GOSPODARKA NARODOWA

3(315)2023, 13–33
DOI: 10.33119/GN/170227
gnpje.sgh.waw.pl
SGH

The Polish Journal of Economics
Istnieje od / Published since 1931

**Sebastian Zalas**
FAME | GRAPE, University of Warsaw, Poland
(*corresponding author*)

**Hubert Drążkowski**
FAME | GRAPE, Warsaw University of Technology, Poland

# The Evolution of the Labour Share in Poland: New Evidence from Firm-Level Data*

## Kształtowanie się udziału płac w wartości dodanej w Polsce. Nowe szacunki z danych jednostkowych

**Keywords:**
missing data, labour share, firm-level data

**JEL classification codes:**
C81, E25, D33

**Article history:**
submitted: December 28, 2022
revised: July 25, 2023
accepted: August 1, 2023

### Abstract

We evaluate the usefulness of non-representative registry data such as Orbis in drawing inferences about economic phenomena in Poland. While firm-level studies of economic phenomena are of key policy relevance, census data and representative samples are scarcely available across countries. We obtain estimates of the labour share for the period 1995–2019. For the overlapping period and samples, we compare our estimates to Growiec [2009], who drew on a census of Polish firms employing 50+ employees. We also refer to OECD STAN data. We demonstrate that time patterns are common across data sources. Additionally, we study the potential for various imputation methods to enrich inference.

**Słowa kluczowe:**
braki danych, udział płac w wartości dodanej, dane jednostkowe

**Kody klasyfikacji JEL:**
C81, E25, D33

**Historia artykułu:**
nadesłany: 28 grudnia 2022 r.
poprawiony: 25 lipca 2023 r.
zaakceptowany: 1 sierpnia 2023 r.

### Streszczenie

Oceniamy przydatność nierepreprezentatywnych danych jednostkowych o firmach (Orbis) do wnioskowania o procesach gospodarczych w Polsce. Reprezentatywne dane jednostkowe nie są w Polsce dostępne do badań naukowych. Korzystając z dostępnych badań Growca [2009], dotyczących udziału płac w wartości dodanej w latach 1995–2008 w firmach zatrudniających ponad 50 pracowników, skupiamy się na tej samej kategorii ekonomicznej. Rozszerzamy zakres badania do 2019 r. oraz poszerzamy grupę przedsiębiorstw o firmy zatrudniające mniej niż 50 pracowników. Nasze oszacowania są podobne do oszacowań Growca [2009]. Wskazujemy także na wzrost udziału płac w wartości dodanej, szczególnie w ostatniej dekadzie oraz w mniejszych przedsiębiorstwach.

* The authors gratefully acknowledge the funding received through Norwegian Financial Mechanism 2014–2021 (grant # 2019/34/H/HS4/00481, within the GRIEG framework of National Science Center). We are grateful to Jakub Growiec for graciously sharing his data. Earlier versions of this study received useful comments from Joanna Tyrowicz, Magdalena Smyk-Szymańska and Lucas Van der Velde. We also thank the participants of the (Ce)^2 Workshop 2022, the 11th Professor Zbigniew Czerwiński Scientific Conference *Matematyka i informatyka na usługach ekonomii* and 11th Summer Workshop on Macroeconomics and Finance for their helpful comments. The remaining errors are ours.

## Introduction

We study the evolution of the labour share in Poland by utilising a novel source of firm-level data, the so-called Orbis data. Poland is notorious for its low and declining labour share [**Dimova, 2019**]. According to Eurostat,[1] the statistical agency of the European Union, Poland ranks roughly 20th in the bloc in terms of labour share changes. **Kónya, Krekó, and Oblath** [**2020**] show that labour shares across Central and Eastern Europe are lower than in Western Europe, with a steady decline in manufacturing and non-monotonous trends in other sectors. These conclusions notwithstanding, a large body of literature warns against the perils of estimating labour shares from macroeconomic aggregates. In particular, self-employment and agricultural employment pose important methodological challenges [**Kónya et al., 2020**].. These are particularly relevant in the case of Poland. Our study draws on the rich and growing literature providing micro-level evidence concerning macroeconomic indicators (e.g., **Cavallo, Rigobon** [**2016**]). We provide labour share estimates obtained from firm-level data.

The Orbis data is readily available for research purposes, which makes it a potentially valuable source in empirical analyses. While used in international studies [**Bruno et al., 2021**; **Kalemli-Ozcan et al., 2022**], Orbis data remain underutilised for the study of the Polish economy. We contribute to the burgeoning literature on the evolution of the labour share. **Karabarbounis and Neiman** [**2014**] demonstrate substantial declines in labour shares worldwide. This trend prevails regardless of the ambiguities regarding the adequate measurement of labour shares from macroeconomic data [**Mućk, McAdam, Growiec, 2018**]. Analysing the case of Poland we find a contrasting trend. We report that the labour share in Poland, after a temporary decline in the mid-2000 s, began rising and reached a level similar to that at the beginning of the 2000s. We also document the labour share in industries and in groups of firms classified by employment size (fewer than 50 and 50+ employees). We find that the labour share is higher in services than in manufacturing. Furthermore, we document that the labour share in firms with fewer than 50 employees is lower in all the analysed years than in firms with 50 and more employees.

Unlike registry data, Orbis data is not constructed as a representative sample, hence its viability for research purposes may be questioned. To tackle this concern, we compare our estimates with the existing literature, notably a study by **Growiec** [**2009**] that utilises firm registry data from the Statistics Poland (GUS) agency for firms employing 50 and more workers. To the best of our knowledge, this registry data is not available for research purposes, except for internal researchers at Statistics Poland and the National Bank of Poland (NBP). Through comparing our estimates with **Growiec** [**2009**], we critically evaluate the usefulness of the Orbis data for studying the Polish economy. We are also able to extend the analysis of **Growiec** [**2009**] by providing estimates for recent years and companies employing fewer than 50 workers. Additionally, we compare Orbis data to aggregate data from the OECD to complete the comparison, since **Growiec** [**2009**] estimates end in 2009. Despite the differences in the labour share levels from Orbis, Growiec and the OECD, we observe a similarity in labour share evolution.

Finally, we discuss and critically evaluate the viability of imputation methods for improving the quality of inference. Although our sample is constructed in such a way that we possess fully observable information on value added and labour costs, which allows for labour share estimation, we have direct information on employment for only about half of the sample. For better comparability of the samples between sources, we perform an imputation study, which also allows us to validate the robustness of our results. We infer that the missingness mechanism is not Missing Completely at Random (MCAR). By proposing imputation methodology, we allow researchers to tackle the problem of non-uniform random gaps in data. We test the methods under a simulation by looking at prediction errors made for the observable years with a scheme using a train

[1] Rank of labour share changes in 2002–2022 from Eurostat.

test split of that data under MCAR and Missing at Random (MAR) missingness mechanisms. Completing our sample by adding observations with imputed observations does not change any of our conclusions.

The paper is structured as follows. The next section describes the relevant literature, with a particular focus on the implications for our analysis. In section three, we describe in detail the features of our data. Section four describes the results using raw Orbis data. We extend our work in section five by presenting alternative data imputation strategies and comparing estimates from the raw data with estimates that also include imputed observations. The paper concludes with key facts about the evolution of the labour share in Poland. We also discuss the implications for researchers intending to use Orbis data for subsequent research.

## Literature

The evolution of the labour share, that is the fraction of gross domestic product allocated to wages (labour), has been widely debated in economic literature in recent years. **Kaldor [1961]** states the stability of the labour share in one of his famous stylised facts of economic growth. Constancy of the labour share is vital for the applicability of the Cobb-Douglas production function in economic theory, as well as for society, since the fraction of the population profiting from economic activity is decreasing. Thus we examine how the labour share changed in Poland.

**Declining labour share.** Literature documenting cross-country evidence on the labour share shows that many countries experienced a decline in the labour share at some point. **Karabarbounis and Neiman [2014]** analyse data on 59 countries from the UN and the OECD between 1975 and 2012 and document that 42 countries experienced a decline in the labour share. **Karabarbounis and Neiman [2014]** observe that the labour share declined among the largest economies, such as the United States, China, Japan and Germany. In these countries, the labour share decreased by 2 to 4 percentage points every 10 years. Likewise, **Dao, Das, and Koczan [2020]** study global changes in the labour share from 1991 through 2014 and confirm the findings from **Karabarbounis and Neiman [2014]**. **Dao et al. [2020]** document that the labour share declined in 29 of the largest 50 economies. In countries featuring decreases, according to **Dao et al. [2020]**, the labour share declined by 2 percentage points after 10 years on average. Later, **Dimova [2019]** reported a decline in the labour share among half of EU countries between 2002 and 2016. In these years, the changes in the labour shares in most countries ranged between –3 to 3 percentage points. However, in four of the new EU countries, **Dimova [2019]** documented significant increases in the labour share, exceeding 4 percentage points. On average, the labour share in the EU declined by around 1 percentage point. **Charpe, Bridji, and McAdam [2020]** present a long-run perspective on the labour shares for France, the United States and the UK, dating back to the 19th century. For instance, **Charpe et al. [2020]** show that the labour share in France declined in the mid-1980 s and then remained stable, while in the United States and the UK it has been gradually diminishing since the 1980s. Although there are many countries with a more pronounced labour share decline, in some countries the labour share declined only temporarily or increased. Several authors focusing on individual countries presented evidence for the stability of the labour share in their studies. In line with **Charpe et al. [2020]**, **Bauer and Boussard [2020]** obtained a labour share for France from both microdata and aggregate data and reported that it has remained stable since the 1990s. In their exploration of a representative sample of firms, **Siegenthaler and Stucki [2015]** report that, for the years studied, the labour share in Switzerland was unchanged. **Kónya et al. [2020]** study the evolution of the labour share focusing on Central and Eastern European EU member states, including Poland. **Kónya et al. [2020]** find no evidence of a systematic decline in the labour share in non-agricultural sectors. **Kónya et al. [2020]** observe differences between sectors and find a sustained fall in the manufacturing labour share, much like **Dimova [2019]** and **Dao et al. [2020]**. Our paper adds updated evidence for Poland.

