de la Vega, Rafael C.

Working Paper

Economic structure and top earnings inequality in South Africa: A firm-level and sectoral perspective

WIDER Working Paper, No. 39/25

Provided in Cooperation with:
United Nations University (UNU), World Institute for Development Economics Research (WIDER)

Suggested Citation: de la Vega, Rafael C. (2025) : Economic structure and top earnings inequality in South Africa: A firm-level and sectoral perspective, WIDER Working Paper, No. 39/25, ISBN 978-92-9256-598-5, The United Nations University World Institute for Development Economics Research (UNU-WIDER), Helsinki, https://doi.org/10.35188/UNU-WIDER/2025/598-1

This Version is available at:
https://hdl.handle.net/10419/322164

39
25

Economic structure and top earnings inequality in South Africa

A firm-level and sectoral perspective

Rafael de la Vega*

May 2025

WIDER WORKING PAPER

wider.unu.edu

United Nations University
World Institute for Development Economics Research

Abstract: Inequality at the top is on the rise, and labour income is a progressively larger contributor to concentration at the top. This paper investigates top earnings inequality in South Africa from a sectoral and firm-level perspective, using matched employer–employee administrative data. We also propose a method for decomposing top shares in within-groups and between-groups components. Despite the clear presence of sectoral heterogeneity, as expected for a middle-income country, we find in the decompositions that the main source of inequality occurs within firms rather than between sectors or between firms in the same sector. Regressions show that larger and more productive firms are associated with greater firm-level inequality, as well as those with a larger share of male workers, while the opposite is valid for listed firms and those in sectors with greater union density. These results are informative for policymaking decisions in support of inclusive processes of structural change.

Key words: structural change, economic structure, top inequality, inequality decomposition, administrative data

JEL classification: D31, O10, L16, C23

Acknowledgements: I am grateful to Bart Verspagen and Tommaso Ciarli for extensive comments on this paper; to participants of the SA-TIED workshop that was held in August 2024, the Southern Centre for Inequality Studies (SCIS) seminar series, and the UNU-MERIT internal conference for valuable comments; and to the entire UNU-WIDER and National Treasury teams whose technical assistance and support made this project possible. I also gratefully acknowledge funding from the SA-TIED programme.

* UNU-MERIT, Maastricht, Netherlands, **consentinodelavega@merit.unu.edu**

This study has been prepared within the UNU-WIDER project **Southern Africa—Towards Inclusive Economic Development (SA-TIED)**.

Information and requests: publications@wider.unu.edu

ISSN 1798-7237 ISBN 978-92-9256-598-1

https://doi.org/10.35188/UNU-WIDER/2025/598-1

United Nations University World Institute for Development Economics Research provides economic analysis and policy advice with the aim of promoting sustainable and equitable development. The Institute began operations in 1985 in Helsinki, Finland, as the first research and training centre of the United Nations University. Today it is a unique blend of think tank, research institute, and UN agency—providing a range of services from policy advice to governments as well as freely available original research.

The Institute is funded through income from an endowment fund with additional contributions to its work programme from Finland and Sweden, as well as earmarked contributions for specific projects from a variety of donors.

Katajanokanlaituri 6 B, 00160 Helsinki, Finland

1 Introduction

Incomes at the top of the distribution have been on the rise in recent decades. One stylized fact from the literature on the long-term behaviour of top incomes in richer countries is that the share of top earners started the 20th century at rather high levels, dropped significantly in the 'great levelling' that took place midway through the century, only to rise again in the final two decades and into the 21st century (Piketty 2013). Another stylized fact about recent years in these countries is that labour incomes have become a progressively larger contributor to concentration at the top (Piketty and Saez 2013; Milanović 2016).

While these trends have been documented in most high-income countries (Atkinson et al. 2011), countries at lower income levels are not immune, even if they show idiosyncratic patterns. In South Africa, top incomes have broadly followed the same long-term trajectory as in richer countries, albeit at higher levels of inequality and with important horizontal disparities across ethnic groups (Alvaredo and Atkinson 2022). Labour incomes also play a central role in shaping top-end inequality in the country, with capital becoming a predominant source of income only for individuals beyond the 99^{th} percentile (Bassier and Woolard 2021).

The literature on top inequality focuses on high-income countries (Roine and Waldenström 2015) and frequently lacks a sectoral or firm-level approach. Middle-income countries with pronounced levels of inequality, such as South Africa or Brazil, are relatively understudied, despite the fact that the processes of change in their economic structures through which they must undergo could make their income distribution even worse. Understanding top inequality from sectoral and firm-level perspectives is crucial to inform inclusive processes of structural change in countries such as these.

In this proposal, we aim to address these gaps by investigating top earnings inequality in South Africa from a sectoral and firm-level perspective. We have three main research questions. First, what role does sectoral heterogeneity play in relation to inequality at the top? Second, is top inequality driven by heterogeneity between sectors, between firms in the same sectors, or within firms in the same sectors? Third, which sectoral and firm-level characteristics are related to the generation of inequality at the top?

To answer these questions, Section 2 reviews the relevant literature, and Section 3 lays out the data and method. Section 4 presents the results, which are discussed in Section 5, while Section 6 concludes. The Appendix is divided into a data section, a mathematical section, a section with additional figures and tables, and a final section covering an unusual relation observed between the two measures of inequality adopted in the paper.

2 Literature review

A growing literature on top earnings has been in development in recent years. We review parts of this literature covering drivers of top income inequality in general and top income inequality in South Africa. We also review empirical papers that perform sectoral or occupational decompositions of inequality measures, not necessarily focusing at the top.

2.1 Drivers of inequality at the top

The literature on top incomes discusses different drivers of concentration, which can be grouped into four categories: (i) sectoral-, trade-, and technology-related drivers; (ii) institutional drivers, particularly the role of taxation; (iii) the specific role of the finance sector and financialization; and (iv) wars and other major political shocks (Bartels and Waldenström 2021; Hager 2021; Medeiros and Souza 2015; Roine and Waldenström 2015).

In the first category, certain characteristics of specific sectors might mean they are more prone to generating super high incomes. Sectors can exhibit superstar or winner-takes-all characteristics, in which not only a few firms dominate the market but, above all, the wage distribution has very long tails (Rosen 1981; Frank and Cook 1996). This is typically the case in cultural sectors, such as audiovisual and sports, but can also be seen in other contexts such as for large innovative firms in very dynamic, innovation-driven sectors, such as ICT (Mahutga and Curran 2022). Trade and globalization play a particular role here: lower transportation costs and greater trade liberalization broaden markets, and effects of economy of scale may lead traditional sectors to exhibit winner-takes-all dynamics (Roine and Waldenström 2015).

Technological change is related to the above as the source of innovation rents in very dynamic sectors. It can also be an enabler of globalization by allowing for cheaper transportation costs. Technology may additionally come into play by shaping the relative demand for different types of labour. The literature on routine-biased technological change explores to which extent recent technological progress has disproportionately substituted routine tasks while complementing cognitive ones (Goos et al. 2014). This would hollow out the middle of the wage distribution and polarize workers at both ends. To the extent that workers on the high end of the distribution are also among top earners in the whole economy, this could additionally propel the detachment of super high incomes (Mahutga and Curran 2022).

The second category of drivers is related to social norms and other institutional aspects. The perceived fairness of top incomes will influence, for instance, the level of top marginal income tax rates (Piketty et al. 2014) or affect the prevalence of wage coordination mechanisms in opposition to individual negotiations—both lower rates and individual negotiations being more conducive to the concentration of income at the top (Piketty 2013). Top incomes might also

benefit from income flows coming from the state—both direct, such as pensions, and indirect, such as corporate subsidies (Medeiros and Souza 2015).

The finance sector and the financialization of the economy receive particular attention in this literature. This is because finance is a sector with superstar-like characteristics and because a more finance-oriented capitalism, with a higher emphasis on maximizing shareholder value, is one that tends to have social norms, institutions, and managerial practices that are more prone to generating super high incomes (Hein 2015). Although financialization is sometimes framed as a standalone driver (Bartels and Waldenström 2021; Mahutga and Curran 2022), it arguably shares the same mechanisms of the two previous categories in being a concentrated sector that affects social norms.

Other drivers are discussed by the literature but are less relevant for our paper, such as wars and other deep political shocks. These drive top income shares through their strong potential for radically rearranging the sectoral composition of economies, driving many businesses to bankruptcy and opening opportunities in other sectors (Bartels and Waldenström 2021; Roine and Waldenström 2015).

2.2 Top incomes in South Africa

Although the literature on top incomes in South Africa is still somewhat emergent, it reproduces two contemporary stylized facts found in many other countries: top income inequality is on the rise and back at the levels of one century ago (Piketty 2013), and labour incomes are increasingly important to explain income at the top (Piketty and Saez 2013; Milanović 2016).

In accordance with these stylized facts, top income shares in South Africa show two different trends through the last 100 years. It falls continuously from 1913 to the 1980s and, after a hiatus in available data, rises back in the 2002–07 period (Alvaredo and Atkinson 2022). This rise continues until 2018 despite a temporary drop following the global financial crisis of 2008 (Bassier and Woolard 2021). It is important to note that even in the period where inequality fell, its level remained high when compared, for example, to other former British colonies, such as Australia, Canada, and New Zealand (Alvaredo and Atkinson 2022).

Labour earnings are shown by the literature to be an important contributor to South African inequality in recent years (Bassier and Woolard 2021; Jacobs et al. 2024). This is true even at the very top, among the top 0.1% or the top 0.01% (Jacobs et al. 2024). In the findings of Bassier and Woolard (2021), salaries and bonuses make up over 80% of income of the top 1%, and it is only for the top 0.01% that capital income accounts for more than half of total income.

On the horizontal aspect of top inequality, while gender gaps have been closing at the bottom of the distribution, they have stagnated or risen at the top, suggesting a glass ceiling effect (Pleace et al. 2023; Mosomi 2019). The racial dimension, although not covered in this paper, appears very saliently for years in which South African tax statistics are published classified by race, particularly in the mid-20th century. Not only are the top earners very overwhelmingly white (always more than 97% across top quantiles through the 1950s and 1970s), but also the differences in levels of inequality with the above-mentioned countries disappear when inequality is measured only among white South Africans (Alvaredo and Atkinson 2022).

The literature, however, does not discuss extensively the role of sectors and firms. Among the few analyses in this direction, Alvaredo and Atkinson (2022) note that the growth of average income for the top 0.1% moves closely together with the value of gold production, and the few non-white individuals found at the top quantiles are more related to public employment than, for example, private employment in manufacturing. At the firm level, in a quantile regression focusing on the top 10%, Bhorat et al. (2017) find that firm size, productivity, profitability, being an importer or importer–exporter, and market concentration ratio are associated with higher individual wages, while firm age, capital intensiveness, and being only an exporter are associated with paying lower wages.

2.3 Sectoral and occupational decompositions of inequality

A segment of the broader literature on structural change and income inequality, not necessarily focused on the top of the distribution, performs decompositions of economy-level inequality into between-groups and within-groups components, with groups defined in reference to economic structure, such as sectors or occupations.[1] This explores, for instance, whether the overall dispersion of incomes is more related with dispersion of typical earnings across these groups or with the dispersion among individuals within these groups. This is done via well-known decomposition equations on levels for different measures of inequality, including generalized entropy (Kim and Sakamoto 2008; Maia 2013; Popli 2010; Ravallion 2022), Gini (Malkina 2019), and the variance of log incomes (Lee 2017; Williams 2013), the latter including at times controls for the effect of covariates (Carvalhaes et al. 2014; Mouw and Kalleberg 2010). The within/between decomposition may also be done on time differentials (Prasad 2002;

[1] Another part of this same literature employs other decompositions that serve the different purpose of exploring drivers of inequality. Methods in this group include extensions of the Oaxaca-Blinder (OB) framework (Beeson and Tannery 2004), notably under the Recentred Influence Function approach (Castellano et al. 2017, 2019, 2021; Firpo et al. 2018; Hinojosa 2021); counterfactual distributions (Molinder 2019; Sologon et al. 2021); and structural decomposition analysis (SDA) in the field of input-output tables (Dweck et al. 2024), among others (Goos and Manning 2007; Kanbur and Zhuang 2013; Mendieta-Muñoz et al. 2021). In this paper, references to decompositions should be understood as referring to within-/between-groups decompositions—not to those that investigate drivers of inequality.

Suen 1995), usually as a shift-share decomposition[2] (Changyuan and Jun 2009; de Serres et al. 2001; Elliott and Murphy 1990; Ibarra and Ros 2019; Kónya et al. 2020; Lopes et al. 2021; Maia et al. 2019; Martorano et al. 2016; Alarco Tosoni 2022; Zhang and Wu 2017). In all cases cited above, groups are either sectors or occupations, at times interacted with elements such as gender or education.

None of these papers, mapped in a systematic review of the literature on structural change and income inequality,[3] perform within/between decompositions at the firm level. Top inequality is also barely addressed. Only Popli (2010) uses a measure that is sensitive to the top, and the author does not discuss this result in depth.[4] At least two more recent papers do, however, adopt a firm-level perspective, even if not focusing at the top, decomposing the variance of log incomes on both sectors and firms (Haltiwanger et al. 2022; de Souza et al. 2023).

Performing a within-/between-groups decomposition of top inequality, having sectors and firms as the groups, is thus a novel contribution as far as we are aware. In any case, it is still relevant to synthesize results from the literature in order to have them as benchmarks to interpret our own. Table 1 details the empirical approaches and gives a broad summary of the findings of decompositions from the literature. Columns show the geographical scope, whether the decomposition is done on levels or on time differentials, and more details on the type of decomposition. Table 1 shows the measure of inequality; the group, or the main unit when there is more than one; the largest component in the findings; and the ratio of the within component to overall inequality, averaged when papers run more than one decomposition.

[2] In shift-share decompositions, an inequality measure is expressed as a weighted sum of group measures, and its change over time is broken down in one component related to changes in the group measures and another related to changes in the weights.

[3] The process is described in de la Vega (2023). Here, we substitute the inclusion criterion of having only papers that run regressions used in that meta-analysis with having only papers that perform decompositions. Searches were performed in November 2022.

[4] It is also unclear at this point how to interpret the results of Popli (2010) because of negative values for some of the between effects of the decompositions on levels of generalized entropy. For these reasons, we do not include these results in the summary done in this section, despite listing the paper in Table 1.

Table 1: Summary of papers that perform inequality decompositions related to structure

Paper	Geographical scope	Type	Detailed type	Measure	Main group	Main effect	Avg. W/O
Kim and Sakamoto (2008)	US	Levels	-	GE(1)	Occupations	Within	-
Popli (2010)	MX	Levels	-	GE(0), GE(2)	Sectors	Within	89.7%, 105%
Maia (2013)	BR	Levels	-	GE(1)	Occupations	Between	-
Ravallion (2022)	CN	Levels	-	GE(0)	Regions	Within	-
Malkina (2019)	RU	Levels	-	Gini	Regions	Within	-
Williams (2013)	GB	Levels	-	Var. of log incomes	Occupations	Within	-
Lee (2017)	KR	Levels	-	Var. of log incomes	Sectors	Within	87.5%
Carvalhaes et al. (2014)	BR	Levels	Also with controls	Var. of log incomes	Occupations	Within	-
Mouw and Kalleberg (2010)	US	Levels	With controls	Var. of log incomes	Occupations	Within	-
Maia et al. (2019)	BR, US	Time differentials	Shift-share	GE(1)	Occupations	Within	-
de Serres et al. (2001)	US, DE, FR, IT, BE, NL	Time differentials	Shift-share	Labour share	Sectors	Within	n/a
Changyuan and Jun (2009)	CN	Time differentials	Shift-share	Labour share	Sectors	Within	76.3%
Beqiraj et al. (2019)	OECD (9 countries)	Time differentials	Shift-share	Labour share	Sectors	Within	79.5%
Ibarra and Ros (2019)	MX	Time differentials	Shift-share	Labour share	Sectors	Within	n/a
Kónya et al. (2020)	EU (24 countries)	Time differentials	Shift-share	Labour share	Sectors	Within	n/a
Lopes et al. (2021)	PT	Time differentials	Shift-share	Labour share	Sectors	Within	69%
Alarco Tosoni (2022)	PE	Time differentials	Shift-share	Labour share	Sectors	Between	0.57%
Elliott and Murphy (1990)	GB	Time differentials	Shift-share	Wage gap	Sectors	Within	138.2%
Martorano et al. (2016)	KR, TW, ID, CN, IN	Time differentials	Shift-share	Wage gap	Sectors	Within	79.36%
Zhang and Wu (2017)	CN	Time differentials	Shift-share	Wage gap	Occupations	Between	-
Suen (1995)	HK	Time differentials	-	Var. of log incomes	Sectors	Within	104.2%
Prasad (2002)	GB	Time differentials	-	Var. of log incomes	Sectors	Within	106%

Source: author's own compilation, building from de la Vega (2023).

Table 1 shows that the within-sectors component is consistently the largest. The ratio of the within component to the overall measure puts in perspective the extent to which the within component dominates the decomposition. This comparison is challenging because papers vary significantly in their choices of inequality measurements, decomposition methods, and units for the decompositions. We then collected these data only for the nine papers that perform between-/within-sectors decompositions. Missing data in Table 1 are due to papers that only report results in figures and, because time differentials may be both positive and negative, the ratio in those cases can be greater than 100%.

Although the sample is small, it is possible to see a rough relation of larger within components for countries with greater income levels. Four papers find ratios above 85%—they cover the United Kingdom, Hong Kong, and South Korea. Another four papers find ratios between 69% and 80%—they cover South Korea, China, India, Indonesia, Taiwan, Portugal, and a group of nine OECD countries. The distinction is definitely not clear-cut, but there is a larger representation of non-high-income countries in the second group. The final paper, a very strong outlier in having an extremely small within component adding up to only 0.57% of the overall effect, performs its decomposition for Peru.

3 Data and method

Our empirical approach employs two measures of top inequality: generalized entropy and the share of earnings above a top quantile threshold. Both are initially measured at the economy level to understand the time trend of top earnings inequality. They are then measured at the sectoral level to explore the presence of sectoral heterogeneity. They are also measured at the firm level in such a way as to allow the decomposition of the economy-level measures in three components: between-sectors, within-sectors-between-firms, and within-sectors-within-firms. Finally, we use regressions to explore patterns between firm-level inequality and firm and sectoral characteristics. The following sections detail the data used in the paper, the measures of inequality and their decompositions, the dependent and independent variables, and the model specifications and estimation techniques used in the regressions.

3.1 Data sources

We use South African administrative data made available at the National Treasury Secure Data Facility (NT-SDF) in Pretoria, in the context of the Southern Africa – Towards Inclusive Economic Development (SA-TIED) programme, supported by the National Treasury of South Africa, the South African Revenue Service (SARS), and UNU-WIDER (National Treasury and

UNU-WIDER 2021a, 2021b; see also Ebrahim et al. 2021; Ebrahim and Axelson 2019; Pieterse et al. 2018). The data, in panel format, range from 2011 to 2017.

We use matched employer-employee data that cover several firm-level variables plus earnings of individual workers. We use SARS income source codes related to wages and different categories of remunerations (e.g., commissions, allowances, benefits), including the remuneration of directors and incomes of independent contractors. We refer to these collectively in this paper as 'earnings'. We deflate earnings and all relevant firm-level variables with the economy-wide gross domestic product (GDP) deflator available in the dataset, which has the first quarter of 2012 as the baseline.

The firm-level data used in this paper are gathered from ITR14 forms (and its predecessor IT14 forms, substituted in May 2013), which must be filled by all businesses resident in South Africa (Pieterse et al. 2018). Data on individual earnings come from job-level reporting done by employers via IRP5 and IT3(a) forms. All employees with earnings above ZAR2,000 must have an IRP5 form submitted on their behalf if employee tax was deducted from their remuneration or an IT3(a) form if no employee tax was deducted (Kerr 2020). The employer-employee match is done at the job level, but we aggregate the data at the level of individual worker identifiers, which are also provided in the dataset. Section A of the Appendix discusses in more detail the data and the process followed to clean it and deal with outliers.

3.2 Measures of inequality and their decompositions

This section presents the two measures of inequality used in the paper: generalized entropy and top earnings shares. The former is a measure of dispersion of earnings from the mean, while the latter is a measure of the share of earnings going to earners above a top quantile threshold. Hereinafter, we use 'quantile' to refer to a threshold of earnings rather than the subset of individuals who lie above this threshold. $P99_t$ is, thus, the smallest amount one needs to earn to be among the top 1% of earners in year t (and so on for PX_t with different values of X).

Generalized entropy offers the advantage of straightforward decomposition in within- and between-groups parcels—an approach that will form a key part of our empirical strategy. However, it has the drawback of being less intuitive to interpret. In contrast, the share at the top is more immediately interpretable but is not typically employed in a decompositions framework. To address this limitation, we propose a method in this paper for decomposing top shares in within- and between-groups components. In the following, we show a summary of equations for the two measures at the economy level, at the sectoral level, and at the firm level, as well as their decompositions. The full equations are shown in Section B of the Appendix.

Generalized entropy

For simplicity in the notation, this section omits the time subscript, but all equations here apply to individual incomes coming from a single year, and all measures are calculated separately for each year. Generalized entropy (GE) indices are such that for n individuals with earnings w_i, $i = 1, \ldots, n$, and mean earnings μ:

$$GE(\theta) = \frac{1}{n} \sum_{i=1}^{n} \varphi_\theta \left(\frac{w_i}{\mu} \right)$$

where the function $\varphi_\theta(x)$ is $\varphi_\theta(x) = \frac{(x^\theta - 1)}{\theta(\theta - 1)}$ (Shorrocks 2013).[5] The greater the parameter θ, the more sensitive $GE(\theta)$ is to changes at the top of the distribution. Typical values for θ are 0, 1, and 2.

Unit contributions are calculated similarly over all earnings paid within the unit. Thus, for the sectoral measure $GE(\theta)_j$, the summation is only done for workers i in sector j, while μ and n are substituted by μ_j and n_j, respectively, the average earnings for the sector, and the number of workers in the sector. The same applies for the firm-level measure of firm k, $GE(\theta)_k$, with μ_k and n_k as the respective counterparts of μ_j and n_j. The equations can be found in Section B of the Appendix.

One main advantage of entropy indices is that they are additively decomposable in group components. We can decompose the economy-level measure in a between-sector component, calculated as the economy-level entropy measure if all individuals in each sector j received the mean wage of the sector (μ_j), and a within-sector component, which is a weighted sum of the sectoral-level measures $\text{GE}(\theta)_j$ (Neves Costa and Pérez-Duarte 2019). Each of the sectoral-level measures may be similarly decomposed in between-firm (B_j) and within-firm (W_j) components. We can then combine these two levels to decompose the overall measure in three effects: between-sectors (B); within-sectors-between-firms (WB); and within-sectors-within-firms (WW):

$$\begin{aligned} GE(\theta) = &\left\{ \left(\frac{1}{n} \right) \cdot \sum_{j=1}^{m} \left[n_j \cdot \varphi_\theta \left(\frac{\mu_j}{\mu} \right) \right] \right\} \quad + \left\{ \sum_{j=1}^{m} \left[\frac{n_j}{n} \cdot \left(\frac{\mu_j}{\mu} \right)^\theta \cdot B_j \right] \right\} \quad + \left\{ \sum_{j=1}^{m} \left[\frac{n_j}{n} \cdot \left(\frac{\mu_j}{\mu} \right)^\theta \cdot W_j \right] \right\} \\ GE(\theta) = &\; B \qquad\qquad\qquad\qquad\qquad + WB \qquad\qquad\qquad\qquad\qquad + WW \end{aligned}$$

[5] For $\theta = 0$, $\varphi_0(x) = \lim_{\theta \to 0} \varphi_\theta(x) = -\ln(x)$. For $\theta = 1$, $\varphi_1(x) = \lim_{\theta \to 1} \varphi_\theta(x) = x \cdot \ln(x)$.

where B_j and W_j are respectively the between-firms and the within-firms components for each sector j, such that:

$$B_j = \left\{ \left(\frac{1}{n_j} \right) \cdot \sum_{k \in j} \left[n_k \cdot \varphi_\theta \left(\frac{\mu_k}{\mu_j} \right) \right] \right\} \qquad W_j = \left\{ \sum_{k \in j} \left[\frac{n_k}{n_j} \cdot \left(\frac{\mu_k}{\mu_j} \right)^\theta \cdot GE(\theta)_k \right] \right\}$$

Share of earnings above a top quantile

For the share of top earnings, the decomposition will be done on the variation of the measure through time, meaning that we cannot omit the time subscript for simplicity in the notation. Once more, we will present an overview of the measure and of its decomposition, while the full equations are shown and demonstrated in Section B of the Appendix.

The share of earnings above a top quantile at the economy level is measured as the sum of all earnings higher than the quantile divided by the sum of all earnings in a given year. We refer to this measure as sPX$_t$, where X is chosen in reference to different top quantiles, namely X $= 95, 99, 999, 9999$ (i.e. the top 5%, the top 1%, the top 0.1%, and the top 0.01%). For each quantile PX$_t$ in year t, we have sPX$_t = T_t / A_t$, where T_t is the sum of the earnings of all workers that earn above PX$_t$ in that year, and A_t is the sum of incomes of all workers in that year.

To calculate this measure at the unit (sector or firm) level, we add a detail that will help us with the decomposition. Instead of recalculating PX$_t$ for the earnings distribution within each sector and/or firm, we consider *the same PX$_t$ threshold valid for the entire economy in each year* to calculate the measure at the unit level.

