to under-declare their real income. The real value of this data is that it provides a credible lower bound estimate on top incomes. Indeed, we observe that despite its flaws, tax data reports considerably higher density for top incomes compared to previously existing estimates. Furthermore, another advantage of working with this data is that it is not part of a stochastic process (as survey-data is). It can thus be treated as including the whole target population, at least for a part of the distribution.

In Chile, personal income tax has two main components: the *Impuesto Global Complementario* (IGC) and the *Impuesto Único de Segunda Categoría* (IUSC).

[6] In addition, we observe that the progressive exclusion of most capital gains from the definition of taxable income around the year 2001 (see Appendix A.1) does not have a substantial impact when comparing top shares that are estimated with and without capital gains (see Figure A.1).

[7] See the difference between income definitions Y_{Rlzd} and $Y_{\mathrm{RlzdNatAcc}}$ in their paper.

The former is the most comprehensive of the two, as virtually every individual resident is required to file it once a year. The latter is the tax paid exclusively by people receiving labor income (wages or pensions). It is generally declared to the tax authorities on a monthly basis by third parties, most of whom are employees of organizations, that is, dependent workers.

Since 1972, individuals receiving labor income from a unique source are not obliged to declare the IGC. This implies that the IGC series of data, for which we have the farthest-reaching statistics (from 1964-2015) and which constitutes our main data source, excludes data for some individuals since 1972. However, we have access to a Consolidated series (2004-2015), which includes income declarations from both the IGC and IUSC taxes without double counting. Hence, the estimates displayed in section 1.3 are built using both the IGC series and the Consolidated data series. Estimates for years prior to 1972 are estimated directly from the IGC series. Estimates for years 1972-2003 are adjusted by the average error that is observed in years where the two series overlap (2004-2015)[8] and estimates for years 2004-2015 are estimated from the Consolidated data series.

Both the IGC and the Consolidated series come in tabulated form. That is, every year there is a table in which the population is arranged by income-intervals. They contain information on marginal tax rates, quantity of people and total income declared at each interval. The information is the same every year, but the level of interval-aggregation differs depending on the year. For instance, for the early years (1962-1981) the IGC data that was transcribed from official publications divides people into a range of 4 to 20 income intervals.[9] The next span in the same series (1990-1995), which was provided as unpublished data by the tax agency, divides people into 15 to 20 intervals. The most detailed period in the series is 1996-2009, which is also unpublished, and separates declarations into 43 to 65 intervals.[10] For the last five years, we use information that is available online on the tax agency's website, where taxpayers are divided into eight intervals. In the Consolidated series (2004-2015), every year the population is divided into eight subsequent intervals. Furthermore, there are missing years in our dataset. Specifically, the year 1977 (1978 tax year) could not be located,

[8]When comparing results for the 10 years that have both tabulations we find a fairly constant error of about 8% (less than one percentage point) of the top 1% share value. This information, along with error estimates for other top shares, is used to adjust estimates that are calculated from the IGC series.

[9]Official publications refer to a report called *Boletin de Estadística Tributaria.*

[10]This series includes information on realized capital gains declared by income-bracket for the period 1998-2009. We use that information to build Figure A.1

even in the headquarters of the tax agency itself, or in any of the major libraries. This punctual discontinuity may be odd, but the disappearance of data covering the 7 years between 1982 and 1989 is even more intriguing. In any case, this kind of situation is to be expected in a dictatorship scenario. After all, tax returns are the only public traces left by the very rich.

Total Income Control

In order to compute top income shares we need to estimate total income for the whole adult population (the denominator), yet our best series of tabulations on fiscal income declarations only cover about 70% of the adult population in most recent years (figure A.4). In our second best series, which is the base of our estimates before 2004, only around 15% of total adults are represented and the share decreases the further we go back in time (figure A.5). We thus need to build an estimate, for every year, that approximates what would be the aggregate amount declared if every resident adult filed a tax declaration. Following Atkinson, Piketty, and Saez (2011), there are basically two ways to build such estimate. The first option is to use the total amount declared by tax filers after adding some income to account for non-filers. The second option is to build an estimate of total household income from National Accounts. It should follow the definition of fiscal income used in tax data as closely as possible. In this chapter we use a combination of both these options, as neither of them would be suitable if used alone.

In Chile, National Accounts are detailed enough to build the second type of estimate for the period 1996-2015. Table 1.2 displays the specific items included in its definition. It is equal to the gross balance of primary income received by households, plus social benefits other than transfers in kind received by households, less social contributions paid by households (which includes those at the expense of both employers and employees), less attributed property income for insurance policy holders, and output for own final use. This latter item mainly consists of imputed rents and the consumption of goods produced within households, both of which do not produce actual income.

Because aggregates from National Accounts are often used as benchmark, one could be tempted to use the definition of income presented in table 1.2 directly as a denominator for top shares. However, the figure that is obtained appears to be excessively high compared to the total income-declarations in tax data (figure A.3). If we were to use it, we would incur in a sizable and unjustifiable

Table 1.2: Total Personal Income in National Accounts

	Total Fiscal Income	
(=)	Balance of Primary Income, received by Households, gross	(B.5g)
(+)	Social Benefits other than Transfers in Kind, received by Households	(D.62)
(−)	Social Contributions paid by Households	(D.61)
(−)	Attributed Property Income for Insurance Policy Holders	(D.44)
(−)	Output for Own Final Use (≈ Imputed Rents + Consumption of own Production by Households)	(P.12)
(−)	Consumption of Fixed Capital, Households	(K.1)

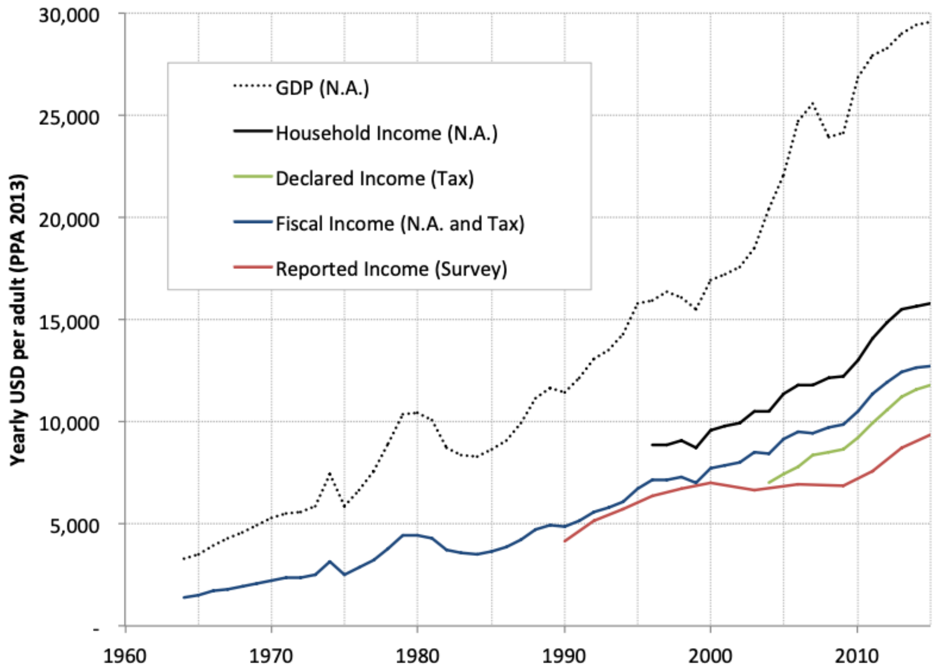
Note: Compiled by the authors

bias. It would be equivalent to imputing the whole difference between what is measured by tax data and National Accounts to the bottom of the distribution, which would result in a considerable underestimation of top income shares. The difference between national accounts aggregates and declared incomes is often interpreted as being due to evasion, avoidance or underreporting.[11] According to Fairfield and Jorratt De Luis (2016), most of the discrepancy comes from two specific items: distributed profits (which they report to be 3 times higher in National Accounts) and independent income (1.5 times). Both of these income types are also found to be highly concentrated at the top of the distribution. Ideally, as is done in Fairfield and Jorratt De Luis (2016), we would scale declared income to fit National Accounts' aggregates. However, due to data constraints, we cannot implement this type of adjustment.

As an attempt to overcome this limitation we proceed to construct, for year 2015, an estimate of total income based on the first option in Atkinson, Piketty, and Saez (2011). That is, we assume that the near 30% of non-filers have a positive but modest income equivalent to 20% of the average declared income (as in Piketty and Saez, 2003). This amount is then added to total income declared to fiscal authorities. The total from year 2015 is then used as a base from which we assume that variations in total income are proportional to the aggregate obtained from table 1.2. For years prior to 1996, again due to data limitations, we consider a third strategy which assumes that total income is a fixed part of GDP, which is

[11]This is in fact the reason behind CEPALSTAT's survey adjustments in the region (Bourguignon, 2015) and the up-scaling adjustments of fiscal income made by Fairfield and Jorratt De Luis (2016).

Figure 1.2: Comparing Income Concepts (per adult)



When data is available, total income declared to tax authorities (green line) appears low compared to the benchmark from national accounts (black line). The total income that is used as a denominator for top shares (blue line) is between both these figures. Its level for 2015 is equal to total fiscal declarations plus a low yet positive income accounting for non-filers. Its evolution is symmetrical to that of the benchmark for years with data. It is a fixed proportion of GDP for years before 1996.

its average value in years with data (42.6% of GDP).[12]

Figure 1.2 shows that, in 2015, Chilean GDP per adult is close to 30 thousand USD, according to National Accounts (dotted black line). Little more than half of that figure corresponds to the aggregate presented in Table 1.2 (black line), which should be the benchmark for the income of the household sector. However, the total amount of income declared to tax authorities in our most comprehensive series corresponds to less than 12 thousand dollars per adult (green line). The estimate of total income we use for computing top income shares (blue line) falls between the benchmark of national accounts and the aggregate total declarations to tax authorities.[13] Furthermore, we can see that disposable income declared in the un-adjusted version of the CASEN survey, which is supposed to represent the whole population, is surprisingly low compared to any of the other figures.

[12]Figure A.3 displays the total fiscal income, the declared income in both series, and the aggregate from SNA, for each year for which information is available.

[13]For a sensitivity analysis of top income shares to total income, see figure A.11

1.2.2 Tax Incentives and Undistributed Profits

Some specific tax incentives should be considered when analyzing the distribution of Chilean personal income. Before 1984, the profit of companies with traded stock was subject to a special tax (the *impuesto adicional*) that was the anticipation of the income tax over distributed profits (Cerda et al., 2014). This setup did not provide major incentives to profit retention by big firms because the income tax was already paid before dividends were actually distributed. However, since 1984, the Corporate tax of virtually all companies operates as a withholding on personal income tax on distributed profits; that is, corporate tax represents a credit against personal income tax. As a result, profits that are retained within the firm are subject only to corporate tax, while distributed profits may be subject to considerably higher marginal tax rates. This is because dividends are part of the personal income tax base (Fairfield, 2010; Fairfield and Jorratt De Luis, 2016). Hence, instead of distributing dividends, the owners of big companies can access less-taxed revenue via the realization of capital gains over stocks, which are mostly exempt of income tax. Furthermore, in response to the data structure, individuals often create investment societies exclusively for tax purposes, generally limiting declared income and using retained revenue indirectly (Jorratt De Luis, 2009).[14].

Although the gap between the corporate tax and the top marginal tax rate has been reduced over the course of the last 25 years, it has remained high throughout the whole period. In 1990, the difference was exactly 40 percentage points, with a corporate tax of 10% compared to a marginal top rate of 50%. However, the gap is progressively being reduced, and during the greater part of the 2000s it stayed at 20 points, with corporate tax of 20% and the top marginal rate of personal tax at twice this figure (Figure A.9).

Alvaredo, Atkinson, Chancel, et al. (2016) define the aggregate amount of pre-tax undistributed profits as the net primary income of the corporate sector in National Accounts (both financial and non-financial). According to this definition, it appears that undistributed profits increase substantially as a share of GDP during the period for which detailed data exists (1996-2015). It increases from around 4-5% during the late 1990s and early 2000s to 8-10% during the past five years. The most significant increase appears to take place around the middle of the 2000s. Figure A.6 displays the evolution of both aggregate

[14]This was partly changed in the 2014 tax reform which is still in the process of coming into effect. Two new tax regimes were created for income tax: a semi-integrated system and an attributed system. In the latter, the incentive diminishes, while in the former it partly remains. However, the income tax system is no longer fully integrated.

undistributed profits and total household income as a share of GDP.[15] Their apparent symmetric progression suggests that there may be a substitution effect, where a part of household income would have been progressively shifted to be recorded as undistributed profit. As corporate ownership is highly concentrated in Chile (Fairfield and Jorratt De Luis, 2016), a substitution effect would likely introduce a noticeable downward bias in the trend of personal income inequality, at least according to measurements of both household surveys and fiscal income data.[16]

In order to address this particular issue, we proceed in section 1.3.2 to the imputation of undistributed profits to the fiscal income distribution. The purpose of this is to check for potential biases to the measured trend of income inequality.

1.2.3 Total Population and Interpolation Method

In order to calculate income shares accurately, we have to determine which individuals will be considered in our total population. The main issue here is to establish whether income declarations are filed on an individual or household basis. Income has been declared individually for the full period under study. Hence, for our estimations the population total will be, as is common in the top incomes literature, individuals over 20 years old. Our source is the World Bank public database. The method we adopt to interpolate between given points in fiscal tabulations is different from the classic Pareto Interpolation and Mean Split Histogram that were generally used in earlier fiscal income studies. Here, we use the Generalized Pareto Interpolation (GPI), which is described in detail by Blanchet, Fournier, and Piketty (2017).

Essentially, the technique allows the income distribution to have a varying Pareto coefficient (average income above a given threshold divided by the threshold itself) that changes across the income distribution, using the information for each income interval of the tabulation. The Pareto coefficient usually follows a U-shape. The GPI is a non-parametric method that has been shown to produce

[15]Table A.2 presents the numbers behind Figure A.6, as well as a comparison between total undistributed profits and our unadjusted total fiscal income (ranging from 7% at the lowest point to 33% at its highest).

[16]The figure for aggregate undistributed profits that is presented in this subsection and imputed in section 1.3.2 is always net of capital depreciation. Moreover, it should be noticed that the income of pension funds is not included in the aggregate. In the latest system of national accounts (SNA 2008), the net primary income of the corporate sector has already been subtracted of the income of pension funds, which is imputed during property income operations (D.4) to the Household sector (more detail on this in paragraph 7.147 of OECD (2008)).

more precise estimates than previous techniques, especially while extrapolating to higher shares of the population. In their paper, the method is compared empirically to previous ones by conducting experiments involving comprehensive tax micro data in parallel with tax tabulations from the United States and France, for the period between 1962 and 2014.

1.3 Results

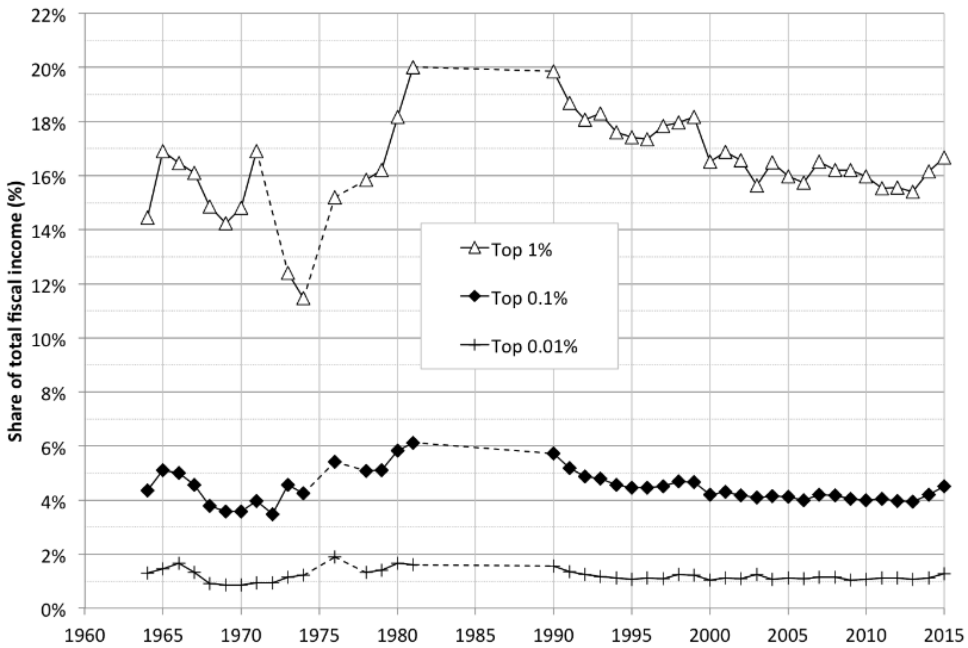
This section comments the evolution of top income shares estimates for both our adjusted and unadjusted series. It should be noted that although these series give relevant information on income concentration trends, they are limited and should not be considered as a satisfying measure of inequality. This is because our estimates systematically overlook the evolution of inequality that happens elsewhere in the distribution (Atkinson, Piketty, and Saez, 2011).

1.3.1 Unadjusted Series (1964-2015)

This subsection provides historical context for our unadjusted estimates on top income shares. Figure 1.3 presents the progression of the top 1%, 0.1% and 0.01% shares of income over the period 1964-2015, while Figure 1.4 provides estimates for the richest 10% of the population for the period 2004-2015. We can observe that the general trend in the 1960s is towards decreasing income concentration. The direction is then inverted towards a steady increase in concentration around the beginning of the military dictatorship in 1973. Regrettably, we cannot comment on the evolution of income concentration over the course of the 1980s, as the information on tax declarations seems to have disappeared for those years. Since the return to democracy in 1990, the unadjusted series shows a generally decreasing trend until 2013, in which we can observe a relatively small but noticeable reversion. Although the decreasing trend that is observed over the last 25 years appears to confirm what is observed in household surveys (Appendix 1.1.1), that information should be treated carefully, because it is observed with a definition of income that excludes retained profits.

Early years (1964-1973) In Chile, as in Latin America and the rest of the world, the 1960s were a time of increasing political polarization. The recent Cuban revolution (1959), combined with decades of increasing demands for justice by workers influenced by socialist philosophy, put social issues at the center of the

Figure 1.3: Top 1%, 0.1% and 0.01% Shares of Fiscal Income (1964-2015)



Authors' calculations using tax data, national accounts and population estimates. Dashed lines connect points between which there is at least one year of missing information.

political debate. At the same time, the building of the Berlin wall (1961), the Cuban missile crisis (1962), the Brazilian military coup (1964), and other ongoing armed conflicts relating to the cold war contributed to levels of tension and anxiety among civilians. In the national political context, two consecutive left-wing presidents governed Chile during this period: E. Frei-Montalva (1964-1970) and S. Allende (1970-1973). The latters' term was brought to an abrupt end by a *coup d'etat* in 1973. Both presidents are recognized for implementing socially oriented policies. Among the most high-profile of their reforms were land reform and the nationalization of the domestic mining industry, and the radical nature of these reforms gradually increased over the course of the decade. Although this chapter does provide some historical context, we do not claim to identify a causal effect of policy reforms on concentration of income.

The tax reform of 1964 sets the starting point for the series displayed in Figure 1.3. This reform introduced, among other things, the first legal definition of income for tax purposes, and raised the top marginal rate from 35% to 60%.[17]

[17]Although there is information available on income declarations for two earlier years (1962 and 1963), we judge them to be inconsistent with the rest of the series, as the reform theoretically affects income received since 1964.

Figure 1.3 shows that the top 1% share increases from 14.5% to 16.9% of total income between 1964 and 1965. However, after 1965, a generally downward trend continues for almost a decade, reaching its lowest point (12.4%) at the end of the period in 1973. Only one discrepancy appears in this trend, in 1971, with a relatively abrupt increase in top shares during that year. Given that this was the first year of the presidency of S. Allende, typified by the implementation of radical socialist reforms, it is difficult to imagine that the richest individuals increased their share of total income. One possible explanation is an increase in enforcement of tax collection, which may have targeted the rich in particular.[18]

There is an extreme lack of data for the year 1972, as the country was going through a large-scale socio-economic crisis.[19] Only 0.3% of the total adult population declared income to the tax agency (Table A.1), which is not enough to be able to estimate the share of the richest 1% of the population. Figures for the top 0.1% and 0.01% shares are thus heavily compromised for that year.[20]

Dictatorship (1973-1990) In the wake of the military coup of September 11th 1973, a government board composed mainly of military generals was created to govern the country. However, A. Pinochet quickly took over power and was named President by a decree passed at the end of 1974. The military dictatorship lasted 17 years. Inspiration for the government's economic policy was closely related to monetarist ideals. The main reforms included the privatization of public firms, budget cuts for social spending, a change of currency, and the liberalization of the labor market. The latter was enforced by violent repression of demonstrations and union activity.

The trend in income concentration during this period is clear and stable, at least according to the available data. The top 1% share increases 8.5 points between 1974 and 1981, rising from 11.5% to 20% over 7 years. We only observe

[18]We exclude the possibility of this increase being due to variations in the denominator of our top income shares, as GDP per capita increased during that year (Larrain and Meller, 1991).

[19]Between 1970 and 1973, a large-scale operation to destabilize the Chilean economy was taking place, coordinated jointly by US officials and the Chilean economic elite. In a report released on September 18th 2000, the CIA describes in detail its activities in Chile intended to prepare the ground for a military coup. These interventions included distribution of propaganda in association with the local press, financing of the political opposition, planning the coup alongside Chilean military officials, providing intelligence, and even offering large sums of money to Allende in exchange for his resignation (`https://www.cia.gov/library/reports/general-reports-1/chile`).

[20]In 1972, the minimum threshold for tax exemption doubled. Moreover, those who perceived wages or pensions from a single source were no longer obliged to declare under the IGC-tax, but rather under the IUSC. The emigration of many wealthy individuals that year may also have contributed to the phenomenon.

a slight decrease in inequality between the first and second years of the period. This rise can be mostly explained by that year's increase in the denominator of top shares: total fiscal income (Figure A.2).

Figure 1.3 does not display top shares for year 1975. This is because we consider estimates from this year to be somewhat inconsistent, perhaps due to an error in the construction of tabulations. Indeed, when that year is included, the top 1% share jumps to an ephemeral 25% of total income for that particular year. However, the increase in total income declared to the tax agency during that year does not correspond to any sizable change in the filing population (Figures A.3 and A.5). The most likely explanation for the phenomenon is that the country was going through one of the most serious economic crises of recent decades. Indeed, real GDP per capita growth was less than negative 10% in 1975, and inflation also reached extreme levels (Figure A.2). Of course, one could expect top incomes to be more resilient to this crisis than lower incomes, which would explain the jump, but the resulting estimates appear exaggerated. Since our estimates of total income are based on a fixed share of GDP for these years, we judge them to be rather sensitive and not sufficiently reliable in this kind of exceptional situation.

Inconveniently, data for the year 1977 (1978 tax year) could not be found. However, what is even more remarkable is the absence of data for the whole period between 1982 and 1989. Tabulations for those years appear to have either disappeared or never existed. It is during the 1980s that Sanhueza and Mayer (2011) document the highest concentration of income, however we are unable to comment on that specific period. Moreover, it is in year 1984 that the most significant tax reform in our series takes place. In the name of boosting savings and investment, incentives for profit retention were introduced, along with the core of the integrated tax system that has prevailed throughout the last 25 years of democracy (see Section 1.2.2 and Appendix A.1).

Return to Democracy (1990-2015) In 1990, Chile returned to democracy in the midst of the most accelerated economic boom of its history.[21] The transition occurred in a relatively peaceful way, as it was organized in a way that ensured political stability as a priority. At the beginning of this period, most of those who had participated in the military government organized themselves into right-wing

[21] The so-called "Chilean miracle" refers to the period of high economic growth rates between 1985 and 1997. It corresponds in part to the fast economic recovery following the economic crisis of 1982, and in part to actual growth relative to the level of GDP per capita in 1981.

political parties.[22] In parallel to this reshuffle, opposition parties were legalized. Furthermore, a succession of four center-left Presidents held office over the next 20 years, followed by a center-right President between 2010 and 2014. The majority of reforms over the period were aimed at the expansion of social security coverage and the reduction of poverty (Contreras and Ffrench-Davis, 2012). Nonetheless, the foundations of the socio-economic model established by the dictatorship remained in place, with reforms in key sectors (e.g., education, health, pensions, housing) were mostly based on private markets.

As Figure 1.3 shows, the concentration of income among the richest 1% of the distribution generally decreases over the democratic period, from 19.9% in 1990 to 16.7% in 2015. This is a fall of 3.2 points over the period. Looking in greater detail, the most accelerated decrease in the span takes place during the first half of the 1990s. Indeed, the 2.6 point decrease in inequality between 1990 and 1996 represents four fifths of the total fall during the democratic period. Furthermore, a slight increase (0.9 points) in top shares can be seen between 1996 and 1999, including at the point where the impact of the Asian crisis was at its most severe in Chile.[23]

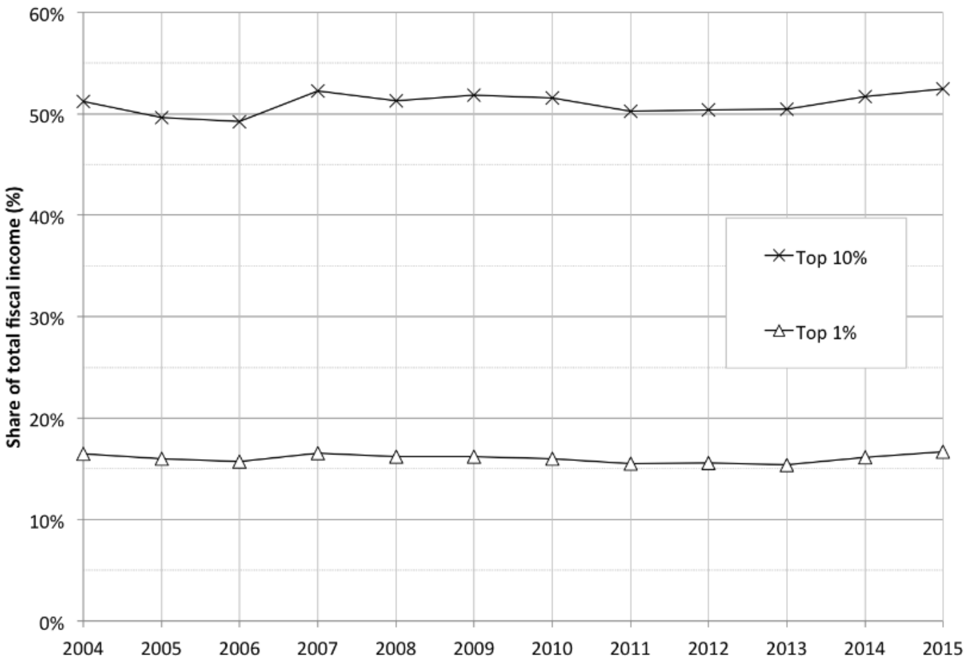
A relatively sizable drop occurs in the top 1% share between 1999 and 2000. This decrease of 1.6 points is the most abrupt recorded since the return to democracy, but its interpretation is not straightforward and should be treated carefully. There is one deduction on the taxable base, intended to enhance economic growth in the housing sector, which could explain at least a part of this phenomenon. Since the end of 1999, and for a limited period of time, people buying new properties with a mortgage were able to deduct a considerable share of their mortgage dividends from their taxable income (Law Nr. 19,622). The benefit was effective until the full value of the mortgage was repaid, presenting an attractive opportunity for investors. The only condition to access the benefit was to buy a new "affordable property", which produced un-taxable income when rented.[24] Over the following years, the top 1% share appears to fall more or less steadily. As mentioned earlier, by the end of the democratic period, the trend had become inverted. Between 2013 and 2015, a considerable increase in the top 1% share is recorded (1.2 points), returning to the same level of inequality that

[22] Only a portion of those who participated directly in ordering human rights violations were tried and imprisoned. Pinochet himself, however, remained as a lifelong senator and retained his post as general commander of armed forces until 1998.

[23] Chilean GDP growth was negative for years 1998 and 1999.

[24] This is a somewhat comprehensive definition. Essentially, a property was considered "affordable" if it comprised less than 140m^2 of usable space.

Figure 1.4: Top 10%, Share of Fiscal Income (2004-2015)



Authors' calculations using tax data, national accounts and population estimates. Estimates for the top 10% share are available for a shorter span, as they are built exclusively using "Consolidated data" (combining declarations for both the IGC and IUSC taxes), beginning in 2004. This is the only series that includes more than 10% of the population over the taxable threshold.

prevailed 10 years previously.[25]

Figure 1.4 displays the unadjusted share of the top decile, which varies between 50% and 53% of total fiscal income over the period. Upper shares – as the top 0.1% and 0.01% – generally follow the same trends described by the top 1%, but with a lesser degree of variability.

1.3.2 Adjusted Series Including Undistributed Profits (1990-2015)

In this section, we build a simple yet straightforward approximation of the trend effects caused by imputing undistributed profits to a relatively long set of estimates on personal income concentration. We impute aggregates from National Accounts by making assumptions based on distributive information found in Fairfield and

[25]It is not clear, however, how this information should be interpreted. We judge that 2 points are not sufficient to consider this a sustained trend. Moreover, a further tax reform introduced in 2012 could be driving this phenomenon, mainly by limiting recourse to special tax regimes (see Appendix A.1).

Jorratt De Luis (2016). Previous works on Chilean top incomes have highlighted the relevance of undistributed profits in the study of local income inequality. This seems to be a priority due to the presence of tax incentives favoring artificial retention of profits within corporations (López, Figueroa, and Gutiérrez, 2016; Fairfield and Jorratt De Luis, 2016). Such a phenomenon is indeed likely to have an impact on both the level and the trend of inequality estimates (see Section 1.2.2).

In order to impute total undistributed profits to individual income distribution, we take estimates on the distribution of the accumulated stock of undistributed profits since 1984 from Fairfield and Jorratt De Luis (2016: Table A.9). They found that in 2005, the richest 1% of the fiscal income distribution owned 75% of that stock (using virtually the same definition of fiscal income as ours). Their next observation – in 2009 – records a lower concentration of 69%. We must then make different assumptions in order to construct upper and lower bound estimates, by conjecturing that flows of undistributed profits follow fairly closely the same pattern of concentration as the stock.[26]

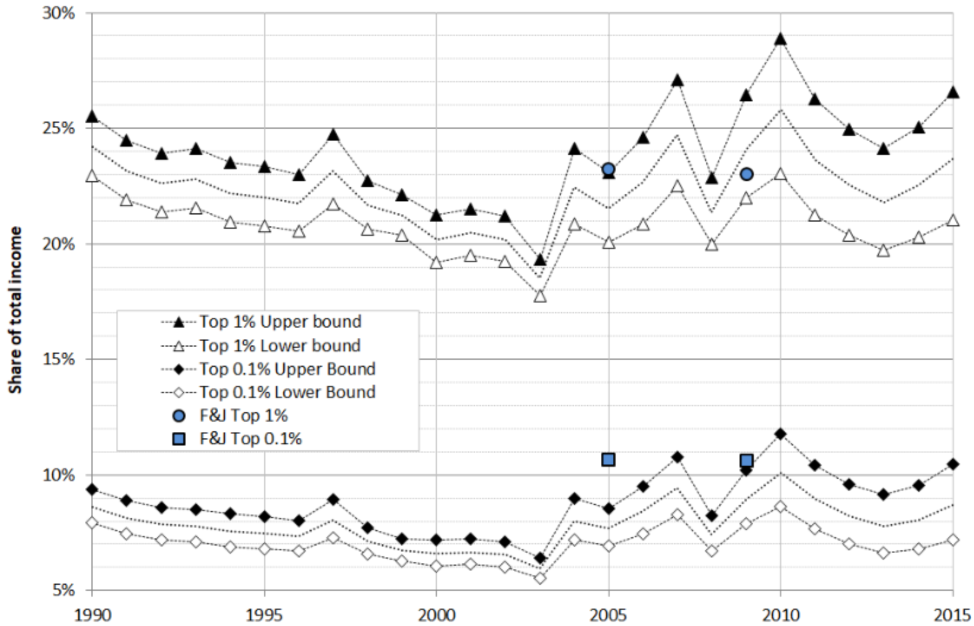
Figure 1.5 displays the adjusted estimates for the top 1% and 0.1% shares of total income, including upper and lower bounds. Both our upper and lower bounds on adjusted top income shares assume that undistributed profits follow a constant pattern of concentration between 1990 and 2005. Between 2005 and 2009, they both mimic the decreasing trend of accumulated profits observed by Fairfield and Jorratt De Luis (2016). However, after 2009, our lower bound estimate assumes that the same linearly decreasing trend will continue until the final year, while the upper bound estimate assumes its constancy throughout the same period. In addition, as one could argue that stocks of undistributed profits may be more concentrated than flows, the lower bound estimate assumes that for the whole period, concentration in flows is two thirds of the concentration in stocks.[27] Our upper bound estimates assume that flows of undistributed profits are concentrated to the same degree as accumulated stock.[28]

[26]Total amounts of undistributed profits are available in Table A.2.

[27]For instance, in 2009 the richest 1% of the fiscal income distribution owned nearly 70% of the stock of undistributed profits. Assuming only two thirds of the concentration would mean that the richest percentile owned nearly 46% of the flow of undistributed profits during the same year.

[28]In their paper, Fairfield and Jorratt De Luis (2016) find that nearly one third of their estimate of accrued profits (the sum of distributed and undistributed profits) is owned by foreigners. They thus exclude that part from the total for imputation. However, we judge that type of adjustment to be unnecessary in our case, because the estimate of pre-tax undistributed profits we use has already been subtracted from reinvested income on foreign direct investment (D43). Furthermore, our definition of undistributed profits takes into account profits held

Figure 1.5: Top 1% Share with Undistributed Profits, Upper and Lower Bounds (1990-2015)



Source: authors' estimates using tax data, detailed National Accounts (1996-2015) and Fairfield and Jorratt De Luis (2016). Note: in each situation, the whole value of undistributed profits is imputed to the fiscal income distribution. Upper bounds assume that yearly flows of undistributed profits are concentrated in top groups to the same degree as the accumulated stock from 1984 (F.U.T.). Lower bounds assume flows to be two thirds as concentrated as stock. The dotted line represents a central tendency, which is estimated as a geometric average of upper and lower bounds. In the absence of detailed National Accounts prior to 1996, the amount of undistributed profits in those years is estimated to be nearly 4.8% of GDP, which is the estimate for 1996. Estimates from Fairfield and Jorratt De Luis (2016) using their definition Y_{AcrdProf} are displayed for comparison.

Perhaps the most striking finding in Figure 1.5 is that despite conservative assumptions, considerable changes in trend directions emerge relative to unadjusted estimates in all cases. Indeed, even lower bounds, which are remarkably conservative, contradict the decrease in income concentration after the year 2000 that is observed in unadjusted estimates. It thus appears reasonable to conclude that income concentration, including undistributed profits, likely follows a U-shape during the last 25 years for which we have data. Income concentration would decrease over the course of the 1990s and then increase fairly steeply after the year 2000.

Moreover, Figure 1.5 displays comparable estimates by Fairfield and Jorratt

abroad by Chilean nationals.

De Luis (2016: Table 1).[29] It appears that their estimates are fairly close to ours in level, as almost all of them fall between our upper and lower bounds, including top 10% shares (Figure A.10). When studying the top 0.01%, however, our adjusted top shares appear to be considerably lower than theirs, as we record a concentration of 1.5% of total income in this group, while their estimates are closer to 5%. The underestimation of this particular part of the population may be due to the fact that their imputation is done using micro data, which allows re-ranking of the distribution after imputation.[30]

The imputation of corporate undistributed profits to individuals implies the acceptance of a definition of income based on an accrual logic. That is to say that we define income as a direct or indirect increase in net wealth, which is not necessarily realized in a market transaction. Following the same rationale, it could also be desirable to take into account accrued gains from the increasing value of dwellings, which are likely to be more equally distributed than business and rental income, thus probably having an equalizing effect. Due to data limitations we are unable to carry such imputation in the frame of this investigation. However, although we could expect income-concentration levels to be displaced with such operation, it is less clear if we should anticipate a noticeable impact on trends.

1.3.3 The Distribution of Income Growth

Unadjusted series (1964-2015) Figure 1.6 shows the evolution of real average income as an index of base 100 in 1964, in different groups of the population: the top 0.1%, the next 0.9% (P99-P99.9) and the rest of the population, which is the bottom 99%. Of course, these groups do not necessarily represent the same people every year, as a certain (but limited) degree of mobility between groups is expected to exist.

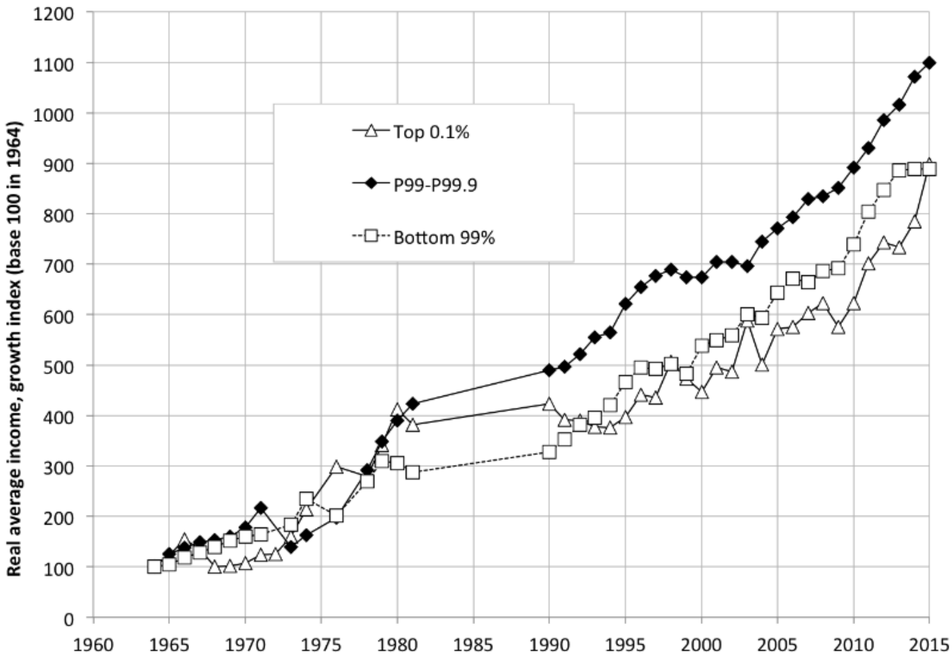
In Figure 1.6, the P99-P99.9 group is the one whose income grew the fastest over the whole period. It had its real income multiplied roughly 11 times, while both the top 0.1% and the bottom 99% saw their income multiplied around 9 times.

Since 1990, it appears that the fastest growing group is actually the bottom

[29] We display results for the definition of income they call Y_{AcrdProf}. Tables A.4 and A.3 display the numbers behind the upper and lower estimates, including the top 0.01%.

[30] Another difference between our adjustment and theirs which could affect trends is the data source we use to estimate undistributed profits. The authors use net accrued profits as declared by businesses to tax authorities, while we use National Accounts aggregates. Aggregates are often judged to be more accurate, although they do not incorporate distributive information.

Figure 1.6: Unadjusted Income Growth. Top 0.1%, Next 0.9% and Bottom 99% (Index base 100 in 1964)

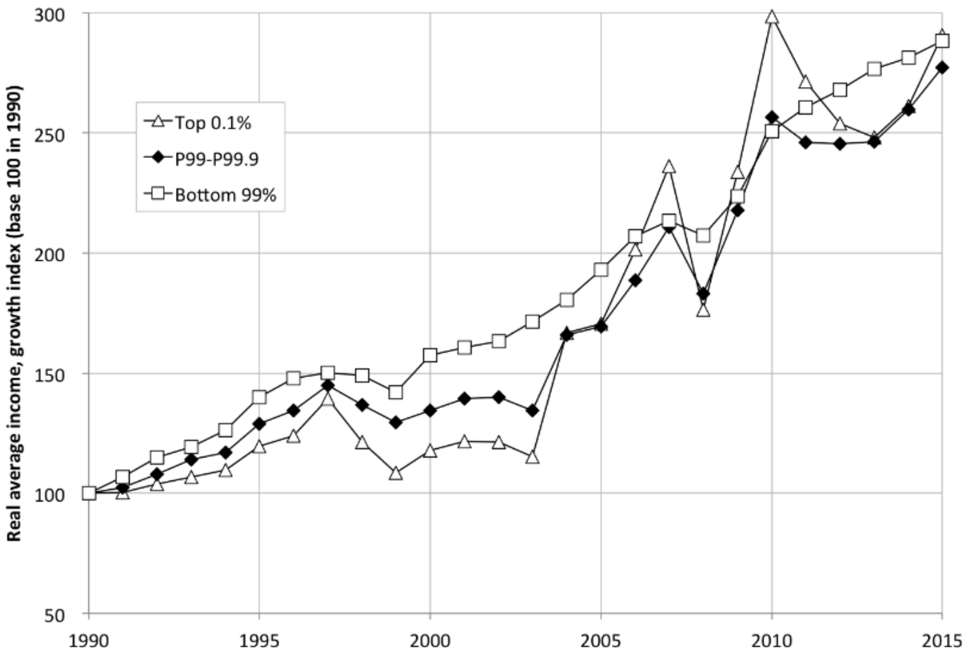


Source: authors' calculations using tax data, national accounts and population estimates. Note: the average income of the bottom 99% of the population is estimated residually using income information for the top 1% (tax data) and total income (National Accounts).