The documented changes in the labour share were not economy-wide. They were driven mainly by manufacturing sectors in the broad sense. For instance, **Dao et al. [2020]** analyse changes in the labour share by

industry and find that the strongest decreases in the labour share occurred in manufacturing, followed by transportation and communication, while some sectors (food and accommodation, and agriculture) experienced an increase. **Dimova [2019]** also observes that the labour share in most EU countries declined strongly in manufacturing and construction but rose in service sectors. In the case of the frequently analysed US labour share, **Kehrig and Vincent [2021]** use US census data to report that the labour share in manufacturing fell by 20 percentage points from 1967 to 2012. According to **Smith et al. [2022]**, the drop in the US labour share between 1987 and 2017 occurred mainly due to an 8-percentage point decline in the manufacturing sector. These findings are in line with global evidence that the labour share decline is most pronounced in manufacturing sectors. Our work also investigates changes in sectors to capture cross-sector heterogeneity.

**Use of microdata.** In empirical studies of the labour share, an important distinction involves the level of analysis. The availability of firm-level data inspired researchers to investigate the causes of the labour share decline. Exploration of microdata revealed the importance of micro-level frictions for shaping macro-level changes in the labour share. In Poland, for instance, **[Growiec, 2009]** exploits a panel of firms and finds that 55% of the observed changes in the labour share in Poland occurred due to within-sector factors, and that reallocation effects account for the remaining proportion. **Böckerman and Maliranta [2011]** show the effects of globalisation on the labour share by exploiting microdata from Finland. **Kehrig and Vincent [2021]** and **Autor et al. [2020]** capture reallocation processes in US manufacturing and empirically investigate the so-called *superstar firm* hypothesis. **De Loecker, Eeckhout, and Unger [2020]** find evidence for a link between rising markups and a declining labour share in the United States on a sample of publicly traded firms from COMPUSTAT. In Germany, **Mertens [2022]** use a 20-year firm-level dataset from manufacturing to study the impact of market power on the labour share. We follow this trend by inspecting the labour share from firm-level data, as we can capture more between-firm heterogeneity. Although we do not propose an explanation for the evolution of the labour share in Poland, we provide researchers with an assessment of a publicly available firm-level database. Access to microdata is vital for enhancing insightful research.

## Data

In this section, we describe our data and the process of creating a final sample. We start with data origins. We subsequently describe variable definitions, sample sizes and the distributions for the variables of interest.

**Data origins.** Orbis data consist of registry data, balance sheets and profit-loss statements submitted by firms to registry courts and local government statistical offices. These data are collected by InfoCredit and subsequently digitised.[2] Given the data collection strategy, only firms subject to mandatory reporting are available in Orbis data. For example, self-employed individuals with low turnover are not subject to mandatory reporting. Among the firms that submitted the reports, especially in the 1990 s and early 2000 s, some of the reports were filled by hand or typewriter and thus digitisation was obscured. The growing popularity of computers gradually increased the share of fully legible reports.

**Firms covered.** We utilise nine editions of Orbis data: 2000, 2002–2004, 2006, 2008, 2010, 2014, 2016 and 2020. Until 2019, each Orbis edition contains firm-level financial information, which can go up to 10 years back. As of 2020, both annual data or the so-called historical samples, which provide the entire information available for a given firm, can be acquired from the provider. The firms are uniquely identified. For Polish firms, the ID is based on the REGON identification number, which permits linking the data with other registries. The data typically cover the period without the most recent year due to data collection before reporting deadlines.

[2] As of 2018, data is submitted to registry courts in electronic form, which permits InfoCredit to obtain new data directly, without the need to digitise paper records. GDPR implementation as of 2019 forced InfoCredit to obtain explicit consent prior to data collection, which poses a challenge to data on owners, board members and other named stakeholders.

## Processing data

The firms report consolidated statements, unconsolidated statements or both. Overall in the Orbis data, most firms report unconsolidated accounts, which is useful for aggregating within sectors, as we do in this study. Then the risk of aggregating the same value added or employment twice is eliminated. Occasionally, the type of reported standards varies within the firm over time. In some years, unconsolidated accounts are not available, but consolidated ones are. For each firm, we count how many annual observations are available for consolidated and unconsolidated statements and select the one which guarantees a longer panel. The problem of multiple reporting due to the presence of consolidated and unconsolidated statements concerns less than 1% of all observations.

Each wave of Orbis data covers a 10-year window. Consequently, it may occur that data for a given financial year are reported in more than one of the available waves. If the values are identical, this redundancy is immaterial. If the values are missing in one wave, but are available in another, we can lengthen the within-firm panel. In case of discrepancies, we select the data from the wave which is the closest to the year at hand.

**Harmonising industries.** The Orbis data reflect the four-digit NACE classification of economic activities. Our data cover the years 1995–2019. During this period, the NACE classification changed twice: Rev. 1 was replaced by Rev 1.1, which was followed by Rev 2.0. This is not an issue in the case of firms observed throughout the entire window. The change in the NACE classification is also immaterial in the case of firms observed only under one classification. However, in some cases, the firm appears in Orbis under the newer classification, but its retrospective data overlap with the period when the older classification was used. For aggregation purposes, we must provide the older NACE codes for the years before a change in classification (s). We apply unique crosswalks whenever they are available. For the cases where crosswalks are many-to-many, we review the area of a firm's activity and assign the adequate classification from among the relevant options. For some firms, the NACE classification was provided at two or three digits rather than the full four-digit classification. In those cases, we assigned the appropriate two-digit code in the older classification.

Our final sample consists of firms in manufacturing (sections 10–43 of NACE Rev. 2) and services (sections 45–99 of NACE Rev. 2)[3].

**Units of observation.** The financial statements in Orbis are reported in USD or EUR, depending on the wave, rounded to thousands. We convert the reported figures to PLN using the exchange rate provided by Orbis. Employment is reported in terms of headcount at the date of reporting, without adjustment for full-time full-year equivalents. Consequently, employment may be overstated in Orbis, relative to the national accounts as well as firm registry.

## Final sample

To measure the labour share, we require payroll and value added. We compute the labour share as the ratio of payroll to value added. After merging nine waves of Orbis, we can obtain value added and payroll for approximately 180,000 firms with nearly 720,000 observations.

For the sake of our analysis and in the interest of comparing our estimates to **Growiec [2009]**, we need to identify firms with 50+ employees. We thus require employment data, which is missing in roughly 52% of the records for which value added and payroll are available. Ultimately, 350,000+ firm-year observations with reported employment are available. The employment data is particularly frequently missing in the 2010–2015 period; see Figure A1 in the Appendix. To contain the role of this data shortcoming in our inference, we use available information to fill in the missing employment data. We classify a firm as having 50 or more employees if a firm in its available history contains employment values equal to or exceeding 50. Otherwise, we classify

[3] We exclude observations featuring the following NACE rev. 2 sections: agriculture, mining, financial and insurance, health, education, public administration and social security, activities of households as employers, and activities of extraterritorial organisations and bodies.

firms as having fewer than 50 employees if the observed number of employees is below this threshold each time. When a company reported employment values both above and below or equal to 50, we only classify those observations for which employment is available.

The final data processing consisted of removing outliers. We drop observations with negative payroll, value added, turnover or employment. We also trim the sample by one percentile from both sides of the capital-to-labour ratio. Next we apply 1% winsorizing procedure in each year to payroll, value added, turnover and total assets and average compensation, calculated as the ratio of payroll and employment. Finally, we keep only those observations for which we can calculate the labour share.

Table 1 summarises the final sample data and across size groups for selected years. In the first part of Table 1, we present descriptive statistics. Initially, we show the number of observations. Our sample contains only 560 observations in 1995, then the size of our sample consequently rises. By 2019, our sample counts over 100,000 observations. We also report how many observations do not possess any information on size. In total, approximately 150,000 observations cannot be assigned to any size groups. In the Imputation section, we describe how to proceed with imputation to attribute the size information to all the available observations. Later we show the means of added value, payroll and employment. As the number of available observations grows, mean employment, value added and payroll decrease due to the influx of small companies. For our analysis, we use a sample consisting of roughly 570,000 observations (with size information), including 118,000 unique firms.