Thus, for each sector j in year t, sPX$_{jt} = T_{jt} / A_{jt}$, where A_{jt} is the sum of earnings of all workers in that sector and T_{jt} is the sum of earnings of all such workers that earn above PX$_t$ (i.e. the economy-level threshold; note the absence of a sectoral subscript). Essentially, the measure at the sectoral level is the share of all earnings paid by the sector that goes to workers who lie at the top for the earnings distribution of the entire economy. Naturally, for a sector that does not have any workers above a top quantile X, sPX$_{jt} = 0$. The measure at the firm level for each firm k in year t is calculated similarly as sPX$_{kt} = T_{kt} / A_{kt}$, where T_{kt} and A_{kt} are the firm-level equivalents of T_{jt} and A_{jt}, such that $T_{jt} = \sum_{k \in j} T_{kt}$ and $A_{jt} = \sum_{k \in j} A_{kt}$.

With these definitions, we can write the economy-level measure as the weighted sum of sectoral-level measures, where weights are the share of each sector in earnings paid: sPX$_t = \sum_j ($sPX$_{jt} \cdot \pi_{jt})$. This allows us to decompose *changes* in sPX across time in two parts. In a simplified equation for this decomposition, we have: ΔsPX$_{t_1} =$ sPX$_{t_1} -$ sPX$_{t_0} = \sum_j \left($sPX$_{jt_1} \cdot \Delta \pi_{jt_1} \right) + \sum_j \left(\Delta$sPX$_{jt_1} \cdot \pi_{jt_0} \right)$. The first addend is a between-sectors component (fixed sector contributions, changing sector sizes), and the second is a within-sectors component (changing sector contributions, fixed sector sizes). In this simplified equation, we fixed sPX$_j$ in t_1 and π_j in t_0,

but we could also have fixed them inversely, i.e. respectively in t_0 and t_1. Indeed, in our decomposition, we actually calculate both expressions and use their average (see Equation 6 in Section B of the Appendix).

We also decompose the change in sectoral-level measures between t_0 and t_1 in between-firms (fixed firm contributions, changing firm sizes) and within-firms (changing firm contributions, fixed firm sizes) parcels. Firms may move across sectors, or in and out of the economy between t_0 and t_1. Because we are doing a decomposition of changes in measures, the calculations for any given sector j are partly done for variables in t_0 and partly for variables in t_1. Firms are only counted, for a given period t, in the calculations of sector j to which they belong in that period. That is, consider a firm that is in sector j in t_0 and moves out of the economy in t_1. When the decomposition for sector j is performed, this firm will appear in the variables related to t_0 but not in those related to t_1. More broadly, since the firm moved out of the economy, it will not appear at all in variables related to t_1, in any sector.

A simplified version of the equation for the decomposition in between-firms and within-firms parcels for sector j is: $\Delta \text{sPX}_{jt_1} = \sum_k \left(\text{sPX}^j_{kt_1} \cdot \Delta\pi^j_{kt_1} \right) + \sum_k \left(\Delta\text{sPX}^j_{kt_1} \cdot \pi^j_{kt_0} \right)$, where both sPX^j_{kt} and π^j_{kt} equal zero if the firm k is not in sector j in time t. Once more, in the actual decomposition, we do the average between this and the alternative expression, which inverts the timing of the fixed elements (see Equation 7 in Section B of the Appendix). Finally, we again combine both levels in a nested decomposition that splits the change of the economy level sPX in a between-sectors effect, a within-sectors-between-firms effect, and a within-sectors-within-firms effect: $\Delta\text{sPX}_t = B_t + WB_t + WW_t$.

3.3 Regressions: dependent and independent variables

After the decompositions, we employ regressions to explore patterns between firm-level measures and firm and sector characteristics. Our dependent variables are the inequality measures of Section 3.2 calculated at the firm level. We have, then, five different dependent variables in this set of regressions: sPX calculated for three different quantiles ($X = 95, 99, 999$) and GE(θ) calculated for two different parameters ($\theta = 1, 2$). We consider $X = 9999$ in the descriptive analyses but not in the regressions. It is also important to mention that GE(1) is not a measure that focuses at top inequality and is included as a robustness check, particularly in contrast with the results of GE(2).

We choose our independent variables in reference to the discussion on drivers of top income inequality from Section 2. Our main specification has four independent variables. First, we account for firm size by including the number of workers in the firm. Calculating the number of workers in a firm in a given year is not straightforward due to the timing of forms, and a number of alternative measures for this variable is available at the CIT-IRP5 panel (Ebrahim et al.

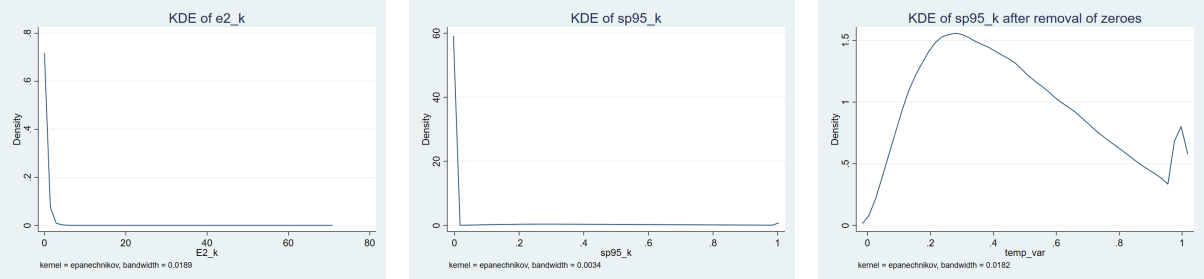
2021). However, to have a measure that maps more closely to our firm-level inequality measures, we use the number of workers from the individual panel who are matched to the firm in a given year ($\texttt{num\_workers}_{kt}$) and who thus served as a basis for the calculation of firm-level inequality. As expected, $\texttt{num\_workers}_{kt}$ has a skewed distribution, so we use its logarithmic transformation $\texttt{log\_num\_workers}_{kt}$ as an independent variable.

Sources of rent other than economies of scale may also allow a firm to remunerate their highest-earning workers so that they lie in the long tail of the distribution. We calculate the labour productivity $\texttt{lprod}_{kt}$ of each firm in each year as the difference between gross sales ($\texttt{g\_sales}_{kt}$) and the cost of production ($\texttt{g\_cos}_{kt}$) divided by the number of workers ($\texttt{irp5\_kerr\_weight\_b}_{kt}$). Here, we use one of the aforementioned variables for the number of workers available at the CIT-IRP5 panel. We use the log transformation $\texttt{log\_lprod}_{kt}$ as a second independent variable. We proxy the discussion on financialization and the wage-setting practices of a firm by including a dummy for whether the firm is listed in a stock exchange ($\texttt{c\_listedcomp\_d}_{kt}$). Unfortunately, this variable is only available for 2019 onwards. We impute the dummy backwards by setting it to one in all years of our analysis (2011–17) if the firm was ever listed in the years for which these data are available (2019–22). Finally, we include a measure of sectoral concentration via the Herfindhal-Hirschman Index (HHI) based on the number of workers on each firm in that sector in a given year. We use market shares as percentages (ranging from 0 to 100) when calculating HHI, meaning that it has an upper bound of 10,000. Because its distribution is also skewed, we again use the log transformation $\texttt{log\_hhi}_{jt}$ in our main set of regressions.

We run four robustness checks that include new controls. In the first one, we further investigate the role of wage-setting mechanisms by including the sectoral union density ($\texttt{uniondens}_{jt}$) in each year as an independent variable. This is calculated as the share of unionized workers in the sector using data from the Quarterly Labour Force Surveys (QLFS). Because data are missing particularly for Section T (related to activities of households as employees and activities for own use) on QLFS, we drop firms allocated in that sector from the analysis of this robustness check. In the second one, we account for geography by including dummies for the province in which the firm is situated ($\texttt{c\_province}$). Geographical data are not very well populated in 2011 and 2012, meaning that we drop those years from this second robustness check. In the third one, horizontal inequality is covered by including the share of male workers ($\texttt{maleshare}_{kt}$) in the company as an independent variable. Data on gender are missing for several individuals, leading to the option of dropping firms in which the gender of more than 10% of workers is unknown from the sample used in this robustness check. Finally, we use data on expenditure in research and development ($\texttt{x\_rd}_{kt}$) of each firm to build a dummy for whether the firm has such an expenditure ($\texttt{x\_rd\_cat}_{kt}$) and include that as an independent variable in this fourth robustness check. Only medium and large firms must report this expenditure, leading to an important amount of missing data for this variable.

We turn to important characteristics of these variables that will justify our choices for estimation techniques. First, the distributions of all firm-level inequality measures are right skewed. Additionally, distributions of all sPX_{kt} have a mass at zero, as the vast majority of firms have zero workers at top quantiles. When removing zeroes in these variables, the remaining distribution is still right skewed but to a much lesser degree. They also have another mass, although also much smaller, at their upper bound of one, with a few firms in which all workers lie at the top. These characteristics are exemplified in Figure 1, which shows kernel density estimations (KDEs) for (a) for GE(2)_{kt}, (b) sP95_{kt}, and (c) sP95_{kt} after removing zeroes.

Figure 1: KDEs for selected dependent variables



(a) GE(2) **(b) sp95_{kt}** **(c) sp95_{kt} without zeroes**

Source: author's own calculations.

Furthermore, variables show significantly more variation across firms than over time. This can be seen in Table 2, which shows summary statistics for dependent and independent variables. It is clear that the standard deviation between units is consistently higher than within units over time, the few exceptions being the share at the top for higher quantiles and unionization. Table 2 shows the number of observations that reflect the patterns of missing data in variables from the robustness checks, as mentioned above.[6]

[6] Table 2 does not show a summary for the province dummies nor for the dummy on R&D expenditure, which will be included in a future version of this paper.

Table 2: Summary statistics

Variable	Variable code	Mean	Std. dev. Between	Std. dev. Within	Std. dev.	Obs.	Panels	Average periods
GE(1)$_{kt}$	e1_k	0.363	0.245	0.216	0.124	669,879	135,558	4.94
GE(2)$_{kt}$	e2_k	0.539	0.679	0.563	0.346	669,879	135,558	4.94
sP95$_{kt}$	sp95_k	0.144	0.244	0.222	0.106	669,879	135,558	4.94
sP99$_{kt}$	sp99_k	0.035	0.125	0.109	0.062	669,879	135,558	4.94
sP999$_{kt}$	sp999_k	0.003	0.038	0.030	0.023	669,879	135,558	4.94
sP9999$_{kt}$	sp9999_k	0.000	0.011	0.009	0.008	669,879	135,558	4.94
Log no. of workers	log_num_workers	2.623	1.296	1.249	0.349	669,879	135,558	4.94
Log labour prod.	log_lprod	12.395	1.043	1.008	0.470	669,879	135,558	4.94
Listed firm dummy	c_listedcomp_d	0.001	0.031	0.031	-	669,879	135,558	4.94
Log HHI (sector)	log_hhi	4.317	1.161	1.076	0.479	669,879	135,558	4.94
Union dens. (sector)	uniondens	0.263	0.132	0.129	0.434	668,870	135,458	4.93
Share male workers	maleshare	0.586	0.277	0.277	0.078	550,597	122,177	4.51

Source: author's own calculations.

3.4 Model specifications and estimation

As mentioned in Section 3.3, we run regressions with five different dependent variables, sPX$_{kt}$ at three different quantiles ($X = 95, 99, 999$) and GE$(\theta)_{kt}$ for the two parameters ($\theta = 1, 2$). In our main regression, we have a model with the following independent variables: log_num_workers$_{kt}$; log_lprod$_{kt}$; c_listedcomp_d$_{kt}$; hhi$_{jt}$.

We perform four robustness checks, each adding a different set of controls. These checks focus on unionization, gender disparity, geography, and research and development (R&D) expenditure. As discussed in Section 3.3, due to missing data, each robustness check uses a separate subsample, exposing results to varying risks of sample selection bias. To explore how the subsampling affects results, and to have grounds for comparison for the newly introduced controls, for each robustness check we first run a model with only the four main independent variables in the respective subsample and then another including the new controls. Then, (a) for unionization, we run one additional specification including sectoral union density (uniondens$_{kt}$); (b) for gender disparity, we alternatively include the share of male workers in the firm (maleshare$_{kt}$); (c) for geography, we include dummies for the nine South African provinces; and (d) for R&D, we include a dummy for whether the firm had positive R&D expenditure[7] (x_rd_cat$_{kt}$).

For the regressions on generalized entropy, we employ linear panel estimation models. One possibility would be to use fixed-effects models, but they would only allow the exploration of the variation over time, ignoring the important variation between firms, while also demeaning away any time-invariant independent variables. These limitations could be overcome with the

[7] Although not reported in detail, we also ran a model with R&D intensity (x_rd_int$_{kt}$) instead of the R&D dummy, which showed very similar results.

use of random-effects models, but these would require the unobserved heterogeneity to be uncorrelated with the independent variables, which could be hard to justify. The framework of correlated random-effects (CRE) models, also known as the Mundlak specification (Wooldridge 2019), offers a solution by explicitly modelling the correlation between the unobserved effect and each independent variable via the within-firm average (i.e. over time) of the independent variable.

Thus, the CRE approach starts by calculating, for each observation and for each independent variable, the average of that independent variable across observations of the same firm over time. These averages are then included as additional independent variables, as done in Mundlak (1978) and elsewhere. When this is done, the coefficient of the non-averaged independent variable reflects how its variation within panels relates to the dependent variable, while the coefficient of its average captures the 'contextual effect' (Bell et al. 2019), which is the difference between the variation over time and variation across panels.

To simplify the interpretation of the coefficients, we would like to be able to isolate the effect of variation over time from the effect of variation across firms in separate coefficients, rather than having them mixed in the 'contextual effect'. This is made possible by a variation of the CRE model known as the within-between random effects (REWB) model (Bell and Jones 2015). We then use REWB models for the regressions related to GE. Here, a random-effects estimation is performed in which each independent variable appears twice: once demeaned/within-transformed, which captures variation within firms over time, and once as its average over time for that firm, which captures variation across firms. For unbalanced panels such as ours, we follow Wooldridge (2019) and calculate the averages over time based only on observations with complete data. Since our robustness checks are conducted on subsamples, the averages must then be recalculated for observations with complete data within each subsample.

As seen in Section 3.2, within components for generalized entropy are weighted sums of the inequality measures calculated at the firm level. On the models above, our generalized entropy dependent variables are always the unweighted ones. We then run an additional variation for these dependent variables switching them for their weighted versions. The weights are the ones shown on Equation 4 in Section B of the Appendix.

Top shares are corner solution variables, with a mass at zero and a (much smaller) mass at one. This motivates the use of a Tobit model. Wooldridge (2010) points out that Tobit models may employ a CRE-like specification to allow for correlation between the unobserved heterogeneity and independent variables and may employ pooled instead of panel estimation to avoid the requirement of serially independent error terms. We follow these recommendations while also using the REWB specification instead of the CRE one. That is, we employ a pooled Tobit estimation in which each independent variable appears twice: once within-transformed

and once as its panel average over time. We refer to this estimation as a within-between pooled Tobit (WBPT). We calculate average partial effects (APEs) following Wooldridge (2010).

All specifications presented above, for all dependent variables, include dummy variables for time and sector at the one-digit level as well as their averages over time, given that the panel is unbalanced. Sector A ('Agriculture, forestry and fishing') and year 2011 are used as baselines. All standard errors are clustered at the firm level, except in the calculation of APEs of the Tobit models.

4 Results

We present the results of our analysis in four subsections. We start with the evolution of the economy-level inequality measures and the sectoral composition of the economy, showing that these move slowly through time and do not exhibit drastic variation within our time range. We then show the clear presence of sectoral heterogeneity in both weighted and unweighted sectoral-level inequality measures. Despite this, we conclude from our decompositions that the main component of overall inequality is not the between-sectors but rather the within-sectors-within-firms. This then motivates the investigation of patterns between firm-level inequality and characteristics of firms and sectors, which we cover in the regressions.

4.1 Economy-level inequality and sectoral transformation: slow movements

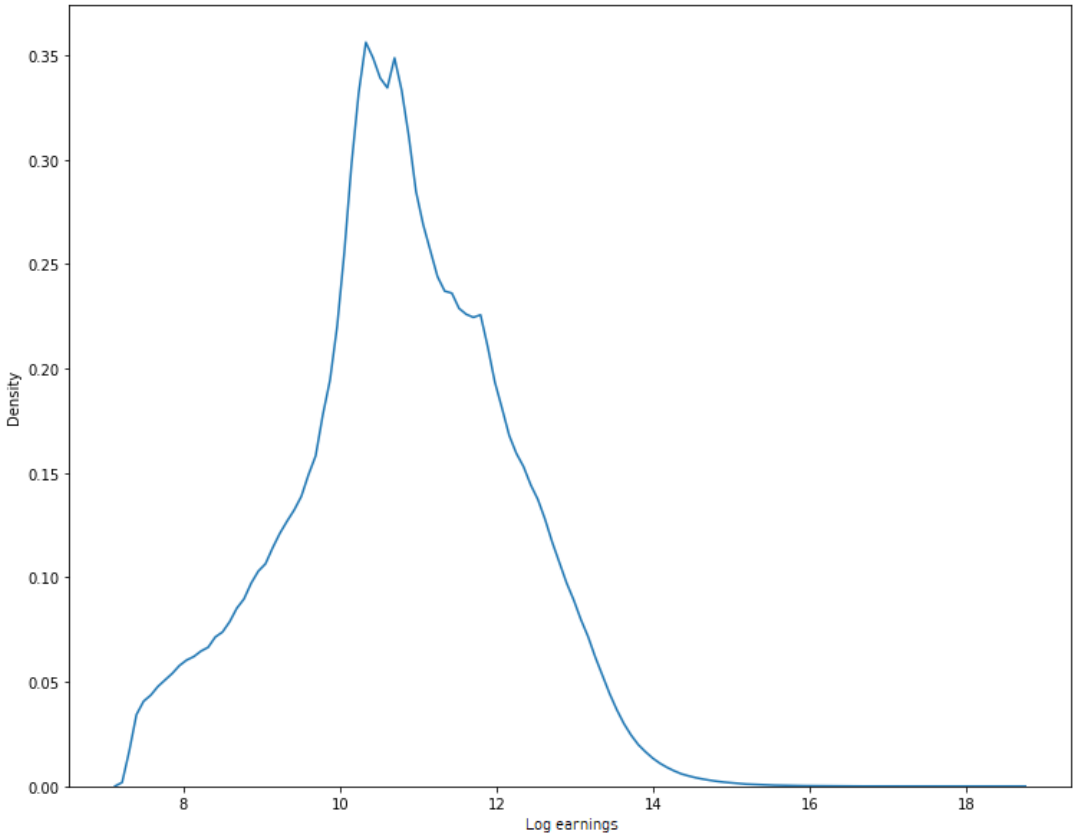
Table 3 shows the time trends for top earnings inequality between 2011 and 2017. Overall, there are not large variations within our rather limited time range. If anything, it can be said that top earnings inequality rises continuously until 2015, followed by a slight reduction in 2016 and/or 2017, and that this trend is more visible the more one focuses at the top (e.g., $X = 999,9999; \theta = 2$). Indeed, for these parameters, the oscillation amounts to over 10% of the initial value—even reaching c. 30% for sP9999—although it is true that the absolute variations over time are a lot less impressive.

Table 3: Economy-level measures of inequality between 2011 and 2017

Year	sP95	sP99	sP999	sP9999	GE(2)	GE(1)
2011	35.91%	14.54%	3.83%	0.98%	2.766	0.857
2012	36.29%	14.80%	4.02%	1.09%	2.711	0.870
2013	36.47%	15.14%	4.27%	1.24%	3.082	0.885
2014	36.59%	15.33%	4.43%	1.31%	3.155	0.890
2015	36.56%	15.53%	4.56%	1.32%	3.351	0.894
2016	36.13%	15.16%	4.35%	1.24%	3.407	0.879
2017	36.21%	15.34%	4.44%	1.24%	3.158	0.880

Source: author's own calculations.

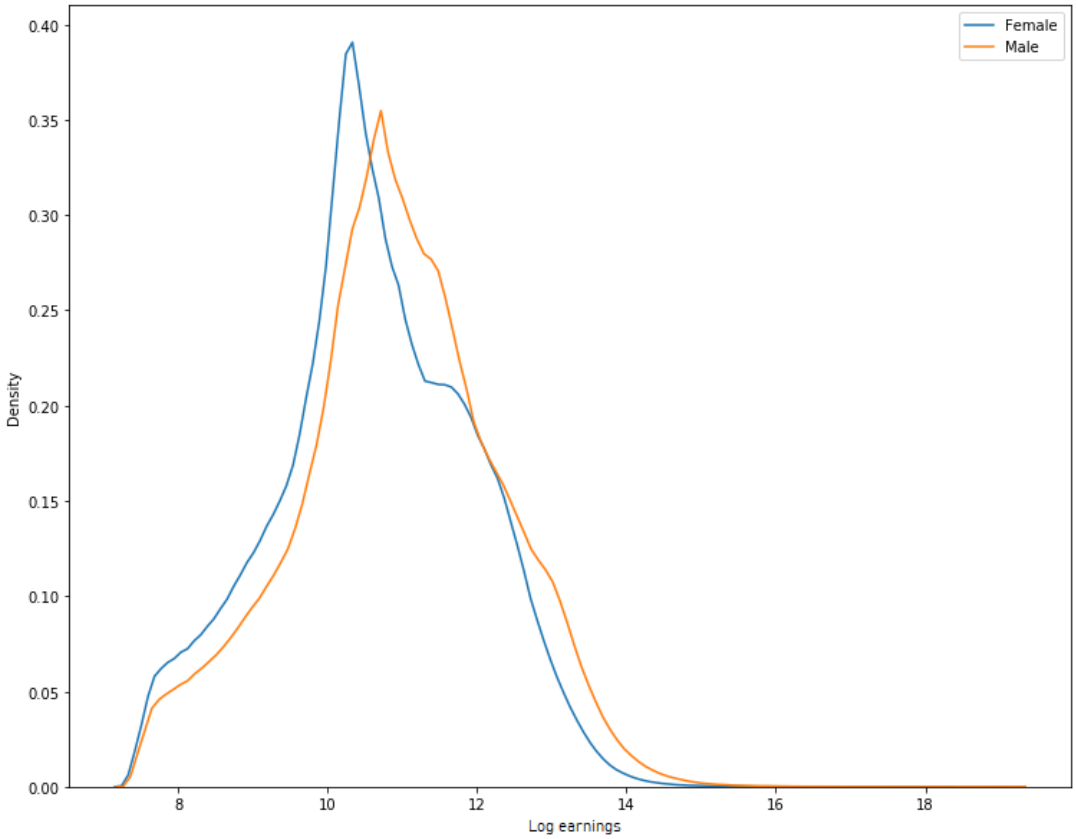
Figure 2: KDE for individual real log earnings in all years



Source: author's own calculations.

Figure 2 shows the KDE for the distribution of individual log incomes when incomes from all years are pooled together. One noticeable aspect of this distribution is that it is bimodal with two distinguishable peaks. These are likely caused by gender gaps. Figure 3 shows the same KDE but broken down by gender, and the peak of each distribution matches the double peaks of the overall distribution. Figure C3 in Section C shows a similar breakdown of the KDE but for each year in our time range, where we once more do not see dramatic movements over time.

Figure 3: KDE for individual real log earnings in all years by gender

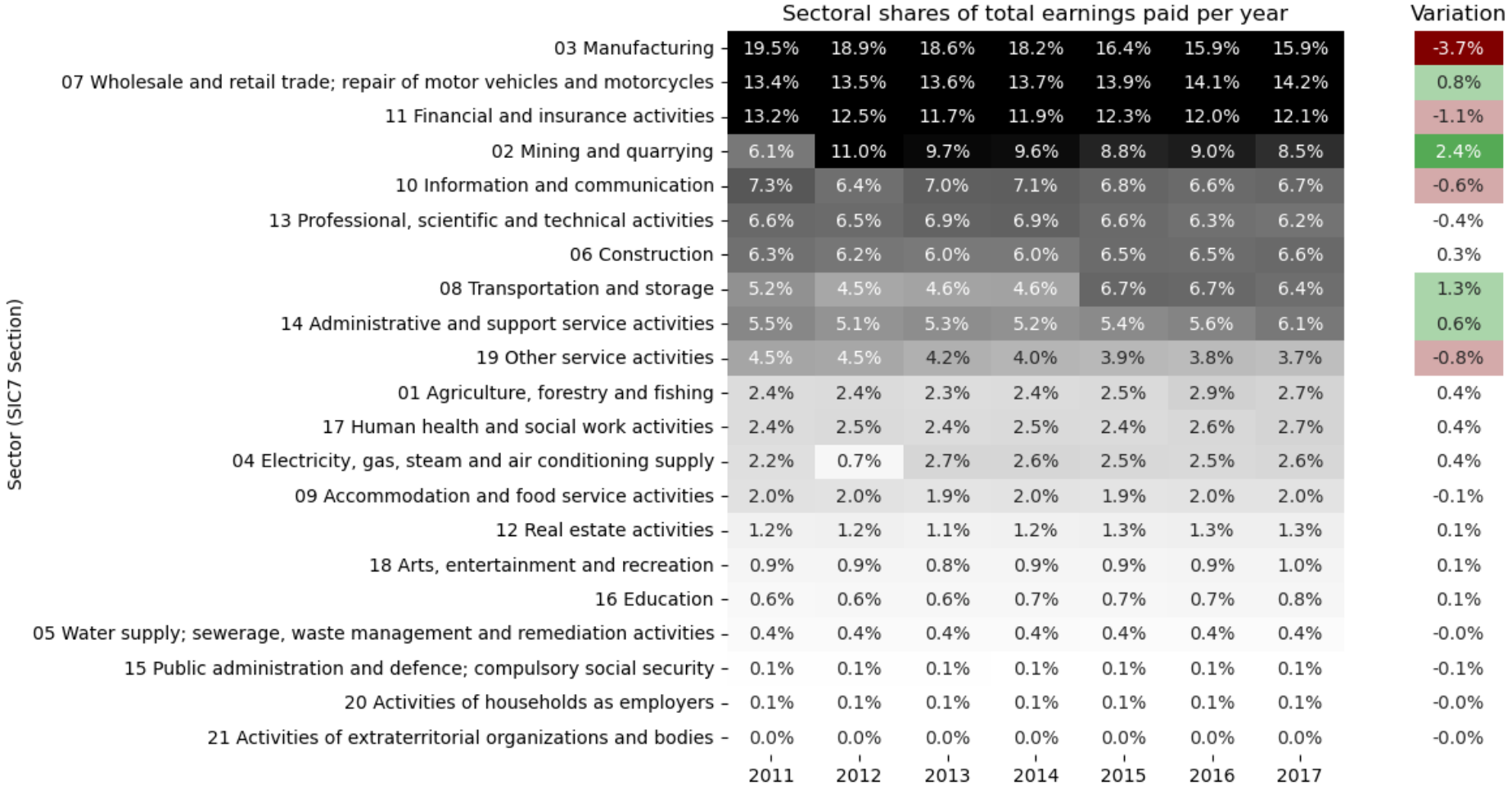


Source: author's own calculations.