99%. Throughout the period, its real income increases about 2.7 times, while for the top 0.1% and the next 0.9% it increases around 2.1 and 2.2 times respectively. This finding is in line with the decreasing inequality that can be observed in surveys conducted since the 1990s (Appendix 1.1.1). Nonetheless, once again, Figure 1.6 does not include undistributed profits, and we therefore consider that it tells an incomplete story.

Series with Undistributed Profits (1990-2015) Figure 1.7 displays the average income of the same groups shown in Figure 1.6, but for a shorter period and including the imputation of undistributed profits, as described in Section 1.3.2. Although these groups have followed different paths over the 25 years, in the end there is no major difference between them in terms of total growth. Indeed, both the top 0.1% share and the bottom 99% have their income multiplied by a factor of roughly 2.9, while the P99-P99.9 group is not far off, with a factor of 2.8. These findings are in line with the U-Shape that is described by the top 1% share in Figure 1.5.

Figure 1.7: Income Growth Including Undistributed Profits. Top 0.1%, Next 0.9% and Bottom 99% (Index base 100 in 1990)



Source: authors' calculations using tax data, national accounts and population estimates. Note: the average income of the bottom 99% of the population is estimated residually, using income information for the top 1% (tax data) and total income (National Accounts).

Before making any conclusive statements about the growth distribution of income, it is worth stating that the bottom 99% is likely to be a somewhat heterogeneous group. Thus, a study of what happens in additional sections of the distribution could be interesting, but is not possible using our tax data due to the fact that it only covers a limited part of the adult population (Figures A.5 and A.4). A reasonable approximation of the median income of our distribution should be provided by the National Socio-Economic Characterization (CASEN) Survey if we assume that median income earners do not pay income tax and do not receive any benefit from undistributed profits.[31] A similar concept to the *base imponible* (Section 1.2.1) may be derived from the survey. When we compare the evolution of the CASEN Survey median (see Figure A.7) with the average income of the top 0.1%, it appears that they too have a very similar end point. The decrease in inequality that can be observed after 1990 is counteracted by a rapid increase from the year 2000 onwards, resulting in more or less equivalent growth.

[31]Figures A.5 and A.4 show that no more than 20% of the adult population has declared taxable income since 1990.

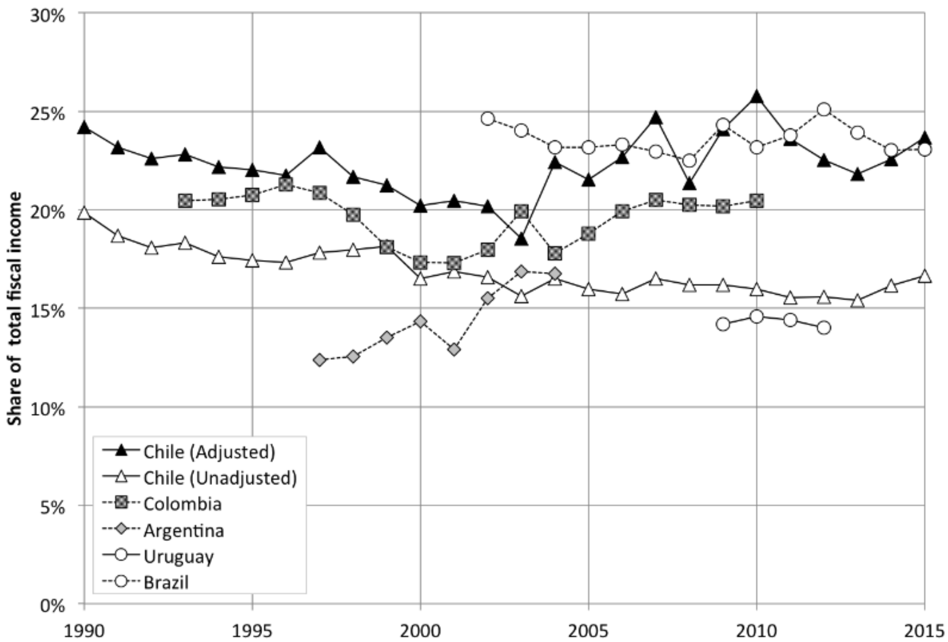
However, it should be noted that the period begins with very high inequality in 1990, and ends in 2015 with similar levels.

1.4 Comparison with Other Estimates

1.4.1 International Comparisons

Figure 1.8 compares both adjusted and unadjusted estimates of the Chilean top 1% share of income to other Latin American countries, which do not have the same incentives to profit retention than Chile.

Figure 1.8: Top 1% Share in Latin America (1990-2015)



Authors' estimates for Chile, Alvaredo (2010) for Argentina, Morgan (2017) for Brazil, Alvaredo and Londoño-Vélez (2013) for Colombia, and Burdín et al. (2014) for Uruguay.

The adjusted estimate, while following a decreasing trend, places Chile as the most unequal country in the region for the period 1990-2001. However, the Brazilian series starts in 2002, with higher levels of income concentration. Yet, the Chilean top 1% seems to catch up quickly during the following four years. From 2007, both countries alternate between the first and second place in the region. When comparing the unadjusted estimate, Chile ranks as the third most unequal country, after Brazil and Colombia, throughout the whole period. Furthermore, there is no distinguishable trend shared by the five countries.

The top 0.1% share of the adjusted series is generally above but relatively close to the level of concentration observed in Colombia (Figure A.8). Brazil leads the ranking with a top 0.1% share of around 11% of total income, which is generally 2-3 percentage points higher than the Chilean estimate. Contrastingly, the Chilean unadjusted estimates are always lower than any other country with comparable data in the region (the only exception being Argentina in 1997-1998). This observation seems odd and unconvincing, especially when compared to Uruguay, which is one of the least unequal countries in the Latin American region, with an official Gini coefficient lower than 0.4. We consider that this underestimation of higher top incomes is likely to be related to a Chile-specific institutional framework that disincentives the distribution of corporate profits in the form of dividends, discussed earlier in this chapter. Again, we interpret this as evidence for the need to take into account undistributed profits, especially in the Chilean context.[32]

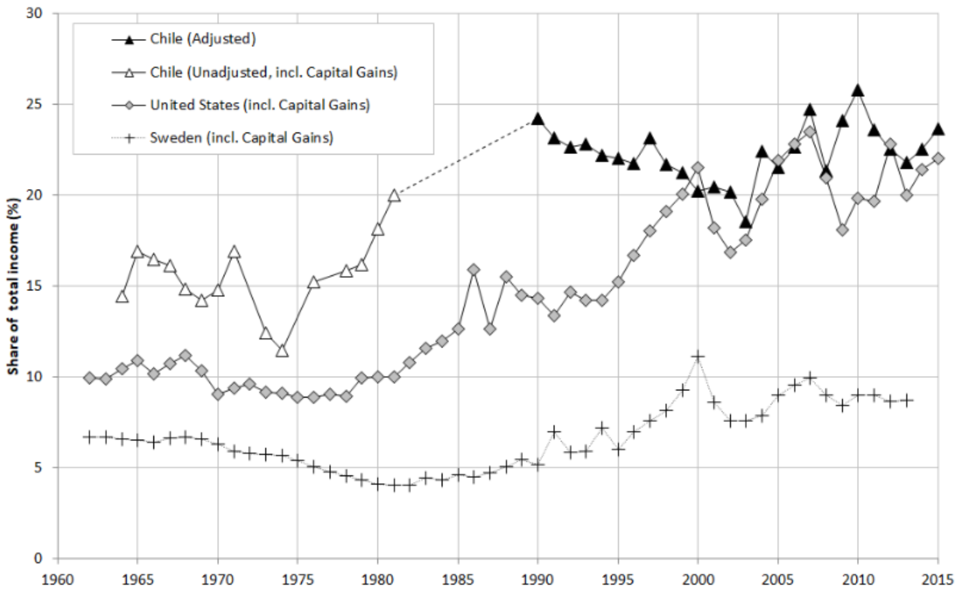
Figure 1.9 compares our estimates of the top 1% share over the long term with estimates from two developed countries: the United States, an icon among unequal countries, with a sizable increase in income concentration since the 1980s; and Sweden, a country with relatively stable and low levels of income inequality. Chile records a higher concentration than both countries, at least between the 1960s and 2000s. Furthermore, it appears that the increase in inequality in the USA in recent years has brought the country close to the level of income concentration that is recorded in the Chilean adjusted series. Both range between 20% and 25% of total income for the richest top 1%. For the years prior to 1990, even the unadjusted series for Chile is considerably higher than both developed countries, with nearly five points distance from the USA and ten from Sweden. Although Sweden experienced an increase in inequality from 1980 onwards, it unsurprisingly reaches levels of concentration that are considerably lower than those of Chile.

1.4.2 Local Surveys

This section measures the bias with which top incomes are underestimated in the most popular local household survey (CASEN). For years with sufficient

[32]Moreover, according to the most recent Forbes list (2017), Chile has the third highest number of billionaires in Latin America, with twelve. The country is only surpassed by Mexico, with 15 billionaires in a population more than seven times larger, and Brazil, with 43 billionaires in a population more than 11.5 times larger.

Figure 1.9: Top 1% Share Compared to Other Countries (1964-2015)



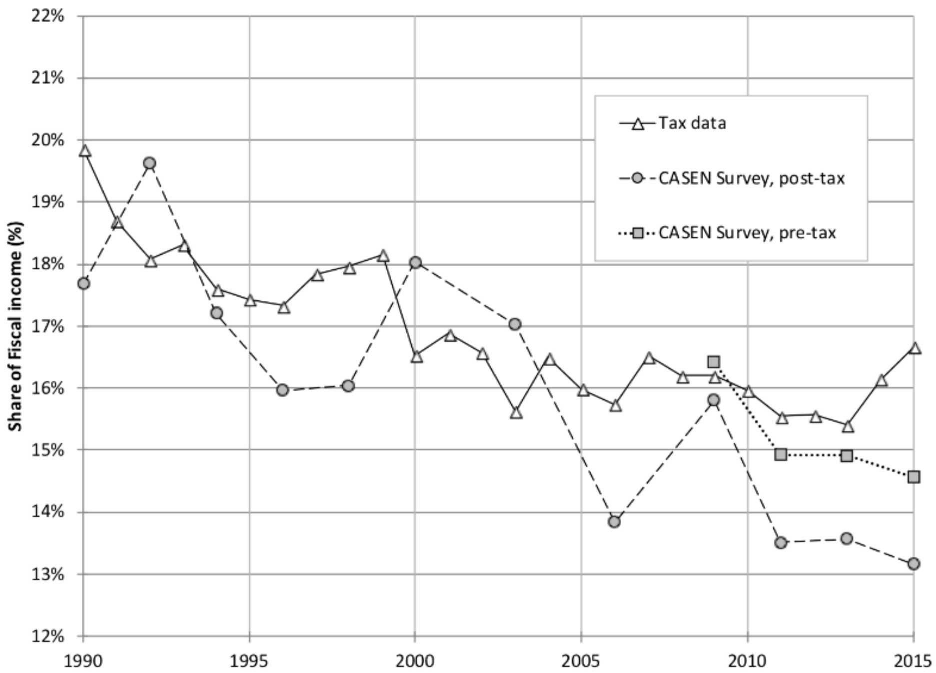
Authors' estimates for Chile, using tax data, National Accounts and population estimates from World Bank. Estimates for other countries come from `www.WID.world`. The Chilean adjusted series includes the imputation of undistributed profits. It corresponds to the central trend that is described in Section 1.3.2. Series for other countries all include realized capital gains.

information (2009-2015), we use the survey to build a definition of personal income that is comparable with that derived from the fiscal data. Perhaps the most important step in this endeavor is to obtain pre-tax income based on post-tax income. This retroactive transformation is non-trivial, as it involves several fiscal rules and different marginal tax rates to be applied. For this purpose we build on a similar work by Martínez-Aguilar et al. (2017). These estimates, along with a longer series with post-tax income, are compared here to our unadjusted tax data series (from Section 1.3.1). Both of our survey estimates are based on CASEN's *orginal* income series.[33]

Figure 1.10 compares top income shares from both survey and tax statistics between 1990 and 2015. As is to be expected, survey data estimates are generally lower and more volatile than those from tax data. However, in some years, survey estimates become close to or even slightly higher than tax estimates. This does

[33]CASEN's datasets included income adjustments to fit aggregated levels of national accounts. Both original and adjusted incomes are publicly available for each year for which data is available (since 2013). Bourguignon (2015) states that this kind of adjustment, applied by the Economic Commission for Latin America and the Caribbean (CEPAL), probably induces considerable biases for the study of the income distribution, and thus should be avoided.

Figure 1.10: Top 1% Share: Tax Data vs. CASEN Survey (1990-2015)



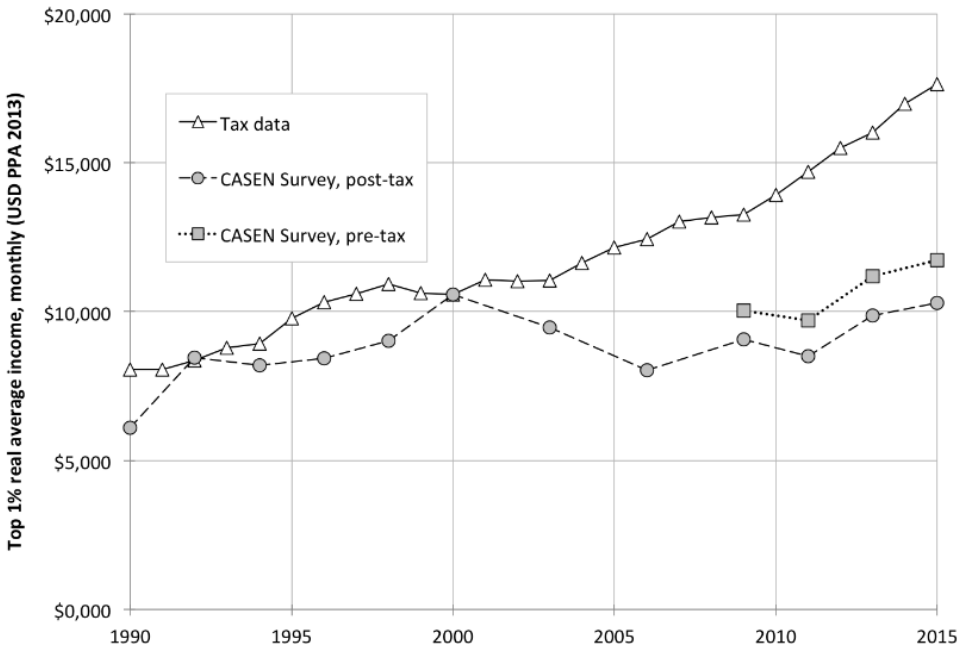
Authors' calculations using unadjusted series from Section 1.3.1 and CASEN Survey *original* income series.

not imply, however, that they are measuring the same phenomenon. There are considerable differences in the structure of the estimates in both the numerator and denominator of income shares. For instance, the total income in the tax series is always higher than in both survey-based estimates (denominator). On average, it is nearly 37% higher than in the pre-tax definition, and 43% higher than in the post-tax definition (between 2009 and 2015). The difference is greater when comparing the income of top groups (numerator). For instance, in the same period, according to the tax data, average income of the richest 1% is nearly 44% higher than in the survey's pre-tax series, and 63% higher than post-tax income. The gap is wider towards the top of the distribution (e.g., top 0.1% or 0.01% shares).

Figure 1.11 displays the evolution of average real income in the top 1% of each series (in 2013 PPA USD).[34] The distance between the tax data series and the survey post-tax series increases throughout the whole period. For the pre-tax series, we can draw the same conclusion, but only for a limited time period. It

[34]Comparing total or average income is virtually the same here, as the adult populations in both distributions are practically identical.

Figure 1.11: Top 1% Share: Tax Data vs. CASEN Survey (1990-2015)



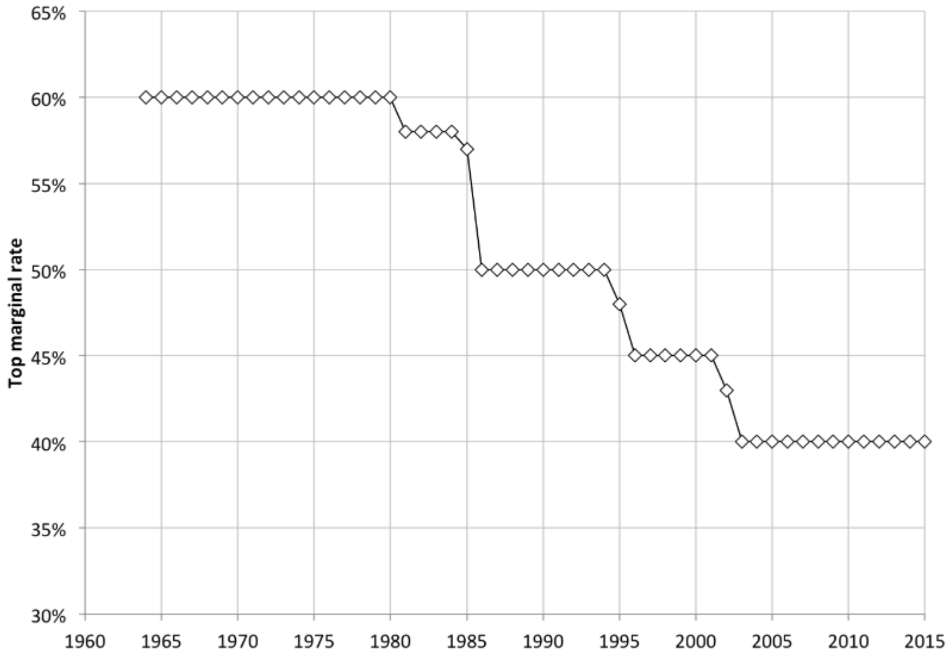
Authors' calculations using unadjusted series from Section 1.3.1 and CASEN Survey *original* income series.

seems that the bias towards lower top incomes in the survey is increasing over the period. Furthermore, survey estimates appear here to be more volatile than their tax data counterpart. This may be due to the sensitivity of survey estimates with respect to extreme observations.

1.5 Trend Robustness

As Alvaredo, Atkinson, Piketty, et al. (2013) state, a strong negative correlation is generally found in previous top income literature between top incomes and the top marginal tax rate. Some interpretations of this correlation are often used as arguments to deny the validity of top income trends. They all expect a negative correlation and try to explain trends as being caused by behavioral responses to tax rates. For instance, one of the arguments claims that a fall in top marginal tax rates offers less incentive to seek tax avoidance strategies, hence a parallel increase in top income shares could be caused by a simple statistical artifact (in the case of the USA, for instance). Figure 1.12 shows the evolution of the top marginal income tax rate for Chile between 1964 and 2015. Contrasting with what is expected in theory, in Chile the coexistence of a constant top marginal

Figure 1.12: Top Marginal Income Tax Rate (1964-2015)

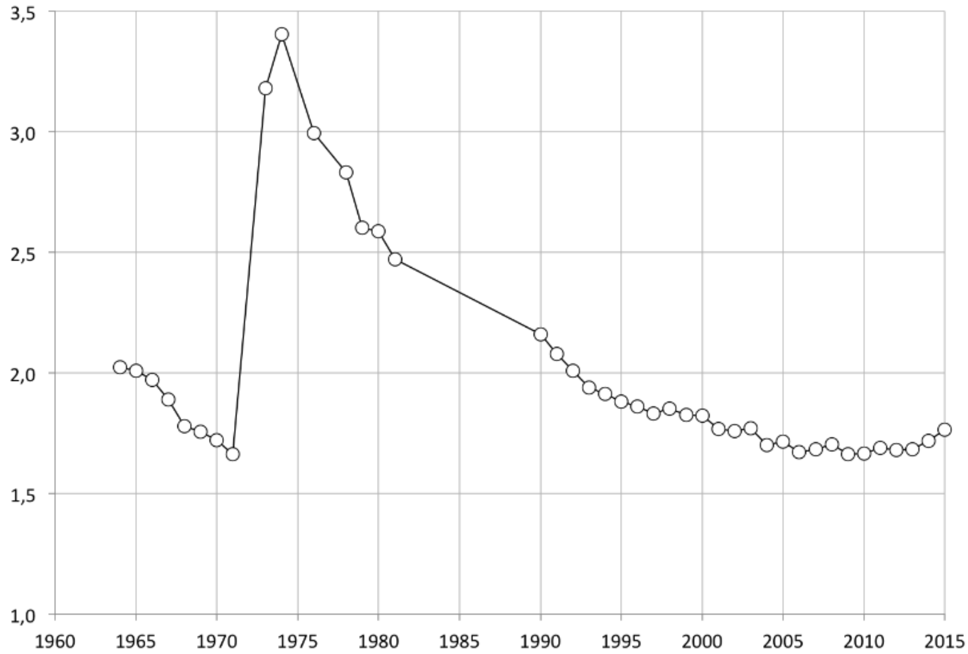


Source: *Servicio de Impuestos Internos* (SII), *Boletines de Estadística Tributaria.*

income tax rate with a period where top shares describe a U-shape (1962-1980), along with the positive correlation between the top marginal income tax rate and top shares over the last 25 years suggests that tax rates are not the main determinants of reported income levels.

Another recurrent criticism of top income studies is that top shares may be markedly sensitive to variations in total personal income. The argument is that the methodology used to calculate income totals from national accounts could be responsible for a major part of what we perceive as top share trends. This would be a problem when dealing with poorly detailed national accounts, as happens in the early years of the study, where total incomes are estimated as a fixed share of GDP. Considering this issue, Figure 1.13 presents the Pareto coefficient of the top 1% share for the whole span being studied. This coefficient is built as the ratio between the average income of the richest 1% divided by its threshold (P99). The main intention here is to look at inequality within the top 1% share independently of total income estimates. Figure 1.13 confirms a generalized decreasing trend of inequality during the 1960s. It then shows a sharp increase in inequality since 1973, followed by a progressive decrease in income concentration within the top 1%. This latter phenomenon occurs at the same time as the increasing overall

Figure 1.13: Inverted Beta Coefficient of Top 1% Share (1964-2015)

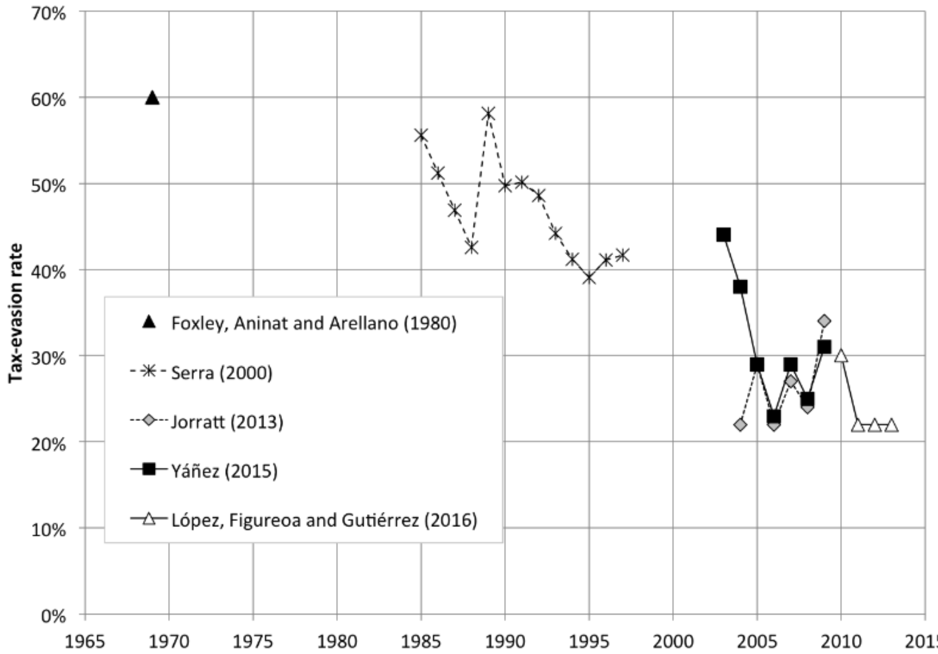


Authors' calculations using tax statistics.

concentration recorded in Figure 1.3. Finally, the democratic period continues with a decreasing concentration, which is interrupted and inverted in 2013.

As has been highlighted, theoretically our trends could be distorted by tax evasion, independently of its causes. Indeed, we should expect to find a negative relationship between tax evasion (or the share of undeclared income) and top income shares. In other words, the less you evade, the more you declare. Figure 1.14 brings together tax evasion estimates found in the literature relating to the "first category tax", which is the tax related to capital income. Globally, estimates seem to draw a downward evolution, especially in the period since 1985, where we have series with comparable estimates. This progression is in parallel to the observed decrease of top income shares since the return to democracy. As happens with marginal tax rates, the contradiction between expected and observed correlation shows that it is highly unlikely that our observed trends are significantly biased by tax evasion trends.

Figure 1.14: Tax Evasion Rates in Literature (1964-2015)



Sources: Foxley, Aninat, and Arellano (1980), Serra (2000), Jorratt (2013), Yáñez (2015) and López, Figueroa, and Gutiérrez (2016). Estimates refer to the *Impuesto de Primera Categoría* (IPC tax): a tax on capital income that for most of the period is integrated into personal income tax.

Conclusions

This chapter aimed to establish personal income concentration levels and trends from a historical perspective, based on the best data available. Our results, which are likely to be biased downward, still rank Chile among the most unequal Latin American and developed countries over the observable period. Chilean income concentration remains high throughout the whole period (1964-2015). Our estimates of top income shares show them to be resistant to changes in top marginal tax rates, to potential flaws in our total income estimates, and most likely to tax evasion trends as well. Furthermore, our fiscal data proves to be consistently better than the CASEN Survey at describing what happens at the top of the distribution. In fact, the gap between survey and fiscal averages of both total and top incomes increased throughout the 25 year period.

Additionally, we find that since the beginning of the 2000s, undistributed profits have been increasing considerably as a share of National Income. The parallel reduction of household income during the same period (% of National

Income) seems to confirm the concern voiced in previous literature that the Chile-specific institutional structure would incentivize retaining corporate profits within firms, while allowing their owners to access them in less detectable and therefore less taxable ways. We go further by finding that not only the level, but also the trend in income concentration may be biased. We question the decreasing trend in income concentration that appears in both survey and fiscal data estimates, at least since the early 2000s. The evolution of undistributed profits most likely played a role in pushing those trends downwards. It is thus crucial to study the joint evolution of corporate and personal income in order to analyze the whole picture and identify more informed inequality trends in the Chilean scenario. Naturally, further research is needed in order to assess whether this change in trend is found when analyzing a corrected version of other more comprehensive measures of inequality.

Chapter 2

Income Under the Carpet: What Gets Lost Between the Measure of Capital Shares and Inequality

This chapter measures the relative underestimation of factor income (i.e., capital and labor) in distributive data, with respect to national accounts' figures. I study a group of countries with harmonized surveys in the Luxembourg Income Studies database, but also tax and Distributional National Accounts (DINA) estimates, from the World Inequality Database, for the US. I find that households receive around only half of national gross capital income, as opposed to private and public corporations, and the trend decreases in most countries over 1995-2015 (panel, 19 countries). Due to heterogeneous non-response and misreporting, household-surveys only capture around 20% of this aggregate, versus 70% of labor income (sub-panel, 13 countries). This structure understates inequality estimates, which become insensitive to changes in the capital share (gross and net estimates) and its distribution. These distortions are weaker in tax data but still present, while DINA estimates are not subject to them by construction. I formalize this system in a novel theoretical framework based on accounting identities. I then use it to compute marginal effects and contributions to changes in fractile shares.

Introduction

During the last 50-60 years most developed countries recorded a substantial growth in their *capital share* of national income (IMF, 2007; Arpaia, Pérez, and Pichelmann, 2009; Piketty and Zucman, 2014; Karabarbounis and Neiman, 2014). That is, the part of macroeconomic income remunerating capital, as opposed to labor, has been growing for decades. This phenomenon occurs in parallel to the increase in personal income-concentration recorded by Atkinson, Piketty, and Saez (2011) and Alvaredo, Chancel, et al. (2018), which has led researchers to explore the relationship between national factor shares (i.e. Capital or Labor shares) and inequality. Fundamentally, the goal of this strand of research is to establish a better connection between macroeconomic aggregates, which are generally used to measure economic progress, and the economic distribution, which is often used to study well-being.

When capital income is more concentrated than labor income, intuitively, one could expect that an increase in the capital share should necessarily provoke an increase in total inequality. However, the relative concentration of factor incomes does not provide sufficient information to define such relation. As Milanovic (2017) puts it:

> A simple high concentration of a given income source will not guarantee that that source contributes to inequality. Unemployment benefits have a Gini which is generally in excess of 90 (since most people receive no unemployment benefits during any given year), but since recipients of unemployment benefits are generally income-poor, an increase in the share of unemployment benefits in total income reduces income inequality.

Since individuals receive income from different sources at the time and recipients of each type of income can be found throughout the whole distribution, one should always take into account the joint distribution of factor incomes before making conjectures. Atkinson and Bourguignon (2000) Atkinson (2009) and Milanovic (2017) contribute with formal demonstrations based on accounting identities, supporting this idea. The first two articles analyze the squared coefficient of variation of a theoretical distribution. They employ it as a measure of inequality, which is defined by the capital share, the coefficient of variation of each factor income and the correlation between labor and capital income. Atkinson

(2009) defines the critical value over which the capital share starts to have a positive impact on inequality.[1] Although this figure can be positive, meaning that at some levels an increasing capital share could in theory result in decreasing inequality, the critical point is expected to be rather low in plausible convex scenarios. Milanovic (2017) depicts a similar framework using the Gini coefficient. The author defines three clear requirements to have a positive relation between capital shares and inequality: first, high saving out of capital income; second, high concentration of assets; third, high correlation of capital income ranking and total income ranking. In real cases, all of these requirements are easily fulfilled.

On an empirical perspective, Bengtsson and Waldenström (2017) use a panel of 21 countries to assess the statistical relation between our variables of interest. They build capital shares using historical national accounts data and then regress them to income-concentration estimates. They use two of them: Top income shares, which they obtain from the World Inequality Database (WID)[2], and Gini coefficients, which they draw from Atkinson and Morelli (2012)[3]. Their estimates, which included country fixed-effects, are in line with what is expected. They find a strong positive marginal effect of the capital share over both inequality measures. When they introduce a set of control variables to the regression, the estimated effect decreases but remains significant.

Another contribution to the empirical strand of this literature, with somehow contrasting results, is Francese and Mulas-Granados (2015). The authors use harmonized household survey data from the Luxembourg Income Study Database (`http://lisdatacenter.org`) to perform a decomposition analysis of the Gini coefficient in 43 countries during the period 1978–2010. They break down the Gini coefficient to its accounting components and then implement a similar regression than Bengtsson and Waldenström (2017), but only using the survey's Gini coefficient as a dependent variable. After analysis they conclude that the capital share plays a negligible role in the evolution of measured inequality, especially relative to the evolution of labor-income inequality, which they judge as the main driver of total inequality. Although in theory the relationship is clearly based on sound accounting identities, empirically it may seem more opaque and

[1] The capital share (π) has to satisfy the following inequality: $\pi > (1-\lambda\rho)/(1-\lambda^2-2\lambda\rho)$, where λ is the ratio of the squared coefficient of variation of capital over that of labor income and ρ is the correlation between capital and labor incomes.

[2] The authors cite the database using its name at the time they were writing: the World Top Incomes Database (WTID).

[3] Atkinson and Morelli (2012) estimate Gini coefficients either directly from popular local household surveys or from well-known international data centers. These estimates are available for a subset of countries and a shorter span in time compared to top shares

unpredictable.

The issue with the existing models is that they only allow for a negative (or null) relation under very restrictive circumstances. For instance, in Milanovic (2017), the only channels trough which it can be achieved, is either by having capital income being more (or equally) concentrated than labor income, by having a negative (or null) correlation between total income and capital income and/or having higher (or equal) saving rates for labor income compared to capital income. These configurations are, of course, rather unrealistic. Therefore, when negative correlations emerge empirically these models do not seem to provide a convincing description of the mechanisms at play. Indeed, when Bengtsson and Waldenström (2017) observe a negative correlation for both Argentina and Canada, they treat it as an anomaly. In a bigger scale, in the Latin American region, Abeles, Amarante, and Vega (2014) report an overall increase of the capital share occurring in parallel to the contradicting and generalized decline of inequality that is recorded in López-Calva and Lustig (2010), which is odd. In order to better understand this kind of complex scenario, we need to add some variables to the equation.

The main intuition of this chapter is a rather simple one: Discrepancies in income-concepts and the quality of data potentially filter-out the impact of capital shares on inequality estimates. Income aggregates from National Accounts, especially capital incomes, are often substantially higher than those reported in most distributive datasets (either survey or tax statistics). At least a part of this phenomenon is due to the fact that National Accounts, which are often used to produce factor shares, have broader definitions of income. However, this is also the case when using harmonized definitions. Given that National Accounts are generally used as benchmark, one may thus consider that at least a part of the 'true' aggregate income is missing in distributive data sets.[4] Although the literature has given some attention to this subject, it has not included it explicitly in theoretical or empirical models. In this chapter, I will consider two main channels through which surveys and tax data could be missing income (i.e., the income under the carpet). First, not all capital income (e.g., dividends, interest, profits) is received by natural persons. At least a part of it goes to private or public corporations. Thus, it makes sense to measure the household's share of capital income systematically. Second, distributive databases often are subject to measurement errors, mostly due to heterogeneous response rates and misreporting. The error with which distributive statistics measure aggregate factor income can

[4]Of course, the recognition of discrepancies among datasets is not new. In fact, Atkinson (2009: Section II) gives a detailed enumeration of pertinent items on this issue.

thus be traced and analyzed.

The first contribution of this chapter is to establish stylized facts on both the household share of capital income and the measurement error in household surveys. Series on the former are built mainly using United Nations' official data on National Accounts for the 43 countries which have sufficient detail in their statistics. The main finding on this issue is a generalized and strong decrease in the share received by the household sector, as opposed to private and public corporations. This finding holds when studying the aggregate income of 19 countries forming a balanced panel during the period 1995-2015. But also at the country level, during the same period, in an unbalanced panel including 43 countries. The longest available series show that trends start falling around 1990 in most cases. Although a decrease of the household share of capital income is likely to have several economic consequences, this study will focus mainly on its impact on the measure of inequality. Furthermore, estimates of measurement error are computed using both National Accounts aggregate data and Household-Survey micro-data from the Luxembourg Income Study Database (LIS) on a balanced panel of 13 countries for the period 1995-2013. Both labor and capital income appear to be undervalued in all countries. Capital income is, in all cases, relatively more underestimated, with only around 20% of its gross figure being recorded in surveys, against 70% for labor income. This relation remains generally stable over the period. For the United States we compare the level of underestimation with that of tax data, which is substantially lower for capital income. Distributional National Accounts (DINA henceforth) estimates are not subject, by construction, to this kind of underestimation.[5]

The second contribution of this study is to introduce a simple theoretical framework which is based on accounting identities following the path from national capital and labor income to household-income shares, as they are recorded in survey or tax data. The product of households' share of capital income and the relative underestimation of capital income acts as a distortion factor, which filters out the effect of capital income. It generally results in the underestimation of levels and trends of inequality and affects the sensitivity of inequality estimates to the capital share and its distribution. This simple and straightforward representation is then used throughout this study in order to explore the empirical sensitivity

[5]Distributional National Accounts is the methodology developed by Alvaredo, Atkinson, Chancel, et al. (2016) as part of a global project that aims to combine surveys, tax data and national accounts to better study the distribution of the whole national income. Through various imputation procedures, the resulting estimates add up to the values of aggregate income present in National Accounts.

of income-concentration estimates to each of the variables in the model across countries and through time. Survey estimates largely understate the influence of capital income and thus seem to follow almost exclusively the distribution of labor income alone. Tax data is relatively more sensitive to the capital share and its distribution, at least during the period 1975-2015. DINA estimates are, again by construction, not subject to these distortions.

This study is organized as follows: Section 2.1 defines stylized facts on the distribution of capital income across institutional sectors and on the measurement-error of factor income in surveys. Section 2.2 presents the theoretical framework that is used to decompose the relation between capital shares and inequality measures. Section 2.3 displays empirical applications of the model, aiming to understand the composition of variations in inequality estimates. The last section discusses the main findings and concludes.

2.1 Stylized Facts

This section starts by describing the concepts and datasources that are used throughout the study. It then briefly displays both the gross and net capital shares that are obtained from national accounts estimates, before proceding to the analysis of stylized facts on both the household's share of capital income and survey's mismeasurment of factor income.

2.1.1 Data and Concepts

National Accounts Estimates in this section are mainly built using United Nation's 'National Accounts Official Country Data', which is publicly available at: `http://data.un.org`. This dataset distributes the whole national income to different institutional sectors. Ideally these are 6: households, non-profit institutions, financial and non-financial corporations, the general government, and the rest of the world. As not all countries build their accounts equally, the level of aggregation among sectors varies. For the sake of clarity and comparability, the main estimates of this chapter aggregate both financial and non-financial corporations in what is referred as 'private corporations' or the 'private sector'. Although the general government, or 'public sector', is partially studied in the present chapter, the evolution of its share of capital income is mostly not commented as it has little economic relevance. This is because its capital income is mainly composed of the profit of publicly held firms and payment on the interest of public debt, both of which are only a part of total public revenue and expenses (income from taxes and most expenses, including payment of interest are excluded). Non-profit institutions are mostly ignored in the analysis, as they always receive a negligible share of capital income. Data on the foreign sector is only used to estimate national income, as opposed to domestic income.

The guidelines of the UN's official System of National Accounts (UN-SNA henceforth) have been re-edited 5 times since its first version in 1953. Every revision included substantial methodological modifications, which often render different series hardly comparable. For that reason, both the aggregate and country-level estimates that are presented here do not mix information from different UN-SNA series. The series that are included in the balanced panel of 19 countries all correspond to the latest existing UN-SNA, which is the 2008 version. However, for long-run analysis and for the inclusion of less developed countries in the unbalanced panel, we also use series based on the 1993 UN-SNA guidelines.

Capital Income Globally, the literature on factor shares defines capital income as the sum of the total Operating Surplus and Net Property Income, plus the share of Mixed Income that is assumed to remunerate capital. While the first two terms of the addition are rather clear and generally accepted, the latter is often considered to be problematic. Indeed, Mixed Income broadly corresponds to the income of the self-employed, who usually combine both labor and capital to produce goods and services. However, the partition of this aggregate between factor incomes always relies on *a priori* assumptions. The literature has developed basically 3 ways to deal with this issue.

The first and more demanding method is to use surveys to estimate the wages that are paid in given economic sectors, then to assume independent workers pay themselves that same wage. In that situation, the capital share of their income is estimated as a residual. A second approach is to estimate the capital share of the private sector, which is the ratio of its Operating Surplus over total Value Added. This procedure depends on the level of detail in the institutional sector accounts, which is generally sufficient only for the last 2 to 3 decades in most countries. The third and simplest approach is the one that consists in assuming that a fixed share of mixed income, which is close to 2/3 or 70% remunerates capital. This and the second approach are used by Bengtsson and Waldenström (2017) and Piketty and Zucman (2014) in the context of limited data availability in long-run historical studies. In the present chapter, main results are obtained using the third method.[6]

Main estimates from national accounts presented in this chapter are gross of fixed capital consumption. Since capital depreciation has increased as a share of GDP in most countries during the last decades (figure B.13), it is important to check whether the broad conclusions of this chapter hold when using net definitions. Appendix B.2 shows that they generally do. However, as the decomposition of capital depreciation by institutional sector is quite uncommon, the number of countries with sufficient data is substantially reduced if we only focus on net estimates.

Sectoral Income Each Institutional sector ($i = 1, \ldots, n$) receives a capital income KI_i, which is defined as the sum of the sector's Operating Surplus (OS_i) and Net Property Income (NPI_i).[7] The only exception is the Household sector, which also receives Mixed Income (MI). As commented in the previous paragraph, a part MI^K of this aggregate is also assumed to remunerate capital. Naturally, the sum of the capital income of every sector KI_i is equal to total national capital income. We thus can define the total capital income of the economy KI in two ways:

$$KI = MI^K + \sum_{i=1}^{n}(OS_i + NPI_i) \qquad \text{and} \qquad KI = \sum_{i=1}^{n} KI_i$$

In the following subsection we mainly study each sector's share of capital income (KI_i/KI). Furthermore, labor income is by definition only affected to a single institutional sector, which is also Households. It simply is the addition of the Compensation to Employees (CE) and MI^L, the share of Mixed Income remunerating labor.

Household Surveys We use survey micro-data from the Luxembourg Income Study Database (LIS) for two empirical applications. First, it is used to assess the measurement error in surveys (Section 2.1.4), and second, to provide an empirical application of the theoretical model (Section 2.3). This particular data source is

[6]This strategy assumes implicitly that independent workers' remuneration has the same composition in both developing and developed countries, which is most probably incorrect. Indeed, we could expect the produce of independent workers in developing countries to be more labor intensive than their developed counterparts. However, this issue should not be of major importance for the purpose of this research. Developing countries are not included in the aggregate analysis and their estimates are not subject to international comparisons but are rather studied as time series. Assuming a different fixed ratio for developing countries would only imply a change in the level of the estimates but not in trends.

[7]In the definition of Net Property Income, the word 'Net' refers to: income received less income paid.

extremely useful as it contains detailed harmonized data sets from many different countries. The income definition that is used corresponds to the LIS variable named 'factor income'. In practice, the definition includes gross yearly income (pre-tax), combining monetary and in-kind revenues. This definition is close to the one used in national accounts as it even includes the virtual income that is generated when households consume goods of their on production (valued at market price). However, it does not include, mainly because of data availability, the imputed income of owner occupiers. That is, the value of the rent that dwelling-owners would receive if they decided to rent their dwellings instead of living in them. The strategy of selection for countries included in the panel (section 2.1.4) is based only on the maximization of countries with common data-points in both UN-SNA and LIS Database's waves.