**Table 1. Summary statistics**

| | 1995 | 2000 | 2005 | 2010 | 2015 | 2019 |
|---|---|---|---|---|---|---|
| I. Descriptive statistics | | | | | | |
| Number of observations | | | | | | |
| all | 563.00 | 4597.00 | 16185.00 | 22108.00 | 71377.00 | 106928.00 |
| 50+ | 504.00 | 3143.00 | 6261.00 | 6469.00 | 7566.00 | 9908.00 |
| 50- | 28.00 | 235.00 | 1677.00 | 3339.00 | 20624.00 | 33273.00 |
| none | 31.00 | 1219.00 | 8247.00 | 12300.00 | 43187.00 | 63747.00 |
| mean Added Value | | | | | | |
| all | 9972.00 | 9094.00 | 4586.44 | 4465.98 | 2418.32 | 2074.18 |
| 50+ | 10231.69 | 11352.93 | 8787.61 | 10182.79 | 10043.42 | 10917.96 |
| 50- | 2541.26 | 3918.28 | 1786.50 | 1857.64 | 1221.89 | 1402.94 |
| mean Payroll | | | | | | |
| all | 4943.54 | 5129.92 | 2429.01 | 2519.62 | 1357.76 | 1200.89 |
| 50+ | 5230.04 | 6876.36 | 4970.90 | 6110.72 | 6171.79 | 7032.04 |
| 50- | 765.24 | 1125.40 | 718.55 | 852.23 | 603.36 | 720.59 |
| mean Employment | | | | | | |
| all | 504.45 | 190.33 | 91.28 | 91.22 | 20.28 | 30.52 |
| 50+ | 530.36 | 245.30 | 172.87 | 161.19 | 161.83 | 124.88 |
| 50- | 31.04 | 24.25 | 20.00 | 20.72 | 6.98 | 12.85 |
| II. Coverage | | | | | | |
| Number of firms: Orbis/Statistics Poland | | | | | | |
| all | no data available from Statistics Poland | | 7.14% | 9.12% | 26.83% | 36.78% |
| 50- | | | 4.56% | 6.63% | 25.06% | 35.43% |
| 50+ | | | 38.21% | 37.67% | 49.24% | 55.93% |

Notes: In the first part, descriptive statistics are computed on Orbis dataset using waves from 2000, 2002, 2004, 2006, 2008, 2010, 2014, 2016 and a historical sample from 2020. Value added and payroll are expressed in thousands of PLN. Employment is expressed as the number of workers. In the Coverage section, we compare the number of firms included in Orbis to the number of firms covered by the annual business census carried out by [**Statistics Poland, 2020b**].

Source: Authors' own elaboration.

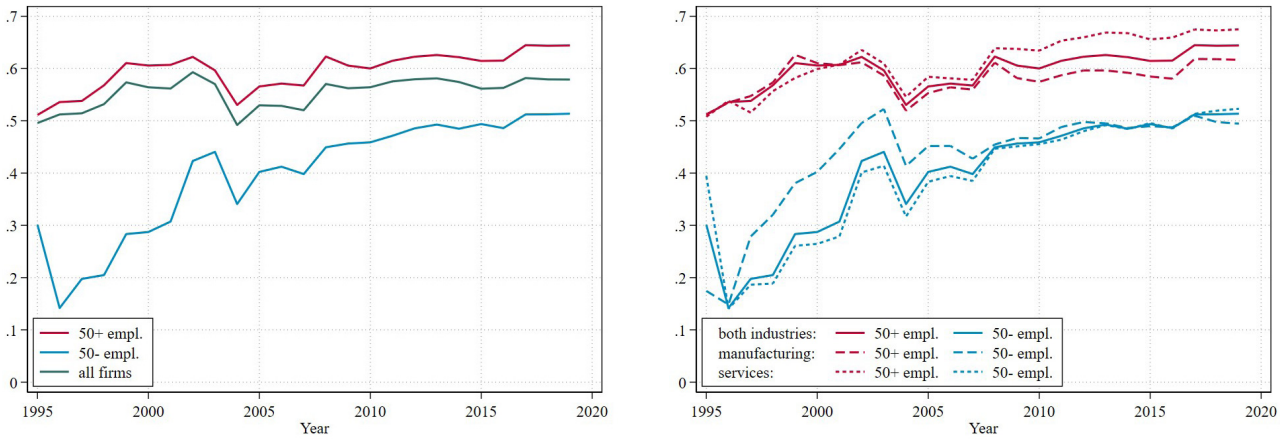
In the second part of Table 1, we present the coverage of our data. Our sample contains notable parts of Polish firms with substantial representation of firms with 50+ employees. We show the number of firms as the percentage of firms included in data collected by Statistics Poland in its surveys. Statistics Poland performs business surveys to collect data on all companies with 50+ employees as well as a substantial portion of firms with between 10 and 49 employees.

Moreover, the Polish statistics office collects data on roughly 10 percent of firms with up to nine employees [**Statistics Poland, 2020a**]. Since we have data on the overall number of firms in each size category, we are then able to evaluate how many firms are included in census surveys. We can compare the size of Orbis and Statistics Poland data only from 2004 and onwards, since earlier data were not available. In general, in the available years, our sample contains 7 to 37 percent of what Statistics Poland collects. For 50+ firms, the percentage of firms included in Orbis relative to Statistics Poland ranges from 30 percent to over 50 percent in the most recent years. For firms with fewer than 50 employees, our sample has between 4 and 35 percent of the number of firms included in official surveys.

## The evolution of labour share

Since we have access to firm-level data across sectors and time, we can contrast the evolution of the averages and the distribution of the labour share. We first report the aggregated labour share. Then we juxtapose our labour share against estimates obtained from industry-level database and other firm-level labour share estimates. Finally, we show some features of the labour share distribution.

**Figure 1. Labour share**




(a) by size (b) by industry

Note: In this figure, the labour share is presented based on size and industry. The labour share is computed as the ratio of the sum of payroll at a given level and the sum of value added at the same level (e.g. in manufacturing).

Source: Authors' own elaboration.

**Labour share from Orbis data.** Figure 1a reports the evolution of the average labour share weighted by the share of value added across the whole economy and for companies with both more and fewer than 50 employees over time. In the beginning of the sample period, the labour share increases and achieves a level of 0.6 by around the year 2000. Then a decline starts and the labour share drops to 0.5 in 2004. Next, after a few years of depression, the labour share slowly rebounds and at the end of the sample it almost reaches levels matching those in the early 2000 s. The labour share for large companies follows the same evolution but its level is higher by 0.3–0.4. The labour share among companies with fewer than 50 employees features a different evolution. First, it has much lower levels in comparison with companies with 50+ employees. Second, from the early years in the sample it rises consistently, excluding a temporary decline around 2004.

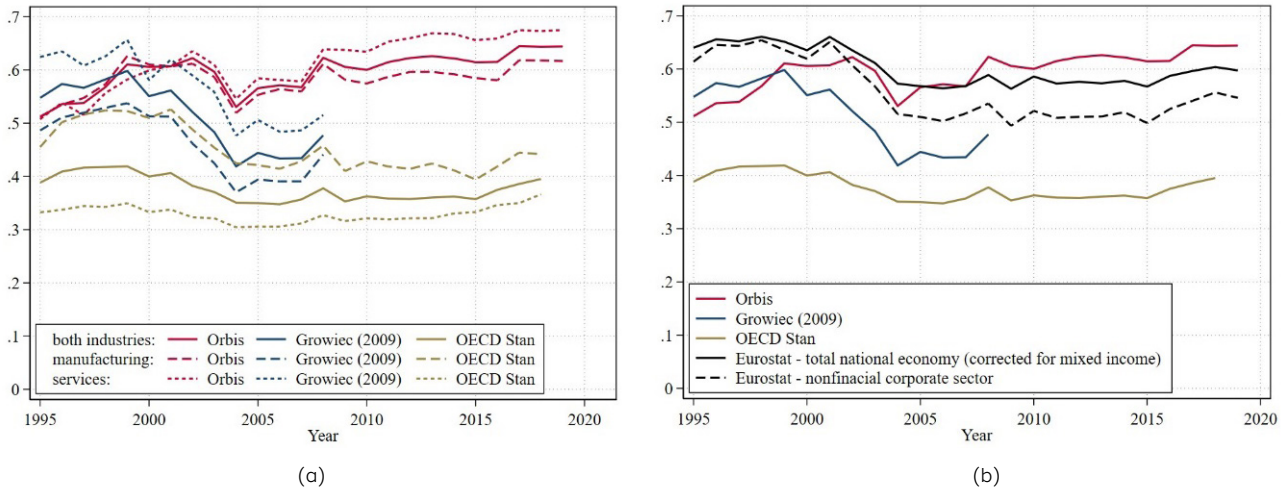
Furthermore, we explore differences in the labour share by size of company and industry, as presented in Figure 1b. Among firms with 50+ employees, the labour share features different behaviour after 2008 depending on industry. In services, the labour share increases and exceeds levels from the early 2000s, while in manufacturing, the labour share increases after a depression in the mid-2000s. It does not, however, rebound to its highest level from the early 2000 s. We also observe differences by industry for companies with fewer than 50 employees. In services, the labour share closely follows the overall labour share for companies with fewer than 50 employees. In manufacturing, the growth of the labour share in the beginning of the sample was more pronounced than the overall index shows. The labour share among manufacturing companies with fewer than 50 employees did not recover after the decline in 2004, although it has been increasing in recent years.

In general, the evolution of the labour share is driven by companies with 50+ employees, as they make up a larger share of the economy in terms of added value. Companies with 50+ employees feature much higher labour share levels than firms with fewer than 50 employees. Industry comparisons show that the overall labour share in services in the last years of the sample is at its highest levels, while in manufacturing the labour share still needs to make up for the decline in the early 2000s.