The same lack of variation over time can be seen for sectoral transformations. Figure 4 shows the sectoral composition of the economy in terms of the sectoral shares in all earnings paid per year (i.e. A_{jt} in the equations of Section 3.2) between 2011 and 2017. Throughout the paper, sectors are identified as Sections (one-digit level) of the Standard Industrial Classification Seventh Edition (SIC7) (Statistics South Africa 2012).

Considering the subset of South African firms that constitute our sample, the main sectors are *Manufacturing*, *Wholesale and retail trade*, and *Financial and insurance activities*, followed by *Mining and quarrying*. The first three account for 42.2% of earnings paid in the South African formal market, rising to 50.7% with the inclusion of mining. The observable sectoral transformation between 2011 and 2017 is mainly away from *Manufacturing* and towards *Mining and quarrying*. It is worth mentioning that this is influenced by an unusually low share for mining in 2011 and that the overall trend for that sector from 2012 onward is actually a negative one. In any case, data suggest somewhat weaker movements away from *Financial and insurance activities*; *Other service activities*; and *Information and communication*—and towards *Transportation and storage*; *Wholesale and retail trade*; and *Administrative and support service activities*.

Figure 4: Sectoral transformation between 2011 and 2017 at the SIC7 sector level



Source: author's own calculations.

4.2 Heterogeneity in sectoral-level measures of inequality

Despite the lack of variation at the economy level through our time range, we still see important differences in top inequality across sectors. As explained in Section 3.2, our two inequality measures can be calculated at the sectoral level for each sector in each year. The respective economy-level measures—or its within-sectors components, in the case of generalized entropy—are weighted sums of the sectoral-level measures, where weights are related to sector size.[8]

Tables 4 and 5 show the averages over time of, respectively, the weighted and non-weighted sectoral-level measures for each sector, as well as their ranks, with the top five sectors in each measure marked in bold. The two tables tell slightly different stories. The weighted measures show which sectors contribute most to top income inequality in South Africa in our time range, which is a combination of how unequal the sectors are in themselves and how large they are in the national sectoral composition. The unweighted measures isolate how inequality behaves at the level of each sector from their sizes—thus, it sheds light onto how inequality could behave in the longer term in the context of structural change.

[8] For all sPX and for GE(1), weights are the sectoral shares in earnings paid. For GE(2), weights are the square of this sectoral share divided by the sectoral share in the number of workers. The weights can be found in Section B of the Appendix, Equation 5 for the share at the top, and Equation 1 (within-sectors component) for generalized entropy.

Table 4: Averages over time of weighted sectoral-level inequality measures and sector rankings

Sector	$sP95_j$	Rank	$sP99_j$	Rank	$sP999_j$	Rank	$sP9999_j$	Rank	$GE(2)_j$	Rank	$GE(1)_j$	Rank
11 Financial and insurance activities	**6.53%**	1	**3.73%**	1	**1.61%**	1	**0.59%**	1	**1.015**	1	**0.107**	3
03 Manufacturing	**6.21%**	2	**2.24%**	2	**0.55%**	2	**0.13%**	2	**0.401**	2	**0.126**	1
07 Wholesale and retail trade; repair of motor vehicles and motorcycles	**3.97%**	3	**1.57%**	3	**0.43%**	3	**0.13%**	3	**0.327**	3	**0.117**	2
02 Mining and quarrying	**3.43%**	4	**1.26%**	5	**0.30%**	5	**0.08%**	5	**0.188**	5	0.048	7
10 Information and communication	**3.39%**	5	1.18%	6	0.25%	6	0.06%	6	0.158	7	0.040	8
13 Professional, scientific, and technical activities	3.26%	6	**1.55%**	4	**0.40%**	4	**0.09%**	4	**0.282**	4	**0.060**	4
06 Construction	2.02%	7	0.73%	7	0.12%	8	0.02%	8	0.110	8	**0.055**	5
08 Transportation and storage	1.38%	8	0.55%	9	0.09%	9	0.01%	10	0.075	9	0.030	10
14 Administrative and support service activities	1.34%	9	0.66%	8	0.21%	7	0.06%	7	0.162	6	0.050	6
04 Electricity, gas, steam, and air conditioning supply	1.12%	10	0.22%	13	0.02%	16	0.00%	18	0.027	13	0.010	14
19 Other service activities	0.98%	11	0.34%	10	0.05%	12	0.01%	13	0.054	10	0.035	9
17 Human health and social work activities	0.80%	12	0.34%	11	0.07%	10	0.00%	14	0.042	12	0.015	12
01 Agriculture, forestry, and fishing	0.59%	13	0.24%	12	0.06%	11	0.02%	9	0.044	11	0.027	11
12 Real estate activities	0.44%	14	0.18%	14	0.03%	13	0.01%	11	0.025	14	0.010	15
09 Accommodation and food service activities	0.29%	15	0.12%	16	0.03%	15	0.01%	12	0.022	15	0.015	13
18 Arts, entertainment, and recreation	0.28%	16	0.13%	15	0.03%	14	0.00%	15	0.019	16	0.008	16
16 Education	0.13%	17	0.04%	17	0.01%	17	0.00%	16	0.007	17	0.005	17
05 Water supply; sewerage, waste management, and remediation activities	0.10%	18	0.03%	18	0.01%	18	0.00%	17	0.005	18	0.003	18
15 Public administration and defence; compulsory social security	0.02%	19	0.01%	19	0.00%	21	0.00%	19	0.001	19	0.001	19
21 Activities of extraterritorial organizations and bodies	0.01%	20	0.00%	21	0.00%	20	0.00%	19	0.000	21	0.000	21
20 Activities of households as employers and activities for own use	0.01%	21	0.00%	20	0.00%	19	0.00%	19	0.001	20	0.000	20

Source: author's own calculations.

Table 5: Averages over time of unweighted sectoral-level inequality measures, and sector rankings

Sector	$sP95_j$	Rank	$sP99_j$	Rank	$sP999_j$	Rank	$sP9999_j$	Rank	$GE(2)_j$	Rank	$GE(1)_j$	Rank
11 Financial and insurance activities	**53.5%**	1	**30.6%**	1	**13.2%**	1	**4.8%**	1	**4.04**	3	**0.88**	4
10 Information and communication	**49.7%**	2	**17.4%**	3	**3.7%**	4	**0.9%**	5	1.13	20	0.59	18
13 Professional, scientific, and technical activities	**49.5%**	3	**23.5%**	2	**6.2%**	2	**1.3%**	2	**3.03**	5	**0.91**	3
04 Electricity, gas, steam, and air conditioning supply	**48.1%**	4	9.6%	14	0.8%	19	0.1%	18	0.60	21	0.46	21
21 Activities of extraterritorial organizations and bodies	**41.2%**	5	12.1%	9	0.7%	20	0.0%	19	1.60	15	0.84	9
02 Mining and quarrying	38.2%	6	14.1%	6	3.4%	6	0.9%	6	1.32	17	0.54	20
12 Real estate activities	35.7%	7	**14.9%**	4	2.8%	9	0.5%	9	1.87	13	0.79	11
03 Manufacturing	35.2%	8	12.7%	8	3.1%	7	0.8%	7	1.91	11	0.72	15
17 Human health and social work activities	32.3%	9	13.5%	7	2.8%	10	0.2%	16	1.23	18	0.61	17
06 Construction	32.0%	10	11.6%	11	1.9%	12	0.3%	12	2.16	10	**0.86**	5
18 Arts, entertainment, and recreation	30.8%	11	**14.4%**	5	**3.6%**	5	0.5%	10	2.38	8	0.85	6
07 Wholesale and retail trade; repair of motor vehicles and motorcycles	28.8%	12	11.4%	12	3.1%	8	**0.9%**	4	**3.10**	4	0.84	8
05 Water supply; sewerage, waste management, and remediation activities	25.5%	13	8.5%	16	1.3%	16	0.3%	11	1.73	14	0.75	13
08 Transportation and storage	25.1%	14	10.0%	13	1.6%	13	0.2%	14	1.13	19	0.54	19
14 Administrative and support service activities	24.5%	15	12.0%	10	**3.8%**	3	**1.1%**	3	**5.53**	1	**0.92**	2
19 Other service activities	24.0%	16	8.4%	17	1.3%	15	0.1%	17	2.21	9	0.85	7
01 Agriculture, forestry, and fishing	23.5%	17	9.6%	15	2.3%	11	0.6%	8	**5.00**	2	**1.10**	1
15 Public administration and defence; compulsory social security	23.2%	18	6.6%	18	0.1%	21	0.0%	19	1.89	12	0.77	12
16 Education	18.8%	19	5.5%	21	1.1%	17	0.2%	15	1.33	16	0.67	16
20 Activities of households as employers and activities for own use	15.9%	20	6.4%	19	0.8%	18	0.0%	19	2.42	7	0.81	10
09 Accommodation and food service activities	14.7%	21	5.9%	20	1.6%	14	0.3%	13	2.55	6	0.74	14

Source: author's own calculations.

We start with the weighted measures of Table 4. We find that the same sectors tend to figure among the top contributors across the different measures: *Financial and insurance activities*; *Manufacturing*; *Wholesale and retail trade*; *Mining and quarrying*; *Professional, scientific, and technical activities*; and *Information and communication*. In particular, finance, manufacturing, and trade are ranked one to three in that order in all top inequality measures, while also being the top three for GE(1). Since weights are related to sector size, and the distribution of sector size is skewed (cf. Figure 4), it is no surprise that the largest sectors appear among those that contribute the most to economy-level inequality. At the bottom of the rankings, there is also very little variation across measures and parameters.

We are able to give more nuance to this analysis by looking at the unweighted measures. A hint at this is that finance is consistently in first place in Table 4, despite being the third-largest sector. Turning attention to Table 5, we see that finance is the most unequal sector in the unweighted share at the top across all quantiles, and it gets progressively detached from other sectors the more we move to the top. It is followed by *Professional, scientific, and technical activities* and *Information and communication*. These three sectors are among the top five in all four quantiles, although information and communication moves down progressively in the rank as we move towards the top. The behaviour of these top sectors, and particularly the detachment of finance from the distribution of unweighted sectoral-level measures, can also be seen in the time trends of these measures, as shown in Figure C2 in Section C of the Appendix. At the bottom, there is a larger pool of low-contributing sectors. Although they switch around the lower parts of the ranking in Table 5, we see a prevalence of *Activities of households as employers* and *Public administration and defence* and, to a lesser extent, *Accommodation and food service activities*; *Electricity, gas, steam, and air conditioning supply*; and *Education*.

This time, however, rankings change a lot for GE(2) in relation to the rankings for top shares. The presence of *Agriculture, forestry, and fishing* among the bottom five for sP95 but the second-most unequal sector for GE(2) is notable. This suggests that agriculture has an internal dispersion of earnings with a very long tail, but individuals in this long tail do not fall in the tail of the earnings distribution of the overall economy. The opposite occurs with *Information and communication*, which has a high rank for the share at the top but a low one for GE(2). This suggests that an important amount of earnings in this sector is paid to people in the top quantiles of the whole economy, but its internal dispersion of earnings is not so significant.

Returning to the relation between size and sector-level inequality, scatterplots of sector size (number of workers) versus the sectoral-level share at the top, particularly for quantiles 99 and above, show that sectors tend to gather around two different slopes. Sectors in the steeper slope are precisely those that are more unequal when measured by unweighted sPX (finance, professional activities, and information and communication). We do not want to overstate this

finding, given the descriptive nature of this relation, but the scatterplots may be seen in Figure C1 in Section C of the Appendix.

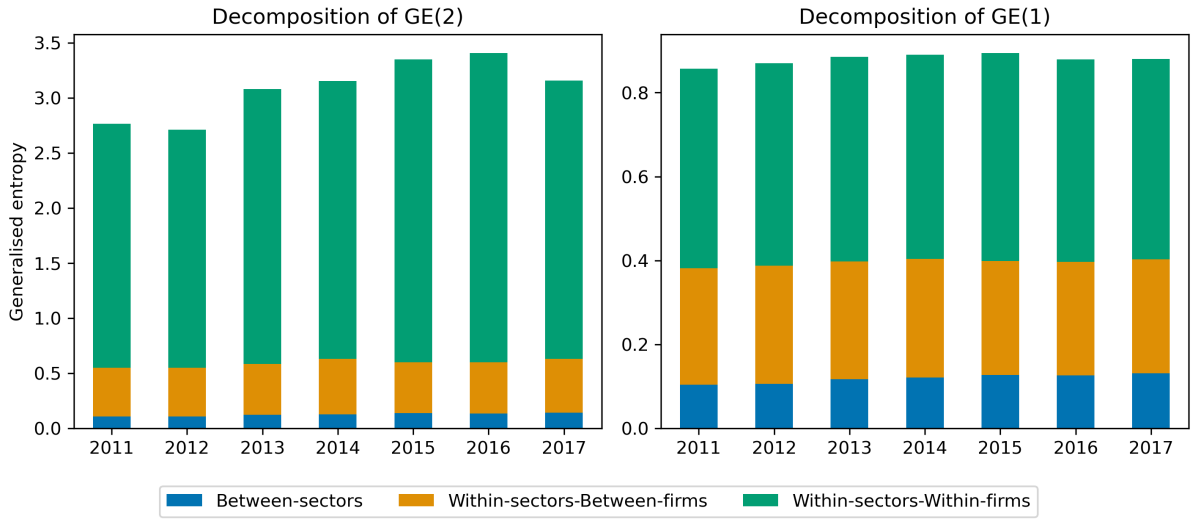
4.3 The predominance of firm heterogeneity: decompositions

Following the equations in Section 3.2, we perform decompositions of our economy-level measures in between-sectors, within-sectors-between-firms, and within-sectors-within-firms effects. Despite the clear presence of sectoral heterogeneity, as shown in Section 4.2, we find that the main component turns out to be the within-firms component rather than the between-sectors.

It is worth remembering that the decompositions for the two measures are different in two important ways. First, the decomposition of GE is done yearly on the economy-level measure of that year, while the decomposition of sPX is done on variations over time between the first and last years in our range. Second, the between components have slightly different meanings. For GE, it measures the variation of average earnings across units, while for sPX, it is related to changes in the relative sizes of units.

We start with the decomposition of GE(2), illustrated in the left panel of Figure 5. It is clearly visible that most of the measure is driven by the within-sectors components and more specifically by the within-sectors-within-firms component. We also show the decomposition for GE(1) in the right panel of Figure 5 for comparison. It shows that the predominance of the within-firms component is even larger when focusing at the top. For GE(2), c. 80% of the economy-level measure is explained by earnings dispersion within firms, rather than between the average earnings of firms and sectors.

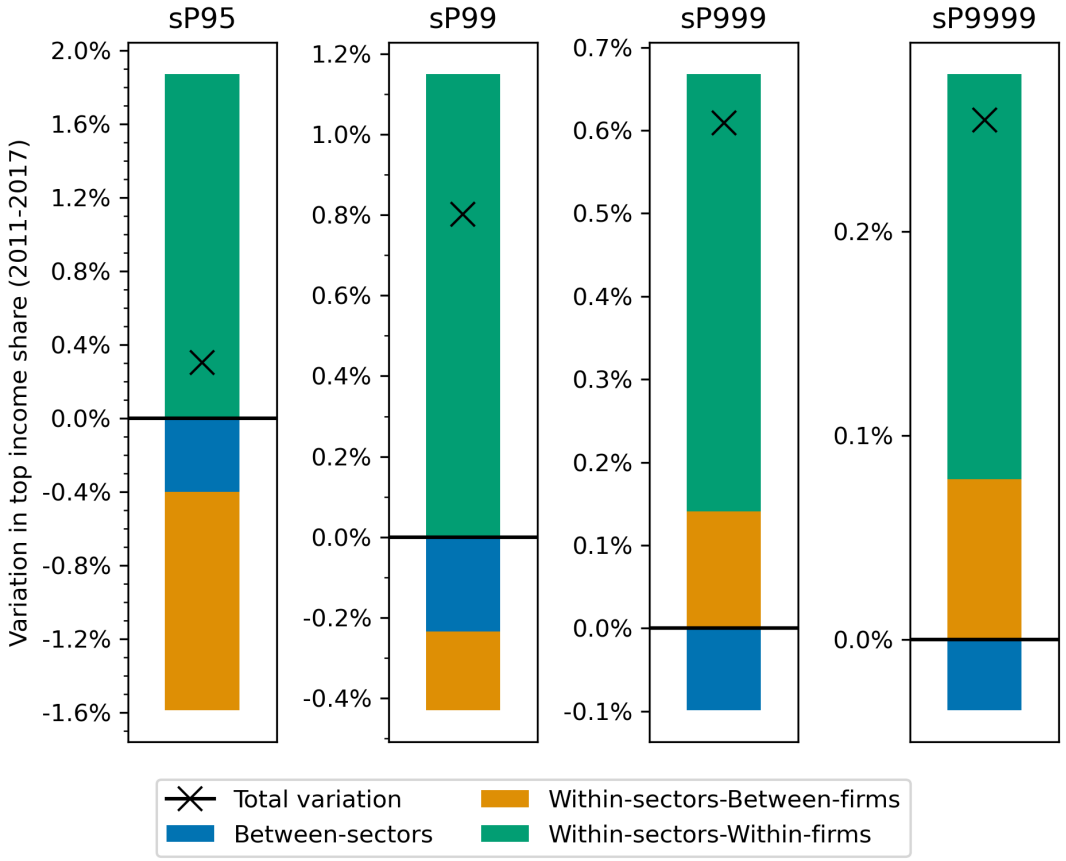
Figure 5: Decomposition of generalized entropy with parameters 2 (left panel) and 1 (right panel)



Source: author's own calculations.

We move to ΔsPX, where results are similar, though less pronounced. The decomposition is shown in Figure 6, where each panel corresponds to a different top quantile ($X = 95, 99, 999, 9999$). Once again, within-firms components dominate—and more so, the more we move up the quantiles—while between-sector components consistently point in a different direction. Results show that, in the covered time range, the rise in top income shares was mostly driven by a greater contribution of firms, while the recomposition of sector sizes counteracted this trend to some extent. The ratios of the within-sectors-within-firms component to the total variation in the top share are, in the order of the quantiles, 664.4%, 155.3%, 103.7%, and 98.1%.

Figure 6: Decomposition of the change (2011–17) in top shares for different quantiles



Source: author's own calculations.

In addition to the decompositions, an analysis of the correlations between movements of workers across firms and sectors, as well as in and out of the top, further highlights the importance of within-firms components. While no causality can be inferred from merely observing the co-occurrence of such movements, we find that workers who do not switch firms or sectors are more likely to simultaneously move into the top (and also out of it) compared to those who do switch, particularly above the top 5% quantile. These numbers are presented in Tables C10 and C11 under Section C3 of the Appendix.

4.4 Investigating patterns in firm heterogeneity: regressions

We have established that, despite the clear presence of sectoral heterogeneity, the within-firms component of top inequality is considerably larger than the between-sectors component or the within-sectors-between-firms component. We then turn to regressions that explore patterns in firm-level inequality and characteristics of firms and their sectors. Table 6 shows results for the main specification for each dependent variable. REWB coefficients are shown for GE(1) and GE(2), while Tobit marginal effects, calculated as APEs, are shown for sP95, sP99, and sP999.

We start with the results for GE(2). Size is shown to consistently have a significant relation with inequality, both for the within-transformed (i.e. demeaned, identified as '(W)') and between-transformed (i.e. averaged for each firm through time, identified as '(B)') coefficients. Because the number of workers is log-transformed, these coefficients represent semi-elasticities: a 1% rise in size over time is related to a 0.00326 increase in GE(2), while a similar variation in size across different firms is related to a 0.00237 increase. For reference, the mean value of GE(2) at the firm level is 0.539. It is perhaps more useful to think in terms of discrete variations in the number of workers. If we take, for instance, a firm with 10 workers (the mean for `num_workers` is 13.8), hiring an additional worker in another period is expected to be associated with an increase of 0.0311 in GE(2), while a firm with an average of 10 workers across periods, when compared with another firm that has 11 workers on average, is expected to be associated with a decrease of 0.0226 in GE(2). If the comparison was made with firm size doubling, the expected variations in GE(2) over time and across firms would be 0.226 and 0.164, respectively.

Productivity also shows a positive relation with GE(2), although only for the variation across firms. The within variation has a negative coefficient but a non-significant one. For HHI, both coefficients are negative, while only the within-transformed one has strong significance. The dummy for listed firms is not found to be significantly related to GE(2)—because this is a time-invariant variable, we do not have its within-transformed version.

The comparison of results for GE(2) with GE(1) can be thought of as an initial robustness check, since the latter is not a measure of top earnings inequality. Size and HHI show a stronger relation with GE(2) than with GE(1), while the opposite is valid for productivity. The listed dummy for GE(1) has statistical significance and a negative coefficient.

Table 6: REWB regression results for GE(1) and GE(2) and WB pooled Tobit APEs for sP95, sP99, and sP999

Dependent variable	GE(1)$_k t$	GE(2)$_k t$	sP95$_k t$	sP99$_k t$	sP999$_k t$
Log number of workers (W)	0.145***	0.325***	0.0353***	0.0128***	0.00173***
	(0.00105)	(0.00401)	(0.000757)	(0.000386)	(0.000124)
Log labour productivity (W)	0.00226***	-0.00244	0.0258***	0.00811***	0.00107***
	(0.000546)	(0.00160)	(0.000572)	(0.000310)	(0.000107)
Listed firm	-0.120***	-0.00325	0.000490	8.37e-06	5.93e-05
	(0.0243)	(0.173)	(0.000483)	(0.000245)	(9.55e-05)
Log HHI (W)	-0.00258***	-0.00586***	0.0617***	0.0240***	0.00305***
	(0.000547)	(0.00134)	(0.000391)	(0.000225)	(7.78e-05)
Log number of workers (B)	0.0864***	0.239***	0.101***	0.0333***	0.00375***
	(0.000548)	(0.00242)	(0.000717)	(0.000403)	(0.000126)
Log labour productivity (B)	0.0135***	0.00801***	-0.0109***	-0.00248	-0.00111*
	(0.000626)	(0.00177)	(0.00329)	(0.00172)	(0.000601)
Log HHI (B)	-0.00223	-0.00902*	-0.103***	-0.0263***	7.29e-05
	(0.00246)	(0.00506)	(0.0177)	(0.00647)	(0.000770)
Observations	669,879	669,879	669,879	669,877	669,515
Panels	135,558	135,558	-	-	-
Estimation	REWB	REWB	WB pooled Tobit	WB pooled Tobit	WB pooled Tobit
Time FE	YES	YES	YES	YES	YES
Sector FE	YES	YES	YES	YES	YES

*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Source: author's own calculations.

Table 7 shows the robustness checks for GE(2). Results for the main specification with the full sample are repeated in the first column of Table 7, identified as 'Main'. Models for the four robustness checks are identified as 'UD' (union density), 'PR' (province dummies), 'SM' (share of male workers), and 'RD' (research and development). Results for the main specification with each subsample are identified, respectively, as UD1, PR1, SM1, and RD1. Specifications including the new independent variables related to each robustness check are identified as UD2, PR2, SM2, and RD2. Table 7 also identifies the sample subset and shows the number of observations; the number of individual firms; the estimation technique; the presence of time, sector, and/or province dummies; and the within, between, and overall R-squared.

We see very little change in the estimates for the main set of independent variables when comparing results for the main regression with those for UD1, PR1, and SM1, suggesting that the subsampling in these cases do not seem to induce significant sample selection bias. The subsampling for R&D (RD1) does seem to cause some alteration, however, particularly in the coefficients for HHI.

Union density, measured at the sectoral level, has a strongly significant negative relation with GE(2), although only for the variation across firms. Results suggest, though, that it takes a sizeable variation in union density to compensate, for instance, for a marginal change in size. In the comparison made above between different firms with 10 or 11 workers, the difference of 0.0226 in inequality is on average offset if the larger firm is in a sector with a union density that is approximately 5.7% greater, everything else remaining constant. The inclusion of the sectoral union density does not seem to have any impact on the estimated coefficient for the listed firm dummy.

Coefficients of geographical dummies are omitted due to spatial limitations, but associations are very weak for GE(2). Compared to the baseline province of the Western Cape, only firms in Mpumalanga and Gauteng show significantly different results, with GE(2) falling on average by 0.0269 and 0.0432, respectively. An increase in the share of male workers by 10 percentage points is associated with a *decrease* of 0.0082 in GE(2) across time and 0.013 across firms. We will come back to this result in the next section in light of the comparison with the results for sPX. Finally, R&D does not seem to be associated with important variations on GE(2). These same additional specifications with GE(1) as a dependent variable can be found in Table C1 in Section C of the Appendix.