Tax data In the case of the United States, we analyze data for a wider time span, that is the period 1975-2015. Piketty and Saez (2003) and relevant updates are used as estimates of total capital and labor income declared to American tax authorities. Furthermore, DINA estimates from Piketty, Saez, and Zucman (2018b) are also used to make empirical comparisons. These estimates combine both survey and tax statistics to distribute the whole national income to the personal income distribution.

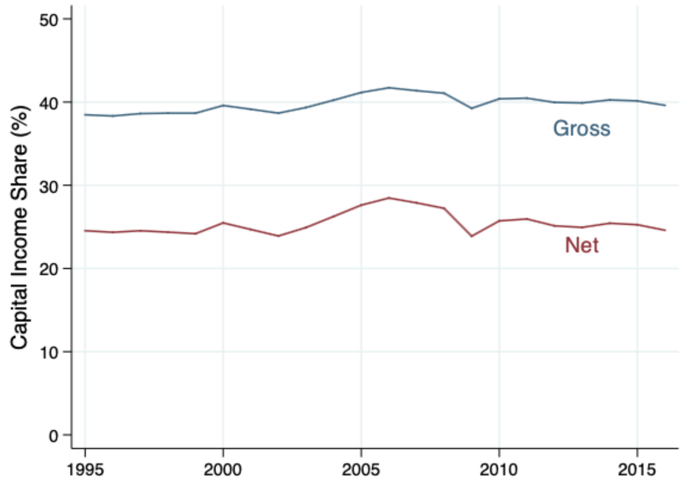
2.1.2 The Capital Share of National Income

The purpose of this investigation is not to establish stylized facts on the levels and trends of national capital income shares. In fact, the estimates that are used here are built in a simpler way, as detailed in the previous section, than those constructed by Piketty and Zucman (2014), which should be treated as a benchmark. They do however replicate rather closely the trends described by their estimates (figure B.2). The main differences between them come from the different treatment in the division of mixed income into its labor and capital components.

Figure 2.1 displays both the gross and net estimate of national capital income shares in a balanced panel of 19 countries. While the gross estimate increases around 2 percentage points through the period, remaining near 40%, the net figure is lower and stable around 25%. At Country level, trends are relatively more dynamic. Gross capital shares range between 45%-30%[8] and net shares are

[8] the only exception is the extreme case of Norway, which records a near 50% gross capital

Figure 2.1: Capital Share of National Income, Balanced Panel (1995-2016)



In a panel of 19 countries, the gross capital share of national income increases near 2 percentage points in 20 decades. The net capital share lower a is relatively more stable. Countries included (see figure 2.1 for country-level shares): Austria, Belgium, Canada, Czech Republic, Denmark, Estonia, Finland, France, Germany, Greece, Italy, Luxembourg, Netherlands, Norway, Poland, Portugal, Slovakia, Sweden, Switzerland and the United Kingdom. Income from different countries is aggregated based on yearly average Market Exchange Rates. The United States is studied separately. Own estimataes from United Nations' National Accounts.

generally between 20-30% (figure B.1). They are however sometimes lower than estimates from Piketty and Zucman (2014), likely due to the fact that the authors use a better method to split mixed income into its capital and labor components, which is probably more capital intensive than the fixed 30% assumption that is made here.

2.1.3 The Household Share of Capital Income

This section provides evidence on the generalized decrease of the household share of capital income. We start by presenting the distribution of capital income among institutional sectors, in a balanced panel of 19 developed countries during the period 1995-2015. The period was selected in such a way that maximizes the quantity of countries in the panel. Indeed, it is from 1995 that most countries provide detailed-enough series of national accounts. An unbalanced panel with 43 countries is also available in the same period. We therefore also analyze its

share

evolution at the country level. In order to explore the historical dimension of the observed phenomenon, we end the section with the study of the 9 countries which report data before 1990.

Balanced Panel Figure 2.2 shows that both the Public Sector and Private Corporations increase their share of gross national capital income between 1995 and 2015. The Public Sector starts with a negative value near -2.8% and ends the period with a low but positive share of 5%. This finding does not have high economic relevance in itself as it does not take into account the full income or expenses of the Government. It only accounts for its Operating Surplus and Net Property Income, which are mostly composed of the profit of publicly held companies and the payment of interest on public debt, respectively.[9] From an accounting point of view, however, it can be interesting to understand that this phenomenon has an impact on the relationship between what we define as the capital share and measured inequality. This idea is further developed in section 2.2.

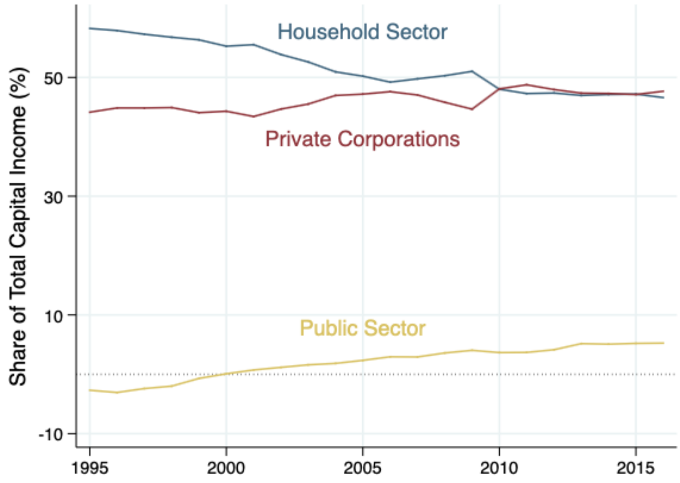
More relevant in figure 2.2 is the trend described by Private Corporations, which actually shows that retained profits represent an increasing part of capital income through the 20 years with data. That is, a bigger part of private profits are held inside corporations instead of being distributed to natural persons. Subfigure 2.2a displays a modest increase of their share by near 3 percentage points. Yet, economically, it makes more sense to study the evolution of this sector by excluding the government, as displayed in subfigure 2.2b. Indeed, value added is generated in corporations. After paying taxes, it is either distributed to natural persons in the form of wages (e.g. Compensation of Employees) or as distributed profits (e.g. dividends), the rest is retained inside private corporations.[10]

Figure 2.2b describes a clear and relatively constant decrease close to 7 percentage points in the household share of gross capital income. It is only interrupted by an ephemeral increase between 2007 and 2009, which is likely to be driven by corporate losses during the financial crisis. This trend corresponds to the aggregate capital income produced in 19 developed countries which form the balanced panel for the period. The weight of each country on this trend is in fact rather unequally distributed, as the first 5 contributors account for near

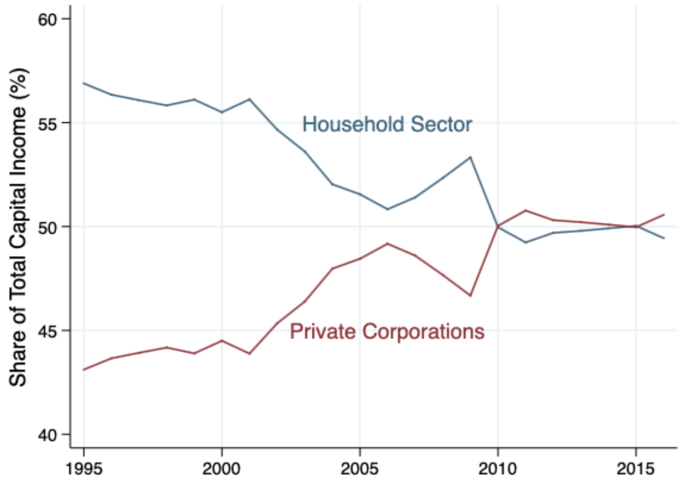
[9]Figure B.5 shows that the trend is mainly driven by the reduction of the expenses related to negative Net Property Income. Again, without taking income from taxes into account.

[10]It is worth noticing that the capital income of private corporations, which is the sum of the Operating Surplus and Net Property Income, is equivalent to the definition of Retained Profits in DINA guidelines (Alvaredo, Atkinson, Chancel, et al., 2016)

Figure 2.2: Decreasing Household Share of Gross Capital Income, Balanced Panel (1995-2016)



(a) Including the Public Sector



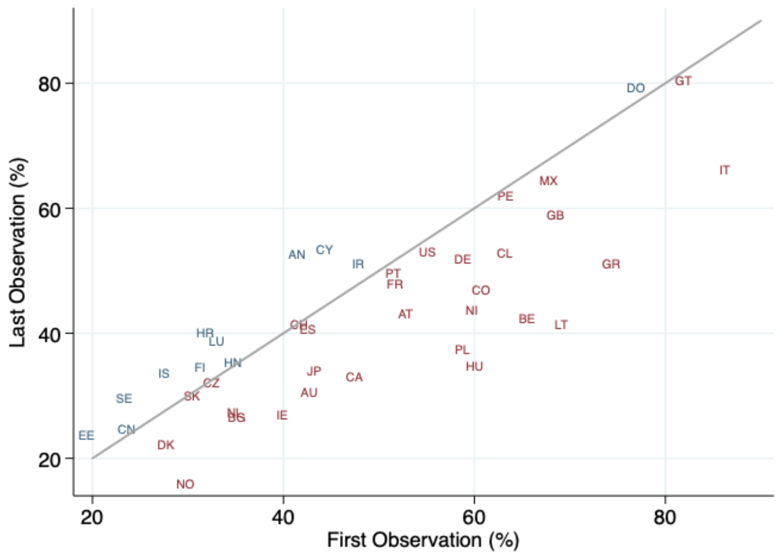
(b) Excluding the Public Sector

Both the Public Sector and Private Corporations increased their share of capital income, while the Household share decreases through the period. In 1995, the household sector received 57% of the capital income produced in a balanced panel of 19 developed countries (excluding the public sector). Two decades later, it receives less than 50%. Countries included: Austria, Belgium, Canada, Czech Republic, Denmark, Estonia, Finland, France, Germany, Greece, Italy, Luxembourg, Netherlands, Norway, Poland, Portugal, Slovakia, Sweden, Switzerland and the United Kingdom. Income from different countries is aggregated based on yearly average Market Exchange Rates. The United States is studied separately.

70% of total capital income in the panel (Table B.1). However, the same general conclusions can be made from the study of un-weighted averages (figure B.4) and net estimates from a sub-panel of 12 countries with available data (figure B.14). Figures B.3 and B.16 display gross and net estimates (respectively) at the country level.

Unbalanced Panel The dynamics displayed in figure 2.2 are not exclusive to developed countries. In total, 46 countries from several continents report detailed-enough data for this period. They do not cover, however, all years in the period. We thus study in figure 2.3 the evolution between the first and the last year recorded by each of these countries (including those in the balanced panel).

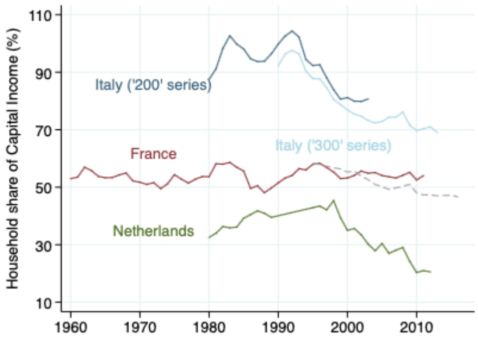
Figure 2.3: Decreasing Household Share of Capital Income, Unbalanced Panel (1995-2015)



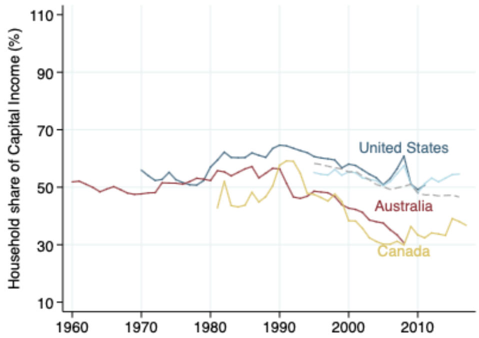
The share of capital income received by households, as opposed to public and private corporations, decreased in 30 out of the 43 cases that have at least 6 observations during the period (excluding the public sector). That is, it decreases in near 70% of cases. The countries that experienced an increase are those which already had relatively low shares to start with.

The red countries below the bisector line in figure 2.3 represent close to 70% of countries which saw a decrease in the household share of gross capital income. Most countries that are present in the balanced panel appear in this side of the plot. Additionally, from the developing world, we can find several Latin American countries (e.g. Chile, Colombia, Guatemala, Mexico, Nicaragua and Peru), and an Eastern Asian one: Japan. We can also find some additional countries from Eastern Europe (e.g. Poland and Lithuania) and Southern Europe as well (e.g. Greece, Spain). The blue dots that are situated above the bisector line gather mostly at the bottom-left corner of the figure. These countries saw an increase in the household share of capital income, yet their relative position shows that they already had low levels to start with, at the beginning of the period. The country that saw the biggest increase is Netherlands Antilles, which gained around 10 percentage points during the period. This case should be noted as a special one because the country is a tax haven (Zucman, 2015). When the public sector is excluded, we get a rather similar picture (figure B.11). For estimates of net capital shares, due to increasing capital depreciation, a relatively lower majority of countries follows a decreasing trend (figure B.19).

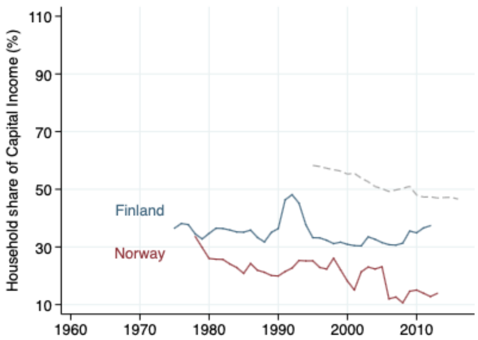
Figure 2.4: Decreasing Household Share of Capital Income, Long-run



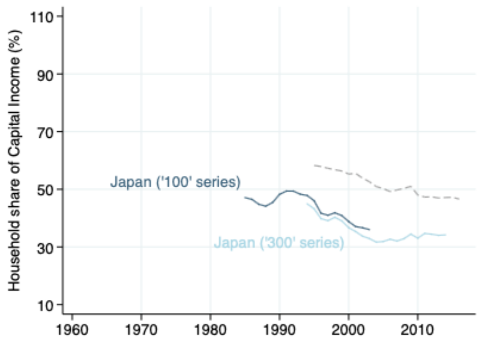
(a) Selected European countries



(b) English Speaking Countries



(c) Scandinavian Countries



(d) Japan

The grey dashed line represents the aggregate tendency in the balanced panel of 19 countries presented in figure 2.2. Most countries with long-run data exhibit a decreasing trend starting before the beginning of the panel, around 1990. Relatively more stable trends are described in previous decades.

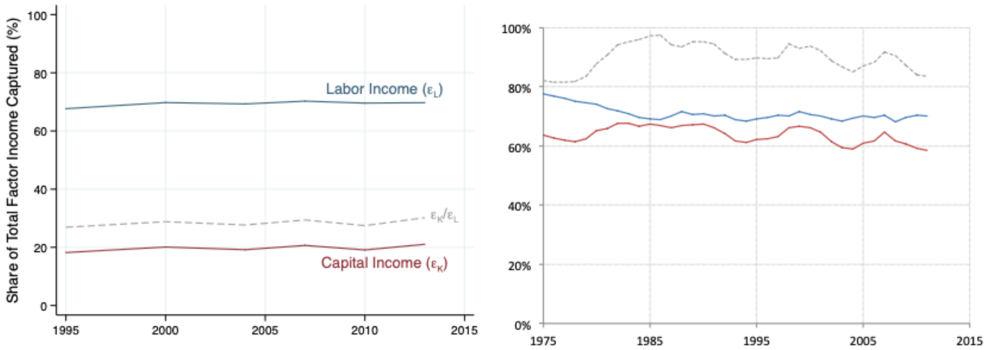
Long-run series Some countries have enough data to estimate the household share of capital income for several decades with consistent UN-SNA definitions. We display their long-run estimates in figure 2.4 as a way to study the starting point of the decreasing trend observed in figures 2.2 and 2.3. Italy is the first country to start a clear decreasing trend in the early 1980s. Others, as the United States, Canada, Australia and Japan start a rather clear decreasing trend around 1990. The countries with the latest starting trend are Netherlands, at the end of the 1990s and Norway, which experiences a big drop after its fiscal reform of 2005 that introduced the permanent taxation of distributed profits (Atkinson and Aaberge, 2010). France and Finland are cases where we do not observe particularly strong or sustained trends, but rather ephemeral dynamics. For instance, France experienced a drop in the household share of capital income at the end of the 1980s. This drop was then counterbalanced during the following years. Finland experienced an ephemeral jump of the estimate at the beginning of the 1980's, which was likely provoked by corporate losses during the Finnish banking crisis of 1991-1993.

The generally decreasing trend that is observed in this subsection probably has an impact on standard measurements of inequality. Common distributive statistics (i.e. tax data and household surveys) only record the income of natural persons. They thus ignore the income that is retained by public or private corporations, which is not the case when we study macroeconomic factor income shares. Figures B.12 and B.17 provide the same analysis than figure 2.4 but excluding the public sector and using net estimates (respectively). In both cases, we also observe a fall of the household share of capital income. This trend can be interpreted as a rise of retained profits over distributed profits.

2.1.4 Measurement Error in Distributive Data

When standard distributive statistics measure the aggregate income of households, they generally provide different figures from those in national accounts. If we assume that the latter are a good benchmark of these aggregates, we can measure the error with which each factor income is measured in different datasets. Figure 2.5 provides empirical estimates based on both surveys and tax statistics. Information from surveys is aggregated in a balanced panel of 13 countries that provide data to both the LIS harmonized survey database and UN's national accounts data.[11] Data from tax declarations is the one used in Piketty and Saez (2003), for the United States, which was updated by the authors to 2015.

Figure 2.5: Unequal Measurement Error in Surveys and Tax Data



(a) Survey Data: Balanced Panel (1995-2013)

(b) Tax Data: United States (1975-2015)

Labor income is better represented in both surveys and tax data, relative to capital income. The difference in the underestimation of each factor income, however, appears to be wider in surveys. Table B.2 displays the weight of each country in the Panel, in terms of aggregate national income. The ϵ_K/ϵ_L ratio is displayed here yet it is only commented in the following sections.

Survey Data Figure 2.5a shows that both labor and capital income are underestimated in surveys relative to national accounts.[12] The level of the underestimation, however, is not equal. While labor income appears to be relatively well represented with near 70% of it being recorded in surveys, the figure remains around 20% for gross capital income during the period. Between 1995 and 2013, the evolution of these estimates appears to be rather stable. At the country level, the relative underestimation also holds in every case (figure B.10), yet there is some variation in levels and trends, without clear patterns (estimates using net capital income at country-level are available in figures B.20 and B.21).

[11]This is a sub-panel of the one presented in figure 2.2. It includes the following countries: Austria, Canada, Czech Republic, Denmark, Finland, Germany, Great Britain, Greece, Hungary, Italy, Netherlands, Poland and Spain.

[12]The definition of income that is used to estimate the aggregate income in surveys correspond to what is referred as 'factor income' using LIS definitions. That is, gross income remunerating labor and capital. This is virtually the same definition used in National Accounts as it even includes, in the case of capital income, the value of goods produced for own consumption (estimated at market value). But, as information is not uniformly available, rent from owner occupiers is excluded from the definition, which explains at least a part of the underestimation of capital income.

Tax data In the case of tax data it is more difficult to find estimates for total factor income. Many studies directly use totals from national accounts to estimate top income shares (Atkinson, Piketty, and Saez, 2011). And those basing their aggregate estimates on fiscal data, usually do not report its composition in terms of factor income. In the case of the United States, most of the adult population fills tax declarations as only around 20% of them do not declare income to the Internal Revenue Service. Piketty and Saez (2003) provides estimates on the decomposition of the aggregate income they use. Figure 2.5b shows that, although capital income remains relatively more underestimated than labor income, the gap is narrower than in survey data for most countries. Indeed, figure B.22 shows that this is also true when we compare tax and survey estimates for the US alone. It is worth noticing that both survey and tax estimates for the United States cover more than 3 decades. During the 1975-2011 period, survey data gets progressively worst at capturing capital income as the share it captures goes from around 30% to 20% during the period (figure B.22). In the case of tax data, the level of underestimation is relatively more stable for both capital and labor income.

The most relevant stylized fact appears to be that although there are some differences in the level of underestimation, capital income is probably more underestimated than labor income in standard inequality databases. From an intuitive perspective, we should expect this to have similar consequences to the estimates presented in subsection 2.1.3, since in both cases we are measuring a part of capital income that is being ignored by distributive data. The following section defines a simple theoretical framework that should help understand the implications of these stylized facts on the measure of income concentration and its relation with the capital share of national income.

2.2 Theoretical Framework

The aim of this investigation is to understand the nuances between variations in the capital share and the income share of each quantile of the distribution (e.g., the richest 10% or the middle 40%). It should be noted that the model presented in the following paragraphs does not claim to describe causality but rather is an accounting framework that sheds light on the structure of estimates and their dynamic behavior.

2.2.1 Setup

Setup To describe the theoretical framework behind this study, we will consider the following setting. Let K and L be two non-negative real random variables whose sum is equal to 1. We will use K to represent the capital share of national income in a given economy. And we will use L to represent its counterpart: the labor share. Both variables are recorded in National Accounts, which divides income by institutional sector $i = (1, \ldots, n)$. Labor income belong integrally to the household sector (h), while capital income is divided in different institutional sectors, which receive a share of total capital income $\Phi_i = (\Phi_1, \ldots, \Phi_h, \ldots, \Phi_n)$, so that $\sum_{i=1}^{n} \Phi_i = 1$. In the following subsections we will focus on the relation between K and common inequality estimates. These estimates are in practice recorded by *distributive statistics* which is either survey or tax statistics. Both data sources use a narrower definition of income than National Accounts. In consequence, we define ϵ_K and ϵ_L as two real numbers that are higher than 0 and lower than 1. They represent, respectively, the share of national accounts' capital and labor income that is present in tax or survey data. We therefore define H, the total household income in the distributive data:

$$H = K\Phi_h\epsilon_K + L\epsilon_L \tag{2.1}$$

2.2.2 Identities

Income Shares We divide the total population in quantiles $q = (1, \ldots, m)$, such that the share of household's income received by each quantile is $S_q = (S_1, \ldots, S_m)$. In the same line, each quantile receives a share of household's capital income $S_q^K = (S_1^K, \ldots, S_m^K)$ and a share of total labor income $S_q^L = (S_1^L, \ldots, S_m^L)$. We then define the share of households' income received by each quantile q as follows:

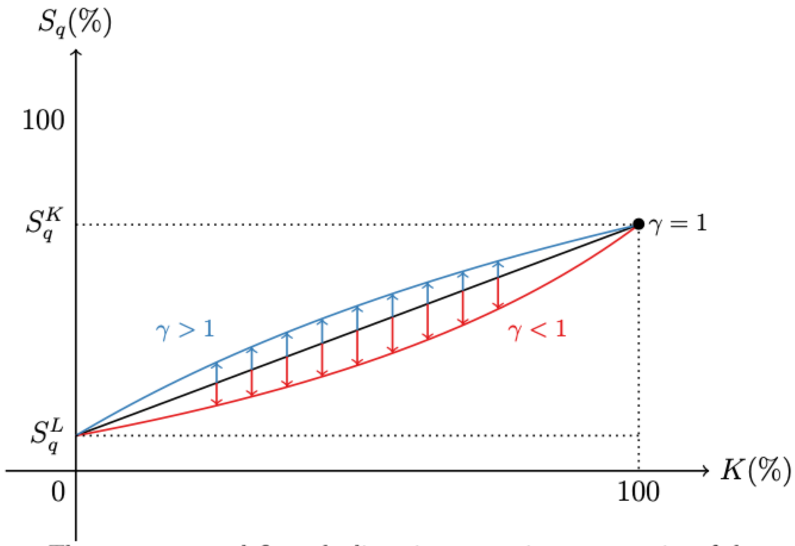
$$S_q = \frac{K\Phi_h\epsilon_K S_q^K + L\epsilon_L S_q^L}{H}$$

In this expression, the share is equal to the sum of both capital and labor income held by quantile q divided by total household income (H), which is defined in equation 2.1. In other words, quantile q receives a given percentage S_q^K of the total capital income recorded in household data, $K\Phi_h\epsilon_K$, plus a share S_q^L of total labor income recorded in the same data, $L\epsilon_L$. Now, the same expression can be rearranged in a less intuitive yet useful way:

$$S_q = \frac{S_q^K K\gamma + LS_q^L}{K\gamma + L} \qquad \text{s.t.} \qquad \gamma = \Phi_h \times \frac{\epsilon_K}{\epsilon_L} \tag{2.2}$$

Equation 2.2 can be translated graphically into figure 2.6, which depicts S_q in the empirically relevant case where quantile q concentrates a relatively higher share of total capital income ($S_q^K > S_q^L$). The function is defined for all possible values of K, keeping other variables as fixed parameters, in 3 different scenarios.

Figure 2.6: From the Capital Share to Top Income Shares



The γ parameter defines the linearity, convexity or concavity of the relation between the capital share and top income shares, while the relative concentration of factor incomes determines the slope and both upper and lower boundaries of top shares.

The black straight line represents the situation were the household sector receives all the capital income and both capital and labor income are estimated with the same error in the distributive dataset ($\gamma = 1$). The red convex line illustrates the most realistic case, where public and private corporations have positive income and/or aggregate capital income is relatively more underestimated than labor income ($\gamma < 1$). Conversely, the blue concave line describes the situation where private and public corporations have capital losses instead of income and/or a bigger part of capital capital income is recorded in the distributive data, compared to labor income ($\gamma > 1$).

Figure 2.6 shows that the γ parameter defines the linearity, concavity or convexity of the function. While the factor income concentration variables (S_q^K and S_q^L) define the sign of the slope and both the upper and lower boundaries for each quantile's income share. The construction of the γ parameter reveals that Φ_h and $\epsilon_K/\ \epsilon_L$ influence income shares in the same way and that they multiply each other. Indeed, both components of the γ parameter operate as filtering-out a part of the capital income from the equation. Furthermore, the γ variable has both a direct and indirect impact on quantile shares. That is, a lower γ not only results in lower S_q for a given K, but it also has an impact on the marginal effect of K over S_q.[13]

2.2.3 Marginal effects

Figure 2.6 gives relevant insight on the sensitivity of income shares to changes in the capital share. In the simpler case (black straight line) a variation ΔK always engenders the same variation ΔS_q that is proportional to the slope of the curve. However, in the convex and concave cases, the marginal effect of ΔK varies with K.

In order to better understand the sensitivity of S_q to every parameter in equation 2.2, Table 2.1 displays the partial derivatives of the model in the cases with and without distortion in the concept of capital income across datasets (Columns 2 and 1 respectively).

Table 2.1: Partial derivatives

f'_x	if γ=1 [1]	if $\gamma \neq 1$ [2]
$S_q(K)'$	$S_q^K - S_q^L$	$(S_q^K - S_q^L) \times \gamma(K\gamma + L)^{-2}$
$S_q(S_q^L)'$	L	$L \times (K\gamma + L)^{-1}$
$S_q(S_q^K)'$	K	$K \times \gamma(K\gamma + L)^{-1}$
$S_q(\gamma)'$		$KL(S_q^K - S_L^K) \times (K\gamma + L)^{-2}$

These are the formulas of partial derivatives, for each variable in the model in equation 2.2. They are used to estimate empirical marginal effects. We compare cases with and without distortions of income concepts across data sets ([2] and [1] respectively). The relevant difference is the multiplication by a 'distortion' factor in [2].

[13]It is worth noticing that the relations described by figure 2.6 and equation 2.2 are based on an underlying assumption whereby individual income rankings are kept unchanged after variations in the capital share. This can be a rather strong assumption, yet if the analysis is restricted to infinitesimal variations, it should not be a problem.

In the distortion-less case (column 1), as mentioned earlier, the marginal effect of the capital share is constant. And it is equal to $(S_q^K - S_q^L)$. In the same line, factor income concentration variables (S_q^K and S_q^L) also have constant marginal effects, which are equal to the value of national factor shares (K and L respectively). Now, when distortions are introduced (column 2), one can see that marginal effects are equal to those of column 1, but multiplied by a given factor. Therefore, some of the effects will be undermined, while others will be exacerbated. In the case of both the capital share (K) and the concentration of capital income (S_q^K), in realistic scenarios (red line in figure 2.6), the marginal impact will be lower in column 2 compared to the corresponding value of column 1. That is because $\gamma(K\gamma + L)^{-2}$ and $\gamma(K\gamma + L)^{-1}$ will both take values between 0 and 1. On the contrary, the impact of variations in the concentration of labor income (S_q^L) will be exacerbated, as it will be multiplied by $(K\gamma + L)^{-1}$, which should take values higher than 1. The marginal effect of γ is only relevant when it is different from 1. It is thus defined only in column 2, yet its interpretation is relatively less intuitive.

The study of partial derivatives implies that in normal 'distorted' cases, we should expect the role of capital-income-related variables to be undermined and those related to labor income should be exaggerated with respect to the distortion-less scenario. In the following section, these derivatives will be calculated empirically for various data-points in different databases in order to study the structural drivers of motions in top income shares.

2.3 Applications

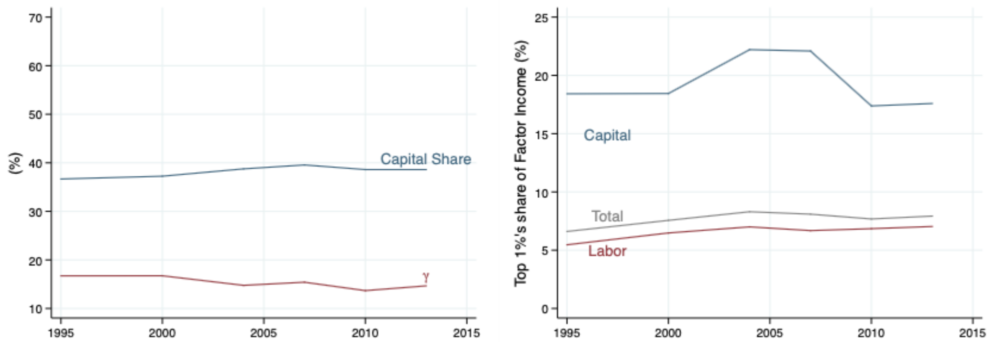
This section exploits the theoretical framework described in section 2.2 to produce empirical estimates based on the data behind the stylized facts in section 2.1.4. The analysis gives further insight on the driving forces of top income shares as estimated by surveys from 14 different countries in the period 1995-2013 but also from tax data and DINA estimates for the United States in the period 1974-2011.

2.3.1 General Trends

Surveys: Balanced Panel (1995-2013) Figure 2.7 depicts the evolution of the relevant variables in the model, using the same balanced panel presented in section 2.1.4 as if it was a single country. Both the capital share and the top 1%

share of total income grow during the period (figure 2.7a and 2.7b respectively).[14] The former gains a bit less than 2 percentage points, while the latter increases 1.3 points. As is to be expected, the top 1% income share is always between the levels of labor and capital income concentration. What appears rather strikingly is that the concentration of total income follows the level of labor-income concentration extremely closely. This is also the case when analyzing every one of the countries in the panel separately (figure B.6). In other words, it seems that the level of accumulation in the top 1% share is relatively insensitive to motions in the capital income distribution as it appears to depend mostly on what happens in the labor income distribution.

Figure 2.7: Evolution of relevant variables, Balanced Panel (1995-2013)



(a) Capital Share and Gamma

(b) Factor Income Concentration (Top 1%)

Both the capital share and the top 1% share of total income increase during the period. The latter estimate follows the concentration of labor income extremely closely. The γ coefficient remains rather low during the period, likely filtering out the influence of both the capital share and its distribution over the top income share.

As is explained further, this phenomenon is related to the low value that is taken by the γ coefficient in most surveys. Through the whole period, the coefficient remains below 20% and even decreases overall. When countries are studied independently, it never takes values above 30% and also decreases in most cases (figure B.7). The decomposition of the γ coefficient at the country level in figure B.8 shows that, although it is common to find that the household share of capital income (Φ_h) is higher than the mismeasurement ratio (ϵ_K/ϵ_L), this is not always the case as there are substantial differences in levels and trends of both

[14]The members of the top 1% here are actually those ranking inside the top 1% percent of each country put together in a single group. The aggregation of incomes is done using average market exchange rates.

estimates accross countries. Furthermore, the γ coefficient can be interpreted as the part of national gross capital income that is taken into account by the distributive data. Under these circumstances, although the capital share and the top 1% income share appear to be positively correlated, we should expect a rather low marginal effect of the former on the latter.

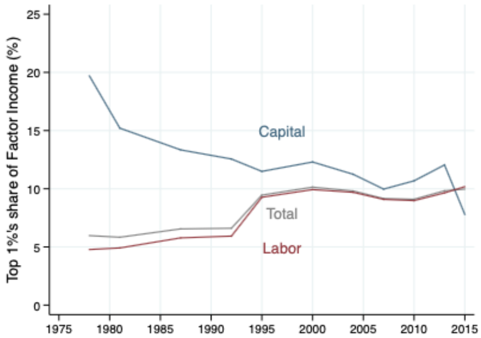
Comparing Datasets: United States (1975-2015) The United States is analyzed separately for two reasons. First, because its aggregate income is approximately equivalent to that of all the countries in the panel put together. Thus, if we were to include it in the panel, it would monopolize trends. Second, the US is one of the few countries which have good quality data for surveys, tax data and Distributional National Accounts (DINA) at the same time.[15] This enables a limited but useful comparison of estimates coming from different databases, and also the study of sensitivity to changes in the capital share.

Figure 2.8 displays estimates of income concentration in the three different datasets for the period 1975-2015. In all cases, the top 1% income share increases substantially through the period. Surveys record a near 4 points increase starting with a 6% share, while the tax data estimate increments around 10 points in the same period and starts higher, at near 8%. The DINA estimate starts at the highest level, near 10%, and shows a similar increase of about 8.5 points during the period. As in figure 2.7b, total income concentration is closer to the estimate of labor income than to the one for capital income in all databases. However, the distance between the red and the gray lines is different in every case. What is even more remarkable is the behavior of the capital income concentration curve. In surveys, it describes a decreasing trend and even reaches a lower level of concentration than the figure for labor income in 2015. In both tax data and DINA estimates the evolution is the opposite. In fact, the level of accumulation of both capital and labor income in the top 1% share increases substantially in both cases. Yet, in the case of tax data, the gap between income concentration in factor incomes gets narrower between years 1990 and 2000.

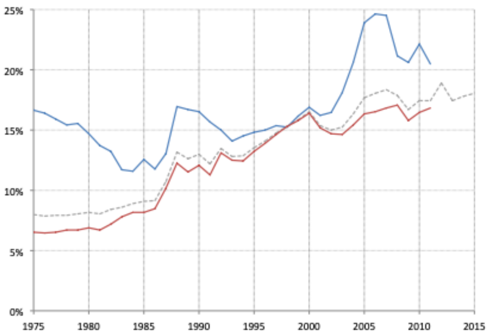
Again, the difference in the γ coefficient, which is the product of the two estimates studied in sections 2.1.3 and 2.1.4, is likely to be crucial to understand this phenomenon. As can be seen in figure 2.9, the gamma coefficient is generally

[15]Survey data for the US is also derived from the LIS Database. What is referred as tax data are the estimates of Piketty and Saez (2003) and subsequent updates that were made by the authors. DINA estimates come from Piketty, Saez, and Zucman (2018b). They correspond to a global project that aims to combine surveys, tax data and national accounts to better study the distribution of the whole national income (Alvaredo, Atkinson, Chancel, et al., 2016).

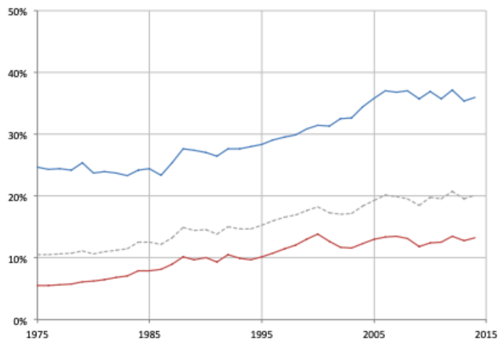
Figure 2.8: Top 1%'s Factor Income Shares, United States (1975-2015)



(a) Survey Data



(b) Tax Data



(c) DINA

Total income concentration increases substantially in all cases. It appears to follow closely the concentration of labor income especially in survey and tax data. Survey estimates are the only that show a decrease in the concetration of capital income. Tax data comes from Piketty and Saez (2003) and updates by the authors. DINA estimates correspond to the personal factor income definition in Piketty, Saez, and Zucman (2018b).

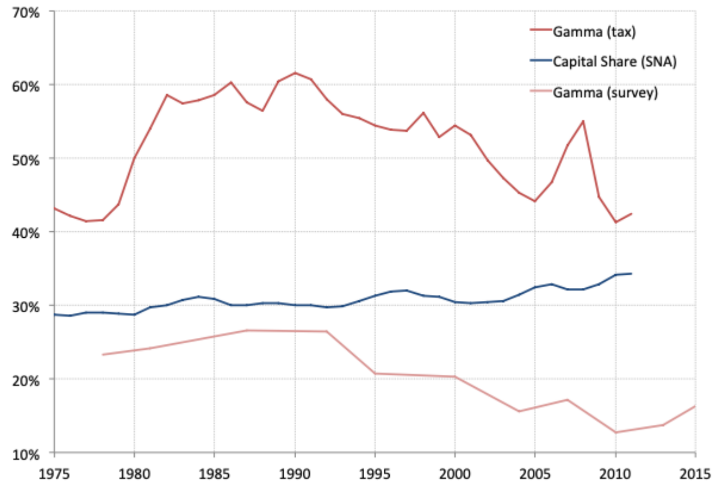
more than twice as high in the tax data compared to the estimate from surveys.

2.3.2 Sensitivity Analysis

Now that all the variables introduced in section 2.2 are defined for case studies, we can produce empirical estimates of the partial derivatives appearing in table 2.1. Table 2.2, presents the average value of estimated marginal effects for each variable in the model, for every country and data source.

Surveys: Balanced Panel (1995-2013) In Table 2.2, the highest marginal effect is that of the concentration of labor income (column 2) for all the countries forming the Panel. In the aggregate scenario, an isolated increase of 1 percentage

Figure 2.9: The Capital Share and the Gamma coefficient, United States (1975-2015)



The capital share increases during the period. Both surveys and tax data account for a progressively decreasing share of national capital income since the 1990's. The figure for tax data is, however, at least twice as high as the one for survey data overall.

point in this variable, we should expect a systematic increase of 0.91 points in the top 1% share, which is rather close to perfect correlation. This is a high value compared to the marginal effect that we would observe if there was not any distortion across distributive data sets and national accounts (if $\gamma = 1$). That is, the marginal effect would be equal to the benchmark labor share estimated using national accounts, which is close to 62% for most of the period (see figure B.7 for country estimates). Instead, here the marginal effect is equal to the labor share that is estimated by the distributive data set. Indeed, accounting identities ensure that the estimate from column 2 is actually equivalent to the labor share measured by surveys, while column 3 is equivalent to its counterpart: the capital share.

Table 2.2: Empirical Estimate of Partial derivatives, Top 1% Share

Country or area	$S_q(K)'$ [1]	$S_q(S^L_{top1\%})'$ [2]	$S_q(S^K_{top1\%})'$ [3]	$S_q(\Phi_h)'$ [4]	$S_q(\epsilon_K/\epsilon_L)'$ [5]	$S_q(\gamma)'$ [6]
Panel (1995-2013):						
- Austria	0.03	0.92	0.08	0.01	0.02	0.05
- Canada	0.03	0.93	0.07	0.02	0.01	0.04
- Czech Rep.	0.03	0.93	0.07	0.02	0.03	0.08
- Germany	0.06	0.91	0.09	0.03	0.05	0.10
- Denmark	0.07	0.94	0.06	0.07	0.03	0.13
- Spain	0.02	0.92	0.08	0.01	0.01	0.03
- Finland	0.10	0.90	0.10	0.08	0.05	0.15
- U.K.	0.02	0.91	0.09	0.01	0.02	0.04
- Greece	0.03	0.85	0.15	0.01	0.02	0.03
- Hungary	0.03	0.94	0.06	0.02	0.02	0.06
- Italy	0.05	0.87	0.13	0.02	0.04	0.06
- Netherlands	0.04	0.93	0.07	0.04	0.02	0.09
- Poland	0.03	0.93	0.07	0.02	0.03	0.06
Total:	0.04	0.91	0.09	0.02	0.04	0.07
U.S. (1974-2011):						
- Survey	0.02	0.92	0.08	0.01	0.01	0.02
- Tax	0.04	0.81	0.18	0.01	0.01	0.02
- DINA	0.19	0.73	0.27	-	-	-

These are empirical estimates of the marginal effects in table 2.1. In Austria, an isolated 1 unit variation in the capital share [1] only results in a variation of 0.03 units. But the same evolution in the concentration of labor income [2] produces a systematic increase of 0.92 units.

This effect is not exclusive to top income shares, as the same reasoning is to be applied to factor income concentration estimates from any other fractile of the population. The miss-measurement of the capital share by surveys should therefore have an impact in measured inequality as a whole, by exacerbating the role of labor income in the distribution. Another way to observe this underestimation is to compare cross-sectional estimates of the capital share produced using both national accounts and survey data. Figure B.24a plots these estimates, not showing any clear correlation between benchmark estimates and those of from surveys. When comparing survey-capital shares to the share of capital income of the household sector in SNA, figure B.24b suggests a clearer correlation but still exhibits an underestimation of at least a half of the value in surveys.