**Comparing labour share estimates in Orbis and other data sources.** The next step we take is to compare our labour share estimates to other available data. The only research which presents labour share estimates from firm-level data is **Growiec [2009]**. Since the estimates presented by **Growiec [2009]** end in 2008, we use industry-level data from OECD STAN[4] data to benchmark later years in our sample. We show this comparison in Figure 2a.

First, we compare our estimates with **Growiec [2009]**. There is a noticeable difference in levels in all the categories shown (manufacturing, services and the overall trend). Still, the labour share estimates from both sources follow a similar course. For instance, despite the difference in magnitudes, Orbis data shows a decline in the labour share between 2001 and 2005, in line with **Growiec [2009]**. Second, because of the absence of data after 2008 in **Growiec [2009]**, we compare the rest of our estimates to OECD STAN data. Again, the time trends for Orbis and OECD STAN are similar although the levels reported by OECD STAN are about 0.2 lower. Moreover, in recent observed years, the labour share from OECD STAN shows a slight but stable increase, which is also observed in the Orbis data.

**Figure 2. Labour share: Orbis vs. Growiec [2009], OECD STAN and Eurostat.**




(a) (b)

*Note:* We compare the labour share estimates from Orbis and **Growiec [2009]** with the indicators from OECD STAN industry-level data and with Eurostat Non-Financial Annual Sector Accounts. In order to make the comparison, the indices presented in Orbis are estimated on a sample of large companies (with 50+ employees), thus matching the census used by **Growiec [2009]**. OECD STAN comprise national accounts and business survey data.

Source: Authors' own elaboration.

4 We compare our estimates to OECD STAN. However, there are other available sources of industry-level data, such as EU KLEMS and Eurostat. These sources give almost identical labour share estimates as OECD STAN. This is documented in Figure A2 in the Appendix.

The differences between the labour shares from Orbis, **Growiec [2009]** and OECD STAN occur perhaps due to Orbis sample properties. As pointed out in **Bajgar et al. [2020]**, Orbis, as compared with nationally representative micro-data, only partially covers firm populations, and the distribution of firms in Orbis is skewed towards specific types of firms. Because of the partial coverage, Orbis has limited ability to reproduce indices computed from official aggregate statistics. In comparison with **Growiec [2009]**, who worked with a census of firms with 50+ employees, our sample underrepresents the population of firms, which should explain the observed differences.

For illustrative purposes, in Figure 2b, we compare the estimates obtained from firm-level data to estimates from macroeconomic aggregates. Note that the macroeconomic aggregates refer to a substantially different unit of the economy. The solid black line refers to the total economy, adjusted for mixed income. The black dashed line refers to the non-financial sector in the national accounts data. Due to data shortages, this estimate cannot adjust for mixed income.[5] The labour share estimates in the non-financial sector decline substantially more than in the case of the total economy in the first half of the 2000s. This decline is bigger when compared to the evolution of the enterprise sector as reported by **Growiec [2009]** as well as in our data. Also, the recovery occurs later and is substantially smaller than in our data. Note, however, that these estimates cannot be directly compared. First, our estimates are limited to the enterprise sector, whereas the national accounts cover the public sector as well. This difference affects the denominator. Public sector employees add to enterprise sector employment, without a commensurate addition in the numerator. Furthermore, our data includes firms with a labour share in excess of 1 due to negative profits, whereas the national accounts aggregate value added and employment before obtaining the ratio. We discuss this issue later.

**The role of aggregation: weighted vs. unweighted.** So far we studied the aggregate labour share, a measure which presents the ratio between the aggregate labour cost and aggregate value added. This measure gives greater weight to the labour share in larger firms in terms of both employment and value added. This measure is not sensitive to several important features occurring at firm level. First, firms that exhibit a loss in a given year may mechanically display a labour share in excess of 1, which is clearly not micro-founded. This is relevant if firms engaged in carry-forward optimisation of profits over years. Second, the standard aggregate measure is not susceptible to structural and cyclical fluctuations in employment, e.g. reallocation of workers between firms with varying levels of efficiency. To address this issue, we exploit the fact that we work with firm-level data and present an *unweighted* average of firm-level measures of the labour share. This measure is juxtaposed with the standard aggregate measure in Figure 3.

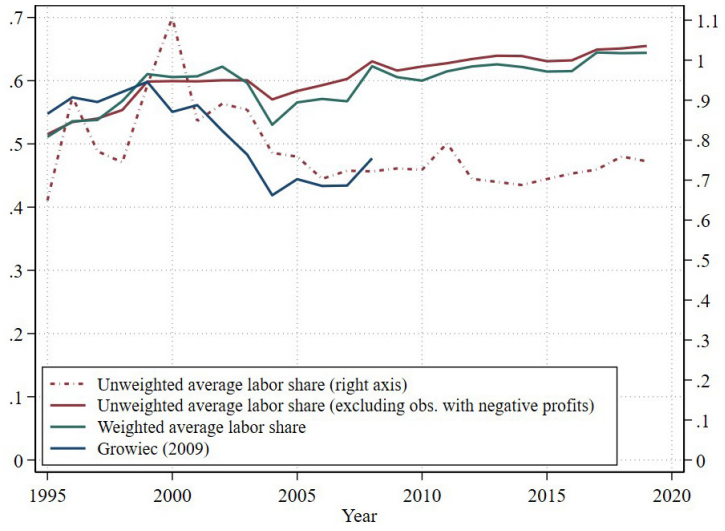
First, we observe that the phenomenon of firms with negative or low profits is prevalent. In 2000, for example, the mean labour share significantly exceeded 1. This measure permanently fluctuates around 0.8, as portrayed by the brown dashed line on the right axis. Once the sample is restricted to exclude observations with negative profits, aggregate (weighted) and unweighted measures become very close and have roughly the same levels and very similar time trends (the green and brown solid lines on the left axis). Interestingly, it is also the case that our results for the first four years from the restricted sample were virtually identical to the unweighted average from **Growiec [2009]**. This is strong evidence that in the first years of Orbis, this sample reflected firms with 50+ employees, while smaller firms became more prevalent in the sample in the late 1990s. A similar trend is observed when we study the manufacturing and service sectors separately (see Figure A3 in the Appendix).

The difference between the weighted (aggregate) and unweighted mean labour shares occurred due to changes in the labour share among the largest firms. This supposition is supported by Figure 4, which shows the labour share across time and some percentiles of the value added distribution. First, there is a striking difference between the labour share in the 25th and 90th percentiles. The labour share across high value-added companies is lower than in low value-added companies. Second, the observed difference between the 25th

[5] Note that the reported STAN data reflect the national accounts without adjusting for mixed income, when restricted to the same sectors as covered by our study.

and 90th percentiles was stable until the beginning of the 2000 s and then expanded in the mid-2000s. This suggests that the difference between the weighted and unweighted averages is explained by the fact that the labour share among the largest firms declined in comparison with smaller companies. In the 2010s, the difference between the labour share in the 25th and 90th percentiles of value added remained steady or diminished in comparison with the 2000s, and in both percentile groups the labour share increased symmetrically. This resulted in a smaller difference between the weighted and unweighted mean labour shares in the 2010s.

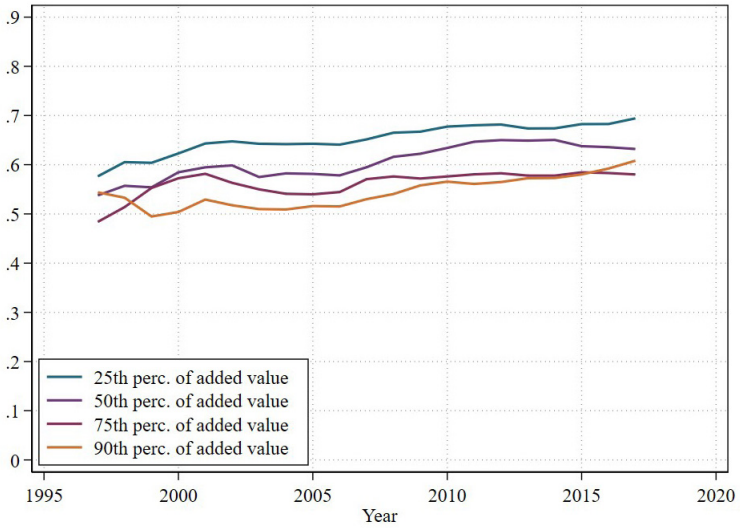
**Figure 3. Weighted vs. unweighted average labour share**




Note: This figure presents the unweighted average labour share, unweighted average labour share excluding observations with negative profits, and the average labour share weighted by the share of value added. These indices from Orbis were computed on a sample of large companies (with 50+ employees). We also add estimates from Growiec [2009].

Source: Authors' own elaboration.

**Figure 4. Labour share by percentiles of added value**




Note: This figure presents the labour shares in the 25th, 50th, 75th and 90th percentiles of value added. All four indices are smoothed with a five-year moving average.

Source: Authors' own elaboration.

Overall, exploring firm-level measures in addition to aggregate labour share measures reveals that aggregation is not necessarily innocuous. On the one hand, aggregate measures are automatically weighted, hence

they mask the importance of firm-level tax optimisation (carry-forward of profits and losses between tax years). On the other hand, aggregate measures understate the role of firm heterogeneity. Careful analysis of micro-data is crucial to explaining the behaviour of the aggregate labour share.

