GRAPE

Group for Research in APplied Economics

GRAPE Working Paper #82

The evolution of labor share in Poland. New evidence from firm-level data

Sebastian Zalas, Hubert Drążkowski

FAME | GRAPE, 2023

Foundation of Admirers and Mavens of Economics
Group for Research in Applied Economics

The evolution of labor share in Poland. New evidence from firm-level data

Sebastian Zalas
FAME|GRAPE and
University of Warsaw

Hubert Drążkowski
FAME|GRAPE and
Warsaw University of Technology

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 labor 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.

Keywords:

labor share, firm-level data, missing data

JEL Classification

C81, E25, D33

Corresponding author

Sebastian Zalas, s.zalas@grape.org.pl

Acknowledgements

This research was supported by National Research Center (grant # 2019/34/H/HS4/00481), whose support is gratefully acknowledged. We are grateful to Jakub Growiec for graciously sharing his data. Earlier versions of this study received useful comments from Joanna Tyrowicz, Magdalena Smyk-Szymanska and Lucas Van der Velde. ´We also thank the participants of the (Ce)^22 Workshop 2022, the 11th Professor Zbigniew Czerwinski Scientific Conference and 11th Summer Workshop on Macroeconomics and Finance for their helpful comments. The remaining errors are ours.

Published by: FAME | GRAPE
ISSN: 2544-2473

Foundation of Admirers and Mavens of Economics
Koszykowa 59/7
00-660 Warszawa
Poland

W | grape.org.pl
E | grape@grape.org.pl
TT | GRAPE_ORG
FB | GRAPE.ORG
PH | +48 799 012 202

I. Introduction

We study the evolution of labor share in Poland utilizing a novel source of firm-level data, the so-called Orbis data. Poland is notorious for its low and declining labor share (Dimova, 2019). According to the Eurostat[1], Poland ranks roughly #20 in the European Union. Kónya, Krekó, and Oblath (2020) show that across the region of Central and Eastern Europe, labor shares are lower than in Western Europe, with a systematic decline of the labor share in manufacturing and non-monotonous trends in other sectors. These conclusions notwithstanding, a large body of literature warns against the perils of estimating labor share from macroeconomic aggregates. In particular, self-employment and agricultural employment pose important methodological challenges (Kónya et al., 2020), both of which 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 estimates of labor share 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, Crescenzi, Estrin, & Petralia, 2021; Kalemli-Ozcan, Sorensen, Villegas-Sanchez, Volosovych, & Yesiltas, 2022), Orbis data remain underutilized for the study of the Polish economy. We contribute to burgeoning literature on the evolution of labor share. Karabarbounis and Neiman (2014) demonstrate substantial declines in labor shares world wide. This trend prevails regardless of the ambiguities regarding the adequate measurement of labor shares from the macroeconomic data (Mućk, McAdam, & Growiec, 2018). Analyzing case of Poland we find contrasting trend of labor share. We report that labor share in Poland after temporal decline in mid 2000s was rising and achieved level similar to beginning of 2000s. We also document labor share in industries and in size group of firms. We find that labor share is higher in services than in manufacturing. Furthermore, we document that labor share in firms with lower than 50 employees is lower in all years than in firms with 50+ 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 the study by Growiec (2009), which utilizes firm registry data from the Central Statistical Office for firms employing 50 workers or more. To

[1] Rank of changes of labor share from Eurostat

the best of our knowledge, this registry data is not available for research purposes (except for internal researchers at Statistics Poland or the National Bank of Poland). 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), providing estimates for recent years and companies employing less than 50 workers. Additionally, we compare Orbis data to aggregate data from OECD to complete the comparison, since Growiec (2009) estimates end in 2009. Despite the differences in labor share levels from Orbis, Growiec and OECD, we observe similarity in labor share evolution.

Finally we discuss and critically evaluate the viability of imputation methods for improving the quality of inference. Although our sample is constructed such that we possess fully observable information on value added and labor costs, which allows for labor share estimation, we have direct information on employment for only circa half of the sample. For better comparability of the samples between sources, as well as validation of robustness of our results, we perform an imputation study. 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 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 III., we describe in detail the features of our data. Section IV. describes the results using the raw Orbis data. We extend our work in section V. 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 labor share in Poland. We also discuss the implications for researchers intending to use Orbis data for subsequent research.

II. Literature

The evolution of the labor share, that is the fraction of gross domestic product allocated to wages (labor), has been widely debated in economic literature in recent years. Kaldor (1961) states the stability of labor share in one of his famous stylized facts of economic growth. Constancy of

labor 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 labor share changed in Poland.

Declining labor share. Literature documenting cross-country evidence on labor share shows that many countries experienced a decline of the labor share at some point. Karabarbounis and Neiman (2014) analyze data on 59 countries from the UN and OECD between 1975 and 2012 and document that 42 countries experienced a decline in labor share. In particular, Karabarbounis and Neiman (2014) observe that labor share declined among the largest economies, such as the US, China, Japan and Germany. In these countries, labor share was decreasing by 2-4 percentage points every 10 years. Likewise Dao, Das, and Koczan (2020) study global changes in labor share from 1991 through 2014 and confirm the findings from Karabarbounis and Neiman (2014). Dao et al. (2020) document that labor share declined in 29 of the largest 50 economies. In countries featuring decreases, according to Dao et al. (2020), labor share was declining on average by 2 percentage points after 10 years. Later, Dimova (2019) reported a decline in the labor share among half of EU countries between 2002 and 2016. In these years, the changes in labor shares in the majority of countries ranged between -3 to 3 percentage points. However, in 4 of the new EU countries Dimova (2019) documented significant increases in labor share, exceeding 4 percentage points. On average, labor share in the EU declined by around 1 percentage point. Charpe, Bridji, and McAdam (2020) presents a long run perspective on labor shares for France, US and UK, dating back to the 19th century. For instance, Charpe et al. (2020) show that decline of labor share in France occurred in the mid 1980s and then remained stable, while in the US and UK it has been gradually diminishing since the 1980s.