As mentioned previously, we have also performed the same regressions with the dependent variables weighted by the weights used in calculating the within-sectors-within-firms effects in the decompositions of Section 4.3. Results can be found in Tables C2 and C3 in Section C of the Appendix. Results are mostly comparable, while some divergence appears for the listed dummy and the within-transformed labour productivity, which become significant and positively

related. On the other hand, HHI is no longer significant, as is the case for sector size, union density, the within-transformed share of male workers, and all the province dummies.[9]

We move to the results for the share at the top, starting with the main regression. As mentioned, Table 6 shows the APEs of the independent variables for the main regression, not the estimates for the CRE pooled Tobit regression. The coefficients from the Tobit regressions can be found in Tables C4, C5, and C6 in Section C of the Appendix.

Results are largely in accordance with those found for GE(2). Across quantiles, size shows a significant positive relation with the sPX. For the same firm with 10 workers doubling its size used as an example above, the expected increases in the shares of earnings going to the top are 1.74 percentage points for the top 5%, 0.7 percentage points for the top 0.1%, and 0.096 percentage points for the top 0.01%. For comparisons across firms, the increases are 3.18, 1.46, and 0.21 percentage points, respectively. This pattern of magnitudes becoming less important the higher we move across quantiles is widely found in the results. Compared to GE(2), labour productivity is now found to have a significantly positive impact for both the within-transformed and the between-transformed variables. HHI, however, is no longer significant, except in the between-transformed case for sP95, where it has a negative APE, as for GE(2). The listed dummy reproduces the findings for GE(1), as it was found to have a significant association with lower sP95 and sP99—sP999 being an exception with a non-significant APE for the listed dummy across most specifications.

APEs for the robustness checks are shown in Tables 8, 9, and 10 for each of the quantiles.[10] The corresponding coefficients of the Tobit estimations can be found in Tables C4, C5, and C6 in Section C of the Appendix, as mentioned. Results for union density and R&D largely follow GE(2), except in the case of sP999, for which neither are significant. Two other divergent results in relation to GE(2) appear across quantiles. First, we now see all regions with significant variation from the baseline for sP95 and sP99, all pointing towards firms from the Western Cape being the ones with greater contribution to the top. We also highlight results for the share of male workers for sP95 and sP99, where now, in accordance to priors, a positive and significant APE is found for the within-transformed variable.

As for GE(2), the magnitudes here are less accentuated than for growth in firm size. For the top 5%, for instance, a firm that doubles in size would require a reduction of 77 percentage points in the share of male workers or an increase of 98 percentage points in the sectoral union density to offset the expected greater contribution to top inequality. For the top 1%, the offsetting increase in union density would have to be 66%, although for the share of male workers,

[9] Coefficients equal to zero in Tables C2 and C3 are due to rounding.

[10] Models C1 and C2 are missing for sP999 in Table 10 due to errors in the calculation of APEs, but the corresponding coefficients may be seen in Table C6.

even a reduction to 0 is not expected to balance out the changes in the contribution. These numbers are mentioned here only for the purpose of comparing magnitudes and not as policy suggestions for neutralizing increases in inequality.

For robustness, we run all the regressions above with sPX as a dependent variable using a REWB model instead of the WB pooled Tobit model, with results shown in Tables C7, C8, and C9 in Section C of the Appendix. The results are largely similar, with the main divergences being a significantly positive effect of listed firms, including to sP99, and the share of male workers having a significant and positive coefficient for the between-transformed variable.

Table 7: REWB estimates for GE(2)

Model	Main	UD1	UD2	PR1	PR2	SM1	SM2	RD1	RD2
Log number of workers (W)	0.325***	0.325***	0.325***	0.334***	0.334***	0.318***	0.318***	0.357***	0.357***
	(0.00401)	(0.00401)	(0.00401)	(0.00479)	(0.00479)	(0.00437)	(0.00439)	(0.00823)	(0.00824)
Log labour productivity (W)	-0.00244	-0.00251	-0.00251	-0.00169	-0.00169	-0.00162	-0.00129	-0.0108***	-0.0108***
	(0.00160)	(0.00161)	(0.00161)	(0.00190)	(0.00190)	(0.00165)	(0.00164)	(0.00325)	(0.00325)
Listed firm	-0.00325	-0.00327	-0.00327	0.0637	0.0635	0.0642	0.0603	0.00777	0.00765
	(0.173)	(0.173)	(0.173)	(0.205)	(0.205)	(0.184)	(0.183)	(0.198)	(0.198)
Log HHI (W)	-0.00586***	-0.00585***	-0.00573***	-0.00456***	-0.00456***	-0.00434***	-0.00433***	-0.0107***	-0.0106***
	(0.00134)	(0.00134)	(0.00136)	(0.00152)	(0.00152)	(0.00147)	(0.00147)	(0.00282)	(0.00282)
Log number of workers (B)	0.239***	0.239***	0.239***	0.241***	0.241***	0.236***	0.237***	0.242***	0.242***
	(0.00242)	(0.00242)	(0.00242)	(0.00257)	(0.00258)	(0.00252)	(0.00253)	(0.00384)	(0.00381)
Log labour productivity (B)	0.00801***	0.00794***	0.00798***	0.00846***	0.00822***	0.00560***	0.00671***	-0.00882***	-0.00884***
	(0.00177)	(0.00177)	(0.00177)	(0.00187)	(0.00188)	(0.00185)	(0.00185)	(0.00290)	(0.00290)
Log HHI (B)	-0.00902*	-0.00896*	-0.00743	-0.0138**	-0.0134**	-0.0139***	-0.0124***	-0.0499***	-0.0498***
	(0.00506)	(0.00505)	(0.00508)	(0.00555)	(0.00555)	(0.00479)	(0.00478)	(0.00802)	(0.00803)
Union density (W)			-0.0250						
			(0.0349)						
Union density (B)			-0.395***						
			(0.120)						
Share of male workers (W)							-0.0819***		
							(0.00916)		
Share of male workers (B)							-0.130***		
							(0.00573)		
R&D expenditure dummy (W)									-0.0145
									(0.0150)
R&D expenditure dummy (B)									0.0168
									(0.0353)
Constant	0.0723**	0.0727**	0.101***	0.0752**	0.0777**	0.116***	0.180***	0.448***	0.449***
	(0.0336)	(0.0336)	(0.0345)	(0.0345)	(0.0346)	(0.0334)	(0.0334)	(0.0490)	(0.0490)
Sample subset	Full	Union	Union	Province	Province	Male share	Male share	R&D	R&D
Observations	669,879	668,870	668,870	516,143	516,143	550,597	550,597	289,795	289,795
Number of firms	135,558	135,458	135,458	133,237	133,237	122,177	122,177	89,765	89,765
Province FE	NO	NO	NO	NO	YES	NO	NO	NO	NO
Within R2	0.106	0.106	0.106	0.100	0.100	0.0988	0.0991	0.0796	0.0796
Between R2	0.289	0.289	0.290	0.276	0.276	0.292	0.295	0.266	0.266
Overall R2	0.244	0.244	0.244	0.238	0.238	0.247	0.250	0.204	0.204

Robust standard errors in parentheses. All models include time and sector FE. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Source: author's own calculations.

Table 8: WB pooled Tobit APEs for sP95

Model	Main	UD1	UD2	PR1	PR2	SM1	SM2	RD1	RD2
Log number of workers (W)	0.0353***	0.0353***	0.0353***	0.0336***	0.0333***	0.0357***	0.0355***	0.0197***	0.0197***
	(0.000757)	(0.000758)	(0.000758)	(0.000881)	(0.000877)	(0.000884)	(0.000882)	(0.00126)	(0.00126)
Log labour productivity (W)	0.0258***	0.0258***	0.0258***	0.0235***	0.0233***	0.0265***	0.0263***	0.0200***	0.0200***
	(0.000572)	(0.000573)	(0.000573)	(0.000649)	(0.000643)	(0.000653)	(0.000651)	(0.000819)	(0.000818)
Listed firm	-0.103***	-0.103***	-0.103***	-0.103***	-0.106***	-0.105***	-0.105***	-0.0418**	-0.0438**
	(0.0177)	(0.0177)	(0.0177)	(0.0186)	(0.0184)	(0.0191)	(0.0191)	(0.0180)	(0.0179)
Log HHI (W)	0.000490	0.000484	0.000597	6.09e-05	3.49e-05	0.000784	0.000774	0.00228***	0.00227***
	(0.000483)	(0.000483)	(0.000488)	(0.000531)	(0.000530)	(0.000544)	(0.000544)	(0.000761)	(0.000761)
Log number of workers (B)	0.0617***	0.0618***	0.0618***	0.0612***	0.0602***	0.0635***	0.0635***	0.0456***	0.0450***
	(0.000391)	(0.000391)	(0.000391)	(0.000384)	(0.000385)	(0.000420)	(0.000421)	(0.000707)	(0.000713)
Log labour productivity (B)	0.101***	0.101***	0.101***	0.0989***	0.0956***	0.102***	0.102***	0.0971***	0.0965***
	(0.000717)	(0.000718)	(0.000718)	(0.000736)	(0.000729)	(0.000790)	(0.000791)	(0.00104)	(0.00104)
Log HHI (B)	-0.0109***	-0.0110***	-0.0102***	-0.00809**	-0.00690**	-0.00747**	-0.00750**	-0.0107***	-0.0104**
	(0.00329)	(0.00329)	(0.00329)	(0.00336)	(0.00332)	(0.00322)	(0.00322)	(0.00409)	(0.00408)
Union density (W)			-0.0248**						
			(0.0103)						
Union density (B)			-0.194***						
			(0.0656)						
Share of male workers (W)							0.0318***		
							(0.00329)		
Share of male workers (B)							0.00403*		
							(0.00240)		
R&D expenditure dummy (W)									0.000964
									(0.00199)
R&D expenditure dummy (B)									0.0416***
									(0.00624)
Sample subset	Full	Union	Union	Province	Province	Male share	Male share	R&D	R&D
Observations	669,879	668,870	668,870	516,143	516,143	550,597	550,597	289,795	289,795

Standard errors in parentheses. All models include time and sector FE. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Source: author's own calculations.

Table 9: WB pooled Tobit APEs for sP99

Model	Main	UD1	UD2	PR1	PR2	SM1	SM2	RD1	RD2
Log number of workers (W)	0.0128***	0.0128***	0.0128***	0.0122***	0.0120***	0.0133***	0.0132***	0.0137***	0.0137***
	(0.000386)	(0.000387)	(0.000387)	(0.000464)	(0.000461)	(0.000446)	(0.000446)	(0.000720)	(0.000719)
Log labour productivity (W)	0.00811***	0.00811***	0.00811***	0.00745***	0.00738***	0.00851***	0.00848***	0.00884***	0.00883***
	(0.000310)	(0.000311)	(0.000311)	(0.000371)	(0.000368)	(0.000359)	(0.000359)	(0.000526)	(0.000525)
Listed firm	-0.0263***	-0.0263***	-0.0264***	-0.0258***	-0.0264***	-0.0259***	-0.0259***	-0.0156*	-0.0161*
	(0.00647)	(0.00647)	(0.00648)	(0.00678)	(0.00671)	(0.00701)	(0.00701)	(0.00890)	(0.00889)
Log HHI (W)	8.37e-06	9.81e-06	6.76e-05	-0.000188	-0.000197	-5.64e-05	-5.72e-05	0.000333	0.000328
	(0.000245)	(0.000245)	(0.000247)	(0.000272)	(0.000272)	(0.000276)	(0.000276)	(0.000471)	(0.000471)
Log number of workers (B)	0.0240***	0.0240***	0.0240***	0.0239***	0.0234***	0.0248***	0.0248***	0.0276***	0.0274***
	(0.000225)	(0.000226)	(0.000226)	(0.000224)	(0.000222)	(0.000246)	(0.000246)	(0.000374)	(0.000378)
Log labour productivity (B)	0.0333***	0.0333***	0.0334***	0.0330***	0.0318***	0.0339***	0.0339***	0.0428***	0.0426***
	(0.000403)	(0.000403)	(0.000403)	(0.000412)	(0.000405)	(0.000443)	(0.000444)	(0.000649)	(0.000653)
Log HHI (B)	-0.00248	-0.00248	-0.00224	-0.00133	-0.000876	-0.00186	-0.00187	-0.00575**	-0.00569**
	(0.00172)	(0.00172)	(0.00171)	(0.00178)	(0.00176)	(0.00166)	(0.00166)	(0.00259)	(0.00259)
Union density (W)			-0.0137**						
			(0.00539)						
Union density (B)			-0.0808**						
			(0.0330)						
Share of male workers (W)							0.00759***		
							(0.00170)		
Share of male workers (B)							0.00100		
							(0.00115)		
R&D expenditure dummy (W)									0.000760
									(0.00119)
R&D expenditure dummy (B)									0.0113***
									(0.00336)
Sample subset	Full	Union	Union	Province	Province	Male share	Male share	R&D	R&D
Observations	669,877	668,868	668,868	516,141	516,142	550,595	550,595	289,795	289,795

Standard errors in parentheses. All models include time and sector FE. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Source: author's own calculations.

Table 10: WB pooled Tobit APEs for sP999

Model	Main	UD1	UD2	SM1	SM2	RD1	RD2
Log number of workers (W)	0.00173***	0.00174***	0.00174***	0.00183***	0.00183***	0.00322***	0.00320***
	(0.000124)	(0.000124)	(0.000124)	(0.000135)	(0.000136)	(0.000262)	(0.000262)
Log labour productivity (W)	0.00107***	0.00107***	0.00107***	0.00120***	0.00120***	0.00198***	0.00198***
	(0.000107)	(0.000108)	(0.000108)	(0.000119)	(0.000119)	(0.000222)	(0.000223)
Listed firm	7.29e-05	7.68e-05	7.50e-05	-6.31e-05	-5.76e-05	0.00131	0.00133
	(0.000770)	(0.000770)	(0.000771)	(0.000817)	(0.000816)	(0.00147)	(0.00148)
Log HHI (W)	5.93e-05	5.82e-05	5.74e-05	3.24e-05	3.21e-05	-8.97e-06	-1.40e-05
	(9.55e-05)	(9.56e-05)	(9.61e-05)	(0.000107)	(0.000107)	(0.000197)	(0.000197)
Log number of workers (B)	0.00305***	0.00305***	0.00305***	0.00310***	0.00310***	0.00549***	0.00550***
	(7.78e-05)	(7.79e-05)	(7.79e-05)	(8.56e-05)	(8.56e-05)	(0.000142)	(0.000144)
Log labour productivity (B)	0.00375***	0.00375***	0.00375***	0.00365***	0.00365***	0.00688***	0.00690***
	(0.000126)	(0.000126)	(0.000126)	(0.000136)	(0.000136)	(0.000244)	(0.000245)
Log HHI (B)	-0.00111*	-0.00111*	-0.00110*	-0.00111**	-0.00111**	-0.00145	-0.00144
	(0.000601)	(0.000600)	(0.000596)	(0.000544)	(0.000544)	(0.00104)	(0.00104)
Union density (W)			0.000212				
			(0.00167)				
Union density (B)			-0.00266				
			(0.0110)				
Share of male workers (W)					-0.000128		
					(0.000600)		
Share of male workers (B)					0.000295		
					(0.000335)		
R&D expenditure dummy (W)							0.00123***
							(0.000436)
R&D expenditure dummy (B)							-0.00183**
							(0.000908)
Constant							
Sample subset	Full	Union	Union	Male share	Male share	R&D	R&D
Observations	669,515	668,510	668,510	550,306	550,306	289,706	289,706

Standard errors in parentheses. All models include time and sector FE. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Source: author's own calculations.

5 Discussion

Some points must be kept in mind while interpreting results. While both our two inequality measures focus at the top, they have different characteristics. Generalized entropy is a measure of dispersion of individual earnings within the firm/sector in relation to the firm/sector average, while the share at the top is the contribution that a firm/sector makes to top inequality as a ratio of all the earnings it pays. Their decompositions are importantly different. For generalized entropy, the decomposition is done on levels. Within components measure dispersion within units, and between components measure dispersion between average earnings across units. For the share at the top, the decomposition is done on the variation between two periods. Within components represent changes of unit contributions over time, and between components represent changes in unit sizes over time.

With these points in mind, we start organizing our findings by the analyses of sectoral heterogeneity. Here, we clearly see the role of particular sectors in shaping inequality at the top. There is wide overlap in the lists of the five most unequal sectors for the sectoral measures of sP95, sP99, sP999, sP9999, and GE(2). Particularly for the weighted versions of these variables, the same set of six sectors appears in all top-five lists.

We remind that the economy-level sPX_t is a weighted sum of the sectoral-level measures sPX_{jt}, where weights are the sector share in total earnings paid in that year (cf. Equation 5 in Section B of the Appendix). Similarly, the within component of economy-level GE(2) is a weighted sum of the sectoral-level GE(2), with weights also related to sector size (cf. Equation 2 in Section B of the Appendix). It is therefore no surprise that top sectors in the weighted sectoral measures are a mix of sectors that are among the largest (*Wholesale and retail trade*; *Manufacturing*) and sectors that are among the most unequal according to the unweighted sectoral-level measures (*Professional, scientific, and technical activities*; *Information and communication*; and *Administrative and support service activities*). In particular, all lists in the weighted measures are led by *Financial and insurance activities*, which is the third-largest sector in earnings paid and the most unequal sector for all non-weighted sPX (and in third place for GE(2)).

There are some divergences in sPX and GE(2), particularly for *Information and communication* (among the top sectors for sPX, among the bottom sectors for GE) and *Agriculture, forestry, and fishing* (vice versa). We interpret this result in light of the difference between sPX and GE(2) alluded to in the beginning of this section. It suggests that agriculture, for instance, is extremely unequal in its internal earnings distribution, particularly with an important long tail. This tail, however, is not long enough so that these high-earning individuals relative to the sector lie at the top quantiles of the overall income distribution. The opposite seems valid for ICT: it creates an important amount of super-earnings but with an internal earnings distribution that is relatively egalitarian.

The weighted measures are the most precise ones to characterize how sectoral heterogeneity shapes economy-wide inequality measures in South Africa in our time range. The unweighted measures, on the other hand, are potentially useful from a policy-making perspective. They signal, to some extent, what could happen with inequality if certain sectors are prioritized in policies that support structural change. For instance, our findings suggest that a development strategy that fosters the financial sector could be one that leads to an even greater concentration among top earners. This is only a rough signal, as we do not have strong grounds to claim what could happen dynamically with sectoral-level inequality as sectors grow and shrink—although, we do know from the regression results discussed below, that, at least at the firm level, unit size is related to even greater unit-level inequality.

Despite the clear presence of sectoral heterogeneity, decompositions demonstrate that top inequality in our data is composed significantly more by variations within firms than those across firms or across sectors. This is in line with what is found for other countries, although the degree to which the effect is concentrated within firms is greater than elsewhere (cf. Table 1). This is particularly relevant considering that the case in point is a middle-income country. One of the main characteristics of the economic structure of these countries is duality, with high heterogeneity and a polarization between dynamic and backwards sectors and between highly productive and laggard firms. What the decomposition shows is that, despite this duality, within-firm variation still is significantly more important for top inequality, both on levels and over time.

It is important to note that our data only cover the formal South African economy. While informality is knowingly relatively low in South Africa, and a mirror image of its high unemployment levels (Kingdon and Knight 2004), it should be expected that the formal economy is less polarized than the overall economy, even for a middle-income country. This might explain why the within component is larger than we could expect. Further analyses exploring the extent to which informality divides explain that structural duality in South Africa would greatly complement our findings.

Having established that, despite the evident sectoral heterogeneity, it is the within-firms component that predominates, we finalize with the results from regressions that explore patterns between firm-level inequality measures and firm- and sector-level independent variables. We find that firm size, measured as the number of workers, is widely correlated with greater firm-level top inequality. The same goes, with slightly lower significance, for labour productivity. Listed firms are associated with smaller shares at the top, although not with smaller GE(2). Sectoral HHI has a surprisingly negative relation with GE(2), although not a significant one with sPX. Further exploration would be needed to understand this latter result, but it is possible that other sectoral aspects are being captured by this variable.

A series of independent variables are added in robustness checks. As expected, sectoral union density shows significant and negative effects on top income inequality. In geographical terms, the Western Cape is found to have more unequal firms on average than Mpumalanga and Gauteng for GE(2), while for the share at the top, this extends to all other provinces as well. The interpretation of these results could profit from an analysis of the geographical concentration of firms in sectors that contribute the most to inequality. R&D-related variables have coefficients that are consistently non-significant across all estimated models.

The share of male workers in the firm has a positive association with inequality for top shares but a negative one for GE(2). The distribution of earnings per gender shows a clear gender gap (cf. Figure 3 in Section C of the Appendix), in line with the literature (Pleace et al. 2023), and the result for the share at the top also reproduces other findings for glass ceilings in South Africa (Mosomi 2019). What the GE(2) result suggests is that firms with fewer men are more polarized—it could be the case that the few men are typically higher-earning and lying at the tail of the distribution of earnings within the firm, while the majority of (female) workers are biased towards lower pay. The GE(2) male share parameters are more negative than the ones for GE(1), providing evidence in favour of this hypothesis.

The regressions with unweighted firm-level inequality as the dependent variable do not map directly to inequality at the economy level, given that the latter is a weighted sum of the former. To address this point, we ran the GE regressions with the weighted firm-level GE as the dependent variable. We did not perform this exercise with sPX because the decomposition of sPX is done on time differences. The story on weighted GE is mainly one of size and productivity, whose robust results remain the same—the latter with even stronger significance. Coefficients for several other independent variables become non-significant, such as for all sectoral variables, for all geographical dummies, and for the within-transformed share of male workers. Also, the coefficient of the listed firm dummy becomes significant but with a positive sign, unlike for sPX.

'Structure' is a polysemic word in the structural change literature (Silva and Teixeira 2008). In our paper, we cover structure both in the sense of firm and sector characteristics—such as for firm size or sectoral concentration—and in the sense of institutions—such as for union density or the listed firm dummy. Even if our results for the number of workers and labour productivity are stronger in the sense of being more consistent across sectors and with more important magnitudes, we are not at all able to disregard the role of the institutional aspects of economic structure. This is relevant since sectoral and institutional drivers are also sometimes framed as competing explanations for the rise in top incomes (e.g., Kaplan and Rauh 2013). Our results suggest that they are complementary rather than competing explanations. This is compatible with the literature, as seen in Section 2, since each of these groups of drivers are understood to affect top inequality through different mechanisms. The discussion on financialization is

the best illustration of this, since the financial sector is understood to act through both sets of mechanisms.

6 Conclusion

In this paper, we investigated the behaviour of top earnings in the formal sector of South Africa through a sectoral and firm-level perspective. We have shown that, despite the clear presence of sectoral heterogeneity, the decomposition of economy-level top earnings inequality measures in between-/within-groups shows a striking dominance of the within-sectors-within-firms component. This motivated further investigation into patterns between firm-level inequality and firm- and sector-level characteristics, which has shown that larger and more productive firms are associated with greater inequality, as well as those with a larger share of male workers. However, firms listed in stock exchanges and those in sectors with greater union density exhibited lower values of within-firm inequality.

Our findings on sectoral heterogeneity suggest that sectors that are crucial for structural change in middle-income countries, such as finance or scientific activities, are also large contributors to top inequality. Policymakers interested in inclusive processes of changes in the economic structure must be mindful not to worsen concentration at the top when fostering such sectors. Since findings at the firm level suggest that concentration at the top is related to large, productive firms, it is possible that policies aimed at mitigating firm-level duality could counteract concentration, such as by improving the competitiveness of small and medium enterprises.

Additionally, we have seen that the largest component of inequality occurs within firms. One important limitation of this paper is the lack of occupational data, which unfortunately makes it difficult to further investigate what is happening with different workers in a firm. However, results suggest that the institutional aspect could offer a relevant path to reduce top inequality. Policies aiming at reducing gender gaps or at improving the negotiation conditions of workers could counteract concentration forces coming from processes of structural change to some extent, although it does not seem like they would be able to fully counteract the effects coming for firm and sector characteristics.

Beyond our findings, we contribute to the literature on top earnings on the methodological aspect by proposing a procedure for decomposing top shares in between- and within-groups components. This is useful for approaching the issue from a sectoral and firm-level perspective, which we hope to have convinced the reader is an important angle to investigate on the issue of top inequality, and in proposing policies that help reduce concentration.