The marginal effect of variations in the capital share of national income (column 1) appears to be rather weak, as its aggregated effect in the panel is only 0.04%. In the un-distorted scenario, this estimate is equal to the difference of concentration in factor incomes. That is, a value near 10% in most years (figure 2.7b), which is likely to be underestimated in surveys due to the non-randomness of the error. To get to the estimate of column 1, this figure is reduced to near a half of its value by being multiplied by $\gamma(K\gamma+L)^{-2}$, due to distortions in income definitions and concepts (see Table 2.1). Furthermore, the concentration of capital income and the gamma coefficient appear to have a relatively low effect on the survey's top 1% share as well. An isolated variation of 1 point in the former variable (column 3) is translated into only a 0.09 point increase of the top share in the aggregate scenario, while the figure for the latter variable (column 4) is 0.07.

Comparing Datasets: United States (1975-2015) In the case of the United States we can compare the same estimates in different data sets. All the comments made in the previous paragraphs on survey estimates also apply to US surveys. Table 2.2 shows that the use of tax data somehow alleviates the exacerbation of the effect of labor income concentration, as the average marginal effect (0.81) gets closer than the survey estimate to the actual value of the labor share, which stays between 65% and 70% through the period (figure 2.9). In the same line, the estimate of capital income concentration has a higher marginal effect (0.18) relative of the one from surveys (0.08). However, although the effect of variations in the capital share is the double in tax data compared to the survey estimate, it remains low, at 0.04. This is not the case for the DINA estimate, which exhibits a marginal effect of 0.19. Of course, this is due to the fact that DINA estimates distribute all the national income to the personal income distribution. This corresponds to the situation where there is no difference in the income definition used to estimate capital shares and inequality estimates.[16]

2.3.3 Estimated Contributions

We can estimate the marginal effects studied in the previous subsection in every country at every data point. This allows us to compute each variable's contribution to change in income shares from one period to the other. If we multiply each variable's yearly variation by its marginal effect, we can analyze, from the perspective of accounting identities, the structure of changes in estimated top shares. Table 2.3 provides estimates for both the balanced panel of survey data and the United States with its different data sources. Column 7 aggregates the total estimated contribution of variables in the model. The difference between the estimated variation of the top share and the real variation (column 8) is due exclusively to the fact that databases only report subsequent snapshots at given points in time. In fact, most of the countries in the balanced panel only report data every 2-3 years, whereas tax data is available on a yearly basis. If we had access to the continuous evolution of these variables, there would not be any error in the estimate. Again, this is because the model is based on accounting identities. In any case, when plotted together, the 'estimated' top share and the real one are indistinguishable at normal scale.

Surveys: Balanced Panel (1995-2013) In surveys, the capital share does not appear to be a relevant driver of trends described by the top 1% income share. In the balanced panel taken as a whole, the top 1% income share increases around 1.3 percentage points during the period. But the parallel 2 points increase in the capital share only explains 0.08 points of such variation (column 1 in table 2.3). In fact, this contribution is completely counterbalanced by the measured variation of capital income concentration (figure 2.7b), which has a negative and modest influence of -0.08 points.

[16]The capital share used in Piketty, Saez, and Zucman (2018b) and therefore in the DINA estiamtes in table 2.2, is slightly different from the one displayed in figure 2.9 because it corresponds to the authors' personal factor income definition of national income.

Table 2.3: Modeled Contribution to Variation in Top 1% Share

Country or Area	Estimated Contribution (%)						Total Variation in period (%)		
	K [1]	S_q^K [2]	S_q^L [3]	Φ_h [4]	ϵ_K/ϵ_L [5]	γ [6] = [4] + [5]	Model Δ [7]=[1] to [5]	Real Δ [8]	Error [7]-[8]
Panel (1995-2013):									
- Austria	0.08	0.30	1.23	-0.09	0.16	0.06	1.68	1.59	0.09
- Canada	0.16	-0.02	1.00	-0.17	0.23	0.05	1.20	1.13	0.07
- Czech Rep.	-0.07	-0.54	0.72	0.05	0.25	0.30	0.41	0.33	0.08
- Germany	0.29	0.17	1.78	-0.13	-0.18	-0.31	1.93	2.23	-0.30
- Denmark	0.40	0.08	1.68	-0.31	-0.29	-0.59	1.57	1.59	-0.02
- Spain	0.12	-0.25	0.28	-0.17	-0.20	-0.38	-0.23	-0.28	0.05
- Finland	-0.07	-0.37	0.50	0.30	-0.04	0.26	0.32	1.02	-0.70
- Utd. Kingdom	-0.19	-0.54	1.84	-0.11	-0.02	-0.13	0.98	1.01	-0.03
- Greece	0.11	0.47	2.71	-0.26	0.03	-0.23	3.06	2.95	0.11
- Hungary	0.11	1.39	-1.28	-0.12	-0.45	-0.58	-0.35	-0.60	0.25
- Italy	-0.02	0.86	0.36	-0.26	-0.30	-0.56	0.65	0.35	0.29
- Netherlands	-0.01	-0.32	1.41	-0.25	0.36	0.12	1.20	2.01	-0.81
- Poland	0.09	-0.67	-1.21	-0.30	0.06	-0.24	-2.03	-1.93	-0.10
Total:	0.08	-0.08	1.43	-0.25	0.10	-0.15	1.28	1.32	-0.04
U.S. (1974-2011):									
- Survey	0.19	-1.00	4.89	0.09	-0.10	-0.01	4.07	4.00	0.07
- Tax	0.30	0.49	8.37	0.18	0.13	0.31	9.46	9.34	0.12
- DINA	1.14	2.69	5.05	0	0	0	8.89	8.95	-0.06

The concentration of labor income appears to be the dominant factor for variations in surveys' top 1% share estimate. The role of the capital share and its distribution is largely undermined. The distortion is weaker in tax data, but still present. DINA estimates are, by definition, not distorted in this sense.

Even though the near 2 points fall in the γ coefficient over the period (figure 2.7a) has the second most influential effect overall (-0.15), it is the concentration of labor income that gets the lion's share of contributions. These conclusions also apply, in general, to the country-level analysis. In surveys, the capital share and the concentration of capital income does not play a significant role in defining total income concentration in surveys. This is most likely explained by the large underestimation of capital income we observe by analyzing the trends and levels of the γ coefficient (figures 2.7a and B.7).

Comparing Datasets: United States (1975-2015) In US surveys, the conclusions are basically the same than with panel data, but with different levels, due to the larger extent of the period under study. The concentration of labor income also explains the largest part of variations in the local top 1% share, with an influence of near 4.9 points. The second largest contribution is an opposing -1 point that is provoked by the spectacular fall of capital income concentration in surveys (figure 2.8a). Furthermore, the decreasing trend described by the γ coefficient in figure 2.9 does not appear to have significant influence, with only a -0.02 contribution. It indeed appears that the low level of γ has a bigger influence than its trend; this, by distorting the marginal effect of other variables. The influence of the capital share is also positive but still relatively weak.

When we analyze contributions using tax data, both the capital share and the concentration of capital income seem to be of higher relevance compared to figures in surveys. However, their aggregate contribution remains modest, adding to less than 10% of the total variation estimated by the model. This is probably due to the fact that even though tax data is better at capturing capital income income than surveys, the γ coefficient associated with tax data oscillates between 40% and 60% during the period (figure 2.9). That is to say, tax data still ignores around the half of capital income produced by the country. It is only with DINA estimates that the capital share and capital income concentration start to play a substantial role in the evolution of the top 1% share. The former explains near 1 point in the total increase of 8.89, while the latter explains 2.7 points. That is an aggregate contribution of almost 40% of the total estimated variation.

Discussion and Conclusion

Stylized facts show that the household share of gross capital income decreases in most countries with data during the last two decades. This chapter however does not investigate on the causes of such trend. For future research on this topic, a possibly relevant clue could be the generalized growth of share buybacks as a way to remunerate shareholders (as opposed to dividends). The financial literature thoroughly documents its explosive growth in the United States since the 1970's, while comparing dividends and buybacks in terms of tax efficiency and signaling, among other aspects (Bagwell and Shoven, 1989; Fama and French, 2001; Skinner, 2008). The size of this phenomenon is remarkable, indeed, Floyd, Li, and Skinner (2015: fig. 3) show that the total amount allocated to share repurchases surpassed in the 2000's that of dividends in the United States. In the case of the European Union, Eije and Megginson (2008) also documents an increasing trend, yet with proportionally lower levels, for 15 countries since the 1990's. The relevance of this topic comes from the fact that the capital gains produced by these operations, whether they are realized or not, are generally not recorded by distributive data. Such shifting in remuneration mechanisms would thus diminish the part of capital income that could be potentially recorded in distributive data.

The general message of this chapter is that survey statistics fail to capture a growing part of national income that remunerates capital. This likely understates inequality levels and trends. The size of this phenomenon renders survey estimates almost completely insensitive to the motion of macroeconomic capital shares. In the light of the evidence presented here, we can better understand the findings of Francese and Mulas-Granados (2015). They also use LIS surveys and, when studying inequality estimates during the last two decades, they find that their evolution is explained almost exclusively by what happens in the labor income distribution. Yet, their conclusions should not be understood as undermining the impact of the capital income distribution on total income concentration, since their estimates simply do not capture most of it. The findings of this research can also help understanding that Bengtsson and Waldenström (2017) find a strong effect of capital shares on top income shares, as measured by administrative data, which is better suited to seize capital incomes than surveys. Moreover, it results somehow surprising that they still find a significant impact of capital shares on survey-based estimates of Gini coefficients.

The fact that surveys represent capital income poorly in most cases, certainly does not imply that they should be discarded. Household surveys are surely the richest and most easily available data source to study income inequality in all its dimensions. This is mainly due to the high amount of covariates that are generally reported by respondents. Moreover, the vast majority of people in high and middle income countries are remunerated via labor income, which is relatively well captured by surveys. It is indeed generally accepted that surveys give valuable information on what happens at both the middle and bottom of the income distribution. Therefore, should we try to adjust surveys to include external and reliable information available in both national accounts and administrative data? Or should we acknowledge limitations and use each different dataset to study specific aspects of income and wealth distributions? On the one hand, the second option avoids the risk of introducing undesired distortions to surveys and resulting estimates. Indeed, depending on the original quality of each dataset and on which datasets would be merged, one may be pushed to make more or less uncomfortable assumptions when applying corrections to surveys. Even if adjustments are done carefully, differences in data quality accross countries could potentially introduce noise to international comparisons. On the other hand, a good reconciliation of datasets would allow to better study the incidence of macroeconomic income growth, which is often emphasized in political discourses as being universally beneficial, but hardly measured. It would also allow, ideally, to base studies of economic inequality and its different dimensions on more reliable and sound data. Furthermore, this kind of adjustment would be especially useful when databases contradict each other in terms of concentration trends.

In any case, one should bear in mind that, despite great efforts to harmonize household surveys accross countries, the goal is at present still far away. Substantial differences are observed in terms of average response rates, income definitions and sampling methods, all of which are potentially substantial sources of bias. Moreover, survey weights are nowadays routinely adjusted mostly using post-stratification techniques and weight-calibration, both of which employ external data to make corrections (often census data). These adjustment techniques, which are rarely questioned by survey-users, aim to adjust for the uneven distribution of response rates among people with different socio-economic characteristics, yet they generally make corrections based on population totals (e.g., age, gender, geographical location) and not on the distribution of variables such as income.

Chapter 3

The Weight of the Rich: Improving Surveys with Tax Data

Tax data show that household surveys generally fail to properly capture the top of the income distribution, and therefore need to be adjusted to estimate inequality correctly. To date, there is no consensus on how to approach this problem. We introduce a method to combine both data-sets that has several advantages over previous ones: it is consistent with standard survey calibration methods; it has explicit probabilistic foundations and preserves the continuity of density functions; it introduces the concept of a 'trustable span' in tax data; it provides an option to overcome the limitations of bounded survey-supports; and it preserves the microdata structure of the survey, maintaining the representativeness of socio-demographic variables. Our procedure is illustrated by applications in five countries, covering both developed and less-developed contexts.

Introduction

For a long time, most of what we knew about the distribution of income came from surveys, in which randomly chosen households are asked to fill a questionnaire. These 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 income at the very top of the distribution.

For this reason, researchers have turned to a different source: 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 the latest 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 (2007, 2010) and served as the basis for the World Inequality Database (`http://wid.world`).

But tax data has its own limitations. It covers only the top of the distribution, and includes at best a limited set of covariates. It is often not available as microdata but rather as tabulations summarizing the distribution, which limits what can be done with them. 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, still have to be calculated from surveys. The use of different — and sometimes contradictory — sources to calculate statistics on the distribution of income and wealth can make it hard to paint a consistent and accurate picture of inequality trends. 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, Atkinson, Chancel, et al., 2016) emphasize the need to look at the entire distribution, harmonize concepts, and where possible to decompose by age and gender. Piketty, Saez, and Zucman (2018a) in the United States, and Garbinti, Goupille-Lebret, and Piketty (2016) in France have used both survey and tax data to construct distributional statistics that account for all of the income recorded in national accounts. But these examples rely in large part on the existence of administrative microdata accessible to researchers, to which information from surveys can be added.

In many countries, both developed and less developed, such direct access is quite rare. Instead, we have tabulations of fiscal income, containing information on the number and declared income of taxpayers by income bracket. The population coverage in the tabulations is often less than the total adult population, and the difference varies with the country studied. In such cases we have 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 this problem, yet the literature has largely failed to converge towards a standard. In this chapter, 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 is based on explicit probabilistic foundations with clear and intuitive interpretations. It also avoids relying, to the extent possible, on *ad hoc* assumptions and parameters. We present a data-driven way to determine where the bias starts in the survey data and beyond which point we merge incomes from tax data into the survey. 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 the behavior of complex statistics such as quantiles or bracket averages, our method makes easily interpretable assumptions at the level of observations. As a result, we can preserve the richness of information in surveys, both in terms of covariates and household structure. By looking at all variables simultaneously, we ensure the representativeness of the survey in terms of income while maintaining its representativeness in terms of age, gender, or any other dimension.

Our method proceeds in two steps, which are 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 first 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. The theoretical foundations of this approach are less explicit and harder to justify.

This first 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 radical example is the maximum income, which is almost always lower in the survey than in the 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 second step, in which we replace observations at the top by a distribution generated from the tax data, and match the survey covariates to it. The algorithm for doing so preserve the distribution of covariates in the original survey, their dependency with income, and the household structure regardless of the statistical unit in the tax data. The result is a dataset where sampling variability in terms 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 age, gender, or any other dimension.

For practical use, We have developed a full-featured Stata command that applies the methodology described in this article. The program works with several types of input, ensuring flexibility for users. Our method may therefore easily be used by researchers interested in analyzing the different dimensions of inequality, for instance those involving gender, education, voting patterns, etc.[2]

The remainder of the chapter is structured as follows. In section 3.1 we relate to the existing literature. In section 3.2 we lay out the theoretical framework of our method. This is followed by a practical guide of the method and its application to specific countries in section 3, before concluding.

[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.

[2]The packages to download are `bfmcorr` for the correction method, and `postbfm` for the postestimation output. Both commands come with a full set of user instructions.

3.1 Related Literature on Correcting Surveys

Numerous studies have sought to combine administrative data and survey data primarily to improve the latter's representativeness or produce a more accurate distribution of income. We identify three distinguishable methodological strands present in this literature. The first strand opts to reweight survey observations. The second strand adjusts the income value of observations through a rescaling approach. Finally, a third strand identifies the need to employ a hybrid procedure by combining reweighting and rescaling.

3.1.1 Reweighting Observations

The papers that focus on reweighting survey observations tailor their approach to remedy the bias of nonresponse. Many studies in this literature rely on parametrically estimating a probabilistic model of response to adjust household survey weights without the use of external data sources on the distribution of incomes. Korinek, Mistiaen, and Ravallion (2006) propose such an adjustment using the inverse of the probability of response for each household, which is estimated using nonresponse rates across geographic areas and the observable characteristics of respondents within regions. This type of approach, while not utilising auxiliary tax data, is sensitive to the degree of geographic aggregation used for inputting response rates into the adjustment. This is an issue explored in more detail by Hlasny and Verme (2017; 2018) for the U.S. and European case respectively, using similar probabilistic models. Depending on the nature of the survey data, greater or less geographic disaggregation on nonresponse rates can be more appropriate to the adjustment at hand. While the parametric models applied in these papers are data intensive, the estimations critically rely on observed survey distributions to adjust household weights given nonresponse rates across regions. Our proposal instead makes use of external administrative data, to guide us in how best to adjust household surveys, given the problem of nonresponse. This approach has the added benefit of indirectly tackling the problem of underreporting as we shall explain further on.

There are a few studies in this literature that combine surveys with external sources to measure inequality. An example of this is Alvaredo (2011), who for his second country-case study, on Argentina, estimates the corrected Gini coefficient 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 fractile and 0 under it. Furthermore, in both of his empirical applications (on the U.S. and Argentina) the merging point 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 type of bias implied by the correction

3.1.2 Rescaling Incomes

The general feature of this type of combination method is that it involves a rescaling of survey incomes with the tax incomes of equivalent rank. Although there is no unified theory or explicit justification behind most of these adjustment methods, they 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) 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; Morgan, 2017). Furthermore, the overall size of the population group whose income is to be adjusted is sometimes chosen arbitrarily, such as the top 20% (Piketty, Yang, and Zucman, 2017), top 10% in the distribution (Burkhauser, Hérault, et al., 2016; Chancel and Piketty, 2017), the top 1% (Burkhauser, Hahn, and Wilkins, 2016), or the top 0.5% of survey observations (DWP, 2015). It is also common to define the size of that group by choosing the point in the distribution beyond which the discrepancy between the average incomes in the two sources starts to become significant (Czajka, 2017; Morgan, 2017). With a somehow different approach, Alvaredo (2011) uses tax data to adjust survey-based Gini coefficients, applying a method inspired from Atkinson (2007a) to the U.S. In constructing the corrected Gini, the top 1% in the income distribution from tax data directly replaces the top 1% from the survey. Thus survey incomes are rescaled accordingly.

Rescaling survey-respondents' declared income has been acknowledged as adjusting for the misreporting bias in surveys (Burkhauser, Hérault, et al., 2018; Jenkins, 2017). In Appendix C.5 we explain why this is only true under very

[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.

strong and unrealistic assumptions, namely that the income rank in the survey distribution and in the true distribution are the same, and that underreporting is a deterministic function of that rank.

3.1.3 Combined Reweighting and Rescaling

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 rescaling. His analysis goes on to study the ways in which these choices impact the adjustments made to the original distribution data. 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 utilizes tax data, guiding users in how to best merge them with surveys to produce more realistic distributions of income.

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-calibrating the census population by intervals above a specified merging point, which is determined by the comparison of total income reported in the tax data and in the Census for the same intervals. The calibrating factors are based on the ratios between the populations in the same intervals of the two income distributions. However, 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. This is an inconsistency our method avoids.

3.2 Theory and Methodology

To describe our method and the methodology behind it, we will consider the following setting. Let X and Y be two real random variables. We will use Y to represent the true income distribution, which we assume is recorded in the tax data.[4] And 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.

3.2.1 Reweighting

In the first step, we adjust the weight of observations in the survey. In doing so, we are effectively adjusting the value of the survey density at different income levels. In this section we start by describing the intuition behind the correction in the simple univariate case. The next section explain how to use the theory of survey calibration to handle more complete settings.

Intuition

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:

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

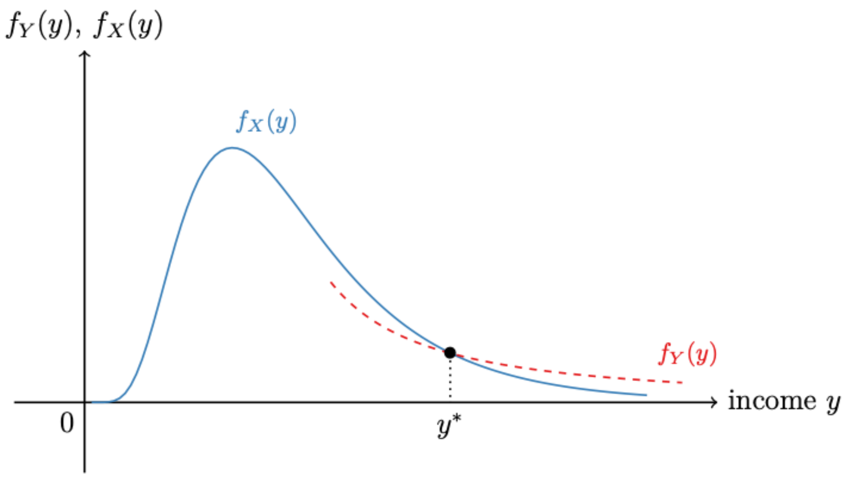
If everyone has the same probability of response, then $\mathbb{P}\{D = 1|Y = y\} = \mathbb{P}\{D = 1\}$, and $\theta(y) = 1$. Hence $f_X(y) = f_Y(y)$ and the survey is unbiased. What

[4] In reality, part of the true income is also missing from the tax data due to non-taxable income and tax evasion. But these issues are beyond the scope of this chapter.

matters for the bias is 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.

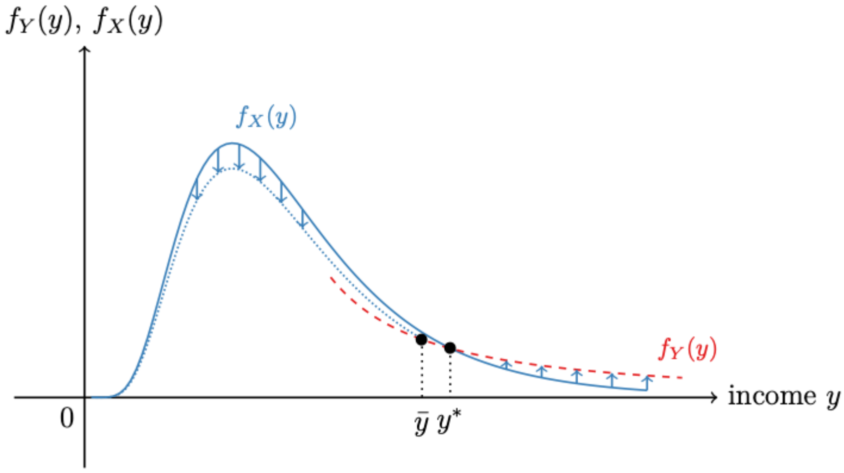
Figure 3.1: A "true" and biased income distribution



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 pivotal point y^*.

Figure 3.1 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 density f_Y. In practice, this means multiplying the weight of any observation Y_i by $1/\theta(Y_i)$.

Figure 3.2: 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.

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 3.2.2, and empirical evidence that we present in section 3.3. Intuitively, it means that there is no problem of representativity 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:

$$\theta(y) = \begin{cases} \bar{\theta} & \text{if } y < \bar{y} \\ f_X(y)/f_Y(y) & \text{if } y \geq \bar{y} \end{cases} \tag{3.1}$$

We call $\bar{y}$ the *merging point*. It is the value at which we start to rely on the tax data. 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 necessary. The proper choice of that point is an important aspect of the method on which we return to in section 3.2.1. For now we will take it as given, and only assume that it is below the pivotal point y^* of figure 3.1. Figure 3.2 shows how the reweighting using (3.2) 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 3.2, 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.

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 is typically wider in developed countries than in less developed ones.

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 call the *merging point* the value $\bar{y}$ at which we start using the tax data. We suggest a simple, data-driven way for choosing that point with desirable properties. In particular, we seek to approximately preserve the continuity of the underlying density function after reweighting. We start from the more simple case where $\bar{y}$ is inside the trustable span $[y_{\text{trust}}, +\infty[$, before moving on to consider cases where the trustable span may be too small.[5]

[5]Other choices of merging point have been suggested by previous works. Most of them are chosen arbitrarily (Burkhauser, Hérault, et al., 2016; Piketty, Yang, and Zucman, 2017; Chancel and Piketty, 2017), others are a more complex, as for instance, chosing the point where quantile functions cross (Morgan, 2017). In any case, these options do not preserve the continuity of density functions nor they are backed by clear economic interpretations.

Merging Point in the Trustable Span Assume that the bias function $\theta(y)$ follows the form (3.2):

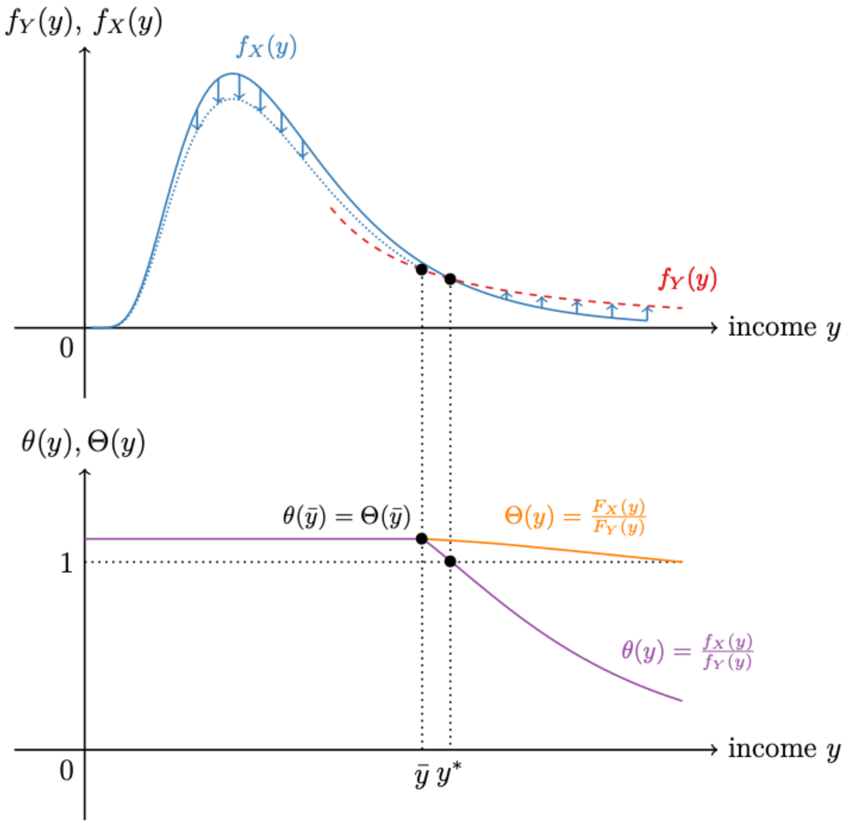
$$\theta(y) = \begin{cases} \bar{\theta} & \text{if } y < \bar{y} \\ f_X(y)/f_Y(y) & \text{if } y \geq \bar{y} \end{cases} \tag{3.2}$$

We introduce a second function, the cumulative bias, defined as:

$$\Theta(y) = \frac{F_X(y)}{F_Y(y)} \tag{3.3}$$

In figure 3.3, we examine the shape of $\theta(y)$ and $\Theta(y)$ in relation to the density functions presented in figure 3.2.

Figure 3.3: Choice of Merging Point when $\bar{y} \geq y_{\text{trust}}$



We have the relationship $\Theta(y)F_Y(y) = \int_{-\infty}^{y} \theta(t) f_Y(t)\, dt$. Given (3.2), for $y \leq \bar{y}$, $\Theta(y) = \bar{\theta}$. As figure 3.3 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 most rescaling approaches: the point at which the quantile functions of the survey and the tax data cross. 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 the merging 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 that point. That choice is the only one that can ensure that the profile of $\theta(y)$, and by extension the income density function, remains continuous.

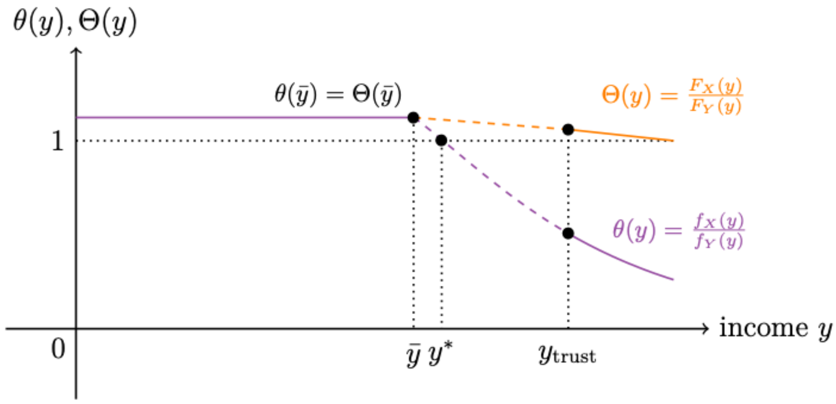
The estimation of $\Theta(y)$ poses no difficulty as it suffices to replace the CDFs by their empirical counterpart in (3.3) to get the estimate $\hat{\Theta}_k$. For $\theta(y)$, however, we have to estimate densities. We define m bins using fractiles of the distribution (from 0% to 99%, then 99.1% 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 $\hat{\theta}_k < \hat{\Theta}_k$.

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 $\Theta(y) = \theta(y)$. That situation is represented in figure 3.4. 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 3.4: Choice of Merging Point when $\bar{y} < y_{\text{trust}}$



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:

$$\log \theta(y) = \gamma_0 - \gamma_1 \log y \tag{3.4}$$

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 (3.4) 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.[6] 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 that problem as:

$$\min_{\gamma_0, \gamma_1} \sum_{i=1}^{m} (\log \tilde{\theta}_k - \gamma_0 - \gamma_1 \log y_k)^2 + \lambda(\gamma_1 - \gamma_1^*)^2$$

[6]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 3.3.2).

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 penalize 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.

3.2.2 Calibration

General Setup

The previous section 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 income, we also need to preserve (or possibly enforce) representativity 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 representativeness-related issues.[7] 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 feature. Second, when the adjustment is made using total income alone (i.e., 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.[8] A statistical tool is

[7]The idea of 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.

[8]One example of this kind of interaction would be: if rich old people are more likely to respond to surveys (say, because they have more free time) than young rich people, then a univariate adjustment will produce an accurate income distribution without solving the

thus presented in the following subsections to introduce this kind of information in the correction when it is possible.

We start by presenting the theory in its general setting below, before explaining how to apply it to the problems at hand.

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. $\sum_{i=1}^n w_i \boldsymbol{x}_i = \boldsymbol{t}$.

That 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:

$$\min_{w_1,\dots,w_n} \sum_{i=1}^n \frac{(w_i - d_i)^2}{d_i} \qquad \text{s.t.} \qquad \sum_{i=1}^n w_i \boldsymbol{x}_i = \boldsymbol{t} \tag{3.5}$$

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 (3.5) leads to:

$$\frac{w_i}{d_i} = 1 + \boldsymbol{\beta} \boldsymbol{x}_i \tag{3.6}$$

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).[9] One

over-representation of old people. A similar rationale can be applied to the issue of income composition.

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

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 There are two interpretations of the procedure. The first one is that of a nonresponse model. In that interpretation, the survey weights are the inverse of the probability of inclusion in the survey sample. That 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 nonresponse. 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 upscaled by 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 sociodemographic 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 that 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:

$$
\begin{aligned}
w_i &= \frac{1}{\mathbb{P}\{D_i = 1\}} \frac{1}{\mathbb{P}\{R_i = 1\}} \\
&= \frac{1}{\mathbb{P}\{D_i = 1\}} \psi(\boldsymbol{U}_i, \boldsymbol{Y}_i) \\
&= \delta_i \phi(\boldsymbol{U}_i) \times \frac{\psi(\boldsymbol{U}_i, \boldsymbol{Y}_i)}{\phi(\boldsymbol{U}_i)} \\
&= d_i \times \frac{\psi(\boldsymbol{U}_i, \boldsymbol{Y}_i)}{\phi(\boldsymbol{U}_i)}
\end{aligned} \tag{3.7}
$$

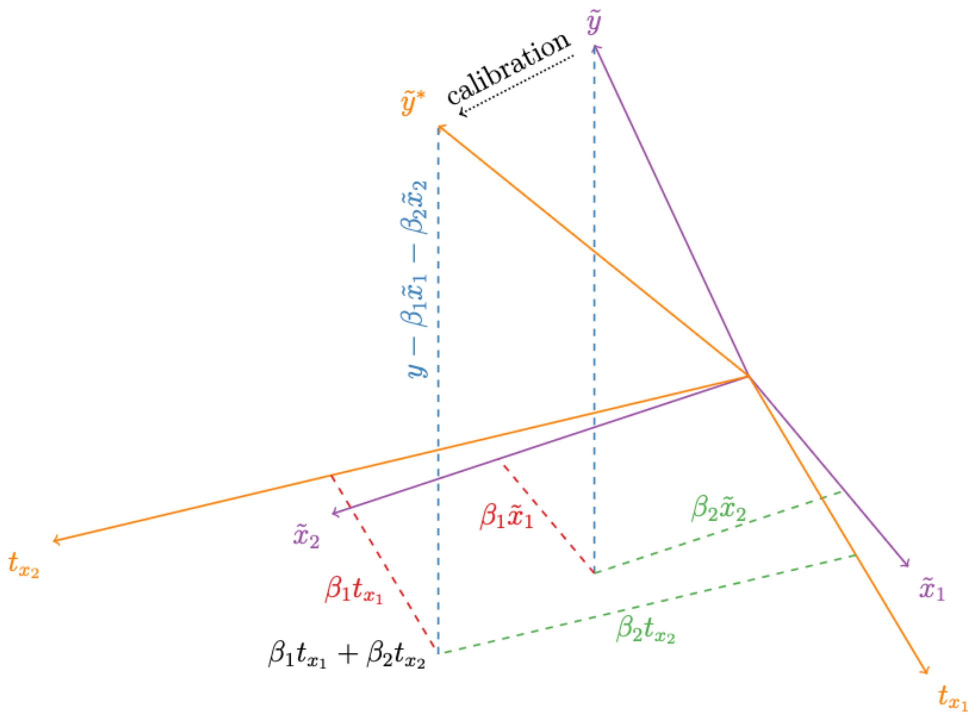
Comparing equation (3.6) with (3.7), we see that the calibration problem suggests both a functional form and an estimation method for $\psi(\boldsymbol{U}_i, \boldsymbol{Y}_i)/\phi(\boldsymbol{U}_i)$. This functional form assumes nonresponse profiles that are as uniform (thus non-distortive) as possible, 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 3.2.1.

The second interpretation is geometrical, and comes from the relationship between (3.5) 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 $x_1, \dots, x_k$, we can use the GREG estimator, whose logic is represented in figure 3.5. The idea is to first use the survey to project the variable of interest y onto the auxiliary variables $x_1, \dots, x_k$ using an ordinary least squares regression. Hence we get a linear prediction $\hat{y}_i = \boldsymbol{\beta}\boldsymbol{x}_i$ of y_i, which corresponds to the part of y that can be explained by the auxiliary variables $x_1, \dots, x_k$. 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 $\tilde{y}^* = \tilde{y} + \boldsymbol{\beta}(\boldsymbol{t} - \tilde{\boldsymbol{x}})$.

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.

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

Figure 3.5: 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 $\tilde{y}^*$.

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 3.2.1. We then match the survey data to their corresponding tax bracket. It is in general 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 brackets to have

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

a partition of the income distribution at the top such that each bracket has at least 5 survey observations. Assume that we eventually get m brackets, with the k-th bracket covering a fraction p_k of the population.

We create dummy variables $b_1, \ldots, b_m$ for each income bracket. If the total population is N and the sample-size is n, then the calibrated weights should satisfy:

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

Since these equations are expressed as totals of variables, they can directly enter the calibration problem (3.5). 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 representativity of the survey along the other dimensions. Additional constraints are also possible, if external information on other variables is available.

For all the observations below the merging point, the dummy variables $b_1, \ldots, b_m$ 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 (3.2) at the bottom of the distribution that we used in section 3.2.1. The calibration, by construction, avoids distorting the bottom of the distribution because it is not necessary to enforce the constraints of the calibration problem.

Our correction procedure also 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 $1/n \leq \theta(y) \leq n$ are automatically grouped into larger brackets. The default limit we choose (which can be changed by users) is n = 5. Thus, in this case, no observation would have their weight multiplied by more than 5 times or less than 0.2 times.

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 can provide 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 d_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 (3.5):

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

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. That 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 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, and the income likely to drive the bias. Yet sometimes only part of the 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, 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 (3.5) 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. That is because the method always favors the least distortive 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 sociodemographic 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. sociodemographic 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:

$$\sum_{i=1}^{n} w_i \boldsymbol{x}_i = \boldsymbol{t} \qquad \text{and} \qquad \forall i \in \{1, \dots, n\} \qquad \frac{w_i}{d_i} = 1 + \boldsymbol{\beta} \boldsymbol{z}_i \tag{3.8}$$

When $\boldsymbol{x}_i = \boldsymbol{z}_i$, the problem (3.8) is equivalent to (3.5). The solution of (3.8) given by Deville (2000) is similar to that of (3.6):

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

We can understand the name "instrument" for the $\boldsymbol{z}_i$ by going back the the GREG estimator (see figure 3.5 and section 3.2.2). 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.

3.2.3 Replacing and Matching

After applying the methods of section 3.2.1, 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 correct estimates of average partial effects in presence of unmodeled heterogeneity of 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 3.2.2) 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 to estimate a continuous income distribution (Blanchet, Fournier, and Piketty, 2017) that reproduces the features of the tax data with high precision. We then match statistically the information in the calibrated survey data with the tax data by preserving the rank of each observation.

First, 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 (i.e. empirical "copulas") 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 surveys' covariate distribution (including the household structure), and limit distortions in the relationship between income and covariates from survey data.

3.2.4 Standard Errors and Confidence Intervals

Once both the re-weighting and replacing steps of the adjustment are realized, researchers should be able to produce standard estimates in the same way they would do using raw datasets (e.g., averages, inequality estimates, regressions, etc.). In the case of inequality estimates, for instance, we recommend the use of commands `ineqdeco` or `svylorenz`, by Stephen P. Jenkins.

The choice of a specific procedure to estimate confidence intervals is, of course, the responsibility of each researcher as it depends mostly on the nature of the estimate that is being produced and the survey-design that was used in each particular case. However, when users do not possess sufficient survey-design information to build satisfying standard errors, the only meaningful way to compute intervals from an adjusted survey, to our knowledge, is the use of `bootstrap`. Indeed, most nationally representative survey-samples are not the result of a purely random selection. Multistage sampling (i.e., clustering and stratification) render the estimation of variances substantially more complex than with pure-randomness.[11] The fact that our method modifies observation-weights and even creates new observations implies that common variance estimates (e.g., linearized standard errors), can mechanically exaggerate the size of intervals where pure randomness is assumed. The increased difference of weights among observations that usually results from the adjustment would be the main driver of this paradoxal phenomenon, where the inclusion of more precise data would be interpreted as a decrease in precision due to wrong assumptions about randomness. Such estimates would thus be meaningless and particularly sensitive to some of the parameters that our program enables users to define.[12]

3.3 Applications

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 it.[13] In order to illustrate how the method operates in practice, we apply it to data from five countries, three developed (France, U.K., Norway) and

[11]For comments on variance estimation in calibrated surveys, see Deville and Särndal (1992).

[12]Due to time limitations, confidence intervals will be provided with the estimates presented in the following section in a future version of this chapter.

[13]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.

two less-developed (Brazil, Chile). Our chosen case studies showcase the wide applicability of the method to both developed countries and less developed ones whose data quality is more challenging.

3.3.1 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.[14] For European countries 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.

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. The latter situation thus requires an imputation of taxes paid to arrive at gross incomes. Appendix C.1 explains how this imputation is done for the Chilean case, as well as the construction of income units in surveys and their approximation with tax data in all countries. For the European countries we work with gross incomes (pre-tax and employee contributions deducted at source) from the SILC database. France is the exception since incomes reported in the tax files are net of social 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

[14]The main purpose of our method is to ensure the representativeness of top incomes in surveys using tax data. Nonetheless, the procedure also preserves the representativeness of other variables for which the survey is assumed to be already representative. These typically concern gender and age variables. Our calibration process leaves the distribution of these variables, or any other specified categorical variables, unchanged.

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 postive 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 20% subject to the personal income tax on taxable income.

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). 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 C.1 and Morgan (2017) for further details).

3.3.2 Empirical Bias and Corrected Population

The Shape of the Bias Our method proposes to find the merging point between surveys and tax data by comparing the population densities at specified income levels, as explained in section 3.2.1. 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.[15]

[15]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

Using the tresholds of these intervals we can construct our key statistics on $\theta(y)$ and $\Theta(y)$.

Figure 3.6 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 what we used in the theoretical formalization presented in Figures 3.3 and 3.4. 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 shape of the empirical bias $\theta(y)$ can be observed for a more comprehensive share of the population, due to their 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 3.6a 3.6b 3.6c, 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.[16] 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 3.6e) provides an example of our program needing to extrapolate the shape of the bias to find the merging point (see Section 3.2.1) For this case we rely on parameters observed for Brazil (specifically, values for elasticity of response to income) above its trustable span as inputs for the Chilean extrapolation.[17] 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 C.2.

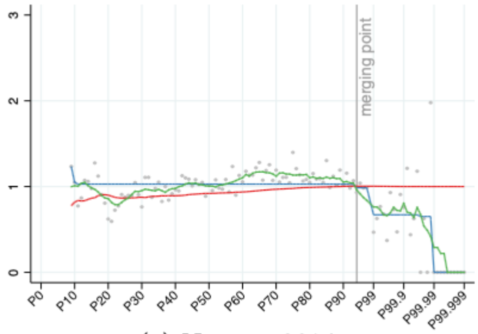
Corrected Population Our program then adjusts the individual weights of survey respondents in line with information from tax data, as described in section 2.1. We provide some summary statistics of the population we correct in Table 3.1, again using the last available year for each country as illustrations (see Appendix C.3 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

precise estimates of the distribution.