## Imputation

In the analyses so far, we worked with observations for which the level of employment was available. In the remainder of this paper, we study the robustness of our results to the act of including observations where the employment level is missing.[6] To this end, we deploy a battery of imputation methods, which we describe in detail in Appendix (part B). First, we test if the missingness mechanism of the employment data is random or systematic. Having identified the missingness mechanism, we select the best performing imputation method based on a within-sample simulation study. Having identified the best performing imputation method within-sample, we deploy it out-of-sample to impute the employment level for those firm-years for which employment data is missing. Thus, we compare the labour share estimates obtained in the sample of 540,000 observations to the full sample of 720,000 observations.

### Missingness mechanism

Missing values are ubiquitous in financial data across different datasets and imputing them is one of the solutions extensively studied in **Bryzgalova et al. [2022]** for the COMPUSTAT, as well as in **White, Reiter, and Petrin [2018]** for the US Manufacturing Census. Missingness is also a feature of the Orbis dataset, as **Bajgar et al. [2020]** reported. Imputation can greatly improve the coverage and strengthen statistical power, as was done for value added in **Gal [2013]** for Orbis. Our work contributes to this thread of research.

In Orbis for Poland, employment data is missing particularly in the years 2010 through 2016 (Figure A1). For firms available in the sample before that period or after it, missingness is less of a problem. For firms which either entered the sample in this period or were observed only during this period, however, missingness can lead to important biases. Imputing information on employment allows us to study the robustness of our inference to this feature of Orbis data.

**Terminology.** The missingness pattern of employment in Orbis is not random, but a systematic one. Complete case analysis or simple unconditional mean imputation could produce biased estimators of population parameters under the Missing at Random (MAR) or Missing Not at Random (MNAR) mechanisms (e.g., **Van Buuren [2018]**). Both technical terms refer to systematic missingness. MAR describes unconfounded missingness, i.e., such that it can be modelled with observed data. MNAR, meanwhile, depends on unobserved values, potentially on unobserved variables or the values themselves being missing. We direct our readers to Appendix (part B) for more details on the missingness mechanisms. **Bajgar et al. [2020]** show that smaller firms are underreported in Orbis in comparison to the population of firms, which in light of our question plays a crucial role and hints at non-uniformly random missings.

**Missingness in Orbis data.** We put the missingness mechanism to the test. The distributions of the variables being tested are approximately normal. Little's test was conducted and the conclusion is that the assumption of the missing completely at random (MCAR) mechanism is rejected (p-value=0.00); **[Little, 1988]**. That was a global test. We have also done multiple hypothesis testing of t-test differences conditional on missingness in employment. Even after the Bonferroni correction, the results strongly imply missing at random (MAR) or missing not at random (MNAR); adjusted p-value=0.00. The t tests consider the differences, conditional on missing employment, in added value, total assets, operational revenue and payroll, which are fully observable in our sample. Such results confirm our hypothesis on the mechanism.

[6] Note that we do not need employment data to obtain labour share measures, but only to classify firms as either small or large.

Small firms have lower values for certain covariates that are observable. The labour cost could be a strong proxy for the value of employment. The Spearman correlation for the two is 0.92 in the observed part. The propensity to miss employment logistic regression achieved an AUC of 0.82 on the whole sample. This suggests the presence of the MAR mechanism since we can explain a significant part of the probability to miss by the observed characteristics. The regression considers sector and year indicators as well as value added, payroll, turnover, fixed assets, other current assets and the value of stocks. Thus, we assume the MAR mechanism is present and next we proceed with imputation.

## Evaluation of imputation method performance

In order to approximate the imputation error and choose the method for final imputation, we design a simulation study. We create a procedure to predict the observed part of the employment vector. In the case where we use the variables described above that are fully visible, the problem is reduced to a one-dimensional imputation problem. Fortunately, the panel setting is in this case quite helpful for modelling since some variables do not change much between sectors or within firms across years. We have considered eight models for imputation. We compare naive imputation using economic identities such as the capital-labour ratio and the average sector-year wage. Further, we compare them to naive production function estimation via Cobb Douglas, as well as linear interpolation of employment between observable years for a given company. Finally, we take a linear regression and decision tree methods such as CART, random forest and XGBoost. We describe the methods in more detail in Appendix (part B).

**Table 2: Raw Mean Square Error of imputation methods**

| | Cobb–Douglas | K-L ratio | Sector wage | Linear Regression | Linear interp. | Random Forest | CART | XGB |
|---|---|---|---|---|---|---|---|---|
| Inside | 1.245e+11 | 1,580.38 | 22.67 | 23.18 | **7.23** | 20.85 | 31.47 | 21.03 |
| Outside | 9.861e+11 | 620.39 | 10.02 | 9.75 | | **9.50** | 19.13 | 9.54 |
| Total | 7.569e+11 | 875.73 | 13.38 | 13.32 | 7.23 | 12.52 | 22.41 | 12.59 |

Notes: The table provides the results of RMSE averaged over 100 simulations for systematic MAR setting. The sample is further divided into inside and outside samples to enable a comparison of linear interpolation with other methods for the variables that lie inside two observable years. The bolded values are the lowest RMSEs in each category of our interest.

Source: Authors' own elaboration.

To test the quality of imputation we have simulated MAR missingness mechanisms. In the MAR setting, for every observation we have drawn a Bernoulli random variable with a probability to ampute, masked to be missing for the simulation purposes, equal to the propensity to miss scores taken from the regression described in the Missingness mechanism subsection. This way we mimic the missingness mechanism observed in the data as closely as possible. We train the methods described in Appendix (part B) and predict the amputed part. In each set we have chosen hyperparameters fitted to training data. We run the simulation scheme 100 times and average the results. Because linear interpolation can only work for observations between two observable years, we further divide the sample into values missing inside two observable years, representing an "inside" sample, and the rest, making up the "outside" sample. For the criterion of quality, we calculate the raw mean squared error (RMSE) on the amputed observations that formed the test set. The results of the simulation are presented in Table 2.

As for a robustness check, we also simulate a MCAR mechanism. Although we argue that our sample missingness mechanism is MAR, with Little's test and the well fitted propensity to miss logistic regression being strong indicators for that fact, we cannot capture fully the process that governs missings appearing in our data. A popular benchmark for imputation methods is simulating uniform, non-systematic missings **Lin and Tsai [2020]**. A similar approach to testing both MCAR and MAR was undertaken by **Bryzgalova et al. [2022]**. In the MCAR design, we have randomly selected four firms from each sector in all the years to be amputed

and to form the test sample. The conclusions about which methods are the best are consistent between the MCAR and MAR scenarios. In turn, in the main text, we focus only on describing MAR simulation results. The MCAR simulation results can be found in Appendix (part B).

The simulation presents a few insights. First, the methods preserve their rank in terms of the quality of imputation regardless of the frame of comparison being inside or outside samples. Second, the best performing methods in terms of RMSE are linear interpolation, random forest, XGB, linear regression, sector wage, CART, the K-L ratio, and Cobb-Douglas. Third, linear interpolation performs better than the alternatives on data to which it can be applied, which means inside two given observed years for a given firm. This confirms the stability of employment in firms. Fourth, the average sector wage performs well in comparison with other methods, suggesting that firms are similar in employment in a given industry in a specific year. However, that is not true in the case of the capital-labour ratio. Finally, the production function estimates employment poorly, showing the scope for potential improvement.

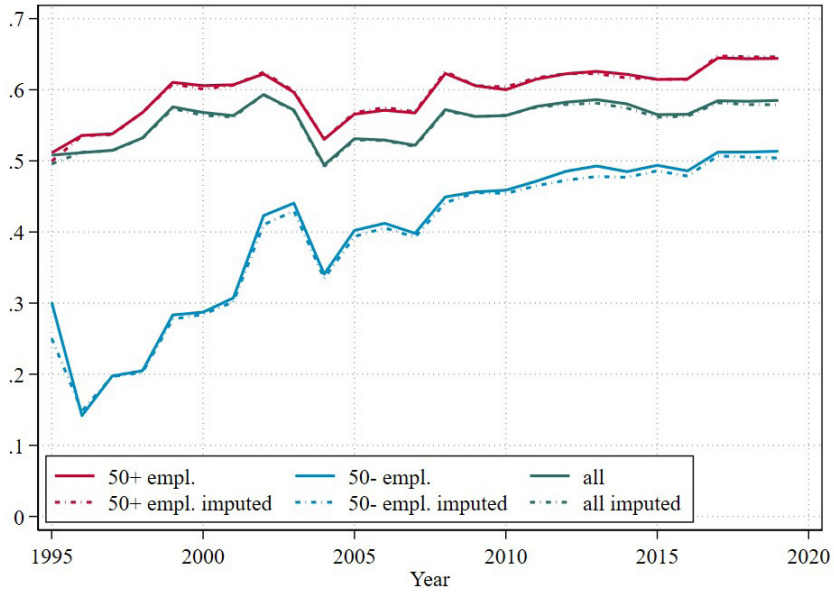
In conclusion, we pursue further imputing the unobserved employment with linear interpolation for gaps between two observed years for a firm and a random forest for the rest of employment missingness. Those two estimators have proven to be the best among the considered methods in terms of sample RMSE in the simulation study.