References

Alarco Tosoni, G. (2022). 'Participación Salarial y Heterogeneidad Estructural en Perú: Diagnóstico y Simulaciones'. *Problemas del Desarrollo. Revista Latinoamericana de Economía*, 53(208): 3–30. **https://doi.org/10.22201/iiec.20078951e.2022.208.69761**

Alvaredo, F., and Atkinson, A. B. (2022). 'Top Incomes in South Africa in the Twentieth Century'. *Cliometrica*, 16(3): 477–546. **https://doi.org/10.1007/s11698-021-00235-4**

Atkinson, A. B., Piketty, T., and Saez, E. (2011, March). 'Top Incomes in the Long Run of History'. *Journal of Economic Literature*, 49(1): 3–71. **https://doi.org/10.1257/jel.49.1.3**

Bartels, C., and Waldenström, D. (2021). 'Inequality and Top Incomes'. GLO Discussion Paper Series No. 959. Essen: Global Labor Organization (GLO). **https://doi.org/10.1007/978-3-319-57365-6_169-1**

Bassier, I., and Woolard, I. (2021). 'Exclusive Growth? Rapidly Increasing Top Incomes Amid Low National Growth in South Africa'. *South African Journal of Economics*, 89(2): 246–73. **https://doi.org/10.1111/saje.12274**

Beeson, P., and Tannery, F. (2004). 'The Impact of Industrial Restructuring on Earnings Inequality: The Decline of Steel and Earnings in Pittsburgh'. *Growth and Change*, 35(1): 21–41. **https://doi.org/10.1111/j.0017-4815.2004.00236.x**

Bell, A., Fairbrother, M., and Jones, K. (2019). 'Fixed and Random Effects Models: Making an Informed Choice'. *Quality & Quantity*, 53(2): 1051–74. **https://doi.org/10.1007/s11135-018-0802-x**

Bell, A., and Jones, K. (2015). 'Explaining Fixed Effects: Random Effects Modeling of Time-Series Cross-Sectional and Panel Data'. *Political Science Research and Methods*, 3(1): 133–53. **https://doi.org/10.1017/psrm.2014.7**

Beqiraj, E., Fanti, L., and Zamparelli, L. (2019). 'Sectoral Composition of Output and the Wage Share: The Role of the Service Sector'. *Structural Change and Economic Dynamics*, 51(-): 1–10. **https://doi.org/10.1016/j.strueco.2019.06.009**

Bhorat, H., Oosthuizen, M., Lilenstein, K., and Steenkamp, F. (2017). 'Firm-Level Determinants of Earnings in the Formal Sector of the South African Labour Market'. WIDER Working Paper No. 2017/25. Helsinki: UNU-WIDER. **https://doi.org/10.35188/UNU-WIDER/2017/249-6**

Budlender, J., and Ebrahim, A. (2020). 'Industry Classification in the South African Tax Microdata'. WIDER Working Paper No. 2020/99. Helsinki: UNU-WIDER. **https://doi.org/10.35188/UNU-WIDER/2020/856-6**

Carvalhaes, F. A. D. O., Barbosa, R. J., Souza, P. H. G. F. D., and Ribeiro, C. A. C. (2014). 'Os Impactos da Geração de Empregos Sobre as Desigualdades de Renda uma Análise da Década de 2000'. *Revista Brasileira de Ciências Sociais*, 29(85): 79–98. **https://doi.org/10.1590/S0102-69092014000200006**

Castellano, R., Musella, G., and Punzo, G. (2017). 'Structure of the Labour Market and Wage Inequality: Evidence from European Countries'. *Quality & Quantity*, 51(5): 2191–218. **https://doi.org/10.1007/s11135-016-0381-7**

Castellano, R., Musella, G., and Punzo, G. (2019). 'Exploring Changes in the Employment Structure and Wage Inequality in Western Europe Using the Unconditional Quantile Regression'. *Empirica*,

46(2): 249–304. https://doi.org/10.1007/s10663-017-9397-z

Castellano, R., Musella, G., and Punzo, G. (2021). 'Wage Dynamics in Light of the Structural Changes in the Labour Market Across Four More Economically Developed Countries of Europe'. *Review of Social Economy*, 79(2): 222–60. https://doi.org/10.1080/00346764.2019.1655163

Changyuan, L., and Jun, Z. (2009). 'Labor Income Shares (LIS) in Economic Development: An Empirical Study Based on China's Sectoral-level Data'. *Social Sciences in China*, 30(4): 154–78. https://doi.org/10.1080/02529200903342750

de la Vega, R. (2023). 'Structural Change and Income Inequality: A Meta-Analysis'. MERIT Working Papers No. 2023-046. Maastricht: United Nations University - Maastricht Economic and Social Research Institute on Innovation and Technology (MERIT).

de Serres, A., Scarpetta, S., and de la Maisonneuve, C. (2001). 'Falling Wage Shares in Europe and the United States: How Important is Aggregation Bias?' *Empirica*, 28(-): 375–401. https://doi.org/10.1023/A:1013922621303

de Souza, P. H. G. F., Maciel, F. B., and Foguel, M. N. (2023). 'Desigualdade Salarial no Setor Formal da Economia Brasileira : A Importância dos Componentes Intrafirma, Entre Firmas e Entre Setores'. Nota Técnica. Brasília: Ipea. https://doi.org/10.38116/bmt76/nt1

Dweck, E., Baltar, C. T., Marcato, M. B., and Krepsky, C. U. (2024). 'Labor Market, Distributive Gains and Cumulative Causation: Insights from the Brazilian Economy'. *Review of Political Economy*, 36(1): 325–50. https://doi.org/10.1080/09538259.2022.2041284

Ebrahim, A., and Axelson, C. (2019). 'The Creation of an Individual Panel Using Administrative Tax Microdata in South Africa'. WIDER Working Paper No. 2019/27. Helsinki: UNU-WIDER. https://doi.org/10.35188/UNU-WIDER/2019/661-6

Ebrahim, A., Kreuser, C. F., and Kilumelume, M. (2021). 'The Guide to the CIT-IRP5 Panel Version 4.0' Vol. 2021. WIDER Working Paper No. 2021/173. Helsinki: UNU-WIDER. https://doi.org/10.35188/UNU-WIDER/2021/113-6

Elliott, R., and Murphy, P. (1990). 'Industry Skill Differentials and the Impact of Changing Industry Structure on Aggregate Skill Differentials in Britain 1970-1982'. *Journal of Economic Studies*, 17(1): 01443589010137059. https://doi.org/10.1108/01443589010137059

Firpo, S. P., Fortin, N. M., and Lemieux, T. (2018). 'Decomposing Wage Distributions Using Recentered Influence Function Regressions'. *Econometrics*, 6(2): 28. https://doi.org/10.3390/econometrics6020028

Frank, R. H., and Cook, P. J. (1996). *The Winner-Take-All Society: Why the Few at the Top Get so Much More than the Rest of Us*. New York, NY: Penguin Books.

Goos, M., and Manning, A. (2007). 'Lousy and Lovely Jobs: The Rising Polarization of Work in Britain'. *Review of Economics and Statistics*, 89(1): 118–33. https://doi.org/10.1162/rest.89.1.118

Goos, M., Manning, A., and Salomons, A. (2014, August). 'Explaining Job Polarization: Routine-Biased Technological Change and Offshoring'. *American Economic Review*, 104(8): 2509–26. https://doi.org/10.1257/aer.104.8.2509

Hager, S. B. (2021). 'Varieties of Top Incomes?' *Socio-Economic Review*, 18(4): 1175–98. https://doi.org/10.1093/ser/mwy036

Haltiwanger, J., Hyatt, H., and Spletzer, J. (2022). *Industries, Mega Firms, and Increasing Inequality*

(SSRN Scholarly Paper No. 4080660). Rochester, NY. **https://doi.org/10.3386/w29920**

Hein, E. (2015). 'Finance-Dominated Capitalism and Re-Distribution of Income: A Kaleckian Perspective'. *Cambridge Journal of Economics*, 39(3): 907–34. **https://doi.org/10.1093/cje/bet038**

Hinojosa, P. G. (2021). 'Skill Prices and Compositional Effects on the Declining Wage Inequality in Latin America: Evidence from Brazil'. *Revista Brasileira de Economia*, 75(2): -. **https://doi.org/10.5935/0034-7140.20210010**

Ibarra, C. A., and Ros, J. (2019). 'The Decline of the Labor Income Share in Mexico, 1990–2015'. *World Development*, 122(-): 570–84. **https://doi.org/10.1016/j.worlddev.2019.06.014**

Jacobs, C., Ebrahim, A., Leibbrandt, M., Pirttilä, J., and Piek, M. (2024). 'Income Inequality in South Africa: Evidence from Individual-Level Administrative Tax Data'. Working Paper. Helsinki: UNU-WIDER. **https://doi.org/10.35188/UNU-WIDER/2024/517-2**

Kanbur, R., and Zhuang, J. (2013). 'Urbanization and Inequality in Asia'. *Asian Development Review*, 30(1): 131–47. **https://doi.org/10.1162/ADEV_a_00006**

Kaplan, S. N., and Rauh, J. (2013). 'It's the Market: The Broad-Based Rise in the Return to Top Talent'. *Journal of Economic Perspectives*, 27(3): 35–56. **https://doi.org/10.1257/jep.27.3.35**

Kerr, A. (2020). 'Earnings in the South African Revenue Service IRP5 data'. WIDER Working Paper No. 2020/62. Helsinki: UNU-WIDER. **https://doi.org/10.35188/UNU-WIDER/2020/819-1**

Kerr, A., Lam, D., and Wittenberg, M. (2019). *Post Apartheid Labour Market Series 1993-2019.* DataFirst. **https://doi.org/10.25828/GTR1-8R20**

Kim, C., and Sakamoto, A. (2008). 'The Rise of Intra-Occupational Wage Inequality in the United States, 1983 to 2002'. *American Sociological Review*, 73(1): 129–57. **https://doi.org/10.1177/000312240807300107**

Kingdon, G. G., and Knight, J. (2004). 'Unemployment in South Africa: The Nature of the Beast'. *World Development*, 32(3): 391–408. **https://doi.org/10.1016/j.worlddev.2003.10.005**

Kónya, I., Krekó, J., and Oblath, G. (2020). 'Labor Shares in the Old and New EU Member States - Sectoral Effects and the Role of Relative Prices'. *Economic Modelling*, 90(-): 254–72. **https://doi.org/10.1016/j.econmod.2020.05.010**

Lee, S. (2017). 'International Trade and Within-Sector Wage Inequality: The Case of South Korea'. *Journal of Asian Economics*, 48(-): 38–47. **https://doi.org/10.1016/j.asieco.2016.11.001**

Lopes, J. C., Coelho, J. C., and Escária, V. (2021). 'Labour Productivity, Wages and the Functional Distribution of Income in Portugal: A Sectoral Approach'. *Society and Economy*, 43(4): 331–54. **https://doi.org/10.1556/204.2021.00013**

Mahutga, M. C., and Curran, M. (2022). 'Micro-Mechanisms and Macro-Effects: How Structural Change and Institutional Context Affect Income Inequality in Rich Democracies'. *Socius: Sociological Research for a Dynamic World*, 8(-): 237802312211245. **https://doi.org/10.1177/23780231221124581**

Maia, A. G. (2013). 'Estrutura de Ocupações e Distribuição de Rendimentos: Uma Análise da Experiência Brasileira nos Anos 2000'. *Revista de Economia Contemporânea*, 17(2): 276–301. **https://doi.org/10.1590/S1415-98482013000200004**

Maia, A. G., Sakamoto, A., and Wang, S. X. (2019). 'How Employment Shapes Income Inequality: A Comparison between Brazil and the U.S.' *Revista de Economia Contemporânea*, 23(3):

e192331. https://doi.org/10.1590/198055272331

Malkina, M. (2019). 'Spatial Wage Inequality and Its Sectoral Determinants: The Case of Modern Russia'. *Oeconomia Copernicana*, 10(1): 69–87. https://doi.org/10.24136/oc.2019.004

Martorano, B., Park, D., and Sanfilippo, M. (2016). 'Catching-Up, Structural Transformation, and Inequality: Industry-Level Evidence from Asia'. *Industrial and Corporate Change*, 26(4): 555–70. https://doi.org/10.1093/icc/dtw039

Medeiros, M., and Souza, P. H. F. (2015). 'The Rich, the Affluent and the Top Incomes'. *Current Sociology*, 63(6): 869–95. https://doi.org/10.1177/0011392114551651

Mendieta-Muñoz, I., Rada, C., and Von Arnim, R. (2021). 'The Decline of the US Labor Share Across Sectors'. *Review of Income and Wealth*, 67(3): 732–58. https://doi.org/10.1111/roiw.12487

Milanović, B. (2016). *Global Inequality: A New Approach for the Age of Globalization*. Cambridge, MA: The Belknap Press of Harvard University Press.

Molinder, J. (2019). 'Wage Differentials, Economic Restructuring and the Solidaristic Wage Policy in Sweden'. *European Review of Economic History*, 23(1): 97–121. https://doi.org/10.1093/ereh/hey005

Mosomi, J. (2019). 'Distributional Changes in the Gender Wage Gap in the Post-Apartheid South African Labour Market'. WIDER Working Paper No. 2019/17. Helsinki: UNU-WIDER. https://doi.org/10.35188/UNU-WIDER/2019/651-7

Mouw, T., and Kalleberg, A. L. (2010). 'Occupations and the Structure of Wage Inequality in the United States, 1980s to 2000s'. *American Sociological Review*, 75(3): 402–31. https://doi.org/10.1177/0003122410363564

Mundlak, Y. (1978). 'On the Pooling of Time Series and Cross Section Data'. *Econometrica*, 46(1): 69. https://doi.org/10.2307/1913646

National Treasury and UNU-WIDER. (2021a). 'CIT-IRP5 Firm-Level Panel 2008–2018 [dataset]'. Version 4.0. Pretoria: South African Revenue Service [producer of the original data], 2019. Pretoria: National Treasury and UNU-WIDER [producer and distributor of the harmonized dataset], 2021.

National Treasury and UNU-WIDER. (2021b). 'Individual Panel 2011–2018 [dataset]'. Version 2019_2. Pretoria: South African Revenue Service [producer of the original data], 2019. Pretoria: National Treasury and UNU-WIDER [producer and distributor of the harmonized dataset], 2021.

Neves Costa, R., and Pérez-Duarte, S. (2019). 'Not All Inequality Measures Were Created Equal - The Measurement of Wealth Inequality, Its Decompositions, and an Application to European Household Wealth'. Statistics Paper Series No. 31. Frankfurt: European Central Bank.

Pieterse, D., Gavin, E., and Kreuser, C. F. (2018). 'Introduction to the South African Revenue Service and National Treasury Firm-Level Panel'. *South African Journal of Economics*, 86(-): 6–39. https://doi.org/10.1111/saje.12156

Piketty, T. (2013). *Le Capital Au XXIe Siècle*. Paris: Éditions du Seuil.

Piketty, T., and Saez, E. (2013). 'Top Incomes and the Great Recession: Recent Evolutions and Policy Implications'. *IMF Economic Review*, 61(3): 456–78. https://doi.org/10.1057/imfer.2013.14

Piketty, T., Saez, E., and Stantcheva, S. (2014). 'Optimal Taxation of Top Labor Incomes: A Tale of Three Elasticities'. *American Economic Journal: Economic Policy*, 6(1): 230–71. https://doi.org/10.1257/pol.6.1.230

Pleace, M., Clance, M., and Nicholls, N. (2023). 'The Gender Wage Gap in South Africa: Insights from Adminsitrative Tax Data'. SA-TIED Working Paper No. 219. Helsinki: SA-TIED.

Popli, G. K. (2010). 'Trade Liberalization and the Self-Employed in Mexico'. *World Development*, 38(6): 803–13. **https://doi.org/10.1016/j.worlddev.2010.02.016**

Prasad, E. S. (2002). 'Wage Inequality in the United Kingdom, 1975-99'. *IMF Staff Papers*, 49(3): 339–63. **https://doi.org/10.2307/3872501**

Ravallion, M. (2022). 'Missing Top Income Recipients'. *The Journal of Economic Inequality*, 20(1): 205–22. **https://doi.org/10.1007/s10888-022-09530-0**

Roine, J., and Waldenström, D. (2015). 'Long-Run Trends in the Distribution of Income and Wealth'. In A. B. Atkinson and F. Bourguignon (eds), *Handbook of Income Distribution* (Vol. 2, pp. 469–592). Elsevier. **https://doi.org/10.1016/B978-0-444-59428-0.00008-4**

Rosen, S. (1981). 'The Economics of Superstars'. *American Economic Review*, 71(5): 845–58.

Shorrocks, A. F. (2013). 'Decomposition Procedures for Distributional Analysis: A Unified Framework Based on the Shapley Value'. *The Journal of Economic Inequality*, 11(1): 99–126. **https://doi.org/10.1007/s10888-011-9214-z**

Silva, E. G., and Teixeira, A. A. C. (2008). 'Surveying Structural Change: Seminal Contributions and a Bibliometric Account'. *Structural Change and Economic Dynamics*, 19(4): 273–300. **https://doi.org/10.1016/j.strueco.2008.02.001**

Sologon, D. M., Van Kerm, P., Li, J., and O'Donoghue, C. (2021). 'Accounting for Differences in Income Inequality across Countries: Tax-Benefit Policy, Labour Market Structure, Returns and Demographics'. *The Journal of Economic Inequality*, 19(1): 13–43. **https://doi.org/10.1007/s10888-020-09454-7**

Statistics South Africa. (2012). *Standard Industrial Classification of All Economic Activities (SIC) Seventh Edition*. Pretoria: Statistics South Africa.

Suen, W. (1995). 'Sectoral Shifts Impact on Hong Kong Workers'. *The Journal of International Trade & Economic Development*, 4(2): 135–52. **https://doi.org/10.1080/09638199500000013**

Williams, M. (2013). 'Occupations and British Wage Inequality, 1970s-2000s'. *European Sociological Review*, 29(4): 841–57. **https://doi.org/10.1093/esr/jcs063**

Wooldridge, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data* (2nd ed ed.). Cambridge, MA: MIT Press.

Wooldridge, J. M. (2019). *Introductory Econometrics: A Modern Approach* (Seventh edition ed.). Boston, MA: Cengage Learning.

Zhang, Z., and Wu, X. (2017). 'Occupational Segregation and Earnings Inequality: Rural Migrants and Local Workers in Urban China'. *Social Science Research*, 61(-): 57–74. **https://doi.org/10.1016/j.ssresearch.2016.06.020**

Appendix

A Data appendix

This section of the Appendix shows further details on the use of data, beyond those already explained in Section 3. We use mainly the firm-level CIT-IRP5 panel version 4.0 (National Treasury and UNU-WIDER 2021a) and the individual panel[11] version 19_2 (National Treasury and UNU-WIDER 2021b), which were accessed at the NT-SDF from 03/06/2024 to 28/06/2024; from 26/08/2024 to 13/09/2024; and from 20/11/2024 to 04/12/2024. We also use data on whether the firm is listed or not from version 5.0 of the firm-level panel. Access was provided under a non-disclosure agreement, and the output was checked so that no firm or individual would be compromised. Results do not represent any official statistics (NT or SARS), and views expressed in the research are not necessarily the views of the National Treasury or SARS. Python was used for data cleaning and merging, and most of the analyses and images; while the regressions were run on Stata. Codes may be made available upon request.

Firms are identified as CIT entities (`taxrefno` or `tax_reference_number_cit`) and their sector (SIC7 Section) is identified by `imp_mic_sic7_1d`, following Budlender and Ebrahim (2020). In the first data cleaning step, dormant firms and those without data on sector were removed. Sectoral-level data, such as sector size and the Herfindahl-Hirschman Index were calculated at this point, based on the sum of workers in the individual dataset related to each firm in a given year (`num_workers`), and before further removal of firms, as explained below.

Individual data was first loaded from the source of income panel. Duplicates with the same `id_d`, `irp5_id`, `source_code`, and `amount` were removed. Codes were then filtered to include only wages and remuneration. We adopted all codes listed in the appendix of Kerr (2020), including 3615 “Director’s remuneration" and 3616 “Independent contractor’s income". The former was included because it is relevant to understand earnings at top quantiles. The latter was included following inspection of the data: some observations were identified in which individuals seemed to drop out of the data from one year to the other, but this was driven by their incomes being turned into 3616 incomes. Data on all sources of income were summed at the `irp5_id` level. The full list of codes is shown in Table A1. Earnings and all relevant firm-level variables were deflated with the economy-wide GDP deflator available in the dataset (`defl_gdp_economywide`), which has 2012 Q1 as the baseline. The deflator is made available as a quarterly price index, which we aggregate via the mean to an yearly measure.

[11] An issue with the 2023 version in the moment of the first visit led to the choice of using an older version of the dataset. We hope in the future to use the most recent version of the individual panel, or to build directly from the IRP5 data.

The employment panel was then loaded to include the firm identifier (tax_reference_number_cit, taxrefno) and gender (gender) in the individual data by merging on irp5_id. An unexpected problem was found: a significant amount of observations in the employment panel (c. 9 per cent of observations in each year) had a duplicated irp5_id, both with the same paye_number but each related to a different CIT identifier (taxrefno). The same taxrefno may be associated with multiple paye_number codes (e.g., for different branches of the same firm) but the opposite should not occur. In most cases (c. 66 per cent), one of the duplicates had missing data on the sector. These would be dropped anyway because firms without sectors had been dropped from the firm-level data – so, only the observation with sectoral data was kept in these cases. For the c. 7 per cent of cases in which both duplicates had missing data on the sector, both observations were dropped. For the remaining ones (which added up to c. 200,000 unique IRP5 IDs, out of the c. 16,000,000 in a given year), the first observation among the duplicates was arbitrarily picked. Furthermore, only observations related with workers in firms present in the filtered firm-level data were kept. On gender, a variable was created with the share of workers per firm with missing data, to avoid including in regressions firms with too imprecise measures of the share of workers per gender.

At this point, the individual data was indexed at the irp5_id level, rather than the id_d level. Aggregating at the latter is needed to keep the panel structure of the data at the worker level. In some cases, a single individual (id_d) had earnings coming from different IRP5 forms but related to the same company. In these cases, earnings were added together. However, when earnings were received from different firms, only the main source of earnings was kept. Earnings coming from ITR12 forms were also dropped, as they could have been received from economic activities in different sectors.

Explorations of the data, particularly of the aggregated inequality measures, showed unusual patterns likely driven by outlying observations. Three further steps were taken to deal with outliers. First, all earnings within a firm k in year t were added up and compared with the value of the firm's sales (g_sales) and labour cost (x_labcost) in that year. If the ratio was above 100 or below 0.01, the firm was dropped in that year. Additionally, many workers would show individual deviations in a given year from their overall trend in ways that did not seem reasonable. In some cases, their earnings would drop to 1 or a very low value. To deal with these, also following Kerr (2020), all earnings below the mandatory reporting threshold of ZAR 2,000 were removed. In other cases, earnings of a given individual would spike in a given year only to come back to the usual trend in the remaining years. We then calculated the ratio of each earning of worker i in time t over the median earning of this same worker over time. Whenever this ratio was higher than 5, that individual observation was dropped.

Finally, firms with a single worker in a given year had their values for generalized entropy set to missing in that year. With a single individual, generalized entropy reaches its lower bound of

zero. This value normally represents perfect equality but this is not meaningful as a measure of earnings dispersion in a group composed of only one single individual.

External data was brought in from the Quarterly Labour Force Surveys (QLFS) to build the union density (`uniondens`) variable. We used the 3.3.1 version of the Post Apartheid Labour Market Series (PALMS) data made available by DataFirst (Kerr et al. 2019) and downloaded on 04/11/2024. We used the new weights produced ahead of the v.3.4 release of the dataset (Kerr et al. 2019). We cleaned the data by removing informal and self-employed individuals, plus those with missing data on unionization. Sectoral data in PALMS/QLFS is made available at the 3-digit level of SIC5, which we needed to convert to 1-digit SIC7 sectors to match the data in our main data source. We used a correspondence table made available in the online documentation of Budlender and Ebrahim (2020), despite the caveat from authors that it should be used at researchers' own risk[12]. Each row in the correspondence table matches one 5-digit SIC5 code to one 5-digit level SIC7 code, and lists the corresponding codes for both systems at other levels. To translate these into a correspondence from 3-digit SIC5 to 1-digit SIC7, we counted the number of rows for each combination of 3-digit SIC5 and 1-digit SIC7 found in the correspondence table. For each 3-digit SIC5 code, we selected the most prevalent 1-digit SIC7 code according to this count, and used that correspondence to convert the sectoral data from PALMS/QLFS to 1-digit SIC7 codes. To aggregate the data by year, we calculated the total number of workers and the number of unionized workers per sector for each quarterly wave. We then took the average of both values separately over the four quarters of each year, and calculated the union density for a sector-year by the ratio of these averages.

Table A1: Income source codes used in the analysis

Code	Description
3601	Income – PAYE
3605	Annual Payment – PAYE
3606	Commission – PAYE
3607	Overtime – PAYE
3615	Director's remuneration
3616	Independent contractors' income
3701	Travel allowance – PAYE
3702	Reimbursed travel allowance – PAYE
3703	Reimbursed travel allowance – IT
3704	Subsistence allowance local travel – IT
3707	Share option exercised – PAYE
3708	Public office allowance – PAYE
3710	Uniform allowance
3711	Tool allowance
3712	Acting allowance

[12] The file used was "SIC edition_5 and SIC edition_7 correspondence table V1.00-1.xls", downloaded from https://github.com/jbudlender/IndustryClassification on 22/11/2024

Table A1: Income source codes used in the analysis

Code	Description
3713	Phone allowance
3714	Other allowances – PAYE
3715	Other allowances – Excl
3716	Subsistence allowance foreign travel – IT
3717	Broad-based employee share plan – PAYE
3718	Other benefits – PAYE
3719	Other benefits – Excl
3751	Foreign bursaries – foreign service income
3752	Reimbursed travel allowance – foreign service income
3753	Travel allowance – foreign service income
3754	Share option exercised – foreign service income
3755	Other allowances – foreign service income
3760	Other non-taxable allowances – foreign service income
3764	BBE share plan – foreign service income
3768	Vesting of equity instruments – foreign service income
3769	Other fringe benefits – PAYE
3801	General fringe benefits – PAYE
3802	Use of motor vehicle acquisition by employer, not lease – PAYE
3803	Use of asset – PAYE
3805	Meals, etc. – PAYE
3808	Accommodation – PAYE
3813	Employee's debt – PAYE
3816	Taxable bursaries or scholarships – PAYE
3818	Use of motor vehicle acquisition by employer by lease – PAYE
3810	Medical aid contributions – PAYE
3813	Medical services costs – PAYE
3814	Non-taxable benefit in respect of NSF pension benefits paid by the employer
3815	Non-taxable bursaries or scholarships – Excl
3816	Use of motor vehicle acquisition by employer by lease – PAYE
3820	Taxable bursaries or scholarships (FE) – PAYE
3821	Non-taxable bursaries or scholarships (FE) – Excl
3852	Use of motor vehicle acquisition by employer, not lease – foreign income
3855	Foreign accommodation
3858	Foreign employee's debt
3860	Foreign other allowances
3861	Medical aid contributions – foreign service income
3863	Medical service costs (PAYE) – foreign service income
3864	Medical service costs – foreign service income
3870	Foreign fringe benefits – PAYE
3880	Non-taxable bursaries or scholarships – foreign services

Source: adapted from Kerr (2020).

B Mathematical appendix

This section of the Appendix shows the full equations for the measures of inequality and their decompositions in within- and between-groups components.