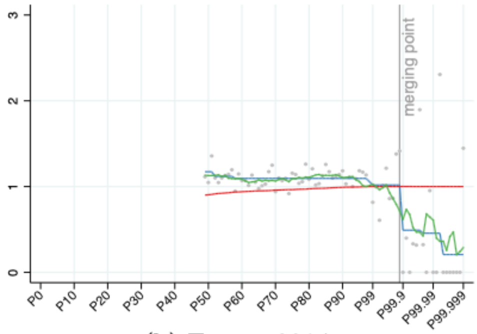
[16]Tax enforcement issues affecting this portion of the distribution could be at play here, as well as the sharp difference in incomes between the top and the rest in these countries leading to higher inequality levels than developed countries.

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

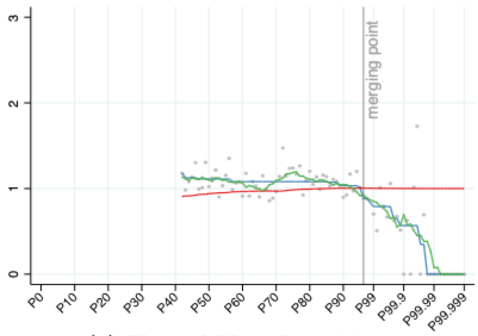
Figure 3.6: Merging Point in 6 Countries, Latest year



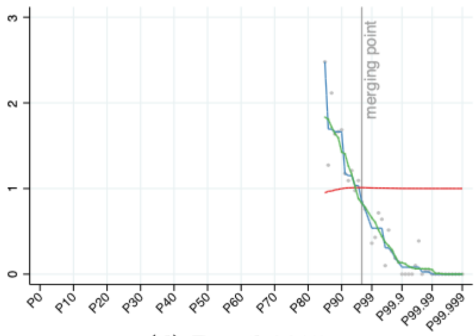
(a) Norway 2014



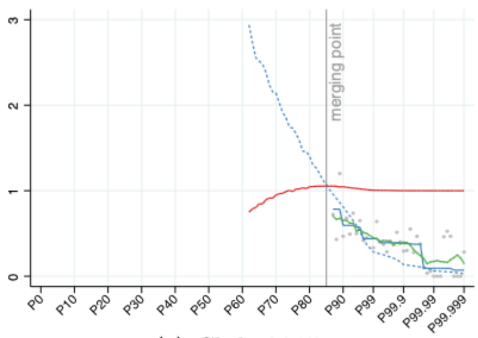
(b) France 2014



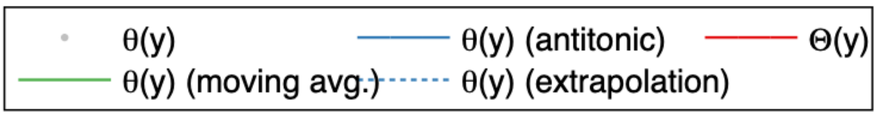
(c) United Kingdom 2014



(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 3.6e,is an extrapolation of the trend described by $\theta(y)$ based on a *ridge* regression. The red line is the ratio of the cumulative densities. For details refer to section 3.2.1.

Table 3.1: Structure of Corrected Population: Latest Year

Country	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]
Chile	14.0%	9.2%	4.8%	99.99%	0.01%
Brazil	3.0%	1.9%	1.1%	98.2%	1.8%
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.

the survey distribution in each country (column [4] of Table 3.1), ranging from 6% in Chile to 0.05% in France for their most recent years.[18] 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 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. In general, this step of the algorithm should be a useful guide for researchers to assess the income coverage of surveys across countries. For instance, it would appear 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.

[18]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 for countries like France and Norway, the goal is not to over-sample the top of the distribution, but rather to improve the precision of responses.

3.3.3 Income Distribution

As detailed above, our method produces an adjusted micro dataset that maintains the survey's original design along a more representative income distribution. We can unveil how this merged distribution changes with respect to the raw survey distribution.

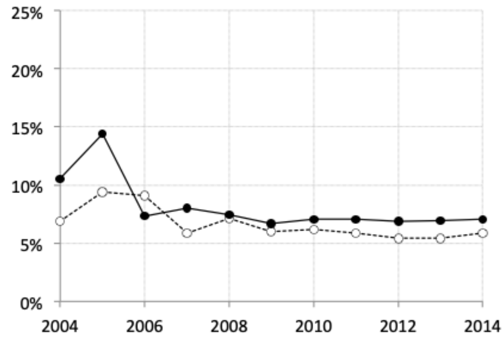
Top Income Shares Our adjustment procedure generally makes significant upward corrections to the shares of income going to the top of the distribution in the surveys. The size of the adjustment, however, varies with countries. Figure 3.7 depicts this for the Top 1% share in 5 countries for all years with data available.[19] Brazil has the most extensive one, with a top 1% share that increases about 10 percentage points every year (Figure 3.7d). Conversely, France and Norway experience relatively smaller adjustments, starting from relatively lower levels of inequality.

The quality of both surveys and tax statistics may have a substantial impact in the size of the adjustment. For instance, in the case of France, several improvements were made to the survey's methodology since 2008. In particular, the matching of individuals across survey and tax statistics allowed the use of tax data as an external source to assess individual income without recourse to self-reporting. Visibly, in Figure 3.7b the gap between raw and corrected estimates is reduced from 2008 because the size of the survey bias was reduced with the methodological novelties. Moreover, when we compare the size of the adjustment in Chile and Brazil (Figures 3.7d and 3.7e 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, 2017), 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 impacts the smaller correction.[20]. Following the same rationale, the inclusion or exclusion of some

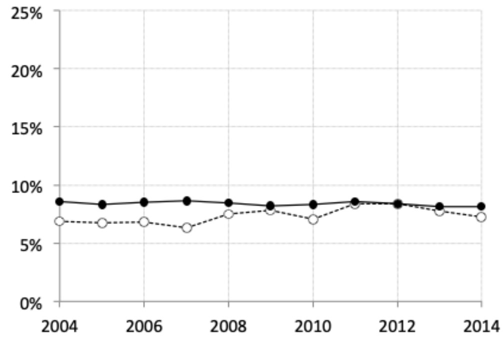
[19]The one exception to this upward correction is Norway in 2006 (see Figure 3.7a). However, this is likely due to a change in the local tax legislation affecting the distribution of business profits (Alstadsæter et al., 2016).

[20]There is also a considerable difference between these countries' tax systems and their respective 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 untaxed, provokes incentives towards the artificial retention of profits that are not as present in Brazil.

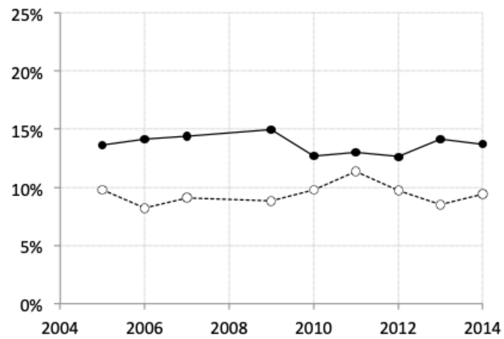
Figure 3.7: The Top 1% Share Before and After Correction



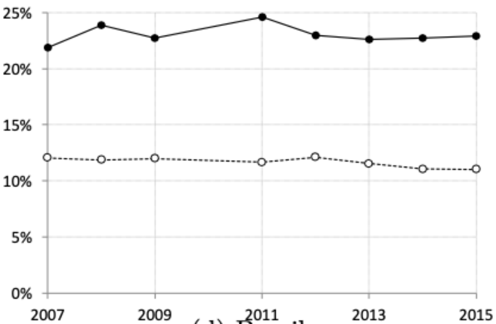
(a) Norway



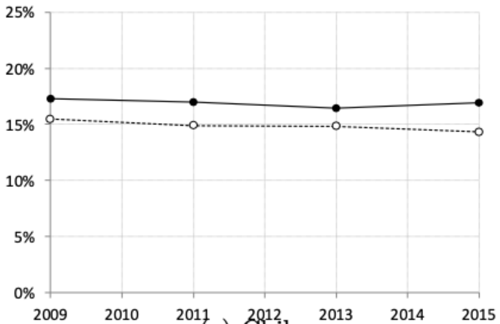
(b) France



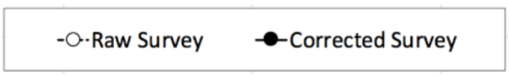
(c) United Kingdom



(d) Brazil



(e) Chile



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 (Atkinson and Aaberge, 2010; Alstadsæter et al., 2016). In Figure 3.7a, 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.

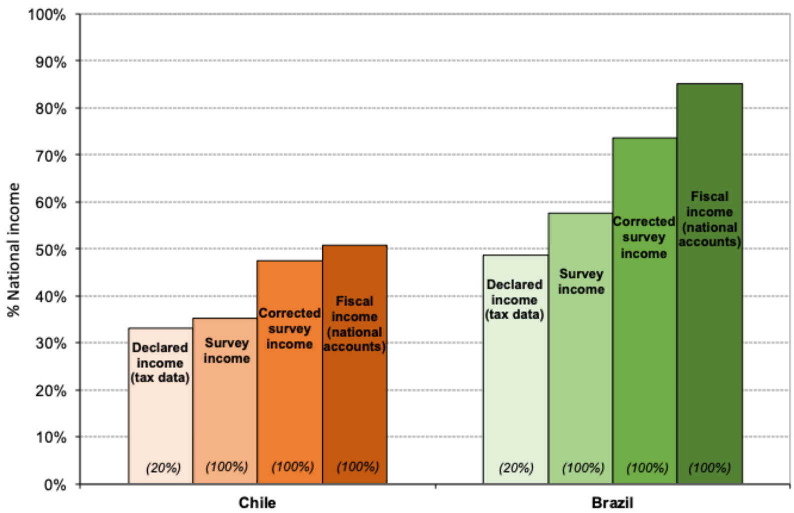
Another potential explanation for the difference in the size of adjustments could be the difference in levels of inequality between countries. This could help explain for instance, why the survey in the United Kingdom receives an adjustment that is higher than the one of both Norway and France, but lower than the one of Brazil (Figure 3.7c). In addition, Brazil offers the clearest illustration of the distinct trends in inequality that can emerge after making a correction to the survey's income representation. While the raw survey depicts falling top income shares, the corrected survey distribution returns stable if not slightly increasing top shares. Distinct trends are also visible, albeit for shorter periods of time, in the other countries.

Detailed Distribution Table 3.2 depicts a more detailed picture of the impact of our adjustment method on the income distribution of our 5 countries. Again, we take the last available year as an illustration. The first point to note is that the average income of the survey is adjusted upwards every time. The extent of the increase, by definition, depends directly on the shape of the bias that is observed in Figure 3.6. 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 3.1) are decisive factors for the size of such an increase.[21] Figure 3.8 presents the impact of our method on total income. For our two country case studies with the largest corrections, we are able to show that the total income

This is why, in Chile, the imputation of undistributed profits to the distribution of personal income appears to be necessary when making international comparisons (Atria et al., 2018). This example emphasizes the importance of the DINA project for cross-country comparisons of inequality (Alvaredo, Atkinson, Chancel, et al., 2016)

[21] 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.

Figure 3.8: 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.

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.

With respect to income shares across the distribution, the main conclusions that are drawn from the analysis of the Top 1% share in previous paragraphs can be generally extended, with more or less intensity, 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. This is consistent with the mechanics of our adjustment, 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. Again, expectations on the scale of the downward correction of these share can be informed via the size of the bias at the top, as depicted in Figure 3.6. A more general picture of what happens in the whole distribution is presented by the Gini coefficients. In all the latest-year

Table 3.2: Income Shares: Raw Survey and Corrected Survey

Raw Survey					
Income groups	**Brazil**	**Chile**	**France**	**Norway**	**UK**
Bottom 50%	16.5%	8.0%	23.4%	25.2%	14.8%
Middle 40%	42.8%	45.2%	47.0%	48.6%	49.6%
Top 10%	40.7%	46.9%	29.6%	26.2%	35.5%
Incl. Top 1%	*11.0%*	*14.3%*	*7.2%*	*5.8%*	*9.4%*
Incl. Top 0.1%	*2.4%*	*3.4%*	*1.5%*	*1.4%*	*2.5%*
Incl. Top 0.01%	*0.6%*	*0.7%*	*0.4%*	*0.3%*	*0.4%*
Incl. Top 0.001%	*0.1%*	*0.2%*	*0.1%*	*0.03%*	*0.04%*
Average income	€8,081	€8,101	€23,367	€37,431	€22,389
Gini	0.53	0.64	0.40	0.37	0.52
Corrected Survey					
Income groups	**Brazil**	**Chile**	**France**	**Norway**	**UK**
Bottom 50%	13.3%	6.6%	23.2%	24.6%	13.9%
Middle 40%	35.2%	39.5%	46.5%	47.7%	46.6%
Top 10%	51.5%	53.9%	30.3%	27.6%	39.6%
Incl. Top 1%	*22.9%*	*16.9%*	*8.2%*	*7.1%*	*13.7%*
Incl. Top 0.1%	*10.5%*	*4.6%*	*2.2%*	*2.2%*	*5.4%*
Incl. Top 0.01%	*5.2%*	*1.3%*	*0.6%*	*0.7%*	*2.1%*
Incl. Top 0.001%	*2.4%*	*0.4%*	*0.2%*	*0.26%*	*0.89%*
Average income	€10,138	€10,949	€23,621	€38,320	€24,081
Gini	0.61	0.69	0.41	0.38	0.55

Notes: The table presents the distribution of pre-tax fiscal income per adult, before the correction and after the correction. Average incomes are expressed in French Euros PPP. Brazil and Chile refer to 2015, while all the European countries refer to 2014.

examples, the Gini increases, which reflects a general increase in total estimated inequality using this composite index.[22] The scale of the increase generally reflects the magnitude of the change in income shares.

Conclusion

The main objective of this chapter is to provide a 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 reweighting 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.

This study can thus be viewed as an attempt to improve survey representativeness by taking the income 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 research motivates the use of auxiliary administrative data sources on the distribution of income, along with other socio-demographic information, to improve the representativeness of the population.

Our procedure has several advantages. First, it is based on an 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 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.

[22] Appendix C.4 presents the trends in country Gini coefficients for the full period.

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, users may also calibrate and correct the survey on covariates of income, in addition to income itself, if reliable statistics exist on their interaction. Ideally, we think that reweighting based on external information on the income distribution could be applied to surveys when employing standard calibration procedures. 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 research reconciling income and wealth distributions in a national accounting framework, as in the Distributional National Accounts project (Alvaredo, Atkinson, Chancel, et al., 2016).

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 practitioners wishing to assess the population coverage of surveys conditional on income. Figure 3.6 and Table 3.1 are examples of the type of information directly computed by our algorithm. With standard survey and tax data at hand, researchers can perform our correction procedure with relative ease. Given that we make the statistical tools openly available, they could provide the seeds for greater collaboration between national statistics institutes and tax administrations in order to improve nationally representative datasets. The combination of survey and register data is already happening in some countries, with the former gradually becoming anchored to the latter in the most developed cases. National statisticians engaged in the production of surveys could make use of our correction method upon having direct access to data on income and other covariates from government ministries. For many countries in which the majority of the population are not included in income tax statistics or social security contributions, our adjustment could make great gains. For more developed economies, researchers who want to continue to make use of their national household surveys can still do so without concerns over distributive representativeness.

Appendix A

Top Incomes in Chile, A Historical Perspective of Income Inequality (1964-2015)

A.1 Changes in tax legislation and income definition.

1964 (Feb): Tax reform, law nº 15,564.

- Defines for the first time what *income* is in legal terms.
- Taxes are declared according to 2 categories instead of 6.

Source: Boletin del Servicio de Impuestos Internos XI(123), February 1964: 3780-3839. This includes both Law 15,564 and the document 'Comentarios e Instrucciones' written by the SII, which compares the new dispositions with previous legislation.

1965 (Aug): Minimum presumed income tax.

- Special and transitory tax which was applied in tax years 1965, 1966 and 1967. It is based on net disposable wealth. It affects natural persons exclusively.

Source: Boletin del Servicio de Impuestos Internos XII(141), August 1965: 4603-4607, 4608-4632.

1972 (Nov): Single law nº 17,828.

- Those who perceive wages or pensions as a single source of income are no longer obliged to declare personal income tax (*Global Complementario*).

1974 (Jan) (Dec): Tax reform (under dictatorship).

- The wealth tax is removed
- Decree Law nº 824: Tax brackets are now defined in terms of Annual Tax Units (UTA) instead of *sueldos vitales*. Tax unity is periodically updated according to variations in the Consumer Price Index.
- Value Added Tax (VAT) is introduced.

Source: Cheyre (1986)

1984 (Jan): Tax reform in favor of savings and investment, law nº 18,293.

- Income declarations by business owners include only distributed profits.
- Corporate taxes can be used as a credit against personal income tax.
- Retained profits are no longer in businesses' taxable base (FUT mechanism. It is a taxable profit fund. According to Fairfield and Jorratt De Luis (2016), it allowed for keeping track of "how much tax credit (corporate tax paid by the firm) owners are due when they eventually withdraw these profits and pay individual income taxes. Total FUT profits reported at the end of 2012 were equivalent to Chile's GDP. FUT funds imputed to taxpayers in our datasets make up 56–61 percent of the total".

1990 (Jun): Tax reform, law nº 18,985.

- Income that has been withdrawn from a company and is reinvested in another one is not subject to personal income tax (art. 1, 2).

1998-2001 (Jul): Transitional article nº 2, law nº 19,578.

- Capital gains made from selling stocks by highly traded corporations can choose to pay an alternative lower tax (Impuesto Unico de Primera Categoría).

2001 (Nov): Capital market reform (MKI), laws nº 19.768 and nº 19,769.

- Capital gains made from selling stocks by highly traded corporations are tax-exempt (for stocks bought before April 2001) (art. 1 – 1 – b).
- Capital gains made from short selling are tax-exempt (art. 1 – 3).
- Stocks of some emergent companies (defined by their growing potential) can be considered as "highly traded". Hence, capital gains made from selling their stocks can be tax-exempt for 3 years (Transitional art. nº 4, law 19,768).
- The range of financial products authorized as voluntary pension savings, which are deducted from the taxable base, is widened. Within a maximum of 48UF, people can chose to invest in AFPs, mutual funds, Investment funds, and life insurances, among other products.

2002 (Apr): Single article (completing MKI) Law nº 19,801.

- Capital gains that became tax-exempt with MKI do not need to be declared.

2007 (Jun): Capital market reform (MKII), law nº 20,190.

- Capital gains made from selling some venture capital shares are tax-exempt (Transitional Art. 1).

2012: Tax Reform, law nº 20,630.

- Access to special regimes of taxation is limited to more strict conditions, especially to *renta presunta*, which was often used to inflate declared costs of companies or professionals (thus, lowering declared profits/revenue).
- Increase in First Category Tax rate (*Impuesto Único de Primera Categoría*): this rate (20%) was a provisional measure decided by the government in the wake of the 2010 earthquake to finance reconstruction. The reform reenacted the provisional measure, establishing it as a permanent rate.
- Reduction of personal income tax rates. The rates dropped to between 20% and 4.10% at each end of the scale. The top tax bracket was not reduced, but the tax burden on the highest incomes decreased, as it is a marginal tax rate.

2014: Tax Reform, law nº 20,780.

- Corporate tax is modified. Companies have to choose between an attributed tax regime and a semi-integrated one. The former is based on a 25% tax rate on profits and firms cannot gain tax credits against the tax paid by business owners. The semi-integrated tax regime has a 27% tax rate, though firms can receive a tax credit that represents up to 65% of the tax payment.

- A rule to tackle avoidance is introduced to give the *Servicio de Impuestos Internos* greater control to enforce and sanction aggressive tax planning.

A.2 Additional tables and figures

Table A.1: Adult population, total and taxable

YEAR	POPULATION SHARES					TOTAL TAX DECLARATIONS (% of adult population)		DECLARATIONS OVER TAXABLE THRESHOLD (% of adult population)	
	Population Total (20 y.o. and over)	10%	1%	0.1%	0.01%	GC Series	GC+IUSC Series	GC Series	GC+IUSC Series
1962	4,140,085	414,009	41,401	4,140	414	1.2%		1.2%	
1963	4,229,695	422,970	42,297	4,230	423	3.8%		1.0%	
1964	4,323,028	432,303	43,230	4,323	432	3.9%		1.1%	
1965	4,419,437	441,944	44,194	4,419	442	4.4%		1.4%	
1966	4,511,536	451,154	45,115	4,512	451	4.8%		1.6%	
1967	4,606,805	460,681	46,068	4,607	461	5.2%		2.0%	
1968	4,706,118	470,612	47,061	4,706	471	6.2%		2.3%	
1969	4,811,300	481,130	48,113	4,811	481	6.6%		2.7%	
1970	4,923,628	492,363	49,236	4,924	492	6.7%		3.5%	
1971	5,033,942	503,394	50,339	5,034	503	7.7%		4.3%	
1972	5,154,466	515,447	51,545	5,154	515	0.2%		0.2%	
1973	5,282,679	528,268	52,827	5,283	528	1.4%		1.2%	
1974	5,414,875	541,488	54,149	5,415	541	1.9%		1.8%	
1975	5,549,289	554,929	55,493	5,549	555	1.7%		1.3%	
1976	5,692,169	569,217	56,922	5,692	569	1.6%		1.6%	
1977	5,834,759	583,476	58,348	5,835	583				
1978	5,979,359	597,936	59,794	5,979	598	2.5%		2.2%	
1979	6,129,161	612,916	61,292	6,129	613	2.5%		2.2%	
1980	6,285,271	628,527	62,853	6,285	629	2.4%		2.3%	
1981	6,443,122	644,312	64,431	6,443	644	2.2%		2.0%	
1982	6,605,725	660,573	66,057	6,606	661				
1983	6,773,640	677,364	67,736	6,774	677				
1984	6,946,805	694,681	69,468	6,947	695				
1985	7,124,933	712,493	71,249	7,125	712				
1986	7,306,223	730,622	73,062	7,306	731				
1987	7,492,910	749,291	74,929	7,493	749				
1988	7,682,445	768,245	76,824	7,682	768				
1989	7,872,732	787,273	78,727	7,873	787				
1990	8,062,479	806,248	80,625	8,062	806	8.8%		2.9%	
1991	8,244,490	824,449	82,445	8,244	824	9.1%		3.0%	
1992	8,431,444	843,144	84,314	8,431	843	10.1%		3.3%	
1993	8,619,650	861,965	86,197	8,620	862	10.4%		3.7%	
1994	8,804,596	880,460	88,046	8,805	880	10.8%		3.9%	
1995	8,984,617	898,462	89,846	8,985	898	12.2%		4.3%	
1996	9,169,195	916,920	91,692	9,169	917	12.4%		6.2%	
1997	9,345,988	934,599	93,460	9,346	935	13.0%		6.6%	
1998	9,519,263	951,926	95,193	9,519	952	13.8%		6.2%	
1999	9,694,975	969,498	96,950	9,695	969	14.1%		6.2%	
2000	9,875,973	987,597	98,760	9,876	988	14.4%		6.5%	
2001	10,054,711	1,005,471	100,547	10,055	1,005	15.8%		7.6%	
2002	10,237,743	1,023,774	102,377	10,238	1,024	15.8%		7.7%	
2003	10,426,873	1,042,687	104,269	10,427	1,043	16.0%		7.9%	
2004	10,623,302	1,062,330	106,233	10,623	1,062	16.3%	62.7%	8.6%	9.9%
2005	10,827,647	1,082,765	108,276	10,828	1,083	16.8%	64.9%	9.1%	10.3%
2006	11,031,536	1,103,154	110,315	11,032	1,103	17.0%	66.6%	7.1%	10.9%
2007	11,245,406	1,124,541	112,454	11,245	1,125	17.0%	67.7%	10.2%	11.6%
2008	11,466,511	1,146,651	114,665	11,467	1,147	17.0%	68.8%	8.2%	11.7%
2009	11,691,082	1,169,108	116,911	11,691	1,169	16.7%	67.8%	8.1%	12.4%
2010	11,916,457	1,191,646	119,165	11,916	1,192	17.5%	69.1%	7.2%	13.4%
2011	12,133,631	1,213,363	121,336	12,134	1,213	18.1%	71.0%	7.7%	14.3%
2012	12,353,033	1,235,303	123,530	12,353	1,235	21.8%	72.4%	10.6%	15.4%
2013	12,573,206	1,257,321	125,732	12,573	1,257	21.6%	72.4%	11.0%	16.7%
2014	12,792,746	1,279,275	127,927	12,793	1,279	21.4%	72.4%	11.4%	16.9%
2015	13,010,453	1,301,045	130,105	13,010	1,301	22.1%	72.6%	11.9%	17.3%

Total adult population from World Bank. Taxable individuals are those who declare income above the minimum taxable threshold

Table A.2: Undistributed Profits

year	as a share of GDP	as a share of total fiscal income	total In current pesos (millions)	per adult (real 2015 pesos)
1990	4.8%	11%	464,747	226,224
1991	4.8%	11%	614,364	240,138
1992	4.8%	11%	778,732	257,859
1993	4.8%	11%	931,033	267,510
1994	4.8%	11%	1,119,313	282,523
1995	4.8%	11%	1,367,304	312,478
1996	4.8%	11%	1,508,727	314,698
1997	5.9%	14%	2,052,014	395,654
1998	4.1%	9%	1,500,796	270,292
1999	3.4%	7%	1,252,789	214,384
2000	3.9%	9%	1,686,499	272,827
2001	3.9%	9%	1,789,975	274,618
2002	3.9%	9%	1,900,493	279,405
2003	3.0%	7%	1,628,456	228,643
2004	6.2%	15%	3,773,292	514,566
2005	5.6%	14%	3,900,090	506,363
2006	6.9%	18%	5,672,819	699,193
2007	8.7%	23%	7,825,294	906,205
2008	5.7%	14%	5,305,391	554,233
2009	9.7%	24%	9,380,210	960,402
2010	12.5%	32%	13,855,518	1,372,429
2011	10.2%	25%	12,318,115	1,159,573
2012	8.9%	21%	11,545,557	1,036,385
2013	8.3%	19%	11,387,809	986,646
2014	8.6%	20%	12,763,526	1,041,105
2015	10.0%	23%	15,739,323	1,209,744

Note: undistributed profits are estimated using National Accounts. The amount is equal to the net primary income of the corporate sector (including both the financial and non financial). National Accounts are detailed enough to estimate undistributed profits since 1996; for previous years we estimate them as a fixed proportion of GDP.

Table A.3: Top shares including undistributed profits (Upper Bound)

Year	Income in 2015 pesos				Share of total income			
	Top 10%	Top 1%	Top 0.1%	Top 0.01%	Top 10%	Top 1%	Top 0.1%	Top 0.01%
1990		56,663,724	207,686,008	357,789,602		25.5%	9.4%	1.6%
1991		57,681,205	209,183,254	336,849,375		24.5%	8.9%	1.4%
1992		60,544,854	217,635,897	339,205,558		23.9%	8.6%	1.3%
1993		63,356,812	223,507,085	332,740,017		24.1%	8.5%	1.3%
1994		65,147,748	230,419,058	334,775,913		23.5%	8.3%	1.2%
1995		71,577,503	251,766,618	355,931,622		23.4%	8.2%	1.2%
1996		74,361,174	260,043,469	389,438,019		23.0%	8.0%	1.2%
1997		81,930,950	295,669,802	401,301,739		24.7%	8.9%	1.2%
1998		74,036,416	251,749,674	427,866,402		22.7%	7.7%	1.3%
1999		68,353,836	222,901,295	391,890,248		22.1%	7.2%	1.3%
2000		72,565,270	245,224,636	384,401,297		21.2%	7.2%	1.1%
2001		75,075,490	252,573,677	420,248,787		21.5%	7.2%	1.2%
2002		75,210,114	252,173,387	415,043,589		21.2%	7.1%	1.2%
2003		71,482,031	236,806,775	479,380,599		19.3%	6.4%	1.3%
2004	22,656,411	96,013,437	356,780,444	473,763,510	56.9%	24.1%	9.0%	1.2%
2005	23,335,900	97,878,894	363,078,919	524,106,807	55.1%	23.1%	8.6%	1.2%
2006	25,742,250	112,804,643	435,317,201	560,479,688	56.2%	24.6%	9.5%	1.2%
2007	28,791,943	129,582,345	514,302,936	613,723,301	60.2%	27.1%	10.8%	1.3%
2008	25,648,048	103,866,745	374,099,124	558,349,399	56.5%	22.9%	8.2%	1.2%
2009	29,803,593	131,651,010	508,925,033	588,196,457	59.8%	26.4%	10.2%	1.2%
2010	34,901,145	163,407,316	666,978,405	693,532,903	61.7%	28.9%	11.8%	1.2%
2011	34,137,056	152,443,661	605,807,128	715,218,708	58.8%	26.3%	10.4%	1.2%
2012	34,291,174	147,808,656	568,363,236	724,930,580	57.9%	25.0%	9.6%	1.2%
2013	34,922,940	146,889,422	557,809,604	709,567,793	57.4%	24.1%	9.2%	1.2%
2014	36,395,915	155,391,823	592,924,873	756,611,693	58.7%	25.1%	9.6%	1.2%
2015	38,532,075	170,318,257	670,753,685	869,040,942	60.2%	26.6%	10.5%	1.4%

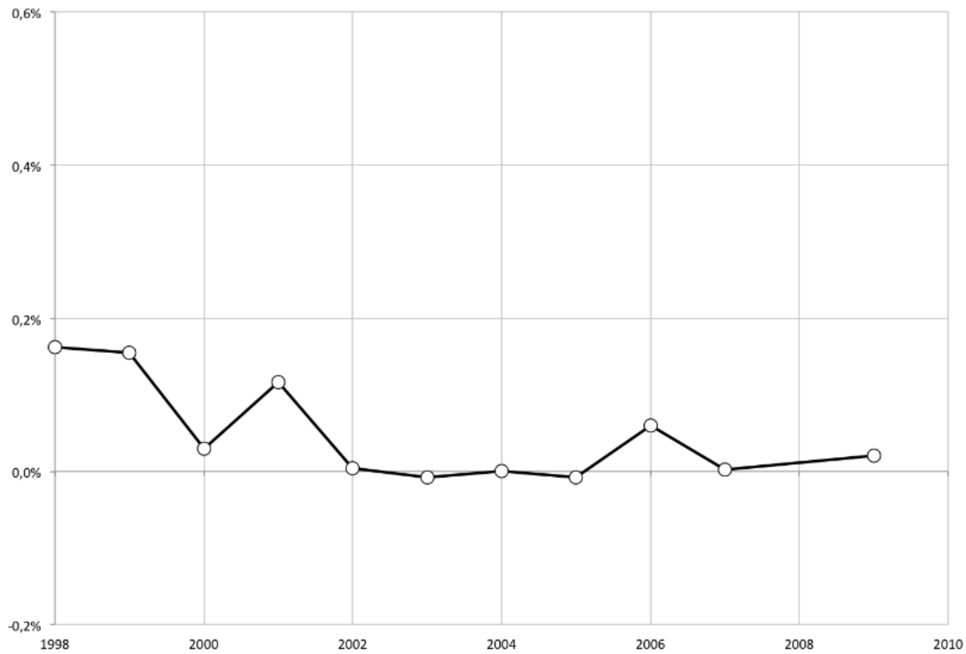
Note: undistributed profits are estimated using National Accounts. The construction of both upper and lower bounds is in Section 1.3.2

Table A.4: Top shares including undistributed profits (Lower Bound)

Year	Average income in 2015 pesos				Adjusted share of total income			
	Top 10%	Top 1%	Top 0.1%	Top 0.01%	Top 10%	Top 1%	Top 0.1%	Top 0.01%
1990		50,957,520	176,404,650	342,375,599		23%	8%	1.5%
1991		51,624,040	175,977,927	320,487,330		22%	7%	1.4%
1992		54,040,704	181,980,205	321,636,087		21%	7%	1.3%
1993		56,609,220	186,516,845	314,512,943		22%	7%	1.2%
1994		58,021,481	191,352,919	315,525,932		21%	7%	1.1%
1995		63,695,664	208,558,446	334,640,638		21%	7%	1.1%
1996		66,423,333	216,528,296	367,995,759		21%	7%	1.1%
1997		71,951,095	240,960,323	374,343,445		22%	7%	1.1%
1998		67,218,645	214,374,710	409,449,753		21%	7%	1.3%
1999		62,946,293	193,257,190	377,283,008		20%	6%	1.2%
2000		65,683,558	207,499,153	365,811,928		19%	6%	1.1%
2001		68,148,615	214,600,604	401,537,418		20%	6%	1.1%
2002		68,162,484	213,538,337	396,006,028		19%	6%	1.1%
2003		65,714,805	205,190,890	463,801,757		18%	6%	1.3%
2004	20,343,249	83,034,197	285,628,363	438,703,064	51%	21%	7%	1.1%
2005	21,059,614	85,106,567	293,061,130	489,605,288	50%	20%	7%	1.2%
2006	22,560,098	95,535,070	341,729,896	514,731,084	49%	21%	7%	1.1%
2007	24,617,062	107,674,903	397,017,312	556,881,391	52%	23%	8%	1.2%
2008	23,063,767	90,758,827	304,820,232	525,084,424	51%	20%	7%	1.2%
2009	25,271,827	109,440,610	393,125,574	533,151,617	51%	22%	8%	1.1%
2010	28,281,674	130,233,258	489,388,120	607,469,237	50%	23%	9%	1.1%
2011	28,422,982	123,202,249	445,527,491	636,247,623	49%	21%	8%	1.1%
2012	29,075,771	120,590,055	415,965,411	648,758,147	49%	20%	7%	1.1%
2013	29,854,671	119,945,448	404,019,045	631,728,490	49%	20%	7%	1.0%
2014	30,939,039	125,872,054	421,458,483	668,859,638	50%	20%	7%	1.1%
2015	32,064,796	134,751,915	460,837,618	760,548,640	50%	21%	7%	1.2%

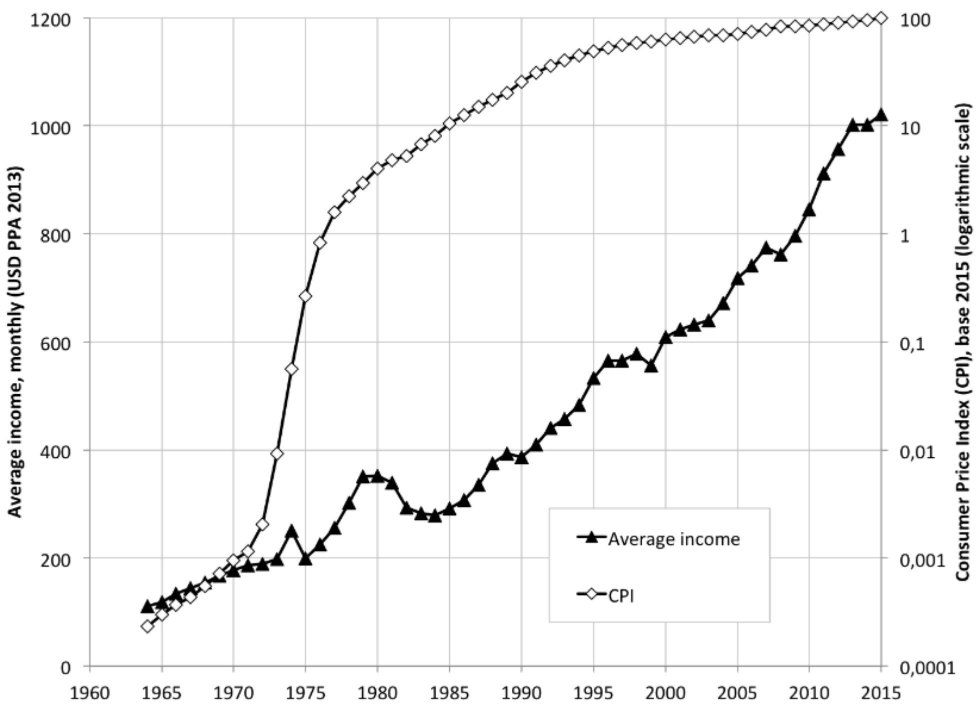
Note: undistributed profits are estimated using National Accounts. The construction of both upper and lower bounds is in Section 1.3.2

Figure A.1: Difference in Top 1% Share, with and without Capital Gains (1998-2009)



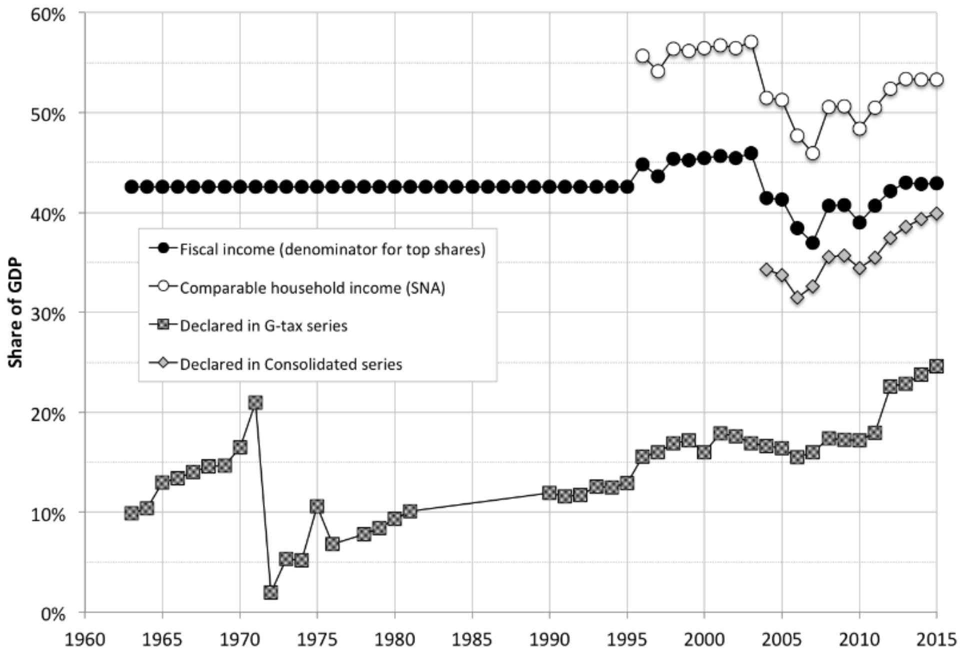
Own estimates based on the short detailed series of IGC tabulations that include capital gains declared by income-bracket (See Section 1.2.1). Reading: in 1998, the top 1% share estimate including capital gains is higher of only 0.16 percentage points compared to the one not including them. After 2002, the difference is generally close to 0. Differences are calculated without applying any adjustment to the series.

Figure A.2: Average Real Income (in 2013 USD PPA) and CPI (base 2015)



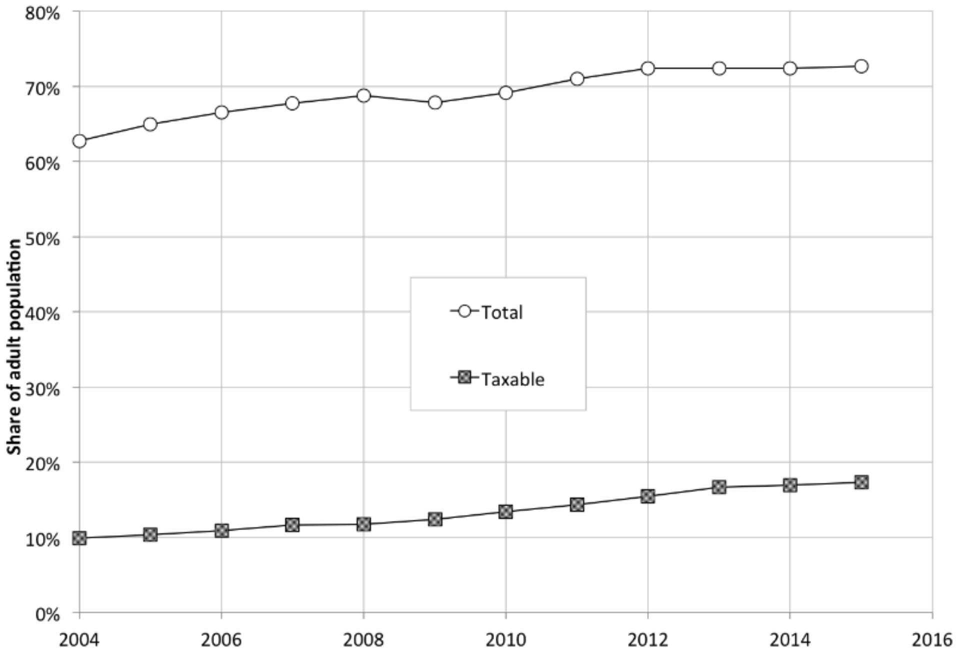
Source: Average real income based on a combination of National Accounts and Tax data (see Section 1.2.1). CPI based on World Bank data

Figure A.3: Comparison of aggregate income concepts



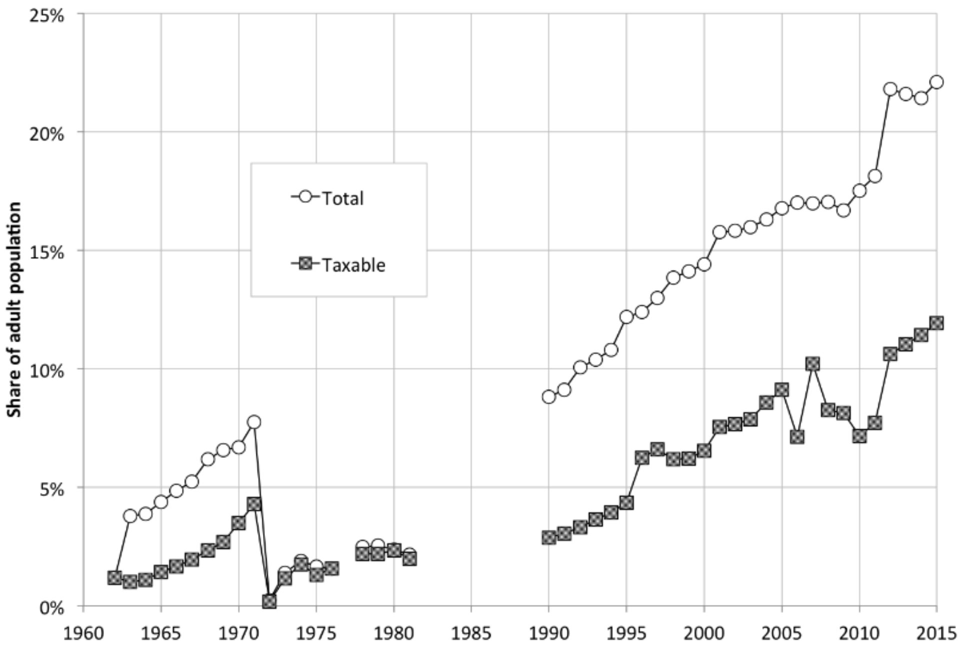
Source: authors' calculations using tax data and national accounts.