## Results of the imputation

In accordance with the ranking presented in Table 2, we continue with linear interpolation for the gaps inside and perform random forest regression for the gaps outside. For the final imputed values, we fit our method of choice to the whole data this time and tune the hyperparameters on five-fold cross-validation.

The gaps in the data are more profoundly present for small companies than for large ones. Of the total of 373,258 observations imputed, small companies account for 318,396 observations (85%) and large ones for 55,820 observations (15%), or 130% and 51% of the original observable samples respectively.

**Figure 5: Labour share: baseline sample vs. imputed sample**




Note: We compare labour share estimates using our baseline sample described in the Final sample subsection and the full sample with size determined by our preferred imputation method: linear interpolation inside two observable years for a given firm and random forest outside.
Source: Authors' own elaboration.

Figure 5 shows the labour share before and after imputation, for all the firms and for those with more and fewer than 50 employees. The labour share levels for firms with 50+ employees before and after imputation

are almost identical. In the case of firms with fewer than 50 employees, the labour share estimates are slightly lower, especially in the 2010–2015 period. Still, this difference does not change the fact that the labour share among firms with fewer than 50 employees increases throughout the considered period. Thus, the lack of any serious differences in the labour share before and after imputation of missing employment suggests that the method of classifying firms based on historically observed employment values yields similar results to more sophisticated imputation methods. In general, our findings described in the Evolution of labor share section remain robust to sample enlargement achieved by the application of the imputation procedure.

## Conclusion

A large body of literature investigated the declining labour share using available aggregate and firm-level micro-data and documented a notable decline in the labour share in many economies and sectors around the world. In this paper, we look at the case of Poland. We construct a new firm-level dataset including 720,000 firm-year observations and covering 25 years from 10 waves of Orbis, which is a non-representative firm-level database. Using this dataset, we document new facts about the labour share in Poland. Before, the only available labour share estimates from firm-level data for Poland were those provided by **Growiec [2009]**.

In general, we show that there was no systematic decline in the labour share in Poland from 1995 to 2019. On the contrary, we provide evidence that the labour share in Poland was quite stable over the timeframe of 20 years. First, we document the evolution of the labour share between 1995 and 2019. In line with findings from **Growiec [2009]**, we also observe a labour share decline during the mid-2000 s. For later years, we find that the labour share rebounded in the late 2010 s. Second, utilising available information on employment, we can distinguish between firms with more than 50 employees and firms with fewer than 50 employees in our data. According to our estimates, the labour share for firms with fewer than 50 employees features stable growth, but its level is lower than the labour share for firms with 50+ employees. We then contrast the aggregate labour share (weighted average) with the unweighted average labour share and analyse the labour share by the distribution of added value. This unveils firm heterogeneity in the labour share in different parts of the added value distribution. Firms with lower value added have a higher labour share, and most firms have an individual labour share higher than the unweighted average. This implies that many companies may be suffering from insufficient employment of capital, which hinders their development. Furthermore, we also benchmark the labour share from Orbis with **Growiec [2009]** when available and for the remaining years with OECD STAN data. In general, the time patterns in the labour share estimates from Orbis are similar to those in other data sources.

Finally, since we do not have data on size for about half of our sample, we deploy a variety of imputation methods to address this problem. The labour share estimates from all the samples are close to those obtained from data with limited size information. Thus, we conclude that all our inferred results remain robust to sample expansion.

## References

Autor D., Dorn D., Katz L. F., Patterson C., Van Reenen J. [2020], The Fall of the Labor Share and the Rise of Superstar Firms, *Quarterly Journal of Economics*, 135 (2): 645–709.

Bajgar M., Berlingieri G., Calligaris S., Criscuolo C., Timmis J. [2020], *Coverage and Representativeness of Orbis data*, OECD Science, Technology and Industry Working Papers.

Bauer A., Boussard J. [2020], Market Power and Labor Share, *Economie et Statistique / Economics and Statistics*, 520–521: 125–146.

Böckerman P., Maliranta M. [2011], Globalization, creative destruction, and labour share change: evidence on the determinants and mechanisms from longitudinal plant-level data, *Oxford Economic Papers*, 64 (2): 259–280.

Breiman L. [2001], Random Forests, *Machine Learning*, 45 (1): 5–32.

Breiman L., Friedman J. H., Olshen R. A., Stone C. J. [2017], *Classification and regression trees*, Routledge.

Bruno R. L., Crescenzi R., Estrin S., Petralia S. [2021], Multinationals, innovation, and institutional context: IPR protection and distance effects, *Journal of International Business Studies*, 1–26.

Bryzgalova S., Lerner S., Lettau M., Pelger M. [2022], *Missing financial data*, SSRN Working Paper.

Cavallo A., Rigobon R. [2016], The billion prices project: Using online prices for measurement and research, *Journal of Economic Perspectives*, 30 (2): 151–78.

Charpe M., Bridji S., McAdam P. [2020], *Labor share and growth in the long run*, European Central Bank.

Chen T., Guestrin C. [2016], Xgboost: A scalable tree boosting system, [in:] *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (785–794).

Chen T., He T., Benesty M., Khotilovich V., Tang Y., Cho H. et al. [2015], Xgboost: Extreme gradient boosting, *R package version 0.4–2*, 1 (4): 1–4.

Dao M. C., Das M., Koczan Z. [2020], Why is labour receiving a smaller share of global income?, *Economic Policy*, 34 (100): 723–759.

De Loecker J., Eeckhout J., Unger G. [2020], The Rise of Market Power and the Macroeconomic Implications, *The Quarterly Journal of Economics*, 135 (2): 561–644.

Dimova D. [2019], *The Structural Determinants of the Labor Share in Europe*, IMF Working Papers.

Gal P. N. [2013], *Measuring Total Factor Productivity at the Firm Level using OECD-Orbis*, OECD Economics Department Working Papers No. 1049.

Growiec J. [2009], Relacja płac do wydajności pracy w Polsce: ujęcie sektorowe, *Bank i Kredyt*, 40 (5): 61–88.

https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Annual_national_accounts_-_evolution_of_the_income_components_of_GDP\#Shares_of_income_components_to_GDP_in_2022.

Kaldor N. [1961], Capital accumulation and economic growth, [in:] Hague D. C. (ed.), *The theory of capital: Proceedings of a conference held by the International Economic Association* (177–222), London: Palgrave Macmillan UK.

Kalemli-Ozcan S., Sorensen B., Villegas-Sanchez C., Volosovych V., Yesiltas S. [2022], *How to construct nationally represen*tative firm level data from the Orbis global database: New facts and aggregate implications, NBER Working Paper No. 21558.

Karabarbounis L., Neiman B. [2014], The Global Decline of the Labor Share, *Quarterly Journal of Economics*, 129 (1): 61–103.

Kehrig M., Vincent N. [2021], The Micro-Level Anatomy of the Labor Share Decline, *The Quarterly Journal of Economics*, 136 (2): 1031–1087.

Kónya I., Krekó J., Oblath G. [2020], Labor shares in the old and new EU member states: Sectoral effects and the role of relative prices, *Economic Modelling*, 90: 254–272.

Lin W.-C., Tsai C.-F. [2020], Missing value imputation: a review and analysis of the literature (2006–2017), *Artificial Intelligence Review*, 53: 1487–1509.

Little R. J. [1988], A test of missing completely at random for multivariate data with missing values, *Journal of the American Statistical Association*, 83 (404): 1198–1202.

Mertens M. [2022], Micro-mechanisms behind declining labor shares: Rising market power and changing modes of production, *International Journal of Industrial Organization*, 81.

Muck J., McAdam P., Growiec J. [2018], Will the "true" labor share stand up? An applied survey on labor share measures, *Journal of Economic Surveys*, 32 (4): 961–984.

Rubin D. B. [1976], Inference and missing data, *Biometrika*, 63 (3): 581–592.

Siegenthaler M., Stucki T. [2015], Dividing the pie: Firm-level determinants of the labor share. *ILR Review*, 68 (5): 1157–1194.

Smith M., Yagan D., Zidar O., Zwick E. [2022], The rise of pass-throughs and the decline of the labor share, *American Economic Review: Insights*, 4 (3): 323–340.

Statistics Poland [2020a], *Activity of enterprises with up to 9 persons employed in 2019*.

Statistics Poland [2020b], *Activity of non-financial enterprises in 2019*.

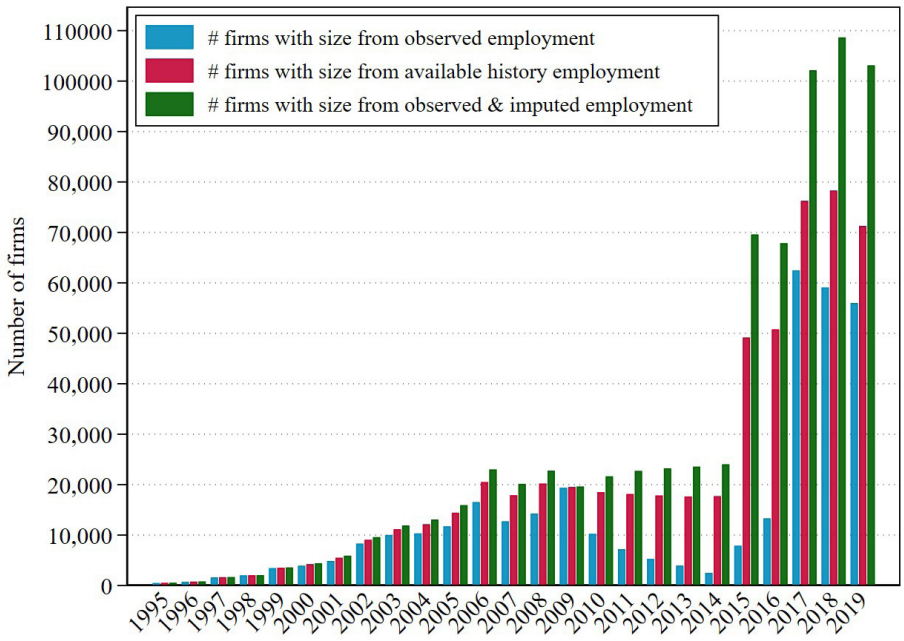
Van Buuren S. [2018], *Flexible imputation of missing data*, CRC Press.