B1 Generalized entropy

For simplicity in the notation, this section omits the time subscript, but all equations here apply to individual incomes coming from a single year, and all measures are calculated separately for each year. Entropy indices are such that for n individuals with earnings w_i, $i = 1, \ldots, n$, and mean earnings μ:

$$GE(\theta) = \frac{1}{n} \sum_{i=1}^{n} \varphi_\theta \left(\frac{w_i}{\mu} \right)$$

Where the function $\varphi_\theta(x)$ is $\varphi_\theta(x) = \frac{(x^\theta - 1)}{\theta(\theta - 1)}$ (Shorrocks 2013)[13]. The greater the parameter θ, the more sensitive $GE(\theta)$ is to changes at the top of the distribution. Typical values for θ are 0, 1, and 2.

Unit contributions are calculated similarly, for wages within the unit. For sector j and for firm k:

$$GE(\theta)_j = \left(\frac{1}{n_j} \right) \sum_{i \in j} \varphi_\theta \left(\frac{w_i}{\mu_j} \right) \qquad GE(\theta)_k = \left(\frac{1}{n_k} \right) \sum_{i \in k} \varphi_\theta \left(\frac{w_i}{\mu_k} \right)$$

For the sectoral measure, the summation is only done for workers i in sector j, μ_j is the average earnings for the sector, and n_j is the number of workers in the sector. The same applies at the firm level, with μ_k and n_k as the respective counterparts of μ_j and n_j.

The main advantage of entropy indices is that they are additively decomposable in group components. We can decompose $GE(\theta)$ in between-and within-sector components for sectors $j = 1, \ldots, m$ as:

$$\begin{aligned} GE(\theta) &= B && + W \\ GE(\theta) &= \left\{ \left(\frac{1}{n} \right) \cdot \sum_{j=1}^{m} \left[n_j \cdot \varphi_\theta \left(\frac{\mu_j}{\mu} \right) \right] \right\} && + \left\{ \sum_{j=1}^{m} \left[\frac{n_j}{n} \cdot \left(\frac{\mu_j}{\mu} \right)^\theta \cdot GE(\theta)_j \right] \right\} \end{aligned} \tag{1}$$

The first addend (B) is the between-sector component, which would be the entropy measure if individuals in each sector j received the mean wage of the sector, μ_j. The second addend (W) is the within-sector component, which is a weighted sum of the sectoral contributions $\text{GE}(\theta)_j$ (Neves Costa and Pérez-Duarte 2019). We may similarly decompose $\text{GE}(\theta)_j$ in between-firm

[13] For $\theta = 0$, $\varphi_0(x) = \lim_{\theta \to 0} \varphi_\theta(x) = -\ln(x)$. For $\theta = 1$, $\varphi_1(x) = \lim_{\theta \to 1} \varphi_\theta(x) = x \cdot \ln(x)$

(B_j) and within-firm (W_j) components:

$$
\begin{aligned}
GE(\theta)_j &= B_j && + W_j \\
GE(\theta)_j &= \left\{ \left(\frac{1}{n_j}\right) \cdot \sum_{k \in j} \left[n_k \cdot \varphi_\theta \left(\frac{\mu_k}{\mu_j}\right) \right] \right\} && + \left\{ \sum_{k \in j} \left[\frac{n_k}{n_j} \cdot \left(\frac{\mu_k}{\mu_j}\right)^\theta \cdot GE(\theta)_k \right] \right\}
\end{aligned} \tag{2}
$$

Finally, we can combine both levels to decompose the overall measure in three effects: between-sectors (B); within-sectors-between-firms (WB); and within-sectors-within-firms (WW). We do so by plugging Equation 2 into Equation 1:

$$
GE(\theta) = \left\{ \left(\frac{1}{n}\right) \cdot \sum_{j=1}^{m} \left[n_j \cdot \varphi_\theta \left(\frac{\mu_j}{\mu}\right) \right] \right\} + \left\{ \sum_{j=1}^{m} \left[\frac{n_j}{n} \cdot \left(\frac{\mu_j}{\mu}\right)^\theta \cdot B_j \right] \right\} + \left\{ \sum_{j=1}^{m} \left[\frac{n_j}{n} \cdot \left(\frac{\mu_j}{\mu}\right)^\theta \cdot W_j \right] \right\} \tag{3}
$$

$$
GE(\theta) = B + WB + WW
$$

As a final note on GE, we saw in Section 3 that regressions are performed with weighted versions of the firm-level measure $GE(\theta)_k$. The weights in this case are the ones implied by the last term (WW) of Equation 3. We can explicitly write them by using the definition of W_j from Equation 2:

$$
\begin{aligned}
WW &= \sum_{j=1}^{m} \left\{ \frac{n_j}{n} \cdot \left(\frac{\mu_j}{\mu}\right)^\theta \cdot \left\{ \sum_{k \in j} \left[\frac{n_k}{n_j} \cdot \left(\frac{\mu_k}{\mu_j}\right)^\theta \cdot GE(\theta)_k \right] \right\} \right\} \\
&= \sum_{j=1}^{m} \sum_{k \in j} \left[\frac{n_j}{n} \cdot \left(\frac{\mu_j}{\mu}\right)^\theta \cdot \frac{n_k}{n_j} \cdot \left(\frac{\mu_k}{\mu_j}\right)^\theta \cdot GE(\theta)_k \right] \\
&= \sum_{j=1}^{m} \sum_{k \in j} \left[\frac{n_k}{n} \cdot \left(\frac{\mu_k}{\mu}\right)^\theta \cdot GE(\theta)_k \right]
\end{aligned} \tag{4}
$$

The weights used in these regressions are, then, $(n_k/n) \cdot (\mu_k/\mu)^\theta$.

B2 Share of earnings above a top quantile

For the share of earnings, as will be clear in this section, the decomposition is done on the variation of the measure through time, meaning that we cannot omit the time subscript for simplicity in the notation.

The share of earnings above a top quantile at the economy level is measured as the sum of all earnings higher than the quantile divided by the sum of all earnings in a given year. We refer to this measure as sPX_t, where X is chosen in reference to different top quantiles, namely

$X = 95, 99, 999, 9999$ (i.e., the top 5%, the top 1%, the top 0.1%, and the top 0.01%). For each quantile PX_t in year t, we have $sPX_t = T_t/A_t$, where T_t is the sum of the earnings of all workers that earn above PX_t in that year and A_t is the sum of incomes of all workers in that year.

To calculate this measure at the unit level, we add a detail that will help us with the decomposition. Instead of recalculating PX_t for the earnings distribution within each sector and/or firm, we consider *the same PX_t threshold valid for the entire economy in each year* to calculate the measure at the unit level.

Thus, for each firm k in year t, $sPX_{kt} = T_{kt}/A_{kt}$, where A_{kt} is the sum of earnings of all workers in that firm and T_{kt} is the sum of earnings of all such workers that earn above PX_t (i.e., the economy-level threshold; note the absence of a sectoral subscript). Essentially, the measure at the firm level is the share of all earnings paid at the firm that goes to workers which lie at the top for the whole earnings distribution. Naturally, for a firm that does not have any workers above top quantiles, $sPX_{kt} = 0$. The measure at the sectoral level for each sector j in year t is calculated similarly as $sPX_{jt} = T_{jt}/A_{jt}$, where T_{jt} and A_{jt} are the sectoral equivalents of T_{kt} and A_{kt}, such that $T_{jt} = \sum_{k \in j} T_{kt}$ and $A_{jt} = \sum_{k \in j} A_{kt}$.

We move to the decomposition. Using the same threshold PX_t across all sectors and firms allows us to decompose *changes* in sPX_t from one year to the other. In the first level, we split the economy-level measure sPX_t in between- and within-sectors components. We start by expressing sPX_t as a weighted sum of the sector-level measures sPX_{jt}, having as weights the sizes of sectors measured by paid earnings, $\pi_{jt} = A_{jt}/A_t$, with $\sum \pi_{jt} = 1$:

$$sPX_t = \frac{T_t}{A_t} = \sum_j \left(\frac{T_{jt}}{A_t} \right) = \sum_j \left(\frac{T_{jt}}{A_{jt}} \cdot \frac{A_{jt}}{A_t} \right) = \sum_j (sPX_{jt} \cdot \pi_{jt}) \tag{5}$$

This lets us decompose the change of sPX from $t-1$ to t in two addends:

$$\begin{aligned}
\Delta sPX_{t_1} &= sPX_{t_1} - sPX_{(t-1)} = \sum_j \left(sPX_{jt_1} \cdot \pi_{jt_1} \right) - \sum_j \left(sPX_{jt_0} \cdot \pi_{jt_0} \right) \\
&= \left[\sum_j \left(sPX_{jt_1} \cdot \pi_{jt_1} \right) - \sum_j \left(sPX_{jt_1} \cdot \pi_{jt_0} \right) \right] + \left[\sum_j \left(sPX_{jt_1} \cdot \pi_{jt_0} \right) - \sum_j \left(sPX_{jt_0} \cdot \pi_{jt_0} \right) \right] \\
&= \sum_j \left(sPX_{jt_1} \cdot \Delta \pi_{jt_1} \right) + \sum_j \left(\Delta sPX_{jt_1} \cdot \pi_{jt_0} \right)
\end{aligned}$$

The first addend is an expression the between-sectors effect (fixed sector contributions, changing sector sizes), and the second is an expression of the within-sectors effect (changing sector contributions, fixed sector sizes). A slightly different algebraic manipulation is possible, which leads to alternative expressions of the two components, respectively, as $\sum_j \left(sPX_{jt_0} \cdot \Delta \pi_{jt_1} \right)$ and $\sum_j \left(\Delta sPX_{jt_1} \cdot \pi_{jt_1} \right)$. Because there is no particular reason to pick one or the other, we use their

averages. The first level of our decomposition is, then:

$$\begin{aligned}\Delta \text{sPX}_{t_1} &= B_{t_1} + W_{t_1} \\ &= \frac{\sum_j \left(\text{sPX}_{jt_1} \cdot \Delta\pi_{jt_1}\right) + \sum_j \left(\text{sPX}_{jt_0} \cdot \Delta\pi_{jt_1}\right)}{2} + \frac{\sum_j \left(\Delta\text{sPX}_{jt_1} \cdot \pi_{jt_0}\right) + \sum_j \left(\Delta\text{sPX}_{jt_1} \cdot \pi_{jt_1}\right)}{2}\end{aligned} \tag{6}$$

We move to the second level of the decomposition. The change over time of the sectoral measure ΔsPX_{jt} can be further decomposed in between-firms and within-firms effects, through an analogous process. Once more, we want to write the sectoral-level measure sPX_{jt} as a weighted sum of the firm-level measures sPX_{kt}, having as weights the firm size shares within the sector. However, firms may move in and out of the economy or across sectors between t_0 and t_1. It is then instrumental to define the firm-level measure and the firm weights in reference to the sector where the firm operates in a given year. We introduce two variables:

$$T_{kt}^{j} = \begin{cases} T_{kt} & \text{if } k \in j \text{ in time } t, \\ 0 & \text{otherwise.} \end{cases}$$

$$A_{kt}^{j} = \begin{cases} A_{kt} & \text{if } k \in j \text{ in time } t, \\ 0 & \text{otherwise.} \end{cases}$$

We then rewrite the firm-level measure for firm k in reference to the sector j where it operates in year t:

$$\text{sPX}_{kt}^{j} = \begin{cases} T_{kt}^{j}/A_{kt}^{j} & \text{if } k \in j \text{ in time } t, \\ 0 & \text{otherwise.} \end{cases}$$

We also write the firm weights in reference to the sector:

$$\pi_{kt}^{j} = \begin{cases} A_{kt}^{j}/A_{jt} & \text{if } k \in j \text{ in time } t, \\ 0 & \text{otherwise.} \end{cases}$$

With these two new variables, we have:

$$T_{jt} = \sum_{k \in j} T_{kt} = \sum_{k} T_{kt}^{j}$$

$$\text{sPX}_{jt} = \frac{T_{jt}}{A_{jt}} = \sum_{k \in j} \frac{T_{kt}}{A_{jt}} = \sum_{k} \frac{T_{kt}^{j}}{A_{jt}} = \sum_{k} \left(\frac{T_{kt}^{j}}{A_{kt}^{j}} \cdot \frac{A_{kt}^{j}}{A_{jt}} \right) = \sum_{k} \text{sPX}_{kt}^{j} * \pi_{kt}^{j}$$

We are now able to decompose ΔsPX_{jt_1} into between-firms (fixed firm contributions, changing firm sizes) and within-firms (changing firm contributions, fixed firm sizes) effects:

$$\begin{aligned}\Delta \text{sPX}_{jt_1} &= \text{sPX}_{jt_1} - \text{sPX}_{jt_0} = \sum_k \text{sPX}^j_{kt_1} * \pi^j_{kt_1} - \sum_k \text{sPX}^j_{kt_0} * \pi^j_{kt_0} \\ &= \left[\sum_k \left(\text{sPX}^j_{kt_1} \cdot \pi^j_{kt_1}\right) - \sum_k \left(\text{sPX}^j_{kt_1} \cdot \pi^j_{kt_0}\right)\right] + \left[\sum_k \left(\text{sPX}^j_{kt_1} \cdot \pi^j_{kt_0}\right) - \sum_k \left(\text{sPX}^j_{kt_0} \cdot \pi^j_{kt_0}\right)\right] \\ &= \sum_k \left(\text{sPX}^j_{kt_1} \cdot \Delta\pi^j_{kt_1}\right) + \sum_k \left(\Delta\text{sPX}^j_{kt_1} \cdot \pi^j_{kt_0}\right) \quad (7)\end{aligned}$$

The summation is done over all firms in the economy. Firms out of sector j in both periods have all parcels equal to zero for the decomposition in reference to that sector. Firms that move from sector j' to sector j between t_0 and t_1 will have their parcels added in two different decompositions: on t_0, the contribution to the top will be counted for j' and the firm size share will be calculated in reference to j'. On t_1, both will be done in reference to j. In other words, in reference to sector j, these firms will have $\text{sPX}^j_{kt_0} = 0$ and $\pi^j_{kt_0} = 0$; and $\text{sPX}^j_{kt_1}$ and $\pi^j_{kt_1}$ will be calculated accordingly. Similarly, in reference to j', such a firm will have $\text{sPX}^{j'}_{kt_1} = 0$ and $\pi^{j'}_{kt_1} = 0$; while $\text{sPX}^{j'}_{kt_0}$ and $\pi^{j'}_{kt_0}$ are calculated accordingly. If a firm moves out of the economy between t_0 and t_1, its parameters are considered in the calculations relative to t_0 (in reference to the sector where it was in t_0) and it is not taken into consideration in the part of the calculation relative to t_1. The analogous applies for firms who moves into the economy.

As in the first level of the decomposition, another algebraic manipulation is possible in which $\Delta \text{sPX}_{jt_1} = \sum_k \left(\text{sPX}^j_{kt_0} \cdot \Delta\pi^j_{kt_1}\right) + \sum_k \left(\Delta\text{sPX}^j_{kt_1} \cdot \pi^j_{kt_1}\right)$. We once more use the averages of the two expressions to have $\Delta \text{sPX}_{jt_1} = B_{jt_1} + W_{jt_1}$ such that:

$$B_{jt_1} = \frac{\sum_k \left(\text{sPX}^j_{kt_1} \cdot \Delta\pi^j_{kt_1}\right) + \sum_k \left(\text{sPX}^j_{kt_0} \cdot \Delta\pi^j_{kt_1}\right)}{2} \quad (8)$$

$$W_{jt_1} = \frac{\sum_k \left(\Delta\text{sPX}^j_{kt_1} \cdot \pi^j_{kt_0}\right) + \sum_k \left(\Delta\text{sPX}^j_{kt_1} \cdot \pi^j_{kt_1}\right)}{2} \quad (9)$$

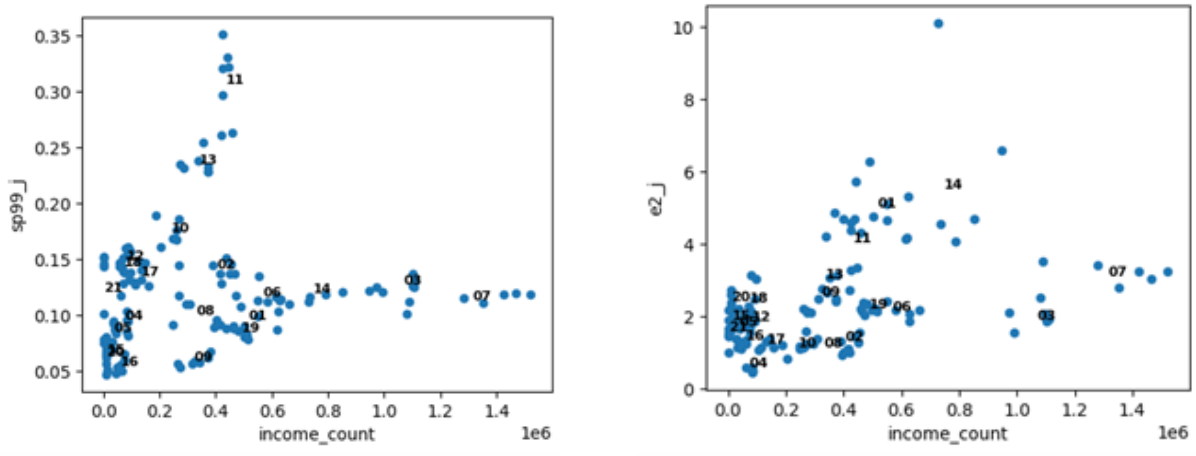
In our final step, we combine both levels of the decomposition. To do so, we use Equations 8 and 9 to plug the expression $\Delta \text{sPX}_{jt_1} = B_{jt_1} + W_{jt_1}$ into Equation 6. As for the generalized entropy measure, this only affects the second parcel of Equation 6. We may then say that we are left with the decomposition of the change over time of the economy-level measure in a between-sectors, a within-sectors-between-firms, and a within-sectors-within-firms effect: $\Delta \text{sPX}_{t_1} = B_{t_1} + WB_{t_1} + WW_{t_1}$

C Additional figures and tables

C1 Figures

The following figures show the sectoral-level scatterplots of size versus sP99 and GE(2) (Figure C1); the time trends of unweighted sectoral-level measures (Figure C2); and KDEs for all individual log earnings per year (Figure C3).

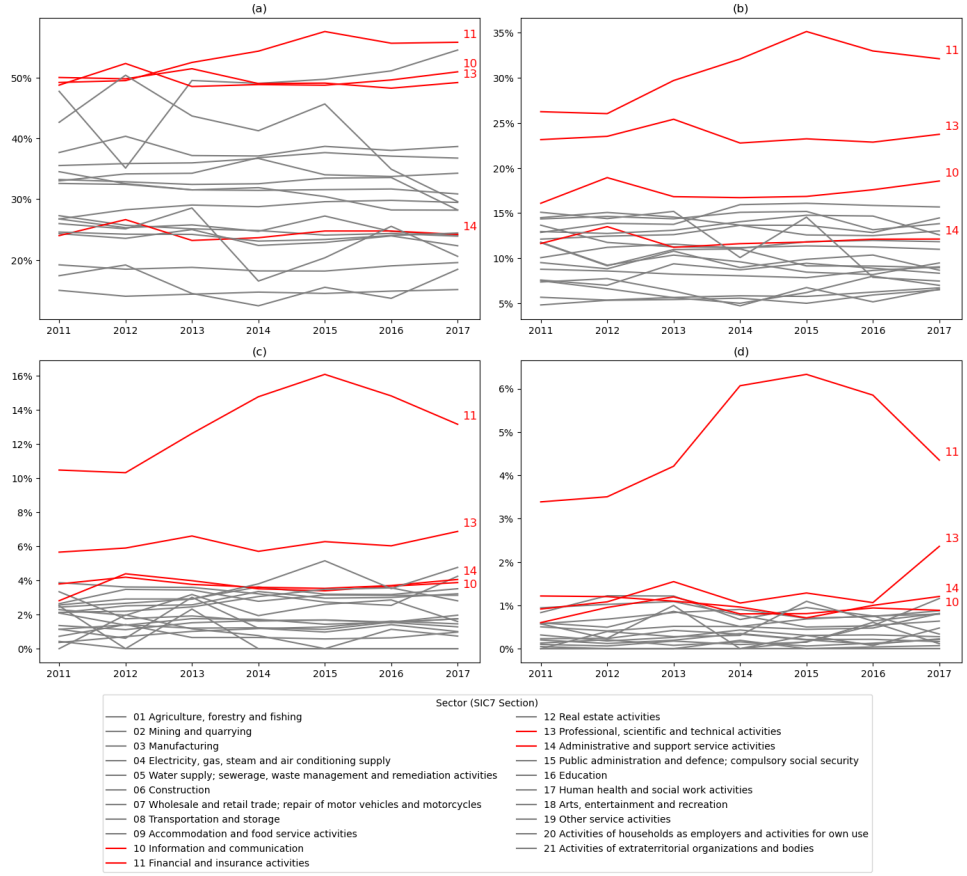
Figure C1: Scatterplots for sector size (number of workers) and sectoral-level sP99 and GE(2)



(a) sP99$_{jt}$ **(b) GE(2)**$_{jt}$

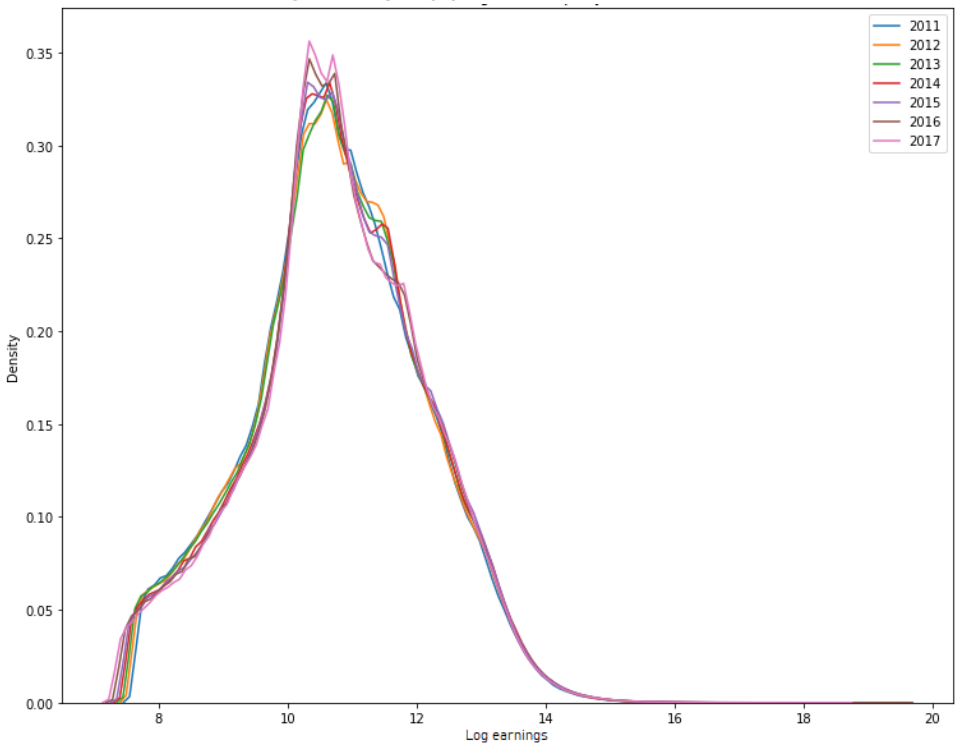
Source: author's own calculations.

Figure C2: Time trends of unweighted sectoral-level top shares: (a) $sP95_{jt}$; (b) $sP99_{jt}$; (c) $sP999_{jt}$; and (b) $sP9999_{jt}$



Source: author's own calculations.

Figure C3: KDE for individual real log earnings by year



Source: author's own calculations.

C2 Tables: full regression results

The following tables present the full regression results not shown in the main text: all REWB results for GE(1) (Table C1); all REWB results for weighted GE(1) (Table C3) and weighted GE(2) (Table (Table C2); all WB pooled Tobit coefficients for sP95 (Table C4), sP99 (Table C5), and sP999 (Table C6); and all REWB results for sP95 (Table C7), sP99 (Table C8), and sP999 (Table C9).

Table C1: REWB estimates for GE(1).