Figure A.4: Individual Tax Declarations as a Share of the Adult Population, *Consolidated* Series (2004-2015)



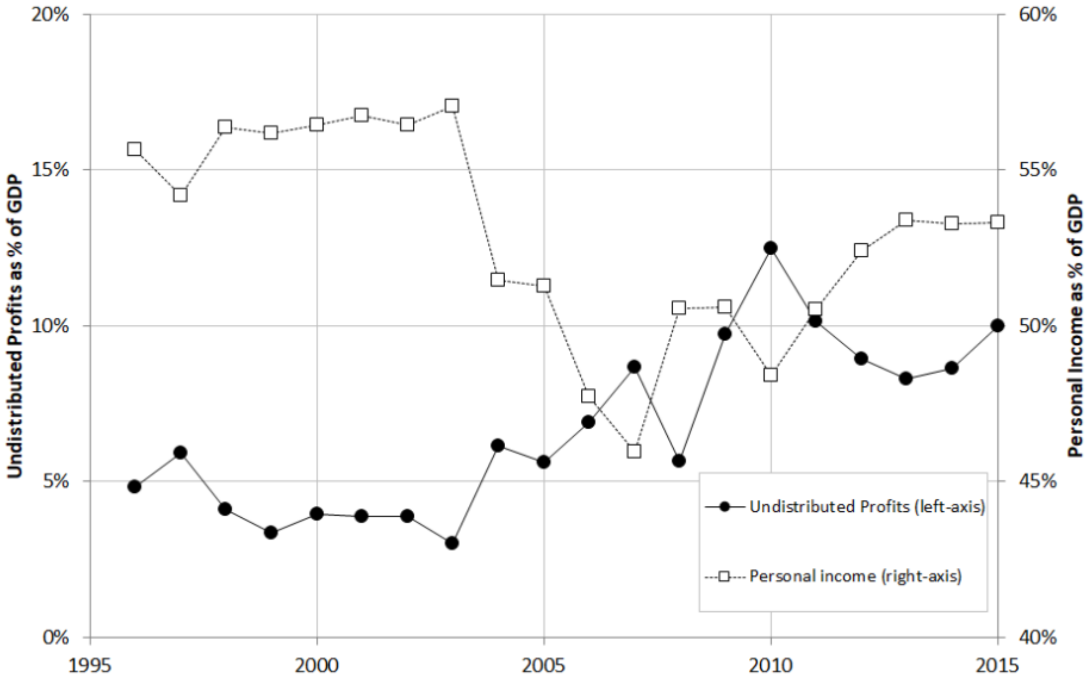
Author's estimates using tax tabulations and population estimates from World Bank. *Taxable* are those declaring income above the minimum taxable threshold. In this series, the 'taxable' population is always above or equal to 10% of the adult population

Figure A.5: Individual Tax Declarations as a Share of the Adult Population, *Global Complementario* Series (1964-2015)



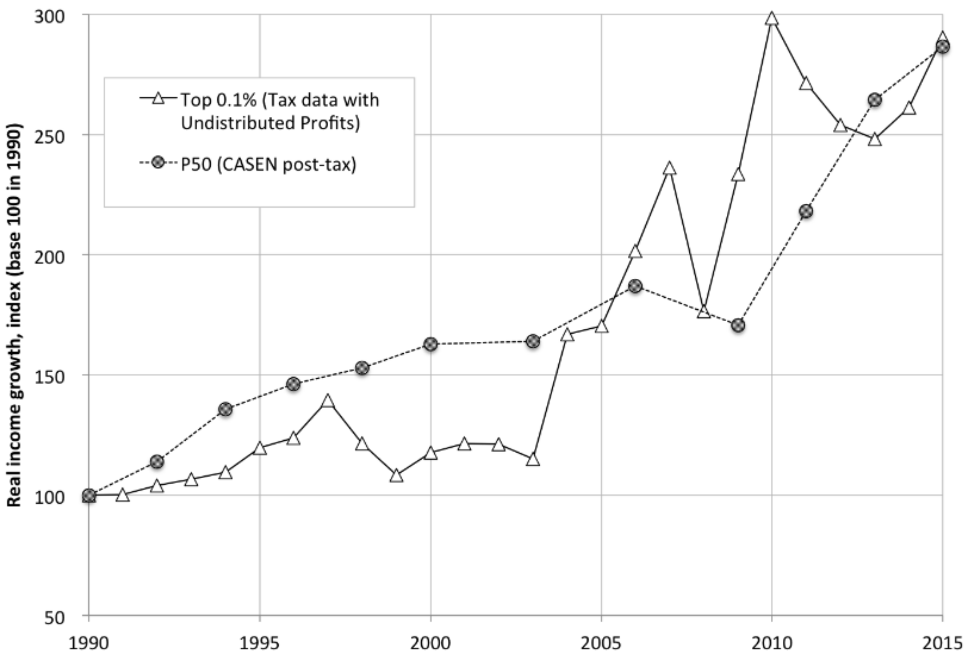
Author's estimates using tax tabulations and population estimates from World Bank. *Taxable* are those declaring income above the minimum taxable threshold.

Figure A.6: Undistributed Profits and Personal Income (1996-2015)



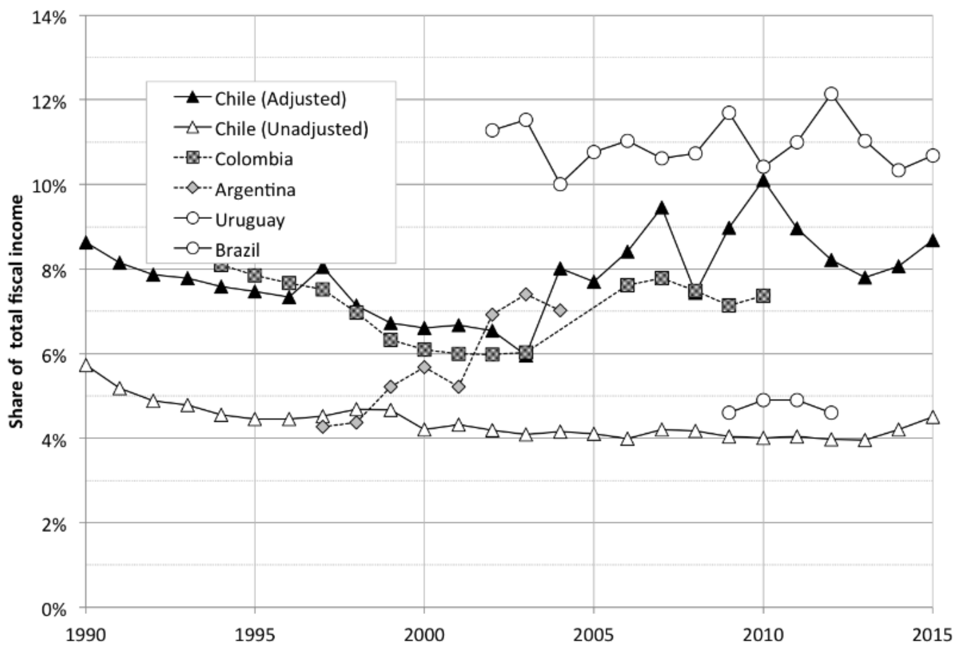
Authors' estimates using National Accounts.

Figure A.7: Real income growth. Top 0.1% in tax data vs. median income in CASEN



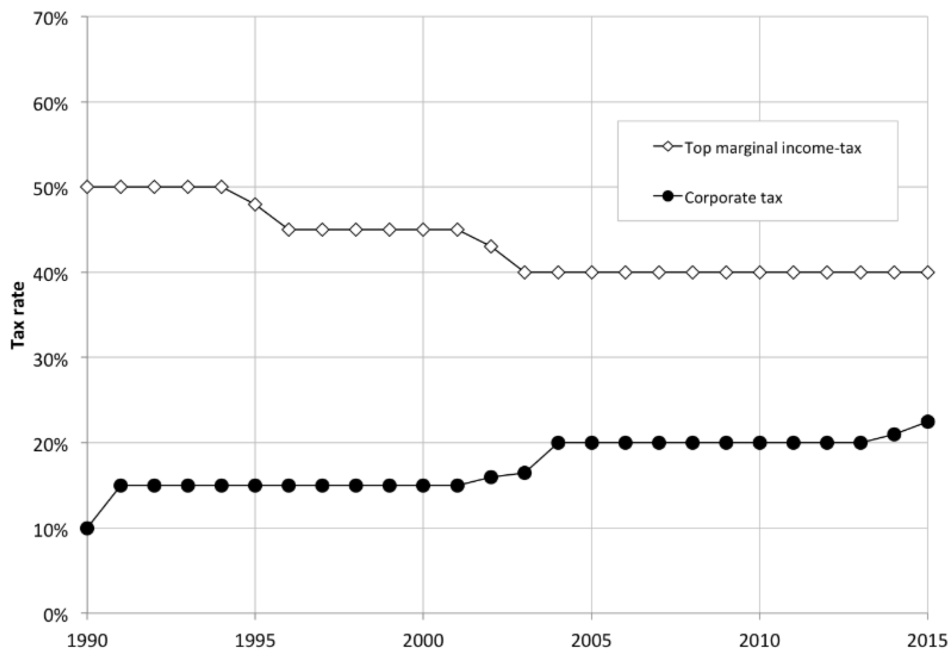
.

Figure A.8: Top 0.1% Share in Latin America (1990-2015)



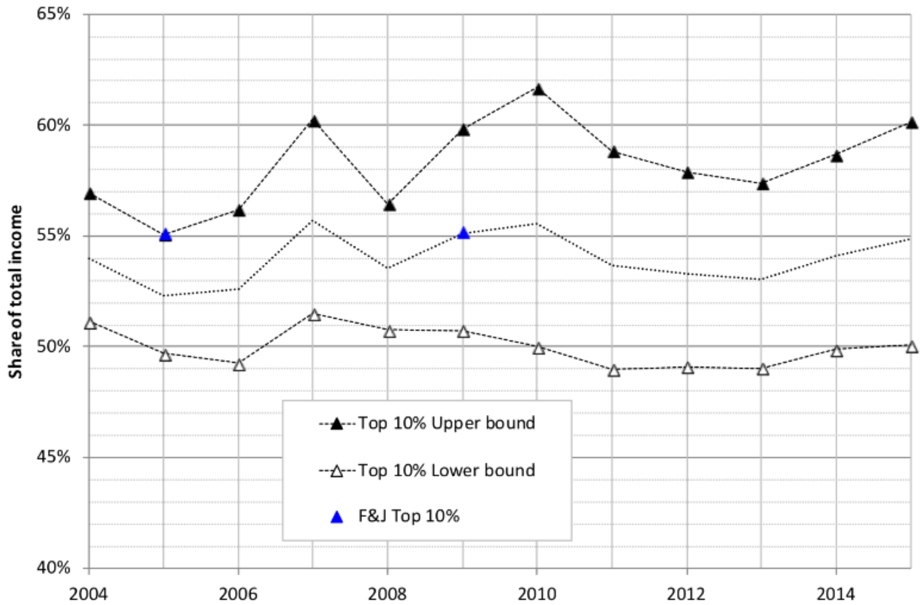
Authors' estimates for Chile, Alvaredo (2010) for Argentina, Morgan (2017) for Brazil, Alvaredo and Londoño-Vélez (2013) for Colombia, and Burdín et al. (2014) for Uruguay.

Figure A.9: Corporate tax rate vs. Top Marginal income tax rate (1990-2015)



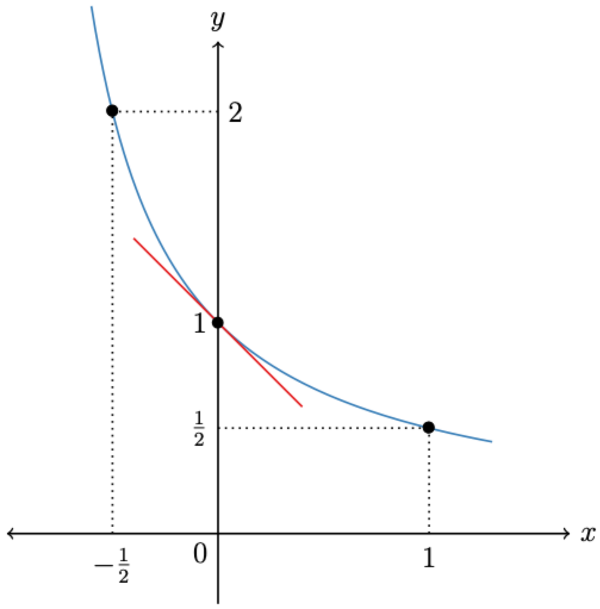
Corporate tax refers to the *Impuesto de Primera Categoría*, which is the tax on Capital income. Thus, it is paid primarily but not exclusively by corporations.

Figure A.10: Top 10% with pre-tax undistributed profits, upper and lower bounds (1990-2015)



Source: authors' estimates using tax data, detailed National Accounts (1996-2015) and (Fairfield and Jorratt De Luis, 2016). Note: in each situation, the total value of undistributed profits is imputed to the fiscal income distribution. Upper bounds assume yearly flows of undistributed profits are as concentrated in top groups as is the cumulated stock from 1984 (F.U.T.). Lower bounds assume flows to be two thirds as concentrated as the stock. The dotted line represents a central tendency, which is estimated as a geometric average of upper and lower bounds. In the absence of detailed National Accounts prior to 1996, the amount of undistributed profits in those years is estimated at nearly 4.8% of GDP, which is the estimate for 1996. Estimates from (Fairfield and Jorratt De Luis, 2016) using their definition Y_AcrdProf are displayed for comparison.

Figure A.11: Sensitivity of Top Income Share to Errors in Total Income



If the actual total income of households is $x\%$ higher than the value we estimate, real top shares are obtained by multiplying their estimate by a scaling factor $y = 1/(1 + x)$. For instance, if real income is twice the estimated value ($x = 100\%$), then the scaling factor to be applied to top shares is $y = 1/2$. For reasonable errors ($-20\% < x < 20\%$), the relation is close to linear. That is, if total income is 10% higher than we estimated, top income shares will be close to 10% lower than estimated.

Appendix B

Income Under the Carpet: What Gets Lost Between the Measure of Capital Shares and Inequality

sectionFigures by Country

Table B.1: Average structure of Total Capital Income in the Balanced Panel by Country, 1995-2015.

Country	Capital Income Share (%)	Cumulated Share (%)
Germany	20.9	20.9
United Kingdom	16.3	37.2
France	14.1	51.3
Italy	13.8	65.1
Canada	6.6	71.7
Netherlands	4.8	76.5
Switzerland	3.4	79.9
Norway	2.8	82.7
Belgium	2.8	85.4
Sweden	2.6	88.0
Austria	2.2	90.2
Poland	1.9	92.2
Greece	1.6	93.8
Denmark	1.6	95.4
Finland	1.5	96.9
Portugal	1.1	98.0
Czech Republic	1.0	99.0
Hungary	0.5	99.5
Slovakia	0.4	99.9
Estonia	0.1	100.0

Lecture: On average, Germany produced near 20% of the total capital income represented in figure 2.2. The first 5 out of 20 countries in the list produced more than 70%.

Table B.2: Average structure of Total National Income in the Balanced Sub-Panel by Country, 1995-2013.

Country	National Income Share (%)	Cumulated Share (%)
Germany	27.3	27.3
United Kingdom	19.6	46.9
Italy	15.2	62.1
Canada	9.9	71.9
Spain	9.1	81.1
Netherlands	5.9	87
Austria	2.7	89.7
Poland	2.6	92.2
Denmark	2.2	94.4
Greece	1.9	96.3
Finland	1.7	98.1
Czech Republic	1.2	99.2
Hungary	0.8	100

Lecture: On average, Germany produced near 27% of the total national income (e.g. capital and labor income) represented in figure 2.5. The first 5 out of 13 countries in the list produced more than 80%.

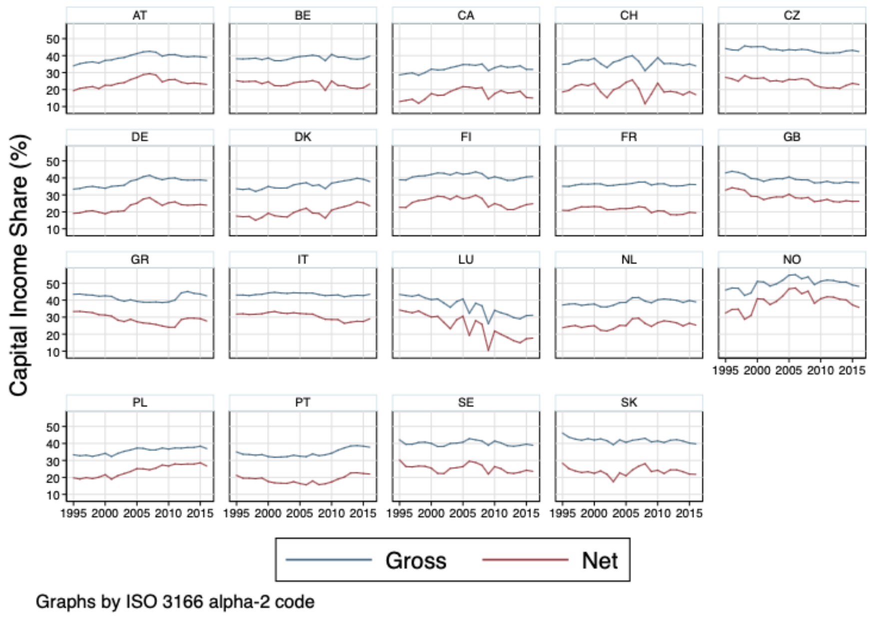
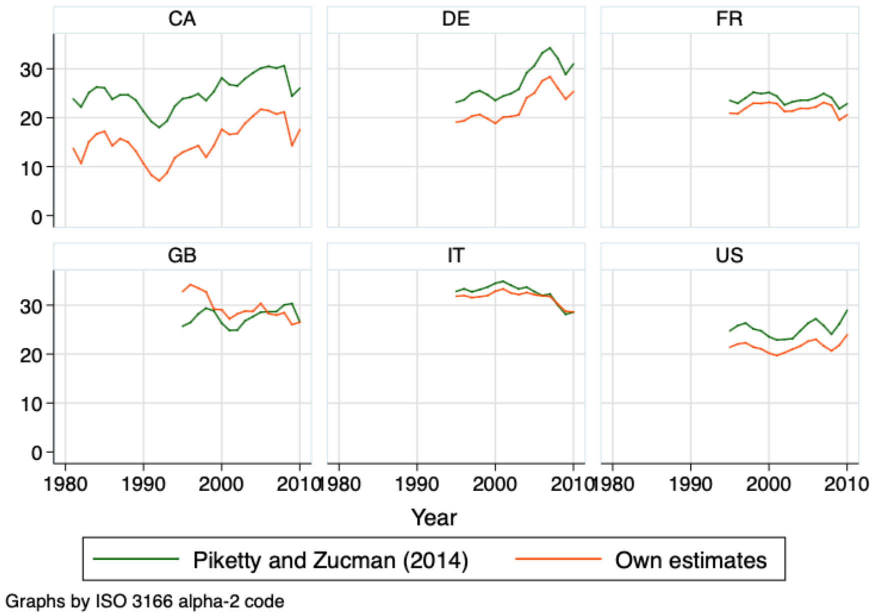
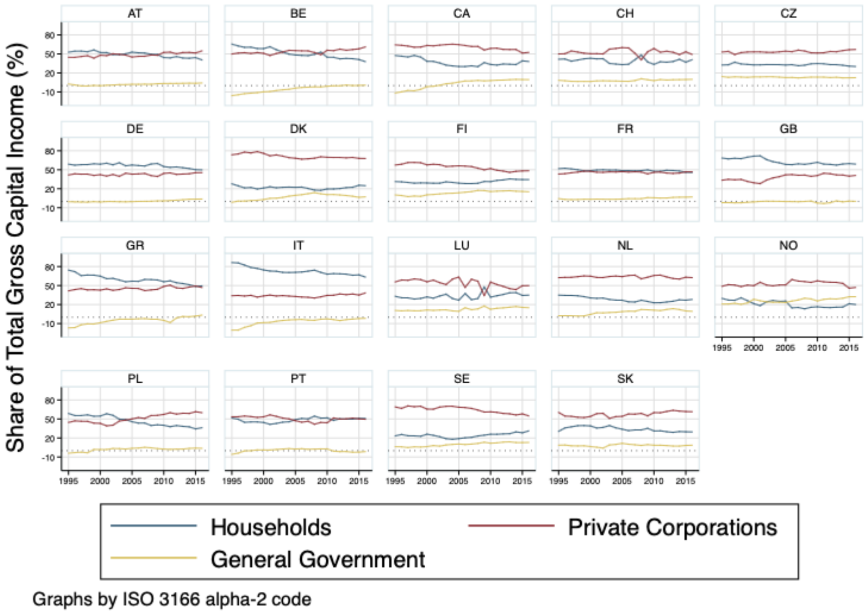


Figure B.1: Capital Share of National Income, by Country, Balanced Panel (1995-2016)

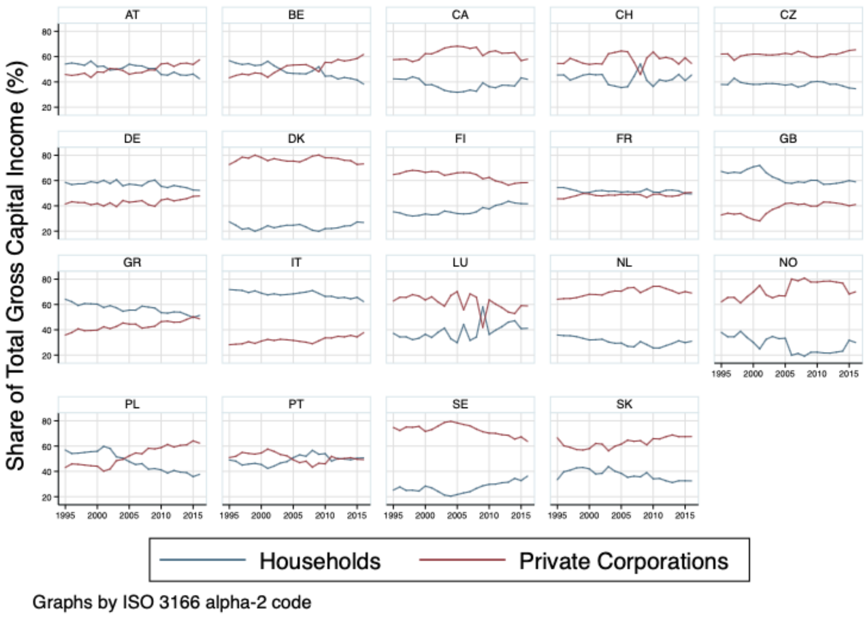


The estimates of capital shares used in this chapter, when comparable, follow the trends described by those of Piketty and Zucman (2014) relatively closely. Piketty and Zucman (2014) estimates exclude government interest.

Figure B.2: Capital Share of National Income, by Country, Balanced Panel (1995-2016)



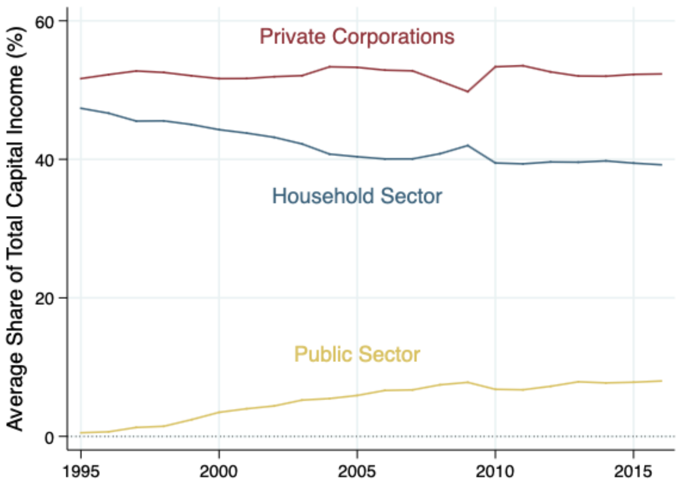
(a) Including the Public Sector



(b) Excluding the Public Sector

Both the Public Sector and Private Corporations increased their share of capital income, while the Household share decreases through the period.

Figure B.3: Household Share of Gross Capital Income in Balanced Panel, by country (1995-2016)

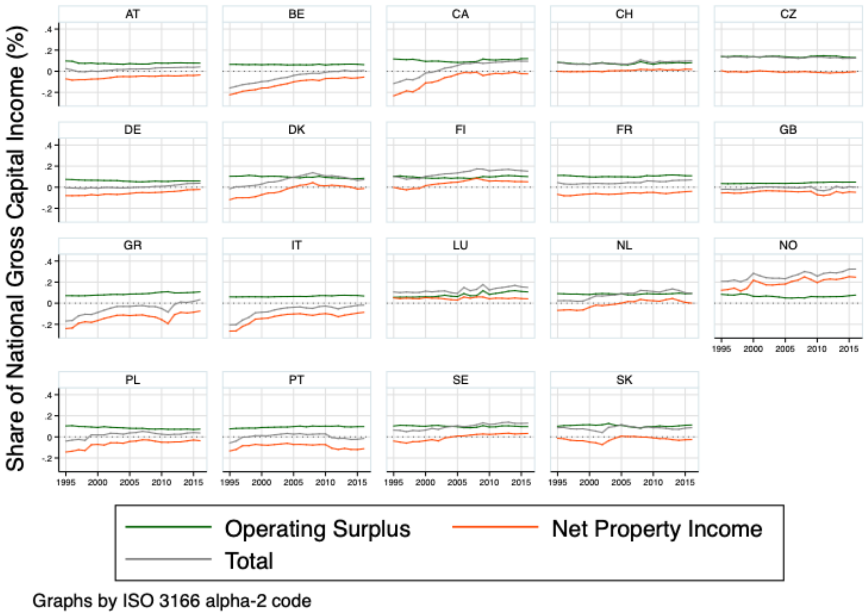


(a) Including the Public Sector

(b) Excluding the Public Sector

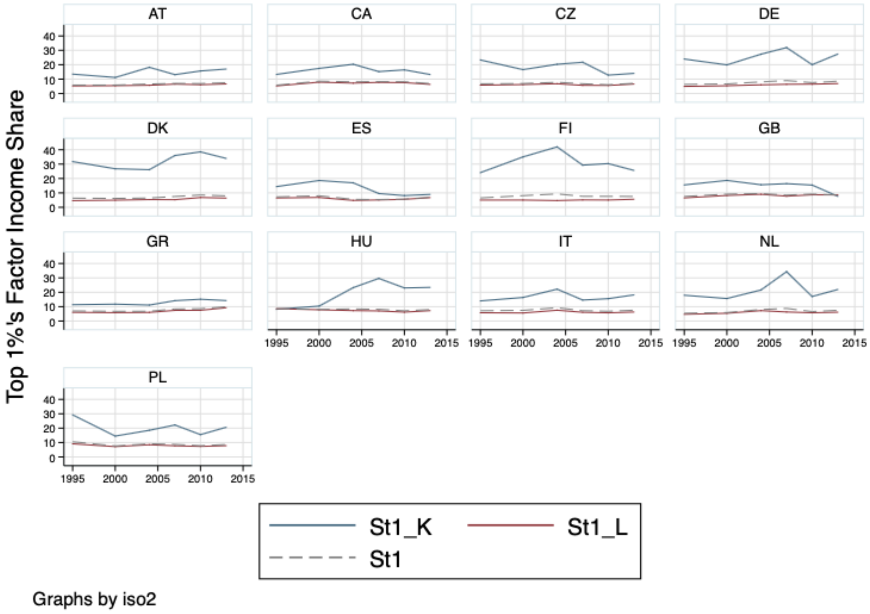
The household average share of gross capital income decreased during the period, while the corporate sector relatively stable and the public sector increases. When the corporate sector is excluded, the household share decreases, while the corporate sector increases. Countries included in the Panel: Austria, Belgium, Canada, Czech Republic, Denmark, Estonia, Finland, France, Germany, Greece, Italy, Luxembourg, Netherlands, Norway, Poland, Portugal, Slovakia, Sweden, Switzerland and the United Kingdom. Income from different countries is aggregated based on yearly average Market Exchange Rates. The United States is studied separately.

Figure B.4: Decreasing Household Share of Gross Capital Income (average), Balanced Panel (1995-2016)



The increasing trend of the public sector's capital income share is mainly driven by the reduction of the expenses related to negative Net Property Income in most cases.

Figure B.5: Decomposition of the Public Sector's Capital Income, Balanced Panel (1995-2014)



The concentration of total income in surveys (gray dashed line) in the top 1% follows extremely closely the concentration of labor income (red line). It appears as rather insensitive to the motion of concentration-estimates for capital income (blue line).

Figure B.6: Top 1% income share in Balanced Panel

The gross capital share (K) appears to grow in most cases, while the γ coefficient usually follows a decreasing trend overall. The latter does not take values above 30% in any case

Figure B.7: Capital Shares and Gamma coefficients in Balanced Panel (1995-2013)

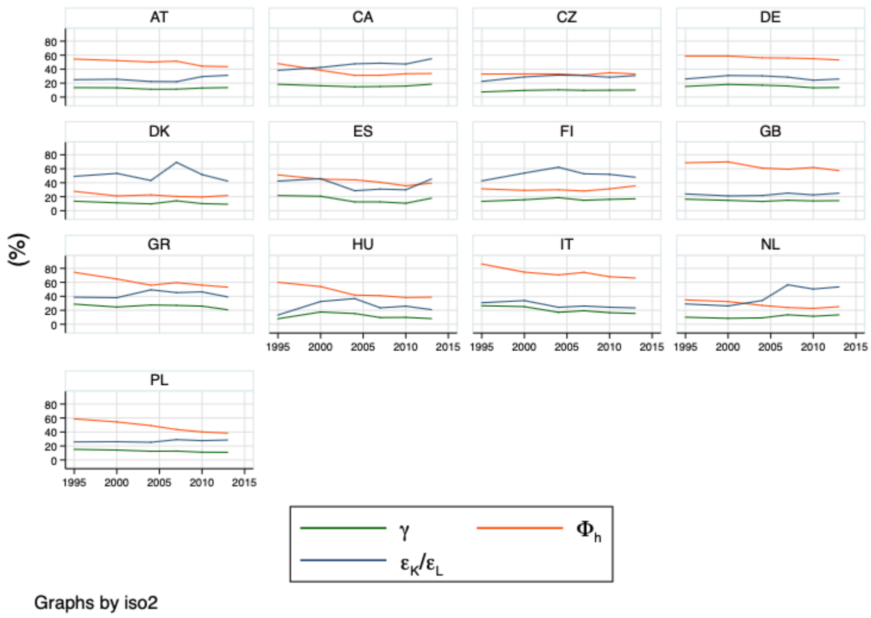
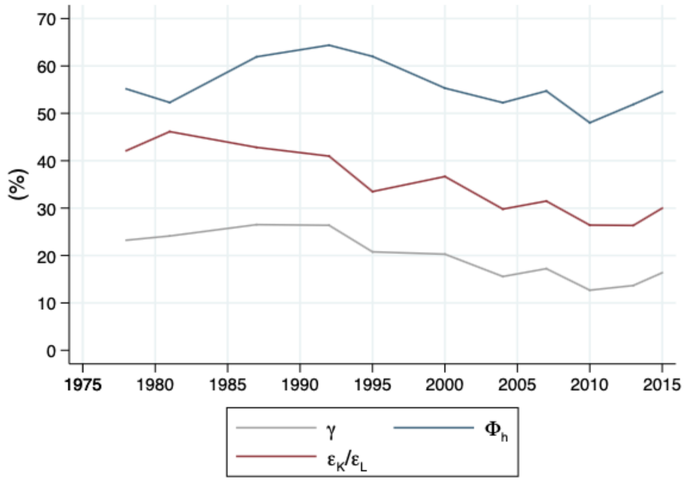
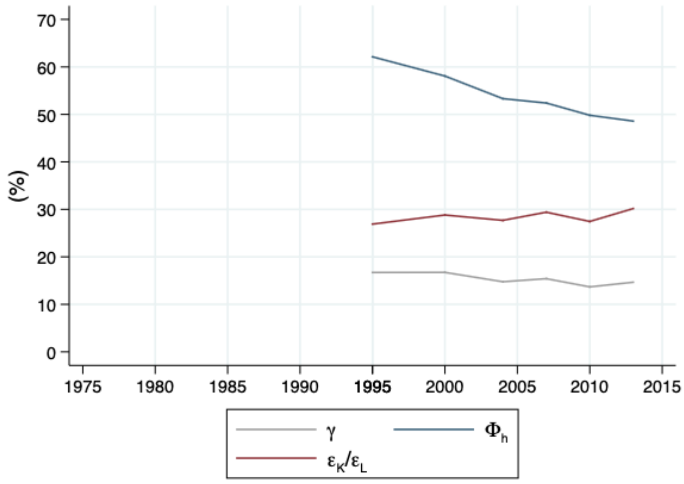


Figure B.8: Decomposition of γ by country (1995-2013)



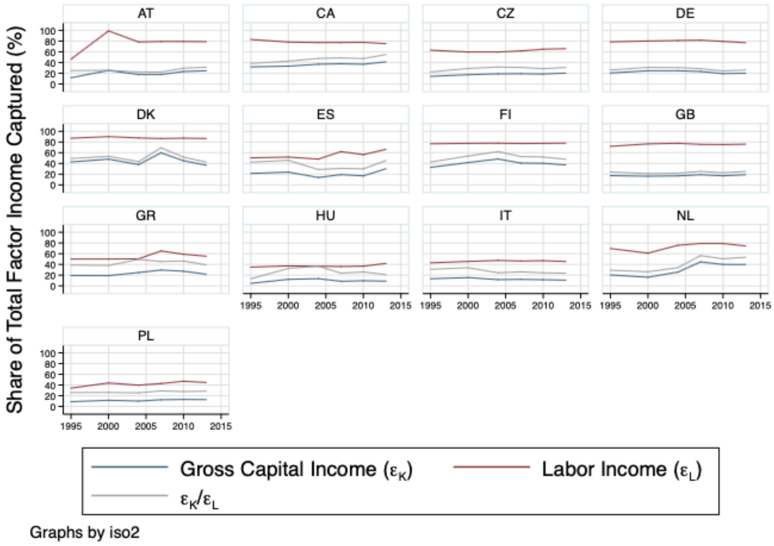
(a) United States



(b) Balanced Panel

In the United states both the ϵ_K/ϵ_L ratio and Φ_h decrease during the period. The former follows an inverted U-shape. In the Balanced Panel, Φ_h decreases rapidly during the 20 years with available data, while the ϵ_K/ϵ_L ratio remains stable. In both cases the γ falls overall.

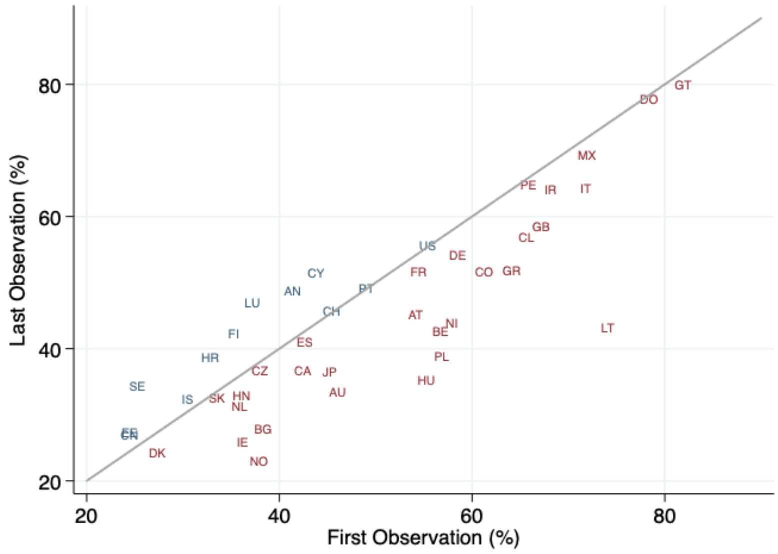
Figure B.9: Decomposition of γ (1975-2016)



Both households' capital and labor income are underestimated substantially by survey estimates (relative to figures in UN-SNA). The former appears to be systematically more underestimated than the later. Trends appear rather stable in general.

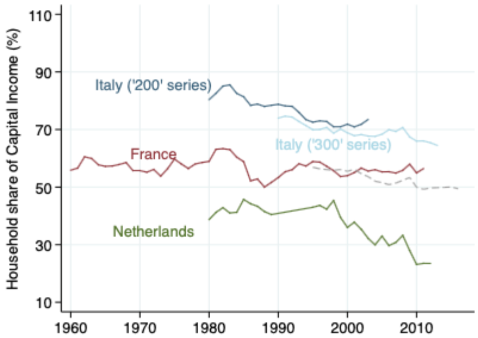
Figure B.10: Unequal Measurement Error of Factor Incomes in Surveys, Individual Countries (1995-2013)

B.1 Excluding the Public Sector

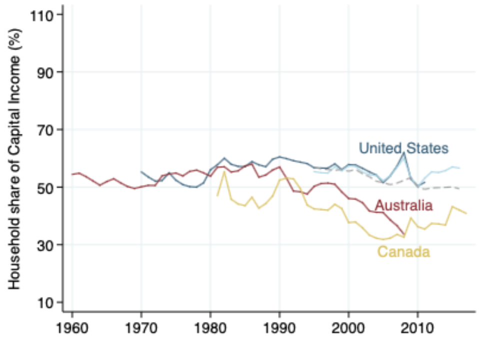


The share of capital income received by households, as opposed to private corporations (excluding the public sector), decreased in 32 out of the 46 cases that have at least 6 observations during the period. That is, it decreases in near 70% of cases. The countries that experienced an increase are those which already had relatively low shares to start with.

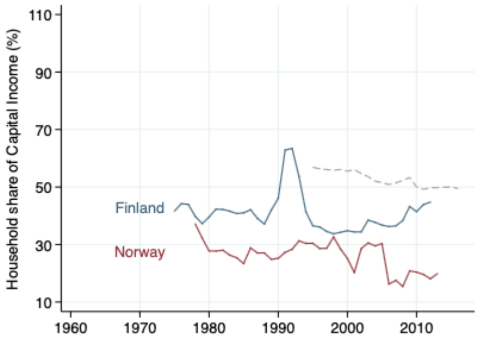
Figure B.11: Decreasing Household Share of Gross Capital Income, Unbalanced Panel (1995-2015)



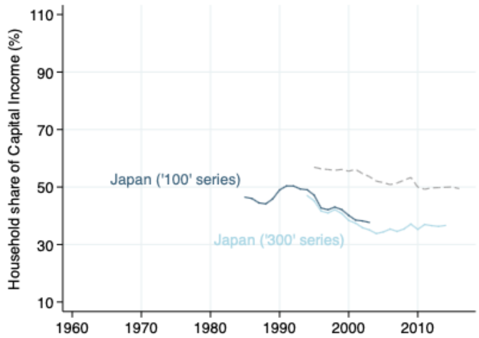
(a) Selected European countries



(b) English Speaking Countries



(c) Scandinavian Countries

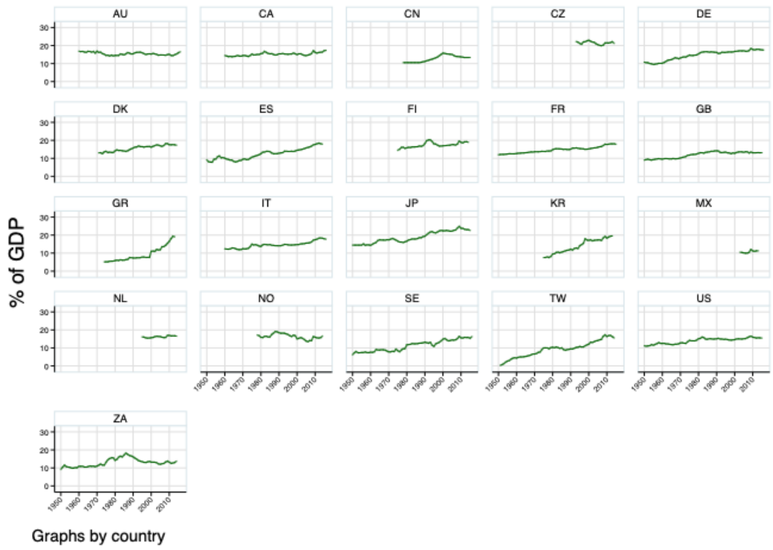


(d) Japan

The grey dashed line represents the aggregate tendency in the balanced panel of 19 countries presented in figure 2.2 (this time excluding the public sector). Most countries with long-run data exhibit a decreasing trend starting before the beginning of the panel, around 1990. Relatively more stable trends are described in previous decades.