White T. K., Reiter J. P., Petrin A. [2018], Imputation in US manufacturing data and its implications for productivity dispersion, *Review of Economics and Statistics*, 100 (3): 502–509.

# Appendix

## A. Data

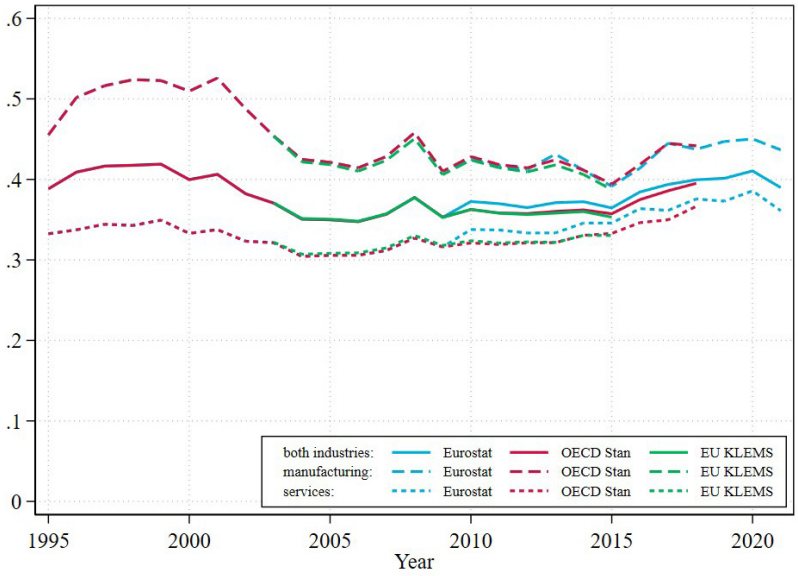
**Figure A1. Sample size by year**




Note: The blue bars show the number of observations (firms) with non-missing value added, employment and the total labour cost. The green bars show the number of observations with non-missing value added and the total labour cost. There is a substantial difference between the green and blue bars, especially between 2010 and 2016, due to missing employment data. The red bars show the number of observations after distinguishing firms with fewer than 50 employees, as discussed in the Final sample subsection.

Source: Authors' own elaboration.

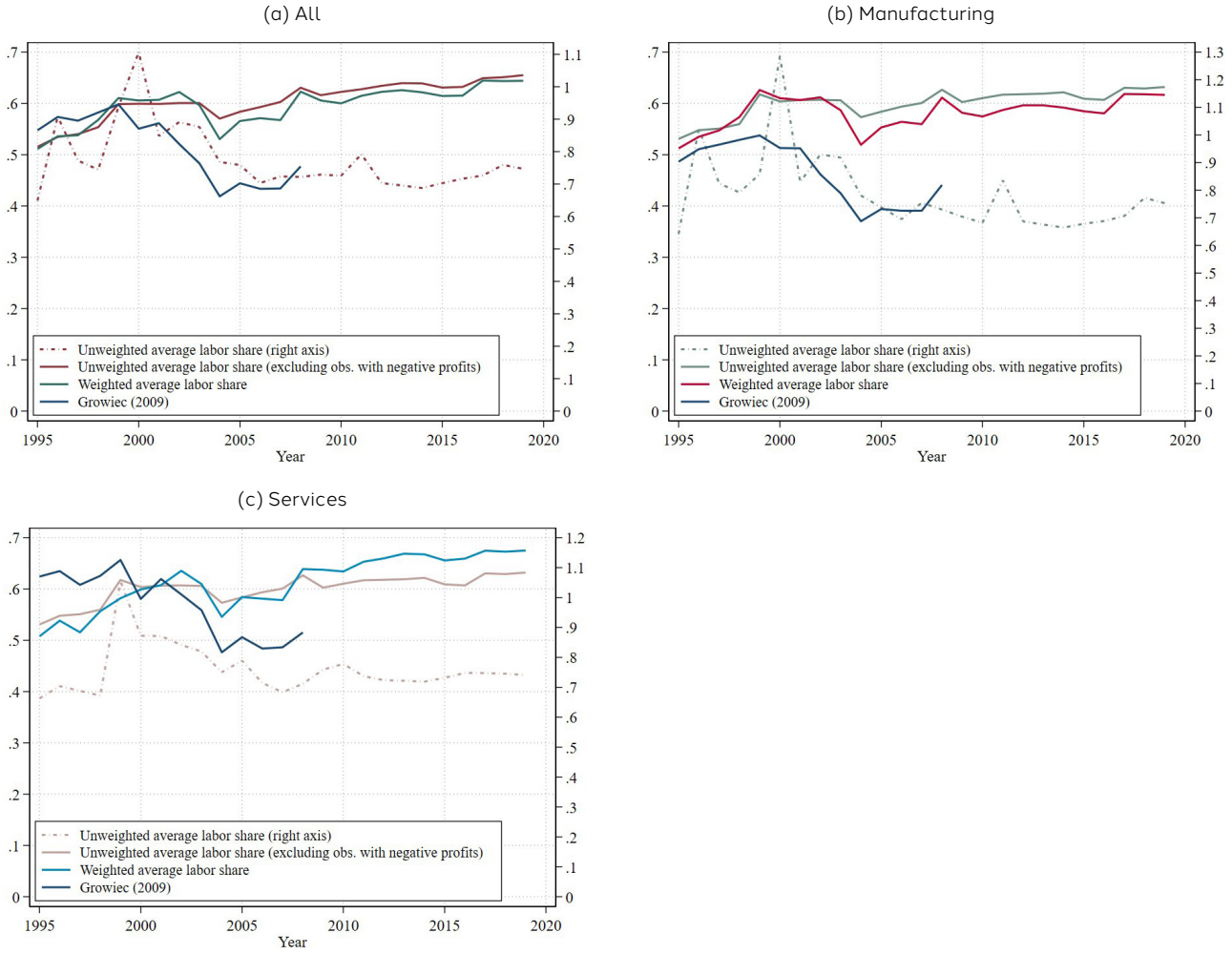
**Figure A2. Comparison of labour share measures from industry-level databases**




Note: In this figure, we compare the labour share measures obtained from three industry-level macroeconomic databases: OECD STAN, Eurostat, and EU KLEMS. The labour share is computed as the ratio of total compensation spending over gross value added. In the first half of the considered period, there is no difference between the sources. In the second half, there is a small upward shift in the Eurostat labour share in services in comparison with EU KLEMS and the OECD. However, this difference seems to be negligible.

Source: Eurostat, EU KLEMS and OECD STAN.

**Figure A3. Mean labour share**




Note: This figure presents the unweighted average labour share, the unweighted average labour share from observations excluding negative profits, and the average labour share weighted by the share of value added among all firms in manufacturing and services. The indices from Orbis were computed on a sample of large companies (those with 50+ employees). We also add estimates from **Growiec** [**2009**].

Source: Orbis, **Growiec** [**2009**].

## B. Imputation

### Missingness mechanisms

We briefly clarify the meaning of different types of missingness mechanisms following **Rubin** [**1976**]. There are three types of missingness mechanism:

1. MCAR (Missing Completely at Random) – The probability to miss depends neither on the values of observed nor on the values of unobserved variables: it is uniform on a given set of characteristics. In causal inference terminology, if we were to interpret the missingness mechanism as an assignment mechanism, where the treatment is the missingness mechanism, we would refer to the regime as a randomised study.
2. MAR (Missing at Random) – The probability to miss depends only on the values observed in the sample and not on unobserved variables. Again, interpreting this scenario in terms of the causal inference framework, as in point (1), we could refer to this situation as a strongly ignorable assignment mechanism.
3. MNAR (Missing Not at Random) – The probability to miss depends crucially on the values of missing variables and/or on unobserved variables. In causal inference terminology, this is close to the confounded assignment mechanism.

We present a more formal description in Table B1 below. Let $X$ be a complete data frame, $X_{obs}$ – observable part of the matrix, $X_{mis}$ – unobserved part of the matrix, $R_{ij}$ – missingness indicator for an element of the data frame, $\psi$ – vector of parameters of the model of the missing data mechanism.

**Table B1. Missingness mechanism**

| Missingness mechanism | Probability to be missing |
|---|---|
| MCAR | $P(R_{ij}=1\|X_{obs},X_{mis},\psi)=P(R_{ij}=1\|\psi)$ |
| MAR | $P(R_{ij}=1\|X_{obs},X_{mis},\psi)=P(R_{ij}=1\|X_{obs},\psi)$ |
| MNAR | $P(R_{ij}=1\|X_{obs},X_{mis},\psi)=P(R_{ij}=1\|X_{obs},X_{mis},\psi)$ |

Notes: The table provides possible reductions of conditional probability given independence on certain parameters. Note that the MNAR case is irreducible.