Model	Main	UD1	UD2	PR1	PR2	SM1	SM2	RD1	RD2
Log number of workers (W)	0.145***	0.145***	0.145***	0.152***	0.152***	0.146***	0.146***	0.121***	0.121***
	(0.00105)	(0.00105)	(0.00105)	(0.00128)	(0.00128)	(0.00120)	(0.00120)	(0.00180)	(0.00180)
Log labour productivity (W)	0.00226***	0.00225***	0.00224***	0.00285***	0.00284***	0.00274***	0.00289***	-0.00213**	-0.00213**
	(0.000546)	(0.000546)	(0.000546)	(0.000632)	(0.000632)	(0.000597)	(0.000597)	(0.000890)	(0.000890)
Listed firm	-0.120***	-0.120***	-0.120***	-0.120***	-0.121***	-0.0997***	-0.101***	-0.109***	-0.109***
	(0.0243)	(0.0243)	(0.0243)	(0.0269)	(0.0269)	(0.0247)	(0.0247)	(0.0246)	(0.0246)
Log HHI (W)	-0.00258***	-0.00258***	-0.00249***	-0.00248***	-0.00247***	-0.00196***	-0.00196***	-0.00283***	-0.00283***
	(0.000547)	(0.000547)	(0.000554)	(0.000593)	(0.000593)	(0.000588)	(0.000587)	(0.000927)	(0.000927)
Log number of workers (B)	0.0864***	0.0864***	0.0864***	0.0865***	0.0863***	0.0872***	0.0877***	0.0780***	0.0780***
	(0.000548)	(0.000548)	(0.000548)	(0.000554)	(0.000555)	(0.000563)	(0.000563)	(0.000792)	(0.000791)
Log labour productivity (B)	0.0135***	0.0135***	0.0135***	0.0136***	0.0133***	0.0125***	0.0129***	0.00625***	0.00627***
	(0.000626)	(0.000627)	(0.000627)	(0.000642)	(0.000646)	(0.000654)	(0.000653)	(0.000859)	(0.000859)
Log HHI (B)	-0.00223	-0.00222	-0.00143	-0.00263	-0.00260	-0.00301	-0.00242	-0.0180***	-0.0180***
	(0.00246)	(0.00246)	(0.00248)	(0.00270)	(0.00269)	(0.00242)	(0.00241)	(0.00313)	(0.00313)
Union density (W)			-0.0170						
			(0.0124)						
Union density (B)			-0.200***						
			(0.0571)						
Share of male workers (W)							-0.0379***		
							(0.00347)		
Share of male workers (B)							-0.0509***		
							(0.00220)		
R&D expenditure dummy (W)									-0.000980
									(0.00294)
R&D expenditure dummy (B)									-0.00171
									(0.00859)
Constant	0.0186	0.0188	0.0329**	0.0179	0.0184	0.0333**	0.0583***	0.178***	0.178***
	(0.0132)	(0.0132)	(0.0137)	(0.0133)	(0.0134)	(0.0130)	(0.0130)	(0.0158)	(0.0158)
Sample subset	Full	Union	Union	Province	Province	Male share	Male share	R&D	R&D
Observations	669,879	668,870	668,870	516,143	516,143	550,597	550,597	289,795	289,795
Number of firms	135,558	135,458	135,458	133,237	133,237	122,177	122,177	89,765	89,765
Province FE	NO	NO	NO	NO	YES	NO	NO	NO	NO
Within R2	0.161	0.161	0.161	0.161	0.161	0.160	0.160	0.130	0.130
Between R2	0.250	0.250	0.250	0.244	0.244	0.260	0.263	0.229	0.229
Overall R2	0.228	0.228	0.229	0.225	0.225	0.239	0.242	0.192	0.192

Robust standard errors in parentheses. All models include time and sector FE. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Source: author's own calculations.

Table C2: REWB estimates for weighted GE(2).

Model	Main	UD1	UD2	PR1	PR2	SM1	SM2	RD1	RD2
Log number of workers (W)	1.14e-05***	1.14e-05***	1.14e-05***	9.08e-06***	9.09e-06***	1.08e-05***	1.08e-05***	2.74e-05***	2.73e-05***
	(1.40e-06)	(1.40e-06)	(1.40e-06)	(1.50e-06)	(1.50e-06)	(1.27e-06)	(1.27e-06)	(4.50e-06)	(4.49e-06)
Log labour productivity (W)	3.48e-06***	3.48e-06***	3.48e-06***	2.70e-06***	2.70e-06***	2.81e-06***	2.81e-06***	8.61e-06***	8.60e-06***
	(8.73e-07)	(8.75e-07)	(8.77e-07)	(7.00e-07)	(7.00e-07)	(8.51e-07)	(8.52e-07)	(2.59e-06)	(2.58e-06)
Listed firm	0.00226***	0.00226***	0.00226***	0.00207***	0.00206***	0.00199***	0.00198***	0.00239***	0.00238***
	(0.000725)	(0.000726)	(0.000726)	(0.000649)	(0.000649)	(0.000755)	(0.000755)	(0.000828)	(0.000829)
Log HHI (W)	-1.29e-07	-1.32e-07	9.10e-09	-7.19e-07	-7.20e-07	-1.56e-07	-1.56e-07	8.37e-08	5.51e-08
	(5.54e-07)	(5.55e-07)	(5.06e-07)	(8.03e-07)	(8.03e-07)	(6.23e-07)	(6.23e-07)	(2.08e-06)	(2.08e-06)
Log number of workers (B)	6.49e-05***	6.50e-05***	6.50e-05***	5.97e-05***	5.97e-05***	6.43e-05***	6.45e-05***	0.000135***	0.000132***
	(9.77e-06)	(9.78e-06)	(9.78e-06)	(1.02e-05)	(1.02e-05)	(1.06e-05)	(1.06e-05)	(2.14e-05)	(2.16e-05)
Log labour productivity (B)	3.77e-05***	3.77e-05***	3.77e-05***	3.32e-05***	3.26e-05***	3.67e-05***	3.69e-05***	8.78e-05***	8.54e-05***
	(5.03e-06)	(5.04e-06)	(5.04e-06)	(5.01e-06)	(4.89e-06)	(5.31e-06)	(5.35e-06)	(1.26e-05)	(1.27e-05)
Log HHI (B)	4.10e-07	4.16e-07	1.91e-07	-7.42e-06	-7.26e-06	6.71e-06	7.06e-06	2.06e-05	2.19e-05
	(9.36e-06)	(9.34e-06)	(9.51e-06)	(1.18e-05)	(1.18e-05)	(5.47e-06)	(5.47e-06)	(1.91e-05)	(1.89e-05)
Union density (W)			-2.92e-05						
			(7.07e-05)						
Union density (B)			6.35e-05						
			(9.47e-05)						
Share of male workers (W)							-1.91e-07		
							(1.42e-06)		
Share of male workers (B)							-2.42e-05***		
							(5.97e-06)		
R&D expenditure dummy (W)									1.01e-05
									(3.08e-05)
R&D expenditure dummy (B)									0.000208**
									(9.45e-05)
Constant	-0.000730***	-0.000731***	-0.000735***	-0.000575***	-0.000562***	-0.000701***	-0.000690***	-0.00149***	-0.00146***
	(0.000125)	(0.000125)	(0.000122)	(9.50e-05)	(9.32e-05)	(0.000118)	(0.000116)	(0.000236)	(0.000238)
Sample subset	Full	Union	Union	Province	Province	Male share	Male share	R&D	R&D
Observations	669,879	668,870	668,870	516,143	516,143	550,597	550,597	289,795	289,795
Number of firms	135,558	135,458	135,458	133,237	133,237	122,177	122,177	89,765	89,765
Province FE	NO	NO	NO	NO	YES	NO	NO	NO	NO
Within R2	2.71e-05	2.72e-05	2.74e-05	2.23e-05	2.42e-05	3.37e-05	3.38e-05	6.44e-05	6.47e-05
Between R2	0.0201	0.0202	0.0202	0.0191	0.0192	0.0255	0.0256	0.0242	0.0251
Overall R2	0.00984	0.00985	0.00985	0.0128	0.0128	0.0145	0.0146	0.0133	0.0135

Robust standard errors in parentheses. All models include time and sector FE. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Source: author's own calculations.

Table C3: REWB estimates for weighted GE(1).

Model	Main	UD1	UD2	PR1	PR2	SM1	SM2	RD1	RD2
Log number of workers (W)	3.61e-06***	3.61e-06***	3.61e-06***	3.20e-06***	3.20e-06***	3.84e-06***	3.85e-06***	7.81e-06***	7.79e-06***
	(1.82e-07)	(1.82e-07)	(1.82e-07)	(1.92e-07)	(1.92e-07)	(2.25e-07)	(2.26e-07)	(4.51e-07)	(4.47e-07)
Log labour productivity (W)	4.45e-07***	4.45e-07***	4.45e-07***	3.82e-07***	3.82e-07***	4.96e-07***	4.98e-07***	1.08e-06***	1.07e-06***
	(9.71e-08)	(9.73e-08)	(9.72e-08)	(6.34e-08)	(6.34e-08)	(1.22e-07)	(1.22e-07)	(2.47e-07)	(2.45e-07)
Listed firm	0.000348***	0.000348***	0.000348***	0.000330***	0.000330***	0.000289***	0.000289***	0.000358***	0.000357***
	(9.44e-05)	(9.44e-05)	(9.44e-05)	(9.17e-05)	(9.17e-05)	(9.18e-05)	(9.17e-05)	(9.99e-05)	(9.99e-05)
Log HHI (W)	3.58e-08	3.58e-08	4.62e-08	-1.77e-08	-1.77e-08	-1.29e-08	-1.28e-08	3.34e-07**	3.14e-07**
	(5.17e-08)	(5.17e-08)	(5.55e-08)	(4.31e-08)	(4.31e-08)	(5.96e-08)	(5.96e-08)	(1.36e-07)	(1.34e-07)
Log number of workers (B)	1.22e-05***	1.22e-05***	1.22e-05***	1.13e-05***	1.13e-05***	1.20e-05***	1.20e-05***	1.93e-05***	1.90e-05***
	(9.40e-07)	(9.41e-07)	(9.41e-07)	(8.90e-07)	(8.91e-07)	(9.56e-07)	(9.61e-07)	(1.63e-06)	(1.58e-06)
Log labour productivity (B)	4.69e-06***	4.69e-06***	4.69e-06***	4.39e-06***	4.28e-06***	4.77e-06***	4.80e-06***	8.91e-06***	8.64e-06***
	(4.19e-07)	(4.19e-07)	(4.20e-07)	(3.99e-07)	(3.91e-07)	(4.43e-07)	(4.46e-07)	(8.47e-07)	(8.08e-07)
Log HHI (B)	1.11e-06	1.11e-06	1.12e-06	5.06e-07	5.25e-07	1.27e-06	1.30e-06	3.75e-06	3.94e-06
	(1.18e-06)	(1.18e-06)	(1.23e-06)	(9.22e-07)	(9.22e-07)	(1.25e-06)	(1.25e-06)	(2.95e-06)	(2.99e-06)
Union density (W)			-2.17e-06						
			(2.22e-06)						
Union density (B)			-2.00e-06						
			(1.40e-05)						
Share of male workers (W)							-4.64e-07*		
							(2.45e-07)		
Share of male workers (B)							-3.19e-06***		
							(6.78e-07)		
R&D expenditure dummy (W)									7.10e-06***
									(2.73e-06)
R&D expenditure dummy (B)									3.01e-05**
									(1.25e-05)
Constant	-9.71e-05***	-9.71e-05***	-9.70e-05***	-8.78e-05***	-8.55e-05***	-9.62e-05***	-9.46e-05***	-0.000167***	-0.000164***
	(1.12e-05)	(1.12e-05)	(1.05e-05)	(9.59e-06)	(9.40e-06)	(1.14e-05)	(1.12e-05)	(2.16e-05)	(2.09e-05)
Sample subset	Full	Union	Union	Province	Province	Male share	Male share	R&D	R&D
Observations	669,879	668,870	668,870	516,143	516,143	550,597	550,597	289,795	289,795
Number of firms	135,558	135,458	135,458	133,237	133,237	122,177	122,177	89,765	89,765
Province FE	NO	NO	NO	NO	YES	NO	NO	NO	NO
Within R2	0.00486	0.00487	0.00487	0.00395	0.00396	0.00516	0.00516	0.00941	0.00990
Between R2	0.0811	0.0811	0.0811	0.0771	0.0774	0.0739	0.0741	0.0969	0.0987
Overall R2	0.0920	0.0921	0.0921	0.0859	0.0862	0.0853	0.0855	0.121	0.122

Robust standard errors in parentheses. All models include time and sector FE. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Source: author's own calculations.

Table C4: WB pooled Tobit estimates for sP95.

Model	Main	UD1	UD2	PR1	PR2	SM1	SM2	RD1	RD2
Log number of workers (W)	0.111***	0.111***	0.111***	0.106***	0.105***	0.108***	0.107***	0.0389***	0.0388***
	(0.00241)	(0.00241)	(0.00241)	(0.00281)	(0.00279)	(0.00271)	(0.00270)	(0.00251)	(0.00251)
Log labour productivity (W)	0.0811***	0.0810***	0.0810***	0.0742***	0.0736***	0.0799***	0.0795***	0.0394***	0.0394***
	(0.00181)	(0.00181)	(0.00181)	(0.00206)	(0.00204)	(0.00198)	(0.00198)	(0.00162)	(0.00162)
Listed firm	-0.324***	-0.324***	-0.324***	-0.326***	-0.333***	-0.316***	-0.316***	-0.0824**	-0.0864**
	(0.0556)	(0.0556)	(0.0556)	(0.0585)	(0.0579)	(0.0578)	(0.0578)	(0.0355)	(0.0353)
Log HHI (W)	0.00154	0.00152	0.00187	0.000192	0.000110	0.00237	0.00234	0.00449***	0.00448***
	(0.00151)	(0.00151)	(0.00153)	(0.00168)	(0.00167)	(0.00164)	(0.00164)	(0.00150)	(0.00150)
Log number of workers (B)	0.194***	0.194***	0.194***	0.193***	0.190***	0.192***	0.192***	0.0900***	0.0888***
	(0.00136)	(0.00136)	(0.00136)	(0.00135)	(0.00135)	(0.00141)	(0.00142)	(0.00148)	(0.00149)
Log labour productivity (B)	0.316***	0.316***	0.316***	0.312***	0.301***	0.309***	0.309***	0.191***	0.190***
	(0.00230)	(0.00230)	(0.00230)	(0.00237)	(0.00235)	(0.00245)	(0.00245)	(0.00211)	(0.00211)
Log HHI (B)	-0.0343***	-0.0343***	-0.0321***	-0.0255**	-0.0217**	-0.0226**	-0.0227**	-0.0210***	-0.0205**
	(0.0103)	(0.0103)	(0.0103)	(0.0106)	(0.0105)	(0.00971)	(0.00971)	(0.00806)	(0.00805)
Union density (W)			-0.0779**						
			(0.0324)						
Union density (B)			-0.608***						
			(0.206)						
Share of male workers (W)							0.0961***		
							(0.00995)		
Share of male workers (B)							0.0122*		
							(0.00726)		
R&D expenditure dummy (W)									0.00190
									(0.00390)
R&D expenditure dummy (B)									0.0820***
									(0.0123)
Constant	-4.578***	-4.577***	-4.532***	-4.645***	-4.462***	-4.604***	-4.610***	-2.947***	-2.934***
	(0.0516)	(0.0516)	(0.0537)	(0.0508)	(0.0503)	(0.0501)	(0.0501)	(0.0400)	(0.0401)
/var(e.sp95_k)	0.214***	0.214***	0.214***	0.216***	0.212***	0.210***	0.210***	0.129***	0.129***
	(0.00136)	(0.00136)	(0.00136)	(0.00142)	(0.00139)	(0.00144)	(0.00144)	(0.000983)	(0.000981)
Sample subset	Full	Union	Union	Province	Province	Male share	Male share	R&D	R&D
Observations	669,879	668,870	668,870	516,143	516,143	550,597	550,597	289,795	289,795
Province FE	NO	NO	NO	NO	YES	NO	NO	NO	NO
Pseudo-R2	0.261	0.261	0.261	0.259	0.268	0.259	0.259	0.277	0.277
Log-lik	-324532	-324202	-324186	-250702	-247539	-271144	-271097	-140002	-139924

Robust standard errors in parentheses. All models include time and sector FE. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Source: author's own calculations.

Table C5: WB pooled Tobit estimates for sP99.

Model	Main	UD1	UD2	PR1	PR2	SM1	SM2	RD1	RD2
Log number of workers (W)	0.130***	0.130***	0.130***	0.124***	0.123***	0.127***	0.126***	0.0690***	0.0689***
	(0.00396)	(0.00396)	(0.00396)	(0.00475)	(0.00473)	(0.00429)	(0.00429)	(0.00365)	(0.00365)
Log labour productivity (W)	0.0825***	0.0824***	0.0824***	0.0760***	0.0754***	0.0814***	0.0812***	0.0444***	0.0444***
	(0.00316)	(0.00316)	(0.00316)	(0.00379)	(0.00376)	(0.00345)	(0.00344)	(0.00265)	(0.00264)
Listed firm	-0.268***	-0.268***	-0.268***	-0.264***	-0.270***	-0.248***	-0.248***	-0.0786*	-0.0812*
	(0.0658)	(0.0657)	(0.0658)	(0.0691)	(0.0685)	(0.0670)	(0.0670)	(0.0447)	(0.0447)
Log HHI (W)	8.51e-05	9.97e-05	0.000687	-0.00192	-0.00201	-0.000539	-0.000547	0.00168	0.00165
	(0.00250)	(0.00250)	(0.00251)	(0.00277)	(0.00277)	(0.00264)	(0.00264)	(0.00237)	(0.00237)
Log number of workers (B)	0.244***	0.244***	0.244***	0.244***	0.239***	0.237***	0.237***	0.139***	0.138***
	(0.00202)	(0.00202)	(0.00202)	(0.00200)	(0.00200)	(0.00207)	(0.00208)	(0.00194)	(0.00195)
Log labour productivity (B)	0.339***	0.339***	0.339***	0.337***	0.324***	0.324***	0.324***	0.215***	0.214***
	(0.00376)	(0.00376)	(0.00376)	(0.00386)	(0.00384)	(0.00392)	(0.00392)	(0.00320)	(0.00322)
Log HHI (B)	-0.0253	-0.0252	-0.0227	-0.0136	-0.00895	-0.0178	-0.0178	-0.0289**	-0.0286**
	(0.0175)	(0.0175)	(0.0174)	(0.0181)	(0.0180)	(0.0159)	(0.0159)	(0.0130)	(0.0130)
Union density (W)			-0.139**						
			(0.0548)						
Union density (B)			-0.822**						
			(0.336)						
Share of male workers (W)							0.0726***		
							(0.0162)		
Share of male workers (B)							0.00961		
							(0.0110)		
R&D expenditure dummy (W)									0.00380
									(0.00591)
R&D expenditure dummy (B)									0.0570***
									(0.0169)
Constant	-5.659***	-5.657***	-5.593***	-5.734***	-5.515***	-5.495***	-5.500***	-3.842***	-3.831***
	(0.0860)	(0.0859)	(0.0896)	(0.0855)	(0.0848)	(0.0822)	(0.0822)	(0.0643)	(0.0644)
/var(e.sp99_k)	0.245***	0.245***	0.245***	0.249***	0.244***	0.231***	0.231***	0.157***	0.157***
	(0.00294)	(0.00294)	(0.00294)	(0.00307)	(0.00302)	(0.00297)	(0.00297)	(0.00204)	(0.00204)
Sample subset	Full	Union	Union	Province	Province	Male share	Male share	R&D	R&D
Observations	669,879	668,870	668,870	516,143	516,143	550,597	550,597	289,795	289,795
Province FE	NO	NO	NO	NO	YES	NO	NO	NO	NO
Pseudo-R2	0.333	0.332	0.333	0.329	0.340	0.334	0.334	0.298	0.298
Log-lik	-146436	-146313	-146301	-113294	-111535	-123910	-123900	-91864	-91838

Robust standard errors in parentheses. All models include time and sector FE. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Source: author's own calculations.

Table C6: WB pooled Tobit estimates for sP999.

Model	Main	UD1	UD2	PR1	PR2	SM1	SM2	RD1	RD2
Log number of workers (W)	0.146***	0.147***	0.147***	0.144***	0.143***	0.146***	0.146***	0.120***	0.120***
	(0.0104)	(0.0104)	(0.0104)	(0.0133)	(0.0133)	(0.0107)	(0.0107)	(0.00977)	(0.00977)
Log labour productivity (W)	0.0901***	0.0902***	0.0902***	0.0841***	0.0828***	0.0957***	0.0957***	0.0740***	0.0740***
	(0.00903)	(0.00904)	(0.00904)	(0.0114)	(0.0113)	(0.00935)	(0.00936)	(0.00830)	(0.00831)
Listed firm	0.00615	0.00648	0.00633	0.00985	0.00459	-0.00501	-0.00457	0.0490	0.0498
	(0.0650)	(0.0650)	(0.0650)	(0.0706)	(0.0701)	(0.0649)	(0.0648)	(0.0552)	(0.0553)
Log HHI (W)	0.00501	0.00491	0.00484	0.00174	0.00134	0.00257	0.00255	-0.000336	-0.000523
	(0.00807)	(0.00807)	(0.00811)	(0.00889)	(0.00889)	(0.00853)	(0.00853)	(0.00737)	(0.00737)
Log number of workers (B)	0.258***	0.258***	0.258***	0.261***	0.257***	0.246***	0.246***	0.205***	0.206***
	(0.00461)	(0.00461)	(0.00461)	(0.00483)	(0.00479)	(0.00481)	(0.00480)	(0.00394)	(0.00399)
Log labour productivity (B)	0.317***	0.317***	0.317***	0.324***	0.316***	0.290***	0.290***	0.258***	0.258***
	(0.00896)	(0.00896)	(0.00894)	(0.00959)	(0.00956)	(0.00922)	(0.00922)	(0.00786)	(0.00791)
Log HHI (B)	-0.0941*	-0.0938*	-0.0929*	-0.106**	-0.102**	-0.0882**	-0.0882**	-0.0543	-0.0540
	(0.0507)	(0.0505)	(0.0502)	(0.0477)	(0.0478)	(0.0431)	(0.0431)	(0.0390)	(0.0390)
Union density (W)			0.0179						
			(0.141)						
Union density (B)			-0.224						
			(0.929)						
Share of male workers (W)							-0.0102		
							(0.0476)		
Share of male workers (B)							0.0234		
							(0.0266)		
R&D expenditure dummy (W)									0.0428***
									(0.0142)
R&D expenditure dummy (B)									-0.0686**
									(0.0339)
Constant	-5.987***	-5.988***	-5.971***	-6.090***	-5.922***	-5.587***	-5.602***	-5.137***	-5.143***
	(0.237)	(0.237)	(0.255)	(0.227)	(0.227)	(0.220)	(0.220)	(0.191)	(0.191)
/var(e.sp999_k)	0.286***	0.287***	0.287***	0.301***	0.298***	0.256***	0.256***	0.227***	0.226***
	(0.0111)	(0.0111)	(0.0111)	(0.0123)	(0.0122)	(0.0110)	(0.0110)	(0.00881)	(0.00881)
Sample subset	Full	Union	Union	Province	Province	Male share	Male share	R&D	R&D
Observations	669,879	668,870	668,870	516,143	516,143	550,597	550,597	289,795	289,795
Province FE	NO	NO	NO	NO	YES	NO	NO	NO	NO
Pseudo-R2	0.420	0.420	0.420	0.416	0.421	0.428	0.428	0.361	0.361
Log-lik	-25990	-25978	-25978	-19854	-19685	-21840	-21839	-22260	-22257

Robust standard errors in parentheses. All models include time and sector FE. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Source: author's own calculations.

Table C7: REWB estimates for sP95.

Model	Main	UD1	UD2	PR1	PR2	SM1	SM2	RD1	RD2
Log number of workers (W)	0.0177***	0.0177***	0.0177***	0.0160***	0.0160***	0.0177***	0.0175***	0.000801	0.000801
	(0.000784)	(0.000785)	(0.000785)	(0.000900)	(0.000900)	(0.000919)	(0.000917)	(0.00136)	(0.00136)
Log labour productivity (W)	0.0239***	0.0239***	0.0239***	0.0212***	0.0212***	0.0244***	0.0242***	0.0193***	0.0193***
	(0.000521)	(0.000521)	(0.000521)	(0.000585)	(0.000585)	(0.000598)	(0.000597)	(0.000812)	(0.000812)
Listed firm	0.0133	0.0132	0.0132	0.0141	0.00834	0.0177	0.0180	0.0339	0.0298
	(0.0248)	(0.0248)	(0.0248)	(0.0260)	(0.0258)	(0.0264)	(0.0264)	(0.0245)	(0.0245)
Log HHI (W)	-0.000540	-0.000543	-0.000301	-0.00157***	-0.00157***	-1.32e-05	-1.92e-05	0.00143*	0.00143*
	(0.000511)	(0.000511)	(0.000515)	(0.000561)	(0.000561)	(0.000573)	(0.000573)	(0.000836)	(0.000836)
Log number of workers (B)	0.0491***	0.0492***	0.0492***	0.0485***	0.0481***	0.0510***	0.0509***	0.0321***	0.0313***
	(0.000430)	(0.000431)	(0.000431)	(0.000436)	(0.000435)	(0.000448)	(0.000449)	(0.000728)	(0.000733)
Log labour productivity (B)	0.102***	0.102***	0.102***	0.101***	0.0988***	0.103***	0.103***	0.0926***	0.0919***
	(0.000815)	(0.000815)	(0.000815)	(0.000842)	(0.000834)	(0.000865)	(0.000865)	(0.00107)	(0.00107)
Log HHI (B)	0.00303	0.00295	0.00259	0.00406	0.00445	0.00171	0.00158	-0.00857**	-0.00813**
	(0.00256)	(0.00256)	(0.00257)	(0.00280)	(0.00278)	(0.00259)	(0.00259)	(0.00342)	(0.00341)
Union density (W)			-0.0502***						
			(0.0123)						
Union density (B)			0.0908						
			(0.0639)						
Share of male workers (W)							0.0313***		
							(0.00335)		
Share of male workers (B)							0.0109***		
							(0.00231)		
R&D expenditure dummy (W)									-0.000201
									(0.00253)
R&D expenditure dummy (B)									0.0822***
									(0.00820)
Constant	-1.305***	-1.305***	-1.312***	-1.304***	-1.258***	-1.329***	-1.334***	-1.173***	-1.167***
	(0.0149)	(0.0149)	(0.0155)	(0.0151)	(0.0150)	(0.0148)	(0.0148)	(0.0178)	(0.0178)
Sample subset	Full	Union	Union	Province	Province	Male share	Male share	R&D	R&D
Observations	669,879	668,870	668,870	516,143	516,143	550,597	550,597	289,795	289,795
Number of firms	135,558	135,458	135,458	133,237	133,237	122,177	122,177	89,765	89,765
Province FE	NO	NO	NO	NO	YES	NO	NO	NO	NO
Within R2	0.0145	0.0145	0.0145	0.0127	0.0128	0.0139	0.0145	0.0143	0.0143
Between R2	0.277	0.277	0.277	0.269	0.280	0.269	0.269	0.279	0.280
Overall R2	0.263	0.263	0.263	0.260	0.270	0.261	0.261	0.283	0.284

Robust standard errors in parentheses. All models include time and sector FE. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Source: author's own calculations.

Table C8: REWB estimates for sP99.