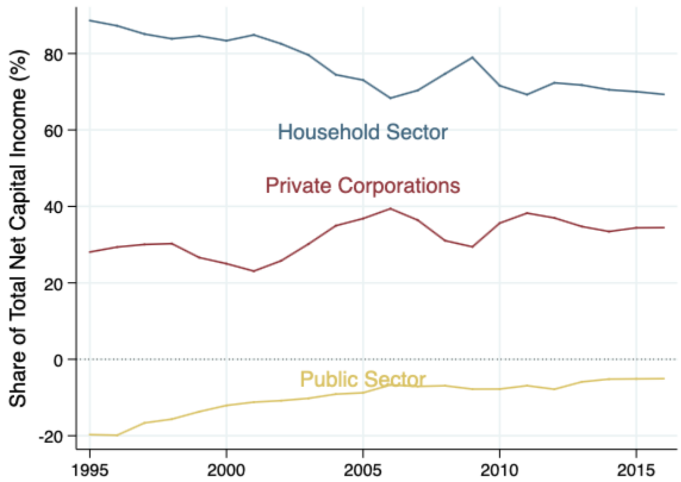
Figure B.12: Decreasing Household Share of Gross Capital Income, Long-run

B.2 Figures using Net Capital Income



Source: `www.wid.world`

Figure B.13: Capital depreciation as a share of GDP, 1950-2015

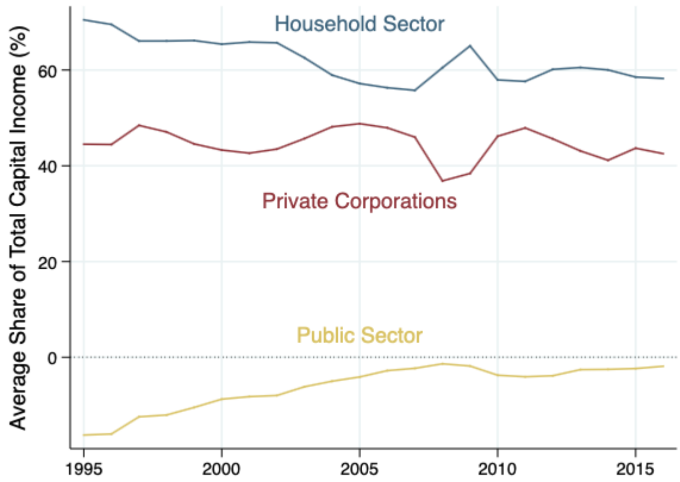


(a) Including the Public Sector

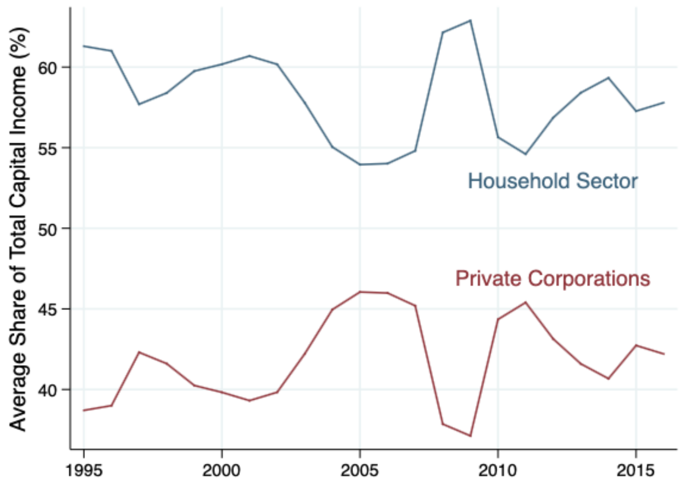
(b) Excluding the Public Sector

Both the Public Sector and Private Corporations increased their share of net capital income, while the Household share decreases through the period. This figure is the same than figure 2.2 but using net capital income instead of gross capital income. Due to data limitations, less countries are included in the Panel (12 instead of 19). Individual countries are available in figure B.16. Observed trends of sector shares are rather similar, yet the levels vary, mostly due to the fact that private corporations are most affected by capital depreciation.

Figure B.14: Decreasing Household Share of Net Capital Income, Balanced Panel (1995-2016)



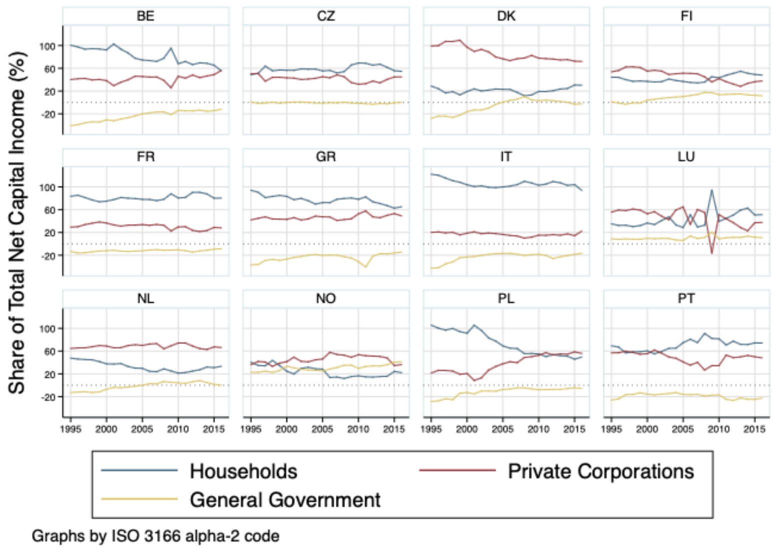
(a) Including the Public Sector



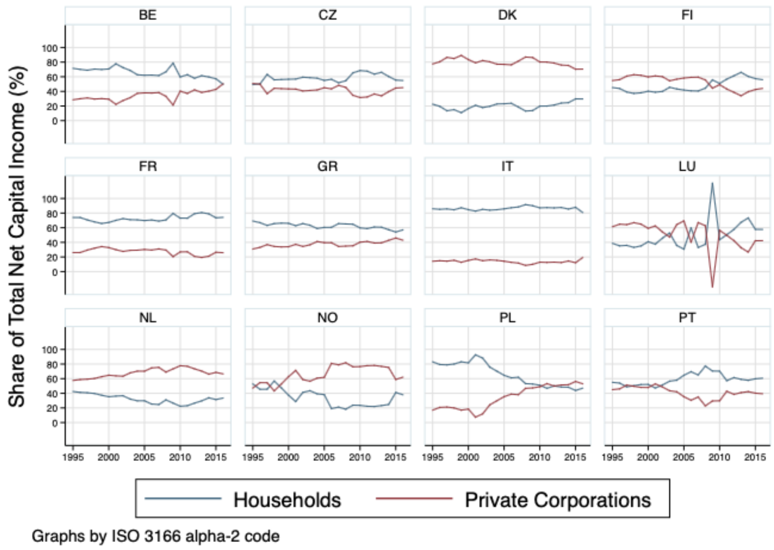
(b) Excluding the Public Sector

On average, the household share of net capital income decreases, while the corporate sector's share remains relatively stable and the public sector's share increases. When the public sector is excluded, in this case, the increase in the share of corporations is somewhat counterbalanced at the end of the period. This figure is the same than figure B.4 but using net capital income instead of gross capital income. Due to data limitations, less countries are included in the Panel (12 instead of 19). Individual countries are available in figure B.16.

Figure B.15: Decreasing Household Share of Net Capital Income (average), Balanced Panel (1995-2016)



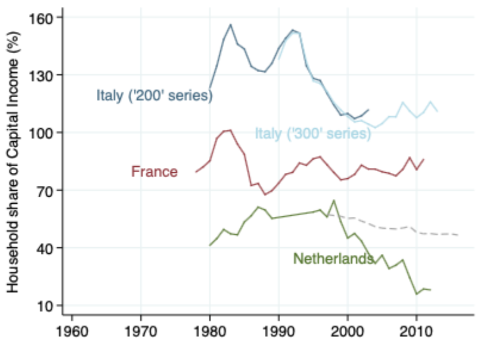
(a) Including the Public Sector



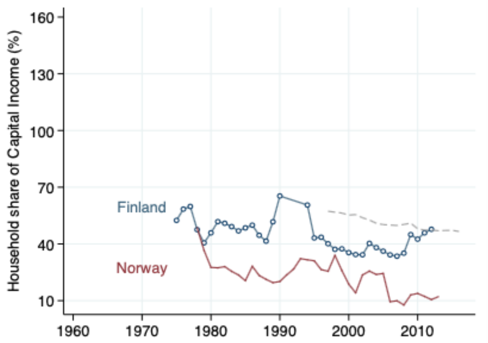
(b) Excluding the Public Sector

Both the Public Sector and Private Corporations increased their share of capital income, while the Household share decreases through the period.

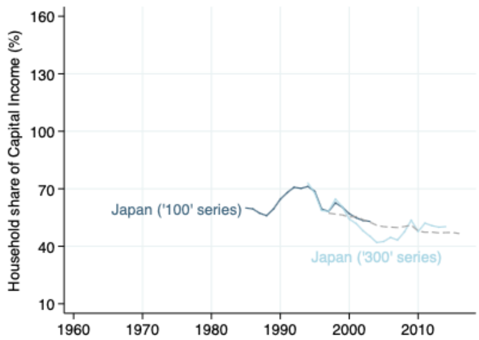
Figure B.16: Household Share of Net Capital Income in Balanced Panel, by country (1995-2016)



(a) Selected European countries



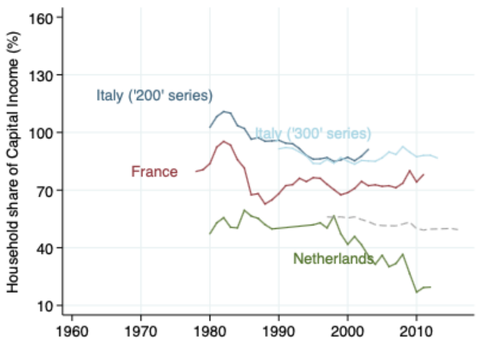
(b) Scandinavian Countries



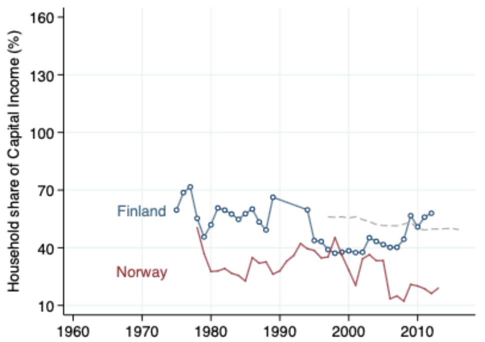
(c) Japan

Most countries with long-run data exhibit a decreasing trend starting before the beginning of the panel, around 1990. Relatively more stable trends are described in previous decades.

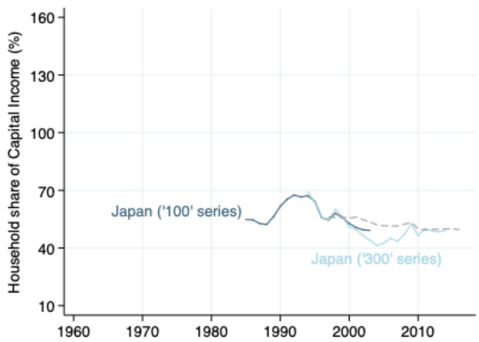
Figure B.17: Decreasing Household Share of Gross Capital Income (incl. Public Sector), Long-run



(a) Selected European countries



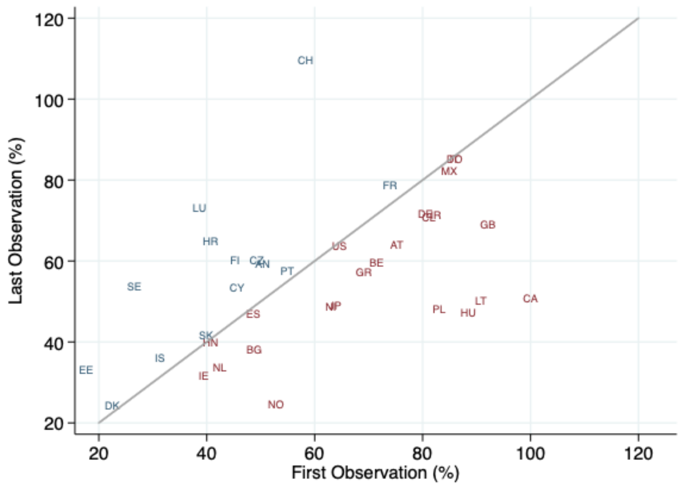
(b) Scandinavian Countries



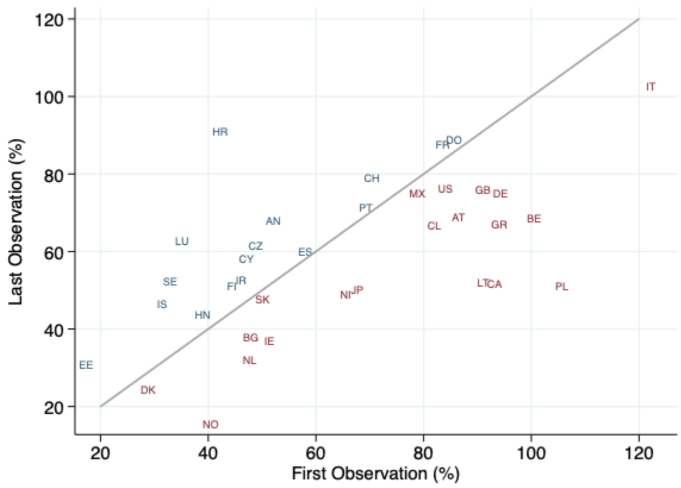
(c) Japan

Most countries with long-run data exhibit a decreasing trend starting before the beginning of the panel, around 1990. Relatively more stable trends are described in previous decades.

Figure B.18: Decreasing Household Share of Gross Capital Income (w/o Public Sector), Long-run



(a) Excluding the Public Sector



(b) Including the Public Sector

In the series figure a) the share of capital income received by households, decreased in 23 out of the 37 cases that have at least 6 observations during the period. That is 62% of cases. When the public sector is included, in figure b), it decreases in 56.7% of cases.

Figure B.19: Household Share of Net Capital Income, Unbalanced Panel (1995-2015)

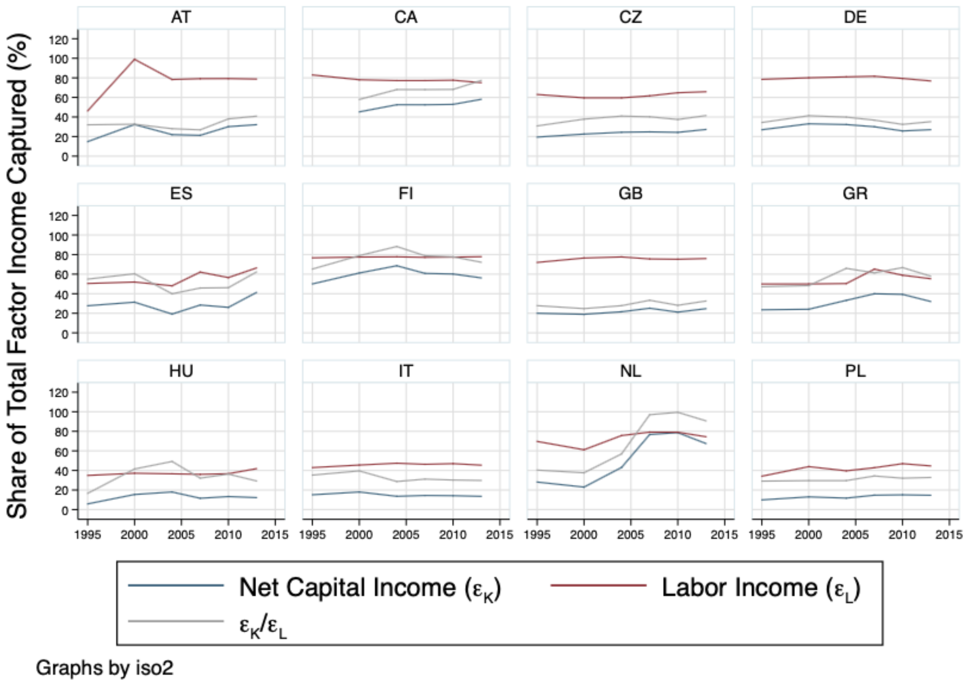


Figure B.20: Unequal Measurement Error of Factor Incomes in Surveys (Net Capital Income), Individual Countries (1995-2013)

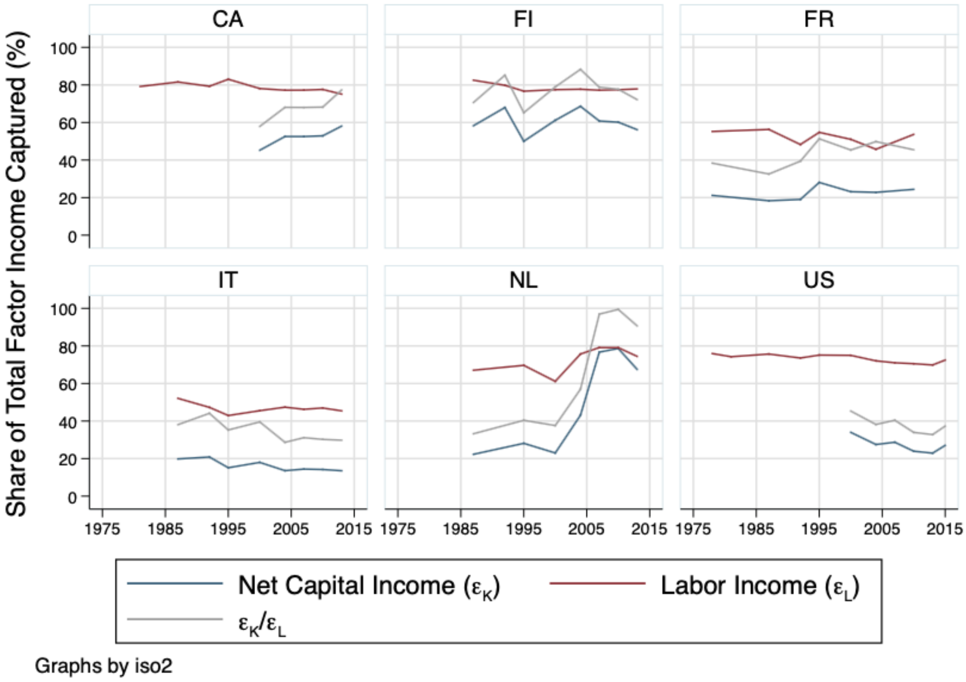
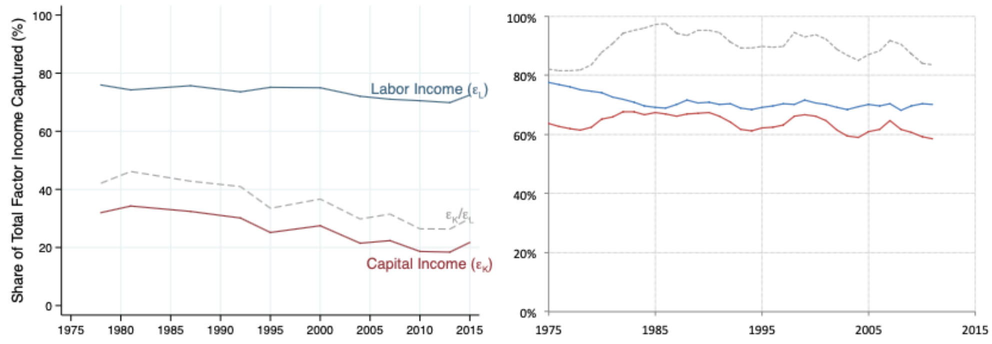


Figure B.21: Unequal Measurement Error of Factor Incomes in Surveys (Long-Run), Individual Countries (1995-2013)

B.3 Additional Figures

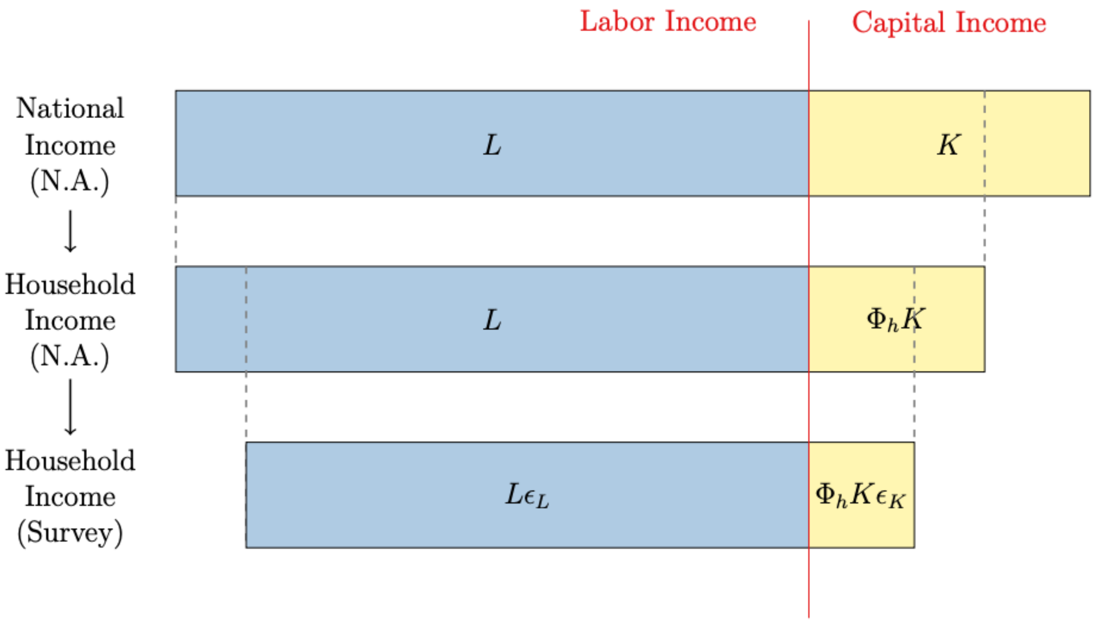


(a) Survey Data

(b) Tax Data (1975-2015)

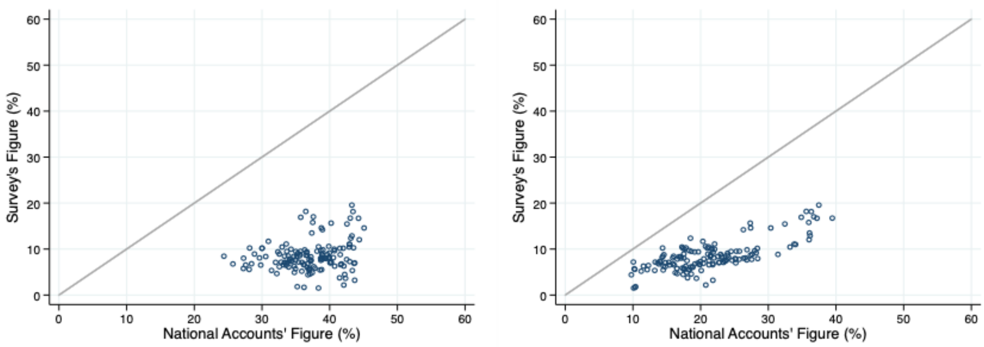
In the case of the US, the same conclusions than figure B.10 are valid for surveys. However, capital income appears to be progressively more underestimated throughout the period. Tax data seems to be approximately as good as survey data to capture labor income, but substantially better at recording capital income.

Figure B.22: Unequal Error of Factor Incomes, The U.S.



Total National Income can be divided in Labor and Capital shares (L and K respectively). Household Income in National Accounts includes the full Labor share, but only a part (Φ_K) of Capital Income. Due to both a narrower definition of income and measurement error, surveys capture only a part of each factor income (ϵ_L and ϵ_K).

Figure B.23: From National Income in National Accounts to Household Income in Surveys



(a) National Capital Share (b) Household's Capital Share

These are, for all data-points in the panel of section 2.1.4, estimates of the capital shares in surveys and UN-SNA. The x-axis of subfigure a) represents the capital share of national income, whereas in subfigure b) it represents the capital share of the household sector. Surveys underestimate substantially the capital share in both cases.

Figure B.24: Unequal Measurement Error of the Capital Share

Appendix C

The Weight of the Rich: Improving Surveys with Tax Data

C.1 Country Specific Income Concepts and Observational Units

C.1.1 Brazil

In reconciling 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 exclude asset transfers are reported in this category, as they are not income flows but transfers of a stock of wealth. These are lump sum payments related to donations and inheritances, as well as the incorporation of company reserves and the disbursement of shares as bonuses.

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) and unemployment insurance. The latter is taken from other income sources declared (V1273) and estimated as income from this source that is reported between 1 and 2 monthly minimum wages. 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 2 monthly minimum wages. 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 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, 2017). 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.

C.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 C.1 gives a comprehensive view on what types of income are assumed to pay taxes or not. For further comments on the definition

Table C.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 income corresponding to tax data, please refer to Atria et al. (2018).

C.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 3.3.2, we interpolate the tabulations using a generalized Parteo interpolation (*gpinter*). 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 of their Data. 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 (2007b) 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.

C.2 Shape of the Bias

Figures C.1-C.5 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.

C.3 Structure of the Corrected Population

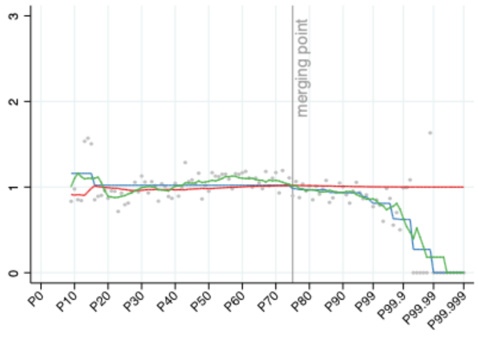
Tables C.2-C.6 show the structure of the corrected population for all years in all sampled countries.

Table C.2: Structure of Corrected Population in Brazil, 2007-2015

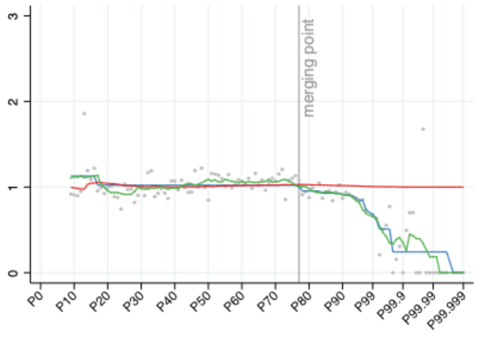
Year	Population over Merging Point (% total population)		Corrected population		
	Tax data [2]	Survey [3]	Total [4] = [2] − [3]	Share inside survey support [5]	Share outside survey support [6]
2007	1.0%	0.7%	0.33%	98.2%	1.8%
2008	1.0%	0.6%	0.44%	97.2%	2.8%
2009	1.0%	0.5%	0.51%	99.3%	0.7%
2011	2.0%	1.4%	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).

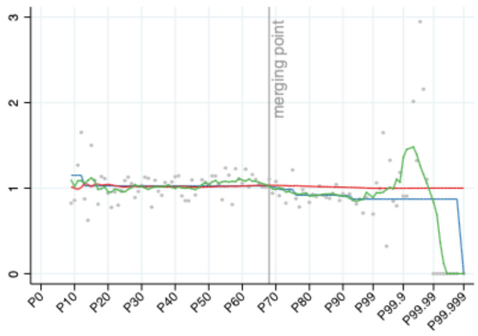
Figure C.1: Merging Points in Norway, 2004-2013



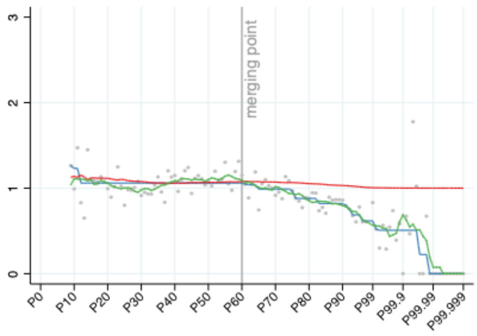
(a) Norway 2004



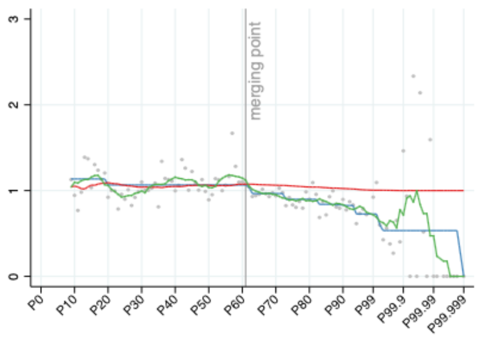
(b) Norway 2005



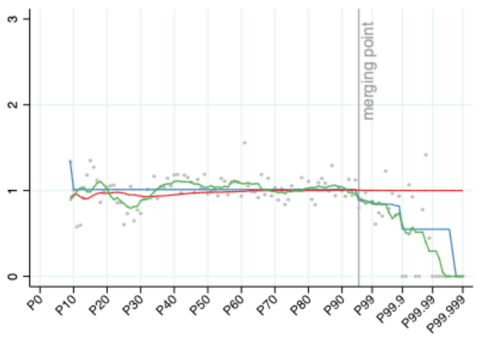
(c) Norway 2006



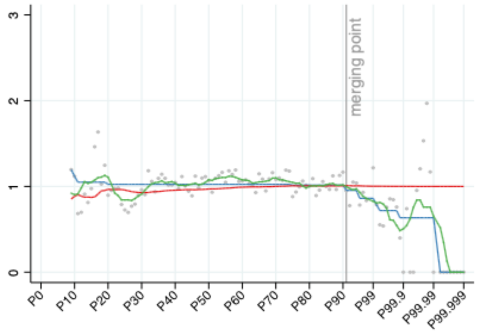
(d) Norway 2007



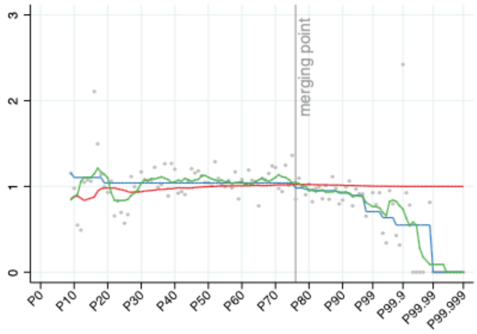
(e) Norway 2008



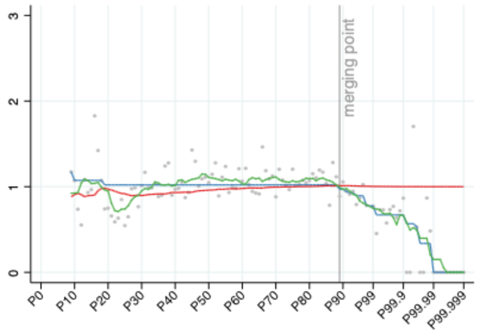
(f) Norway 2009



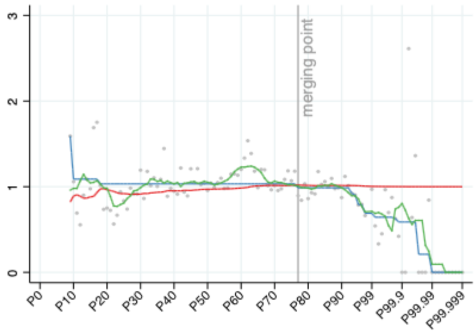
(g) Norway 2010



(h) Norway 2011



(i) Norway 2012



(j) Norway 2013

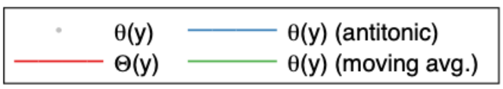
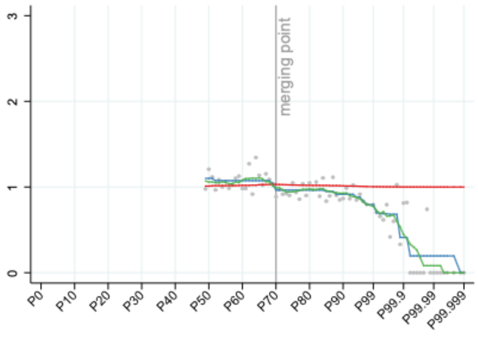
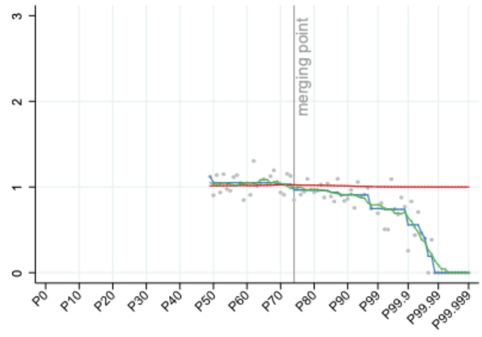


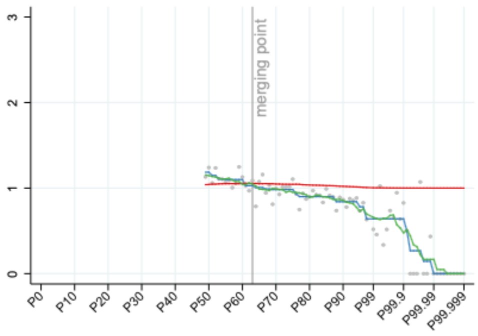
Figure C.2: Merging Points in France, 2004-2013



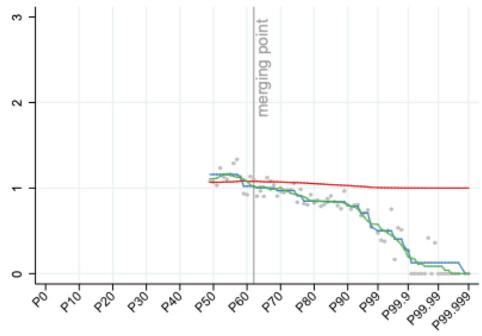
(a) France 2004



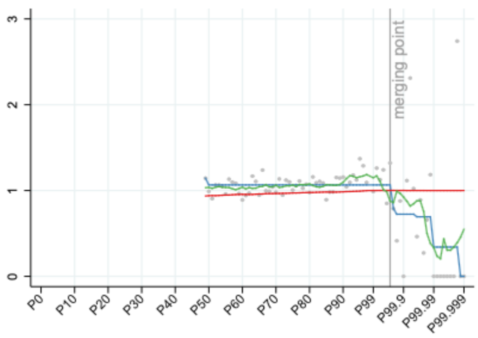
(b) France 2005



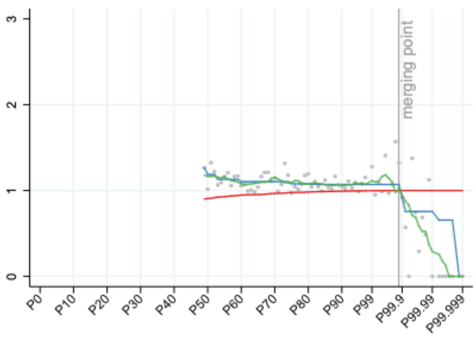
(c) France 2006



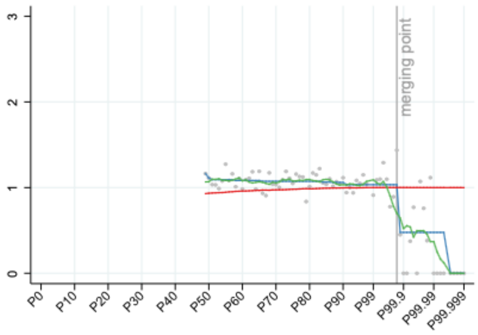
(d) France 2007



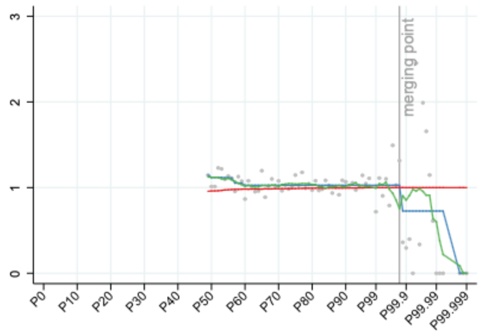
(e) France 2008



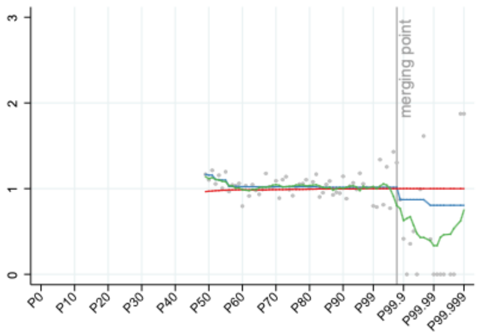
(f) France 2009



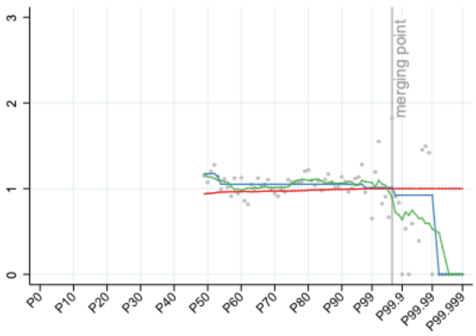
(g) France 2010



(h) France 2011



(i) France 2012



(j) France 2013

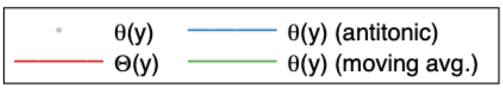
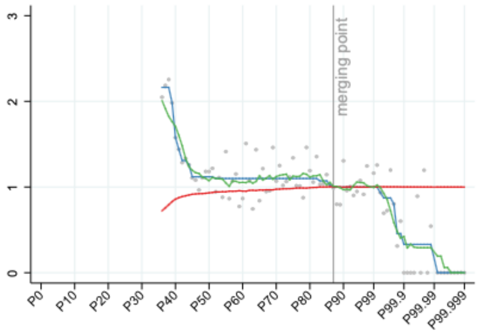
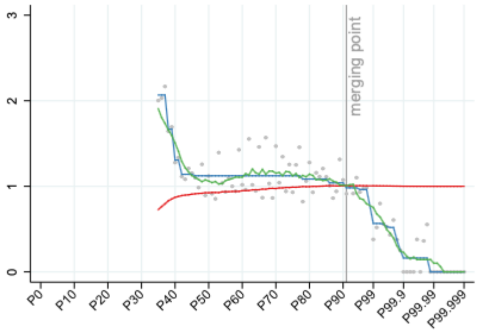


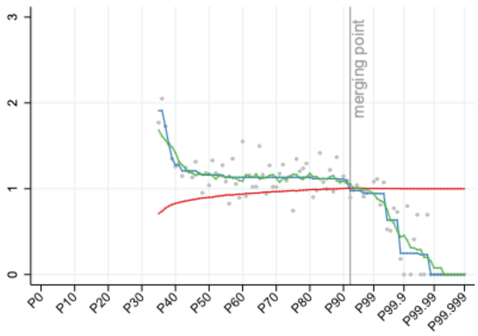
Figure C.3: Merging Points in United Kingdom, 2005-2013