## Panel missingness

Analysing the panel structure we can infer valuable information. The heterogeneity between companies in terms of employment is larger than within companies. The between-firms standard deviation is 107 and within-firms is 47 for employment. For the observable part of employment, the lagged value of employment is linearly correlated at 0.97. Furthermore, we construct a missingness index, which is the number of missing employment records divided by all records in the database for different years for a given company. A missingness index of 0 would mean that we observe employment for all years. The results are that 141,215 observations do not have any information and 78,543 have all information. Half of the sample have between [0, 0.5] missingness index, 0.75 of the data have between [0, 0.8] missingness index. Thus we note that there is a possibility to take the year structure of firms information. Unfortunately, only 82,500 missing observations are inside of two observable years for a given company. This is summarised in Table B2.

**Table B2. Quantiles of missingness index distribution**

| Quantile | Missingness index |
|---|---|
| 0.25 | 0.25 |
| 0.50 | 0.52 |
| 0.75 | 0.8 |

Notes: These are the quantiles of the distribution of the missingness index, which show what proportion of years for a given company is missing, non-observable.

Source: Author's own calculation using Orbis data.

## Simulation

For completeness, we have conducted both MCAR and MAR scenario simulations. In the MCAR setting, we have randomly selected four firms from each sector in all the years to be amputed and form the test sample. In the MAR setting, for every observation we have drawn a Bernoulli random variable with the probability to ampute equal to the propensity to miss scores. The miss score was the predicted outcome taken from a logistic regression that included added value, labour costs, NACE-2 sector indicators, year indicators, turnover, fixed assets, stock value and other current assets where all of them were present, so for 661,416 observations. The remaining 60,125 observations were filled with a logistic regression of added value labour costs, operational revenue, sector indicators and year indicators estimated on the whole sample. The AUC of the regression was 0.82. After the amputation, on each generated train – test split the hyperparameters, discussed in the method sections, were tuned and the test error measured via RMSE was estimated. The final RMSE was averaged over 100 trials. For each method, we do hyperparameter tuning each time a new train test sample

is provided, so we test the method class in a sense and not the specific method instance with specific hyperparameters in mind. Nevertheless, we optimised over a small space of hyperparameters, so the two inferences remain connected. The results of the simulations are presented in Table B3.

**Table B3. Raw Mean Square Error of imputation methods**

| | Cobb–Douglas | K-L ratio | Sector wage | Linear Regression | Linear interp. | Random Forest | CART | XGB |
|---|---|---|---|---|---|---|---|---|
| | Random missingness (MCAR) | | | | | | | |
| Inside | 134.48 | 3431.18 | 50.50 | 65.69 | **13.92** | 46.73 | 73.22 | 46.71 |
| Outside | 212.85 | 2706.34 | 50.23 | 62.66 | | **48.31** | 79.49 | 48.94 |
| Total | 175.15 | 3055.08 | 50.36 | 64.12 | 13.92 | 47.07 | 47.07 | 47.87 |
| | Systematic missingness (MAR) | | | | | | | |
| Inside | 1.245e+11 | 1,580.38 | 22.67 | 23.18 | **7.23** | 20.85 | 31.47 | 21.03 |
| Outside | 9.861e+11 | 620.39 | 10.02 | 9.75 | | **9.50** | 19.13 | 9.54 |
| Total | 7.569e+11 | 875.73 | 13.38 | 13.32 | 7.23 | 12.52 | 22.41 | 12.59 |

Notes: The table provides the results of RMSE averaged over 100 simulations for MCAR and MAR settings respectively. The sample is further divided into inside and outside samples to compare linear interpolation with other methods for the variables that lie inside two observable years. The bolded values are the lowest RMSE in each category of interest.

Source: Authors' own elaboration.

## Methods

Below we present a description of methods used for imputation for missing employment observations. In our notation, *i* stands for individual firm, *s* for industry and *t* for year. We simplify our notation such that when we write an industry subscript for variable *X*, we mean that

$$X_{s,t} = \frac{1}{|s|}\sum_{i \in s} X_{i,t},$$

where $|s|$ denotes the cardinality of set s.

**Capital-labour ratio imputation.** From firm-level data we compute the capital-labour ratio:

$$klratio_{i,t} = \frac{total\ assets_{i,t}}{employment_{i,t}}$$

Then we create the average capital-labour ratio over two-digit NACE codes and a specific year, denoted as $klratio_{s,t}$. We can back out employment by dividing firm-level payroll and the average capital-labour ratio:

$$\widehat{employment}_{i,t} = \frac{payroll_{i,t}}{klratio_{s,t}}$$

given $i \in s$.

**Wage imputation**. In this method, we use data on payroll and employment to impute for missing employment. We compute the wage for each firm:

$$wage_{i,t} = \frac{payroll_{i,t}}{employement_{i,t}}$$

Then we compute the average firm-level wage over two-digit NACE industry codes and year. Next we use those industry-level means to impute for missing observations at the firm level within a specific sector and year:

$$\widehat{employment}_{i,t} = \frac{payroll_{i,t}}{wage_{s,t}}$$

given $i \in s$.

**Cobb-Douglas production function.** Here we start from an assumption that firms produce according to the Cobb-Douglas production function widely used in economic literature of the following form:

$$Y = AK^{\alpha}L^{1-\alpha}.$$

We can calculate employment with the observed total assets and value added, which are our proxies for capital and production. However, we must calculate $\alpha$ and TFP. We obtain those in the following way. First, we assume constant returns to scale and observe that with the Cobb-Douglas production function $\alpha$ is indeed a factor share parameter. Hence we can back out $\alpha$ as 1 diminished by the ratio of payroll and value added, both separately summed over two-digit NACE industries:

$$\widehat{\alpha_{s,t}} = 1 - \frac{employment_{s,t}}{added\ value_{s,t}}$$

We also need to find the TFP term, so we derive it from the production function and calculate it with industry-level variables and factor share $\alpha$, in the step before:

$$\widehat{A_{s,t}} = \frac{added\ value_{s,t}}{total\ assets_{s,t}^{\widehat{\alpha_{s,t}}}\ employment_{s,t}^{1-\widehat{\alpha_{s,t}}}}.$$

Finally, with the obtained $\hat{\alpha}_{s,t}$ and $\widehat{A_{s,t}}$ and the observed total assets and added value, we can back out firm-level employment:

$$\widehat{\text{employment}_{i,t}} = \left( \frac{\text{added value}_{i,t}}{\widehat{A_{s,t}}\ \text{total assets}_{i,t}^{\widehat{\alpha_{s,t}}}} \right)^{\frac{1}{1-\widehat{\alpha_{s,t}}}},$$

given $i \in s$.

**Individual regression imputation.** We formulate the following equation and estimate with OLS:

$$\ln\left(\text{employment}_{i,t}\right) = \beta_0 + \beta_1 \ln\left(\text{added value}\right)_{i,t} + \beta_2 \ln\left(\text{total assets}\right)_{i,t} + \beta_3 \ln\left(\text{payroll}\right)_{i,t} + \delta_s + \delta_t + \epsilon_{i,t},$$

where $\delta_i$, $\delta_t$ are fixed effects for two-digit NACE industries and years respectively. With the estimated equation, we predict employment and use it to impute the missing parts.

**Linear interpolation imputation.** We also use the linear interpolation between the two closest observations for a given company. So if we choose company *i*, and there are missing observations between *t* and *t* + *n*, we determine them according to the following formula:

$$\forall_{t<k<t+n} employment_{i,\,t+k} = employment_{i,t} + (t+n-k)\frac{employment_{i,t+n} - employment_{i,t}}{n}.$$

**CART.** The decision tree is a widely used method developed by **Breiman et al.** [**2017**]. It was used in the case of productivity estimation by **White et al.** [**2018**]. The tree implementation is described in https://cran.r-project.org/web/packages/rpart/rpart.pdf (date of access 1.12.2022). Decision trees are outlier robust, scale invariant, nonlinear, with naturally existing interaction methods. Thus they are widely recommended. The tuning of the hyperparameters was done on the parameter of the maximum depth of a tree; the rest were set default. We included the same variables as in the linear regression above when creating the tree.

**Random forest.** The random forest is a possible improvement over CART **Breiman** [**2001**]. The method we used is implemented in https://cran.r-project.org/web/packages/ranger/ ranger.pdf (date of access 1.12.2022). The random forest is a more stable technique with reduced variance and a natural weapon against overfitting due to bagging technique in comparison to CART. The hyperparameter tuned was the "mtry" from the R pack-

age, so the number of variables to possibly split at in each node. The rest of the parameters were set to default. We included the same set of variables as in the linear regression above when creating the trees.

**XGBoost.** The XGBoost algorithm has consequently outperformed even deep neural networks for tabular data. It is a combination of two techniques of ensembling algorithms, bagging and boosting, as in **Chen and Guestrin [2016]**. We use the implementation of **Chen et al. [2015]** https://cran.r-project.org/web/packages/xgboost/xgboost.pdf (date of access 1.12.2022). We optimised over the maximum depth of a tree and the number of rounds setting the rest of the parameters to default. We included the same set of variables as in the linear regression above when creating the trees.