Model	Main	UD1	UD2	PR1	PR2	SM1	SM2	RD1	RD2
Log number of workers (W)	0.00655***	0.00655***	0.00655***	0.00575***	0.00575***	0.00681***	0.00676***	0.00276***	0.00276***
	(0.000462)	(0.000462)	(0.000462)	(0.000544)	(0.000544)	(0.000531)	(0.000531)	(0.000941)	(0.000940)
Log labour productivity (W)	0.00823***	0.00823***	0.00823***	0.00746***	0.00746***	0.00861***	0.00859***	0.00916***	0.00916***
	(0.000306)	(0.000307)	(0.000307)	(0.000355)	(0.000355)	(0.000344)	(0.000343)	(0.000572)	(0.000572)
Listed firm	0.0698***	0.0698***	0.0698***	0.0751***	0.0731***	0.0753***	0.0753***	0.0736***	0.0723***
	(0.0196)	(0.0196)	(0.0195)	(0.0207)	(0.0207)	(0.0205)	(0.0204)	(0.0202)	(0.0202)
Log HHI (W)	-5.18e-05	-5.16e-05	0.000106	-0.000426	-0.000426	2.29e-05	2.16e-05	3.33e-05	3.25e-05
	(0.000289)	(0.000289)	(0.000293)	(0.000318)	(0.000318)	(0.000321)	(0.000321)	(0.000595)	(0.000595)
Log number of workers (B)	0.0216***	0.0216***	0.0216***	0.0213***	0.0212***	0.0220***	0.0220***	0.0161***	0.0159***
	(0.000244)	(0.000244)	(0.000244)	(0.000246)	(0.000245)	(0.000250)	(0.000251)	(0.000425)	(0.000428)
Log labour productivity (B)	0.0384***	0.0384***	0.0384***	0.0381***	0.0372***	0.0380***	0.0380***	0.0384***	0.0382***
	(0.000504)	(0.000505)	(0.000504)	(0.000517)	(0.000513)	(0.000523)	(0.000522)	(0.000659)	(0.000662)
Log HHI (B)	0.00115	0.00113	0.000837	0.00121	0.00133	0.000713	0.000677	-0.00317*	-0.00302
	(0.00127)	(0.00127)	(0.00128)	(0.00139)	(0.00138)	(0.00128)	(0.00128)	(0.00185)	(0.00185)
Union density (W)			-0.0327***						
			(0.00753)						
Union density (B)			0.0743**						
			(0.0322)						
Share of male workers (W)							0.00650***		
							(0.00174)		
Share of male workers (B)							0.00314***		
							(0.00117)		
R&D expenditure dummy (W)									0.000305
									(0.00197)
R&D expenditure dummy (B)									0.0262***
									(0.00523)
Constant	-0.516***	-0.516***	-0.521***	-0.512***	-0.497***	-0.516***	-0.517***	-0.512***	-0.510***
	(0.00841)	(0.00841)	(0.00861)	(0.00852)	(0.00846)	(0.00828)	(0.00834)	(0.0105)	(0.0105)
Sample subset	Full	Union	Union	Province	Province	Male share	Male share	R&D	R&D
Observations	669,879	668,870	668,870	516,143	516,143	550,597	550,597	289,795	289,795
Number of firms	135,558	135,458	135,458	133,237	133,237	122,177	122,177	89,765	89,765
Province FE	NO	NO	NO	NO	YES	NO	NO	NO	NO
Within R2	0.00546	0.00545	0.00552	0.00458	0.00459	0.00542	0.00549	0.00701	0.00701
Between R2	0.154	0.154	0.154	0.149	0.154	0.150	0.150	0.164	0.165
Overall R2	0.140	0.140	0.140	0.138	0.143	0.140	0.140	0.162	0.162

Robust standard errors in parentheses. All models include time and sector FE. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Source: author's own calculations.

Table C9: REWB estimates for sP999.

Model	Main	UD1	UD2	PR1	PR2	SM1	SM2	RD1	RD2
Log number of workers (W)	0.00100***	0.00101***	0.00101***	0.000867***	0.000865***	0.00110***	0.00110***	0.00163***	0.00162***
	(0.000168)	(0.000168)	(0.000168)	(0.000219)	(0.000219)	(0.000178)	(0.000179)	(0.000384)	(0.000384)
Log labour productivity (W)	0.00113***	0.00114***	0.00114***	0.000978***	0.000977***	0.00123***	0.00124***	0.00219***	0.00218***
	(0.000121)	(0.000121)	(0.000121)	(0.000152)	(0.000152)	(0.000123)	(0.000123)	(0.000269)	(0.000269)
Listed firm	0.0503***	0.0503***	0.0503***	0.0511***	0.0509***	0.0500***	0.0500***	0.0478***	0.0478***
	(0.0110)	(0.0110)	(0.0110)	(0.0118)	(0.0118)	(0.0112)	(0.0112)	(0.0114)	(0.0114)
Log HHI (W)	9.28e-05	9.21e-05	9.68e-05	5.40e-05	5.36e-05	7.09e-05	7.10e-05	-2.99e-05	-3.45e-05
	(0.000123)	(0.000123)	(0.000124)	(0.000129)	(0.000129)	(0.000137)	(0.000137)	(0.000282)	(0.000282)
Log number of workers (B)	0.00344***	0.00344***	0.00343***	0.00335***	0.00334***	0.00337***	0.00337***	0.00401***	0.00402***
	(9.91e-05)	(9.92e-05)	(9.92e-05)	(0.000101)	(0.000100)	(0.000100)	(0.000101)	(0.000166)	(0.000167)
Log labour productivity (B)	0.00533***	0.00533***	0.00533***	0.00529***	0.00520***	0.00496***	0.00496***	0.00736***	0.00737***
	(0.000192)	(0.000192)	(0.000192)	(0.000199)	(0.000198)	(0.000190)	(0.000190)	(0.000282)	(0.000285)
Log HHI (B)	9.36e-05	9.04e-05	5.85e-06	-9.37e-06	5.71e-06	5.15e-05	4.89e-05	0.000474	0.000470
	(0.000337)	(0.000336)	(0.000342)	(0.000345)	(0.000345)	(0.000366)	(0.000366)	(0.000627)	(0.000627)
Union density (W)			-0.000972						
			(0.00266)						
Union density (B)			0.0221**						
			(0.00952)						
Share of male workers (W)							-0.000102		
							(0.000508)		
Share of male workers (B)							0.000222		
							(0.000305)		
R&D expenditure dummy (W)									0.00161*
									(0.000889)
R&D expenditure dummy (B)									-0.00222
									(0.00181)
Constant	-0.0735***	-0.0736***	-0.0751***	-0.0731***	-0.0715***	-0.0700***	-0.0701***	-0.104***	-0.104***
	(0.00296)	(0.00297)	(0.00301)	(0.00298)	(0.00296)	(0.00295)	(0.00296)	(0.00439)	(0.00440)
Sample subset	Full	Union	Union	Province	Province	Male share	Male share	R&D	R&D
Observations	669,879	668,870	668,870	516,143	516,143	550,597	550,597	289,795	289,795
Number of firms	135,558	135,458	135,458	133,237	133,237	122,177	122,177	89,765	89,765
Province FE	NO	NO	NO	NO	YES	NO	NO	NO	NO
Within R2	0.000664	0.000655	0.000656	0.000493	0.000524	0.000761	0.000761	0.00138	0.00140
Between R2	0.0424	0.0423	0.0423	0.0384	0.0391	0.0403	0.0403	0.0473	0.0473
Overall R2	0.0310	0.0310	0.0310	0.0299	0.0304	0.0307	0.0307	0.0421	0.0421

Robust standard errors in parentheses. All models include time and sector FE. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Source: author's own calculations.

C3 Worker level movements: switching firms and sectors, and moving at the top

Because of the panel nature of our data also at the individual level, we are able to track workers to analyse how their patterns of movement across firms and sectors correlate with their patterns of movement in and out of top quantiles. We are of course not able to claim any causality in the analysis that follows, but we do observe a co-occurrence of mobility at the top and lack of movement across firms or sectors which reinforces the predominance of within components.

We start with firms. In each observation, and in reference to the previous period, a worker may be in four different states in terms of their mobility at a top quantile: joining the top; leaving the top; staying in the top; staying out of the top. Similarly, there are five possible simultaneous stats in terms of firm mobility: staying in the same firm; switching firms; moving into the formal economy; moving out; or staying out. We focus on workers who stay in the formal economy (staying in the same firm or switching) to understand correlations of firm mobility and mobility at the top.

Each row of Table C10 shows, for a given quantile and a given state of firm mobility, the proportion of workers in each state of mobility at the top. We are interested in differences between the two groups for each quantile. For the top 5%, workers are more likely to move in or out of the top at the same time that they move across firms in comparison to staying in the same firm. This is also true for the top 1%, although for a much smaller margin. Particularly, the likelihood of *joining* the top 1% is already more associated with staying in the same firm rather than with switching firms. And this pattern becomes even stronger when looking at the top 0.1% or the top 0.01%

Table C10: Worker mobility across firms and top quantiles

Quantile	Firm mobility	Moves at top	*Joins top*	*Leaves top*	Stays at top	Stays out of top	Total
95	Stays in firm	1.708%	*0.989%*	*0.719%*	5.605%	92.688%	100%
95	Switches firms	2.458%	*1.267%*	*1.191%*	1.910%	95.632%	100%
99	Stays in firm	0.484%	*0.273%*	*0.211%*	1.058%	98.458%	100%
99	Switches firms	0.506%	*0.246%*	*0.260%*	0.292%	99.202%	100%
999	Stays in firm	0.078%	*0.043%*	*0.035%*	0.090%	99.832%	100%
999	Switches firms	0.052%	*0.021%*	*0.031%*	0.019%	99.930%	100%
9999	Stays in firm	0.010%	*0.006%*	*0.004%*	0.008%	99.982%	100%
9999	Switches firms	0.005%	*0.002%*	*0.004%*	0.001%	99.993%	100%

Source: author's own calculations.

Table C11 shows results for an analogous exercise for sectors. It reaches similar conclusions, although switching sectors seems less associated with movement at the top than switching

firms across all quantiles – indeed, we now see a greater proportion of workers who stay in the same sector while moving at the top already at the top 1%, both for joining and leaving the top.

Table C11: Worker mobility across sectors and top quantiles

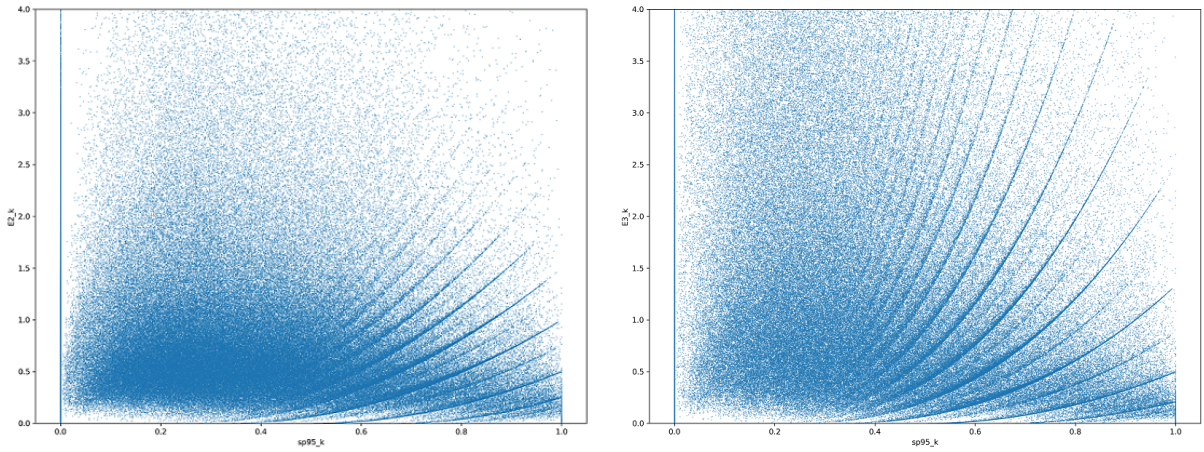
Quantile	Sector mobility	Moves at top	*Joins top*	*Leaves top*	Stays at top	Stays out of top	Total
95	Stays in sector	1.790%	*1.026%*	*0.764%*	5.493%	92.717%	100%
95	Switches sectors	1.970%	*1.038%*	*0.932%*	2.230%	95.800%	100%
99	Stays in sector	0.498%	*0.279%*	*0.220%*	1.036%	98.465%	100%
99	Switches sectors	0.408%	*0.202%*	*0.206%*	0.347%	99.246%	100%
999	Stays in sector	0.079%	*0.044%*	*0.036%*	0.089%	99.832%	100%
999	Switches sectors	0.037%	*0.015%*	*0.022%*	0.020%	99.943%	100%
9999	Stays in sector	0.010%	*0.006%*	*0.005%*	0.008%	99.982%	100%
9999	Switches sectors	0.004%	*0.001%*	*0.002%*	0.001%	99.995%	100%

Source: author's own calculations.

D Relations between firm-level measures: sPX$_{kt}$ and GE$(\theta)_{kt}$

When exploring the relations between firm-level inequality measures, an unexpected pattern was found. Figure D1 shows scatterplots for sP95$_{kt}$ and $GE(2)_{kt}$ on the left panel; and for sP95$_{kt}$ and $GE(3)_{kt}$ on the right panel[14], each dot representing one firm in one year. Although a cloud of points is visible, there is also a tendency for firms to gather in a family of curves.

Figure D1: Scatterplots for firm-level top share and generalized entropy: selected parameters



(a) sP95$_{kt}$ and GE$(2)_{kt}$ **(b) sP95$_{kt}$ and GE$(3)_{kt}$**

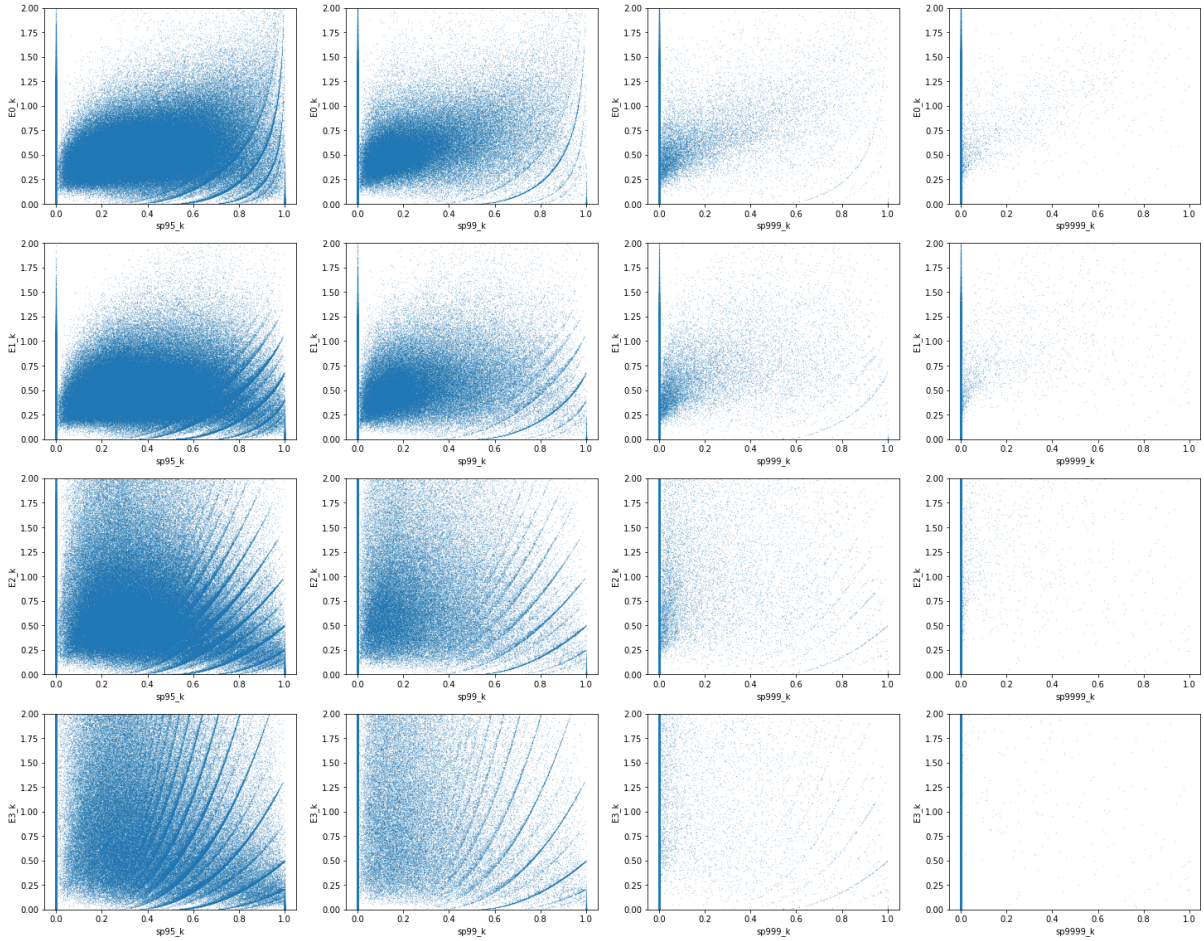
Source: author's own calculations.

The shapes of the curves are related uniquely to the parameter of the generalized entropy measure and independent from the quantile at which the top share is measured. This can be seen in Figure D2, which shows the scatterplots for combinations of $X = 95, 99, 999, 9999$ (one

[14] GE with parameter 3 was subsequently dropped from our analysis

in each column) and $\theta = 0, 1, 2, 3$ (one in each row). Both the cloud and the curves become less visible the greater the quantile because there are more firms with top share equal to zero.

Figure D2: Scatterplots for firm-level top share and generalized entropy: all parameters.



Source: author's own calculations.

We were able to show that the family of curves is related to the number of workers in the firm who lie at the top quantile[15]. Figure D3 shows the scatterplot for sP95 and GE(2) only for firms with 2 workers in total and in which one of these lie at the top 5%. We see that this allows us to isolate one of the curves seen in the full scatterplot. Additionally, by varying the number of total workers, and the number of workers at the top, we are able to draw additional curves from the full scatterplot, as shown in Figure D4. Also, if we filter firms such that the number of workers at the top is fixed but the total number of workers vary, we are further able to draw multiple curves at the same time, as shown in Figure D5.

[15] This hypothesis was originally formulated by Bart Verspagen.

Figure D3: Scatterplots of sP95 and GE(2) only for firms with 2 workers in total, and 1 at the top.

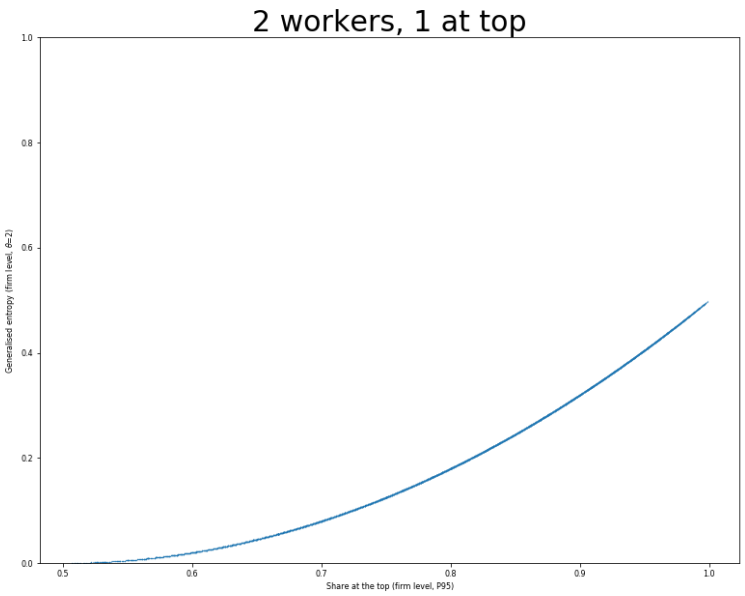


Figure D4: Scatterplots of sP95 and GE(2) filtered for specific numbers of total workers, and numbers of workers at the top.

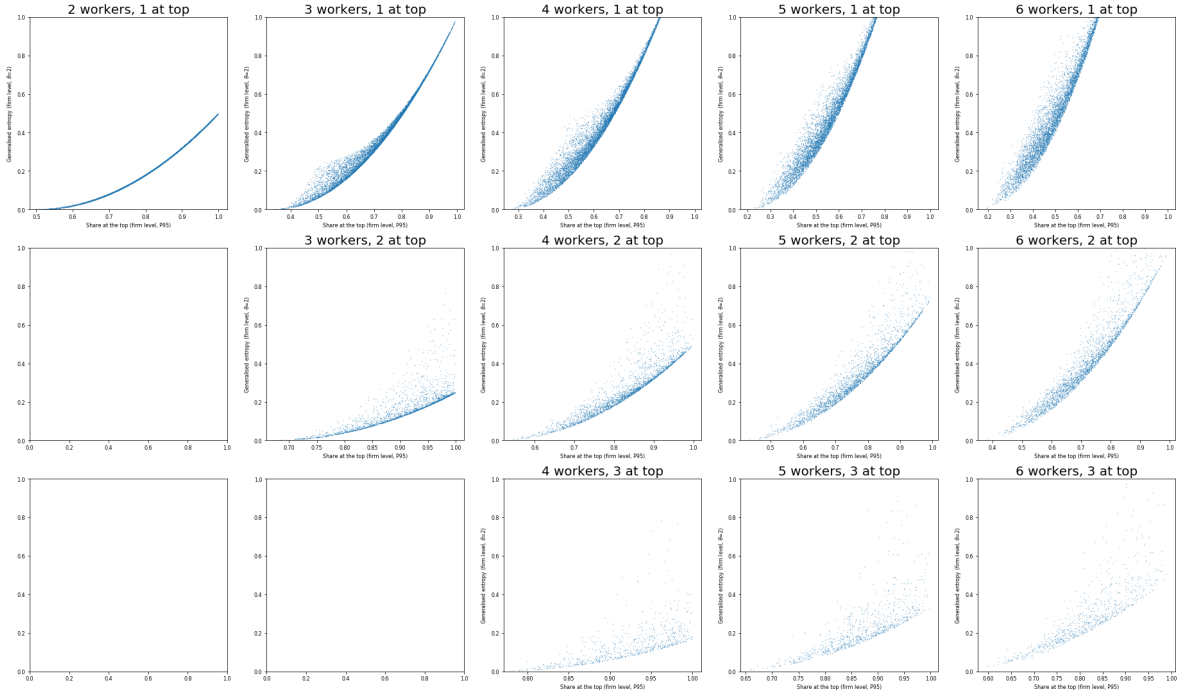
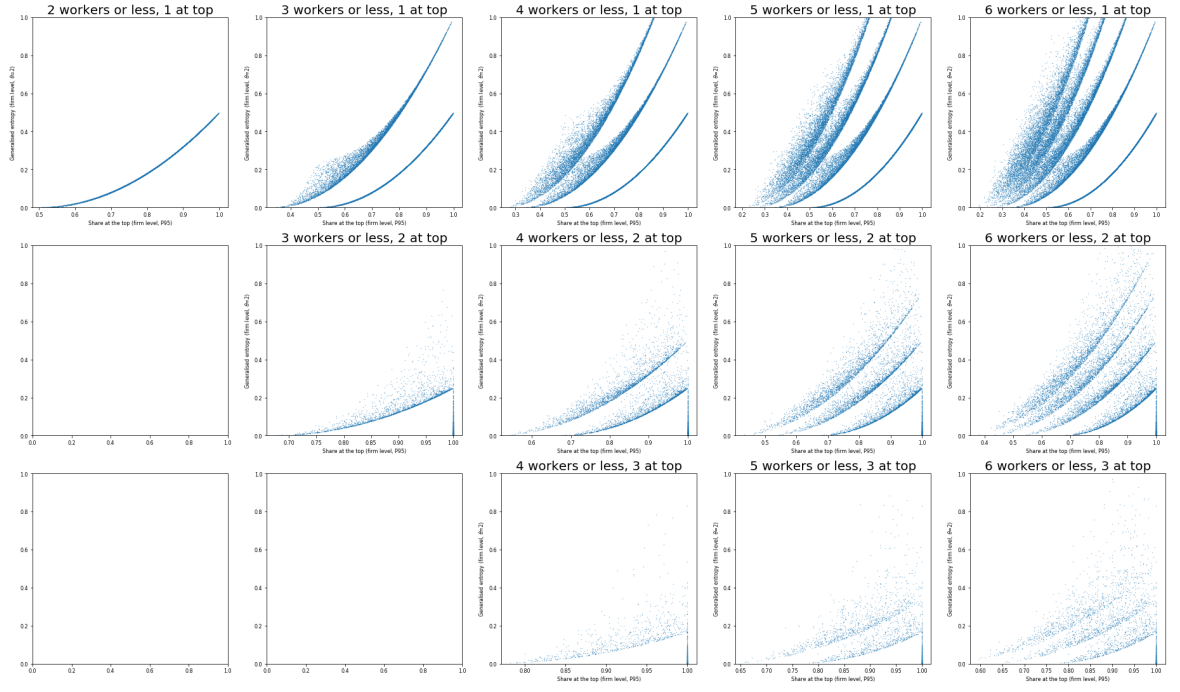


Figure D5: Scatterplots of sP95 and GE(2) filtered for specific upper bounds in total workers, and numbers of workers at the top.



Finally, we set up simulations drawing individual incomes according to different long-tailed distribution functions, and randomly allocating workers in firms with skewed distributions of firm sizes. We were able to observe that the occurrence of such curves in the scatterplots is contingent on the chosen distribution function and its parameter: in some combinations the curves are formed and in other the curves are not there. Further exploration is needed to understand systematically which combinations of functions and parameters give rise to the curves in the scatterplots of these within-firm inequality measures, and to understand the details of why this is so. While this is out of the scope of this work, we believe this finding could be used to help in the selection of functions to model the distribution of individual earnings. If such a pattern is found in empirical data, only combinations of distributions and parameters that give rise to the pattern should be chosen to model the data.