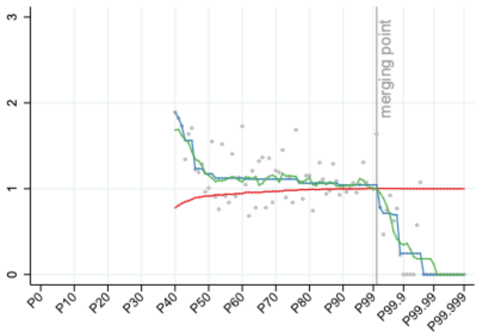
(a) United Kingdom 2005



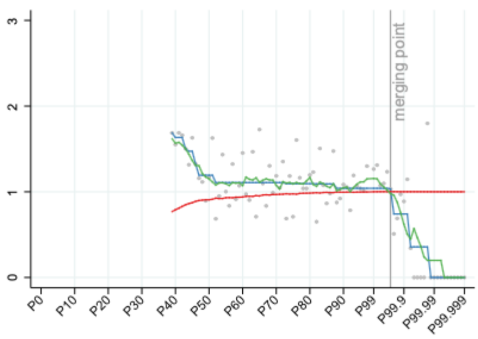
(b) United Kingdom 2006



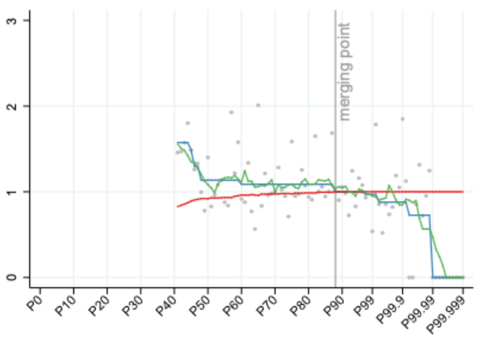
(c) United Kingdom 2007



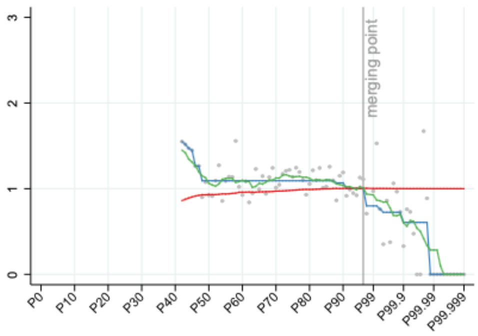
(d) United Kingdom 2009



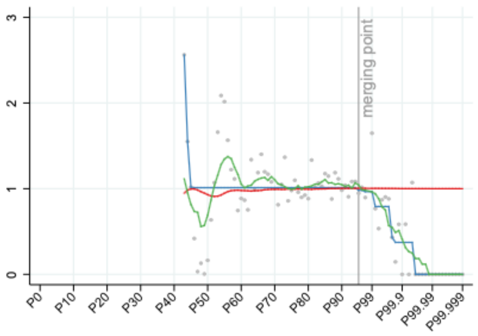
(e) United Kingdom 2010



(f) United Kingdom 2011



(g) United Kingdom 2012



(h) United Kingdom 2013

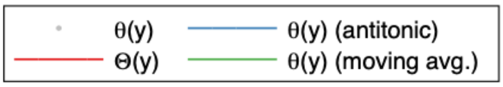
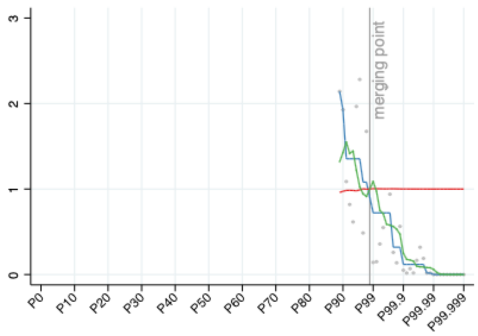
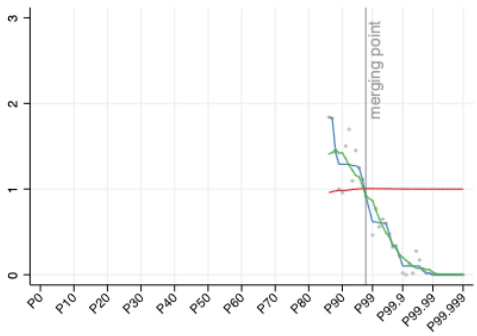


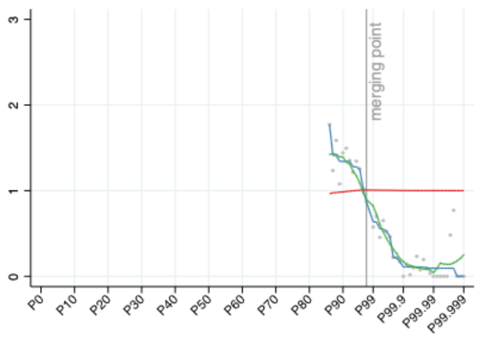
Figure C.4: Merging Points in Brazil, 2007-2014



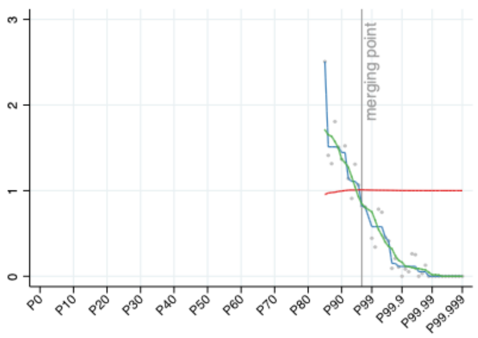
(a) Brazil 2007



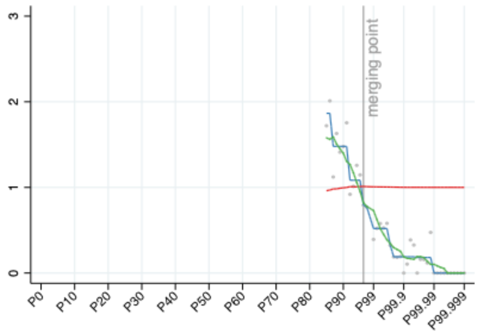
(b) Brazil 2008



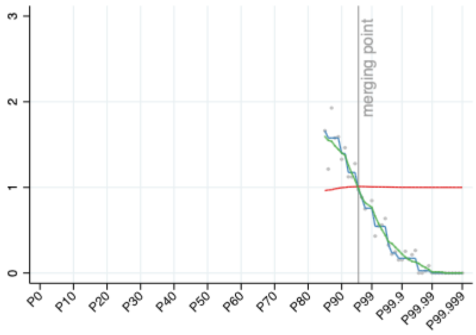
(c) Brazil 2009



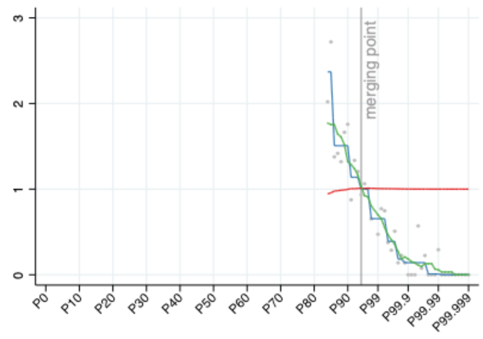
(d) Brazil 2011



(e) Brazil 2012



(f) Brazil 2013



(g) Brazil 2014

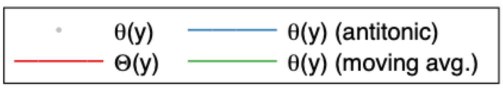
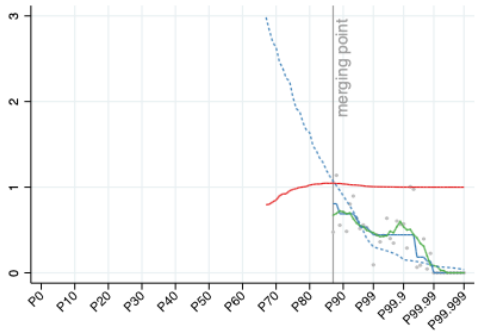
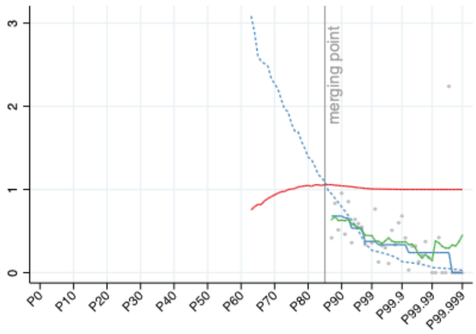


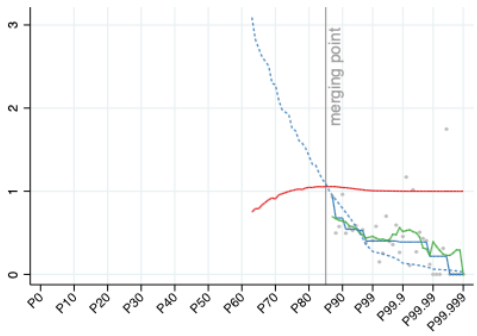
Figure C.5: Merging Points in Chile, 2009-2013



(a) Chile 2009



(b) Chile 2011



(c) Chile 2013

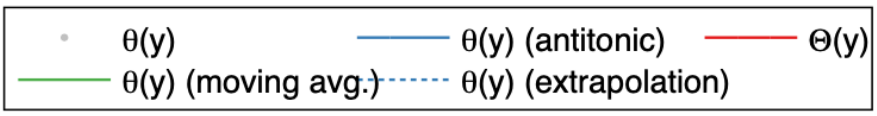


Table C.3: Structure of Corrected Population in Chile, 2009-2015

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]
2009	12.0%	7.7%	4.28%	99.7%	0.3%
2011	14.0%	8.9%	5.10%	99.9%	0.1%
2013	14.0%	9.0%	4.98%	99.9%	0.1%
2015	14.0%	9.2%	4.83%	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 C.4: Structure of Corrected Population in France, 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	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 C.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 C.6: Structure of Corrected Population in United Kingdom, 2005-2014

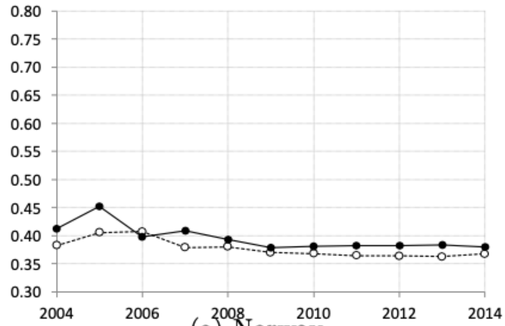
Year	Population over Merging Point (% total population)		Corrected population		
	Tax data [2]	Survey [3]	Total [4] = [2] − [3]	Share inside survey support [5]	Share outside survey support [6]
2005	12.0%	11.7%	0.26%	99.5%	0.5%
2006	8.0%	7.3%	0.72%	96.9%	3.1%
2007	7.0%	6.5%	0.53%	95.5%	4.5%
2009	0.8%	0.5%	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).

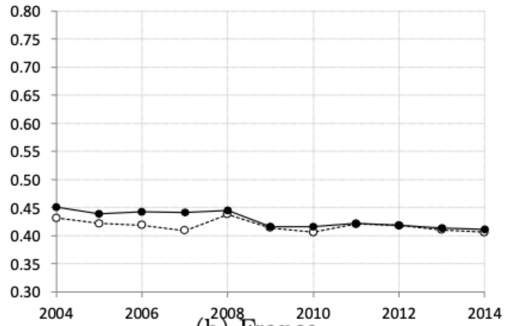
C.4 Gini Coefficients

Figure C.6 shows graphs of the Gini coefficients of our 6 country case studies before and after the correction for all available years.

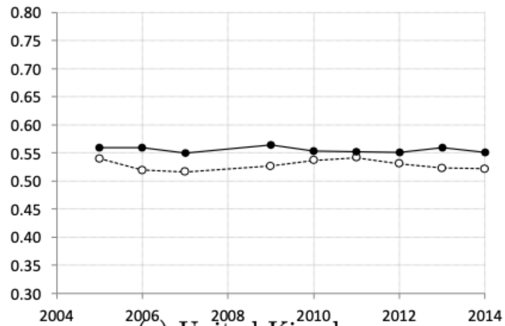
Figure C.6: Gini Coefficients, Before and After Correction in 6 Countries



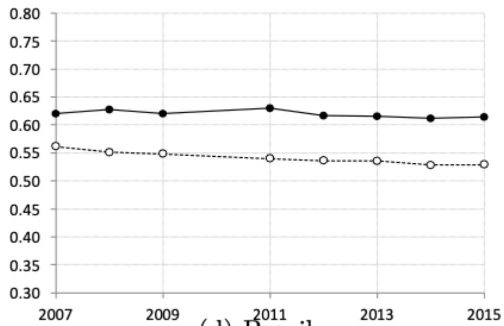
(a) Norway



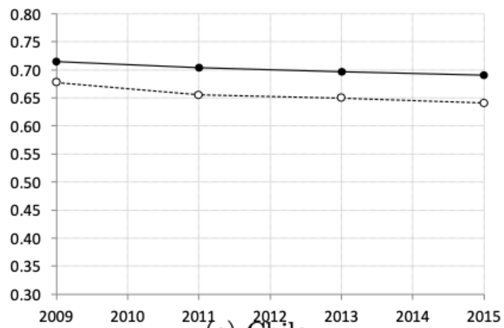
(b) France



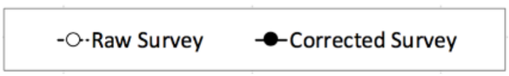
(c) United Kingdom



(d) Brazil



(e) Chile



C.5 The Impact of Misreporting

Our adjustment procedure is based on the interpretation of the whole difference between tax and survey densities as being due solely to nonresponse. However, there are at least some cases in which another co-existing bias is detected. Misreporting in surveys tends to have a negative correlation with income. That is, on average, the poor are more likely to overreport while the rich tend to underreport true income. It is thus fair to ask: what are the consequences of such behavior in our analytical framework, and what can we learn about misreporting from our experience?

To define the misreporting bias, let $f_M(y)$ be the distribution of misreported income, $p(y)$ the probability of misreporting for a given level of income and $\bar{p}$ its average. Then we define f_Z as the distribution including both the nonresponse and misreporting biases:

$$f_Z(y) = f_Y(y)\theta(y)(1 - p(y)) + f_M(y)\theta(y)\bar{p} \tag{C.1}$$

The left half of the sum stands for those who report income correctly with a given (relative) probability of response ($\theta(y)$). The right-hand side of the sum accounts for those declaring misreported income equal to y, again taking into account the nonresponse bias. In this situation, the over or under-estimation of f_Z with respect to the true distribution (f_Y) can also be formulated as the ratio of the two distributions.

$$\frac{f_Z(y)}{f_Y(y)} = \theta(y)(1 - p(y)) + \frac{f_M(y)}{f_Y(y)}\theta(y)\bar{p} \tag{C.2}$$

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.

The probability of misreporting is likely to be higher in both ends of the distribution, yet we usually do not have explicit information on the actual shape of the misreported distribution. In order to better understand the potential impact that different distributions could cause, it can be useful to analyze a simplified situation where misreporting operates alone. In that case we would have:

$$\frac{f_Z(y)}{f_Y(y)} = 1 - p(y) + \frac{f_M(y)}{f_Y(y)}\bar{p} \tag{C.3}$$

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 understand that both the nonresponse and misreporting biases have a rather similar impact and that we are unable to tell them appart *ex post*. Indeed, in that case both biases, either working alone or together, can perfectly describe a profile as the one in Figure 3.3. If $f_M \neq f_Y$, under some circumstances we can get a similar result. If both densities are of the same type with different parameters (e.g. if both are log-normal with a slightly different mean and standard error) the bias profile would likely have a form similar to Figure 3.3 but with strong or slight perturbations near the mode of each distribution (with unimodal densities). As shown in Section 3.3.2 our empirical estimate of the θ coefficient, which should be capturing both biases if they exist, describes a rather flat shape through most of the distribution and only falls in the high end of the income distribution where data is sufficient (Norway, France and the United Kingdom). Such a shape implies that, if misreporting has a significant impact on the distribution of survey-income, the differences between f_M and f_Y are not big enough to cause perturbations that are easily distiguishable from noise while observing the θ coefficient. In any case, as far as we know, it is not possible to measure the relative size of both the nonresponse and misreporting biases without access to individual matching across datasets.

In a purely distributive perspective, our reweighting method is able to correct both the nonresponse and misreporting biases together. Symetrically, we can also find an algorithm that theoretically reproduces the same adjusted distribution *via* modifying individual income. Such an algorithm, would correct for both biases too. Nonetheless, despite having virtually the same distributive results, the reweighting algorithm should be prefered when we aim to use other variables in the survey, because it preserves the internal consistency of each observation. On the contrary, replacing incomes, at least as it has been implemented to date, assumes implicitly a deterministic form of misreporting. That is, everybody underreports increasingly from a given level of income. A correction method that is based on this conception of misreporting most likely modifies the income of those who report correctly, thus, worsening the representativeness of each observation in terms of covariates.

Appendix D

List of Figures and Tables

List of Figures

1.1 Gini coefficient by CEPALSTAT and World Bank (1987-2015) . . 20
1.2 Comparing Income Concepts (per adult) 29
1.3 Top 1%, 0.1% and 0.01% Shares of Fiscal Income (1964-2015) . . 33
1.4 Top 10%, Share of Fiscal Income (2004-2015) 37
1.5 Top 1% Share with Undistributed Profits, Upper and Lower Bounds (1990-2015) 39
1.6 Unadjusted Income Growth. Top 0.1%, Next 0.9% and Bottom 99% (Index base 100 in 1964) 41
1.7 Income Growth Including Undistributed Profits. Top 0.1%, Next 0.9% and Bottom 99% (Index base 100 in 1990) 42
1.8 Top 1% Share in Latin America (1990-2015) 43
1.9 Top 1% Share Compared to Other Countries (1964-2015) 45
1.10 Top 1% Share: Tax Data vs. CASEN Survey (1990-2015) 46
1.11 Top 1% Share: Tax Data vs. CASEN Survey (1990-2015) 47
1.12 Top Marginal Income Tax Rate (1964-2015) 48
1.13 Inverted Beta Coefficient of Top 1% Share (1964-2015) 49
1.14 Tax Evasion Rates in Literature (1964-2015) 50

2.1 Capital Share of National Income, Balanced Panel (1995-2016) . . 61
2.2 Decreasing Household Share of Gross Capital Income, Balanced Panel (1995-2016) 63
2.3 Decreasing Household Share of Capital Income, Unbalanced Panel (1995-2015) 64
2.4 Decreasing Household Share of Capital Income, Long-run 65
2.5 Unequal Measurement Error in Surveys and Tax Data 67
2.6 From the Capital Share to Top Income Shares 70
2.7 Evolution of relevant variables, Balanced Panel (1995-2013) 73
2.8 Top 1%'s Factor Income Shares, United States (1975-2015) 75

2.9 The Capital Share and the Gamma coefficient, United States (1975-2015) . . . 76

3.1 A “true” and biased income distribution . . . 92
3.2 The intuition behind reweighting . . . 93
3.3 Choice of Merging Point when $\bar{y} \geq y_{\text{trust}}$ . . . 95
3.4 Choice of Merging Point when $\bar{y} < y_{\text{trust}}$ . . . 97
3.5 Geometrical Interpretation of Linear Calibration . . . 102
3.6 Merging Point in 6 Countries, Latest year . . . 112
3.7 The Top 1% Share Before and After Correction . . . 115
3.8 Discrepancy of income across datasets in Chile and Brazil: 2015 . 117

A.1 Difference in Top 1% Share, with and without Capital Gains (1998-2009) . . . 128
A.2 Average Real Income (in 2013 USD PPA) and CPI (base 2015) . . 129
A.3 Comparison of aggregate income concepts . . . 130
A.4 Individual Tax Declarations as a Share of the Adult Population, *Consolidated* Series (2004-2015) . . . 131
A.5 Individual Tax Declarations as a Share of the Adult Population, *Global Complementario* Series (1964-2015) . . . 132
A.6 Undistributed Profits and Personal Income (1996-2015) . . . 133
A.7 Real income growth. Top 0.1% in tax data vs. median income in CASEN . . . 133
A.8 Top 0.1% Share in Latin America (1990-2015) . . . 134
A.9 Corporate tax rate vs. Top Marginal income tax rate (1990-2015) 134
A.10 Top 10% with pre-tax undistributed profits, upper and lower bounds (1990-2015) . . . 135
A.11 Sensitivity of Top Income Share to Errors in Total Income . . . 136

B.1 Capital Share of National Income, by Country, Balanced Panel (1995-2016) . . . 140
B.2 Capital Share of National Income, by Country, Balanced Panel (1995-2016) . . . 141
B.3 Household Share of Gross Capital Income in Balanced Panel, by country (1995-2016) . . . 142
B.4 Decreasing Household Share of Gross Capital Income (average), Balanced Panel (1995-2016) . . . 143

B.5 Decomposition of the Public Sector's Capital Income, Balanced Panel (1995-2014) . . . 144
B.6 Top 1% income share in Balanced Panel . . . 145
B.7 Capital Shares and Gamma coefficients in Balanced Panel (1995-2013) 146
B.8 Decomposition of γ by country (1995-2013) . . . 146
B.9 Decomposition of γ (1975-2016) . . . 147
B.10 Unequal Measurement Error of Factor Incomes in Surveys, Individual Countries (1995-2013) . . . 148
B.11 Decreasing Household Share of Gross Capital Income, Unbalanced Panel (1995-2015) . . . 149
B.12 Decreasing Household Share of Gross Capital Income, Long-run . 150
B.13 Capital depreciation as a share of GDP, 1950-2015 . . . 151
B.14 Decreasing Household Share of Net Capital Income, Balanced Panel (1995-2016) . . . 152
B.15 Decreasing Household Share of Net Capital Income (average), Balanced Panel (1995-2016) . . . 153
B.16 Household Share of Net Capital Income in Balanced Panel, by country (1995-2016) . . . 154
B.17 Decreasing Household Share of Gross Capital Income (incl. Public Sector), Long-run . . . 155
B.18 Decreasing Household Share of Gross Capital Income (w/o Public Sector), Long-run . . . 156
B.19 Household Share of Net Capital Income, Unbalanced Panel (1995-2015) . . . 157
B.20 Unequal Measurement Error of Factor Incomes in Surveys (Net Capital Income), Individual Countries (1995-2013) . . . 158
B.21 Unequal Measurement Error of Factor Incomes in Surveys (Long-Run), Individual Countries (1995-2013) . . . 159
B.22 Unequal Error of Factor Incomes, The U.S. . . . 160
B.23 From National Income in National Accounts to Household Income in Surveys . . . 161
B.24 Unequal Measurement Error of the Capital Share . . . 161

C.1 Merging Points in Norway, 2004-2013 . . . 168
C.2 Merging Points in France, 2004-2013 . . . 170
C.3 Merging Points in United Kingdom, 2005-2013 . . . 172
C.4 Merging Points in Brazil, 2007-2014 . . . 173

C.5 Merging Points in Chile, 2009-2013 175
C.6 Gini Coefficients, Before and After Correction in 6 Countries . . . 179

List of Tables

1.1 General Income Definition in Tax Data 24
1.2 Total Personal Income in National Accounts 28

2.1 Partial derivatives 71
2.2 Empirical Estimate of Partial derivatives, Top 1% Share 77
2.3 Modeled Contribution to Variation in Top 1% Share 80

3.1 Structure of Corrected Population: Latest Year 113
3.2 Income Shares: Raw Survey and Corrected Survey 118

A.1 Adult population, total and taxable 125
A.2 Undistributed Profits 126
A.3 Top shares including undistributed profits (Upper Bound) 127
A.4 Top shares including undistributed profits (Lower Bound) 127

B.1 Average structure of Total Capital Income in the Balanced Panel by Country, 1995-2015. 138
B.2 Average structure of Total National Income in the Balanced Sub-Panel by Country, 1995-2013. 139

C.1 From Post-Tax to Pre-Tax Income in Chilean Surveys 165
C.2 Structure of Corrected Population in Brazil, 2007-2015 167
C.3 Structure of Corrected Population in Chile, 2009-2015 176
C.4 Structure of Corrected Population in France, 2004-2014 176
C.5 Structure of Corrected Population in Norway, 2004-2014 177
C.6 Structure of Corrected Population in United Kingdom, 2005-2014 177

Appendix E

References

Bibliography

Abeles, Martín, Verónica Amarante, and Daniel Vega (2014). "Participación del ingreso laboral en el ingreso total en América Latina, 1990-2010". In: *Revista CEPAL* 114.December, pp. 31–52. URL: https://repositorio.cepal.org/handle/11362/37435.

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 (2010). *The Rich in Argentina over the Twentieth Century, 1932-2004.* Ed. by A. B. (Anthony Barnes) Atkinson and Thomas Piketty. Top income. Oxford University Press, pp. 253–298. ISBN: 0191558230.

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

Alvaredo, Facundo, Anthony B. Atkinson, Lucas Chancel, et al. (2016). "Distributional National Accounts (DINA) Guidelines : Concepts and Methods used in WID.world". URL: http://gabriel-zucman.eu/files/AACPSZ2017.pdf.

Alvaredo, Facundo, Anthony B. Atkinson, Thomas Piketty, et al. (2013). "The Top 1 Percent in International and Historical Perspective". In: *Journal of Economic Perspectives* 27.3, pp. 3–20. DOI: 10.1257/jep.27.3.3. URL: http://pubs.aeaweb.org/doi/10.1257/jep.27.3.3.

Alvaredo, Facundo, Lucas Chancel, et al. (2018). *World Inequality Report 2018.* Belknap Press of Harvard University Press.

Alvaredo, Facundo and Juliana Londoño-Vélez (2013). "High Incomes and Personal Taxatio in a Developping Economy: Colombia (1993-2010)". URL: http://www.commitmentoequity.org/publications_files/Colombia/CEQWPNo12%20HighTaxationDevEconColombia1993-2010_19March2013.pdf.

Arpaia, Alfonso, Esther Pérez, and Karl Pichelmann (2009). "Understanding labour income share dynamics in Europe". URL: https://mpra.ub.uni-muenchen.de/15649/1/MPRA_paper_15649.pdf.

Atkinson, Anthony B. (2007a). "Measuring Top Incomes: Methodological Issues". In: *Top Incomes over the Twentieth Century: A Contrast Between Continental European and English-Speaking Countries.*

— (2007b). "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* 1, pp. 82–140.

— (2009). "Factor shares: the principal problem of political economy?" In: *Oxford Review of Economic Policy* 25.1, pp. 3–16. DOI: 10.1093/oxrep/grp007. URL: https://academic.oup.com/oxrep/article-lookup/doi/10.1093/oxrep/grp007.

Atkinson, Anthony B. and Rolf Aaberge (2010). "Top Incomes in Norway". In: *Top incomes : a global perspective.* Ed. by A. B. (Anthony Barnes) Atkinson and Thomas Piketty. Oxford University Press, p. 776. ISBN: 0191558230. URL: https://books.google.fr/books?hl=fr&lr=&id=eGaVAwAAQBAJ&oi=fnd&pg=PA448&ots=weNqqVKWAc&sig=Wx-ayJHI7MgXWMF5JCKCbB978Kw#v=onepage&q&f=false.

Atkinson, Anthony B. and François Bourguignon (2000). "Introduction: Income Distribution and Economics". In: *Handbook of Income Distribution.* Ed. by Anthony B. Atkinson and François Bourguignon. Geneva: Elsevier. Chap. Introducti, pp. 1–53.

Atkinson, Anthony B. and Salvatore Morelli (2012). "The Chartbook of Economic Inequality".

Atkinson, Anthony B. and Thomas Piketty (2010). *Top incomes: a global perspective.* Oxford University Press, p. 776. ISBN: 9780199286898. URL: https://global.oup.com/academic/product/top-incomes-9780199286898?cc=fr&lang=en&#.

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. ISSN: 0022-0515. DOI: 10.1257/jel.49.1.3.

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.

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

Bagwell, Laurie Simon and John B Shoven (1989). "Cash Distributions to Shareholders". In: *Journal of Economic Perspectives* 3.3, pp. 129–140. DOI: 10.1257/jep.3.3.129. URL: http://pubs.aeaweb.org/doi/10.1257/jep.3.3.129.

Bengtsson, Erik and Daniel Waldenström (2017). "Capital shares and income inequality: Evidence from the long run". URL: http://www.uueconomics.se/danielw/Research_files/BengtssonWaldenstrom_Capitalshares_long.pdf.

Blanchet, Thomas, Juliette Fournier, and Thomas Piketty (2017). "Generalized Pareto Curves: Theory and Applications". URL: https://wid.world/document/blanchet-t-fournier-j-piketty-t-generalized-pareto-curves-theory-applications-2017/.

Bourguignon, François (2015). "Appraising income inequality databases in Latin America". In: *The Journal of Economic Inequality* 13.4, pp. 557–578. DOI: 10.1007/s10888-015-9304-4. URL: http://link.springer.com/10.1007/s10888-015-9304-4.

— (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. DOI: 10.1214/aoms/1177728420. URL: https://doi.org/10.1214/aoms/1177728420.

Burdín, Gabriel et al. (2014). "Inequality and Top Incomes in Uruguay: A Comparison Between Household Surveys and Income Tax Micro-Data". URL: http://stonecenter.tulane.edu/uploads/CEQWPNo21_IneqTopIncomesUruguay_May_2014-1405720326.pdf.

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. ISSN: 14755890. DOI: `10.1111/1475-5890.12158`.

Cerda, Rodrigo et al. (2014). “El Fondo de Utilidades Tributables (FUT): Elementos para la discusión”. In: *Estudios Públicos* 135.invierno, pp. 39–87. URL: `https://www.cepchile.cl/cep/site/artic/20160622/asocfile/20160622111229/revista_estudios_publicos_135.pdf#page=40`.

Chancel, Lucas and Thomas Piketty (2017). “Indian income inequality, 1922-2014: From British Raj to Billionaire Raj ?” URL: `https://econpapers.repec.org/paper/cprceprdp/12409.htm`.

Cheyre, Hernán (1986). “Análisis de las Reformas Tributarias en la década 1974-1983”. In: Summer, pp. 1–48. URL: `https://www.cepchile.cl/cep/site/artic/20160303/asocfile/20160303184330/rev21_HCheyre.pdf`.

Contreras, Dante and Ricardo Ffrench-Davis (2012). “Policy regimes, inequality, poverty and growth : the Chilean experience, 1973 - 2010”. URL: `https://www.econstor.eu/handle/10419/81078`.

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. DOI: `10.1214/aoms/1177731829`. 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. ISBN: 978-3-642-57678-2. DOI: `10.1007/978-3-642-57678-2{\_}6`. 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. DOI: `10.1080/01621459.1992.10475217`. URL: `http://www.tandfonline.com/doi/abs/10.1080/01621459.1992.10475217`.

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.

Eije, Henk von and William L. Megginson (2008). "Dividends and share repurchases in the European Union". In: *Journal of Financial Economics* 89.2, pp. 347–374. DOI: 10.1016/J.JFINECO.2007.11.002. URL: https://www.sciencedirect.com/science/article/pii/S0304405X08000949.

Fairfield, Tasha (2010). "Business Power and Tax Reform: Taxing Income and Profits in Chile and Argentina". In: *Latin American Politics and Society* 52.2, pp. 37–71. ISSN: 1531426X. DOI: 10.1111/j.1548-2456.2010.00081.x.

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.S1, S120–S144. DOI: 10.1111/roiw.12196. URL: http://doi.wiley.com/10.1111/roiw.12196.

Fama, Eugene F and Kenneth R French (2001). "Disappearing dividends: changing firm characteristics or lower propensity to pay?" In: *Journal of Financial Economics* 60.1, pp. 3–43. ISSN: 0304-405X. DOI: 10.1016/S0304-405X(01)00038-1. URL: https://www.sciencedirect.com/science/article/pii/S0304405X01000381.

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.

Floyd, Eric, Nan Li, and Douglas J. Skinner (2015). "Payout policy through the financial crisis: The growth of repurchases and the resilience of dividends". In: *Journal of Financial Economics* 118.2, pp. 299–316. ISSN: 0304-405X. DOI: 10.1016/J.JFINECO.2015.08.002. URL: https://www.sciencedirect.com/science/article/pii/S0304405X15001415.

Foxley, Alejandro, Eduardo Aninat, and José Pablo Arellano (1980). *Las desigualdades económicas y la acción del Estado*. D.F. Mexico: Fondo de Cultura Económica (CIEPLAN).

Francese, Maura and Carlos Mulas-Granados (2015). "Functional Income Distribution and Its Role in Explaining Inequality". URL: https://www.imf.org/en/Publications/WP/Issues/2016/12/31/Functional-Income-Distribution-and-Its-Role-in-Explaining-Inequality-43415.

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.

Gutiérrez, Pablo, Ramón López, and Eugenio Figueroa (2015). “Top income measurement and undistributed profits”. In: *Economics Letters* 134.September, pp. 138–140. ISSN: 0165-1765. DOI: 10.1016/J.ECONLET.2015.07.013. URL: https://www.sciencedirect.com/science/article/pii/S0165176515002864.

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. DOI: 10.3390/econometrics6020030.

Hoerl, Arthur E and Robert W Kennard (2000). “Ridge Regression: Biased Estimation for Problems Nonorthogonal”. In: *Technometrics* 42.1, pp. 80–86. ISSN: 0040-1706. DOI: 10.1080/00401706.2000.10485983.

IMF (2007). “The Globalization of Labor”. In: *World Economic Outlook.* Washington: International Monetary Fund (IMF). Chap. nº5, pp. 161–189. ISBN: 9781589066267. DOI: 10.2501/JAR-51-4-564-570.

Jenkins, Stephen P. (2017). “Pareto Models, Top Incomes and Recent Trends in UK Income Inequality”. In: *Economica* 84.334, pp. 261–289. DOI: 10.1111/ecca.12217. URL: http://doi.wiley.com/10.1111/ecca.12217.

Jorratt De Luis, Michel (2009). *La tributación directa en Chile: equidad y desafíos.* Tech. rep. Santiago de Chile: CEPAL.

Jorratt, Michel (2013). “Gastos tributarios y evasión tributaria en Chile: evaluación y propuestas.” In: *Tributación para el Desarrollo: Estudios para la reforma del sistema chileno.* Ed. by José Pablo Arellano and Vittorio Corbo. Santiago de Chile: CEP/CIEPLAN.

Karabarbounis, Loukas and Brent Neiman (2014). “The Global Decline of the Labor Share”. In: *The Quarterly Journal of Economics* 129.1, pp. 61–103. DOI: 10.1093/qje/qjt032. URL: https://academic.oup.com/qje/article-lookup/doi/10.1093/qje/qjt032.

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. ISSN: 15691721. DOI: 10.1007/s10888-005-1089-4.

Kuznets, Simon (1953). *Shares of Upper Income Groups in Income and Savings.* NBER. ISBN: 087014054X. DOI: 10.2307/2343040. URL: http://www.jstor.org/stable/10.2307/2343040?origin=crossref.

Larrain, Felipe and Patricio Meller (1991). *The Socialist-Populist Chilean Experience, 1970-1973.* Ed. by Sebastian Edwards and Rudiger Dornbusch. January, pp. 175–221. ISBN: 0226158438.

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

López, Ramón, Eugenio Figueroa, and Pablo Gutiérrez (2013). "La Parte del León: Nuevas Estimaciones de la Participación de los Súper Ricos en el Ingreso de Chile". URL: http://www.econ.uchile.cl/uploads/publicacion/306018fadb3ac79952bf1395a555a90a86633790.pdf.

— (2016). "Fundamental accrued capital gains and the measurement of top incomes: an application to Chile". In: *The Journal of Economic Inequality* 14.4, pp. 379–394. DOI: 10.1007/s10888-016-9337-3. URL: http://link.springer.com/10.1007/s10888-016-9337-3.

López-Calva, Luis Felipe and Nora Lustig (2010). *Declining inequality in Latin America : a decade of progress?* Ed. by Luis Felipe López-Calva and Nora Lustig. Washington: Brookings Institution Press., p. 253. ISBN: 0815704445. URL: https://books.google.fr/books?hl=fr&lr=&id=KGFCdQVGAaYC&oi=fnd&pg=PP1&dq=lustig+lopez-calva&ots=Xa31DhE5md&sig=Xz19j6e07s5UjmAMDj5nEor2xGw#v=onepage&q=lustig%20lopez-calva&f=false.

Martínez-Aguilar, Sandra et al. (2017). "The Impact of Fiscal Policy on Inequality and Poverty in Chile". URL: http://www.commitmentoequity.org/wp-content/uploads/2017/08/CEQ_WP46_MartinezAguilar-et-al_ImpactFiscalPolicyChile_June2017.pdf.

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.

Milanovic, Branko (2017). "Increasing Capital Income Share and Its Effects on Personal Income Inequality". In: *After Piketty.* Ed. by H. Bouchey, J. DeLong, and M. Steinbaum. Harvard University Press. Chap. III. Dimen, pp. 235–258.

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.

OECD (2008). *The System of National Accounts 2008*. ISBN: 978-92-1-161522-7. DOI: 10.1057/ukna.2008.3.

— (2015). *Why Less Inequality Benefits All*. Tech. rep. Paris: OECD. DOI: 10.1787/9789264235120-en. URL: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.737.3637&rep=rep1&type=pdf.

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. ISSN: 01677152. DOI: 10.1016/S0167-7152(00)00242-X.

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

Piketty, Thomas (2001). *Les hauts revenus en France au XXème siècle*. Hachette. ISBN: 9782246616511.

— (2003). "Income Inequality in France, 1901–1998". In: *Journal of Political Economy* 111.5, pp. 1004–1042. DOI: 10.1086/376955. 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* 118.1, pp. 1–39. ISSN: 00335533. DOI: 10.1162/00335530360535135. URL: https://academic.oup.com/qje/article-lookup/doi/10.1162/00335530360535135.

Piketty, Thomas, Emmanuel Saez, and Gabriel Zucman (2018a). "Distributional National Accounts: Methods and Estimates for the United States". In: *Quarterly Journal of Economics* 133.May, pp. 553–609. DOI: 10.1093/qje/qjx043.Advance.

— (2018b). "Distributional National Accounts: Methods and Estimates for the United States*". In: *The Quarterly Journal of Economics* 133.2, pp. 553–609. DOI: 10.1093/qje/qjx043. URL: https://academic.oup.com/qje/article/133/2/553/4430651.

Piketty, Thomas, Li Yang, and Gabriel Zucman (2017). *Capital Accumulation, Private Property and Rising Inequality in China, 1978-2015*. Tech. rep. Cambridge, MA: National Bureau of Economic Research. DOI: 10.3386/w23368. URL: http://www.nber.org/papers/w23368.pdf.

Piketty, Thomas and Gabriel Zucman (2014). "Capital is Back: Wealth-Income Ratios in Rich Countries 1700–2010". In: *The Quarterly Journal of Economics* 129.3, pp. 1255–1310. DOI: 10.1093/qje/qju018. URL: https://academic.oup.com/qje/article/129/3/1255/1818714.

Sanhueza, Claudia and Ricardo Mayer (2011). "Top Incomes in Chile using 50 years of household surveys: 1957-2007". In: *Estudios de economía* 38.1, pp. 169–193. ISSN: 0718-5286. DOI: 10.4067/S0718-52862011000100007.

Serra, Pablo (2000). "Evasión Tributaria: ¿Cómo Abordarla?" In: *Estudios Públicos* 80.Spring, pp. 194–228. URL: https://www.cepchile.cl/cep/site/artic/20160303/asocfile/20160303184323/rev80_serra.pdf.

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

Skinner, Douglas J. (2008). "The evolving relation between earnings, dividends, and stock repurchases". In: *Journal of Financial Economics* 87.3, pp. 582–609. ISSN: 0304-405X. DOI: 10.1016/J.JFINECO.2007.05.003. URL: https://www.sciencedirect.com/science/article/pii/S0304405X07002334.

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. ISSN: 0022-166X. DOI: 10.3368/jhr.50.2.301. 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. DOI: 10.1016/J.PHYSA.2015.02.038. URL: https://www.sciencedirect.com/science/article/pii/S0378437115001429.

Yáñez, José (2015). "Evasión tributaria: atentado a la equidad". In: *Revista de Estudios Tributarios* 13, pp. 171–206. URL: https://revistas.uchile.cl/index.php/RET/article/view/39874/41444.

Zucman, Gabriel (2015). *The hidden wealth of nations : the scourge of tax havens.* The University of Chicago Press, p. 129. ISBN: 022624542X. URL: https://books.google.fr/books?hl=fr&lr=&id=tpIpCwAAQBAJ&oi=fnd&pg=PR5&dq=zucman+hidden+wealth&ots=dxYkb-HeEd&sig=opV51z9h9CZ9J-gVxB_YC786OpU#v=onepage&q=zucman%20hidden%20wealth&f=false.

Summary The 1st chapter presents historical series of Chilean top income shares over a period of half a century, mostly using data from tax statistics and national accounts. The study contradicts evidence based on survey data, according to which inequality has fallen constantly over the past 25 years. Rather, it changes direction, increasing from around the year 2000. Chile ranks as one of the most unequal countries among both OECD and Latin American countries over the whole period of study. The 2nd chapter measures the underestimation of factor income in distributive data. I find that households receive only half of national gross capital income, as opposed to corporations. Due to heterogeneous non-response and misreporting, Surveys only capture 20% of it, vs. 70% of labor income. This understates inequality estimates, which become insensitive to the capital share and its distribution. I formalize this system based on accounting identities. I then compute marginal effects and contributions to changes in fractile shares. The 3rd chapter, presents a method to adjust surveys. These generally fail to capture the top of the income distribution. It has several advantages over previous ones: it is consistent with standard survey calibration methods; it has explicit probabilistic foundations and preserves the continuity of density functions; it provides an option to overcome the limitations of bounded survey-supports; and it preserves the microdata structure of the survey,

Resumé Le 1er chapitre présente une série de 50 and sur les hauts revenus chiliens basée sur des données fiscales et comptes nationaux. L'étude contredit les enquêtes, selon lesquelles les inégalités diminuent les 25 dernières années. Au contraire, elles changent de direction à partir de 2000. Le Chili est parmi les pays les plus inégalitaires de l'OCDE et l'Amérique latine. Le 2ème chapitre mesure la sous-estimation des revenus factoriels dans les données distributives. Les ménages ne reçoivent que 50% des revenus du capital brut, par opposition aux firmes. L'hétérogénéité des taux de réponse et autres problèmes font que les enquêtes ne capturent que 20% de ceux-ci, contre 70% du revenu du travail. Cela sous-estime l'inégalité, dont les estimations deviennent insensibles à la *capital share* et sa distribution. Je formalise à partir d'identités comptables pour ensuite calculer des effets marginaux et contributions aux variations d'inégalité. Le 3ème chapitre présente une méthode pour ajuster les enquêtes. Celles-ci capturent souvent mal le sommet de la distribution. La méthode présente plusieurs avantages par rapport aux options précédentes : elle est compatible avec les méthodes de calibration standard ; elle a des fondements probabilistes explicites et préserve la continuité des fonctions de densité ; elle offre une option pour surmonter les limites des supports d'enquête bornées; et elle préserve la structure de micro données en préservant la représentativité des variables sociodémographiques. Notre procédure est illustrée par des applications dans cinq pays, couvrant à la fois des contextes développés et moins développés.