Although there are many countries with a more pronounced labor share decline, there are countries with temporary decline or increase of labor share. Several authors focusing on individual countries presented evidence for the stability of labor share in their respective studies. In line with Charpe et al. (2020), Bauer and Boussard (2020) obtained labor share both from microdata and aggregate data for France and report that since the 1990s labor share has remained stable. In their exploration of a representative sample of firms, **?** also report that, for the years studied, labor share in Switzerland was unchanged. Kónya et al. (2020) studies the evolution of labor 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 labor share in non-agricultural sectors. Kónya et al. (2020) observe differences between sectors and they find a sustained fall in

the manufacturing labor share, similarly to Dimova (2019) and Dao et al. (2020). Our paper adds updated evidence for Poland.

The documented changes in labor share were not economy-wide, but driven mainly by manufacturing sectors, broadly understood. For instance Dao et al. (2020) analyze changes in labor share by industry and find that the strongest decreases in the labor share occurred in manufacturing, followed by transportation and communication, while some sectors (food and accommodation, agriculture) experienced an increase. Dimova (2019) also observe that labor share in the majority of EU countries declined strongly in manufacturing and construction but rose in service sectors. In the case of the frequently analyzed US labor share, Kehrig and Vincent (2021) use US census data to report that labor share in manufacturing fell by 20 percentage points between 1967-2012. According to Smith, Yagan, Zidar, and Zwick (2022), decline of labor share in the US 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 labor share decline is most pronounced in manufacturing sectors. Our work also investigates changes in sectors to capture cross-sector heterogeneity.

Use of microdata. Concentrating on empirical studies of the labor share, an important distinction involves the level of analysis. Availability of firm-level data inspired researchers to investigate causes of labor share decline. Exploration of microdata revealed the importance of micro-level frictions for shaping macro-level changes in labor share. In Poland, for instance, (Growiec, 2009) exploits a panel of firms and finds that 55% of observed change in labor share in Poland occurred due to within-sector factors, and that reallocation effects account for the remaining change in labor share. Böckerman and Maliranta (2011) show effects of globalization on labor share by exploiting microdata from Finland. Kehrig and Vincent (2021) and Autor, Dorn, Katz, Patterson, and Van Reenen (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 declining labor share in the US on a sample of publicly traded firms from COMPUSTAT. In Germany, Mertens (2022) use a twenty year firm-level dataset from manufacturing to study the impact of market power on labor share. We follow this trend by inspecting labor 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 labor 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.

III. 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 the firms to registry courts and local government statistical offices. These data are collected by InfoCredit and subsequently digitized.[2] Given this 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 those firms which submitted the reports, especially in the 1990s and early 2000s, some of the reports were filled by hand or a typing machine and thus digitization was obscured. The growing popularity of computers gradually increased the share of fully legible reports.

Firms covered We utilize 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 reach 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 the Polish firms, the ID is based on REGON number, which permits linking these data with other registries). The data typically cover the period without the most recent year due data collection occurring before the reporting deadlines.

III.A. Processing data

The firms report consolidated statements, unconsolidated statements or both. Overall in the Orbis data, the vast majority of firms report unconsolidated accounts, which is useful for aggregating within sectors, as we do in this study – 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

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

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 presence of consolidated and unconsolidated statements concern less than 1% of all observations.

The waves of Orbis data each cover a ten-year window. Consequently, it may occur that the 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 wave, we are able to 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.

Harmonizing industries The Orbis data report NACE classification at four-digits. Our data cover the years 1995-2019. During this period NACE classification has 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 NACE classification is immaterial also in the case of firms which were observed only under one classification. However, in some cases the firm appears in Orbis under a newer classification, but its retrospective data cover periods of older classification. For the aggregation purposes, we have to 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 firms activity and assign the adequate classification from among the relevant options. For some firms, NACE classification was provided at two or three-digits rather than full four-digits classification. In those cases we assigned the adequate two-digit 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 in 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.

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

III.B. Final sample

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

For the sake of our analysis and in the interest of comparing our estimates to Growiec (2009), we need to identify the 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+ thousand firm-year observations with reported employment are available. The employment data is particularly frequently missing in the period 2010-2015, see Figure A.1 in the Appendices. 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 firms as having less than 50 employees if observed number of employees is below this threshold each time. When a company has reported employment values both above and below or equal to 50, we classify only those observations for which employment is observed.

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 capital-to-labor ratio. Next we apply 1% winsorizing procedure in each year to payroll, value added, turnover and total assets and average compensation, calculated as ratio of payroll and employment. Finally we keep only those observations for which we can calculate labor share.

Table 1 summarizes the final sample data and across size groups for selected years. The first part of Table 1 we present descriptive statistics. Initially we show 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 thousand observations. We also report how many observations do not posses any information on size. In total, approximately 150 thousands observations cannot be assigned to any size groups. In section V. we describe how to proceed with imputation to attribute the size information to all available observations. Later we show means of added value, payroll and employment. As number of available observations grows, mean employment, value added and payroll are decreasing due to the influx of small companies. For our analysis we use a sample consisting finally from roughly 570 thousand observations (with

size information), including 118 thousand 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 availabe 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 PLN. Employment is expressed as a 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) .

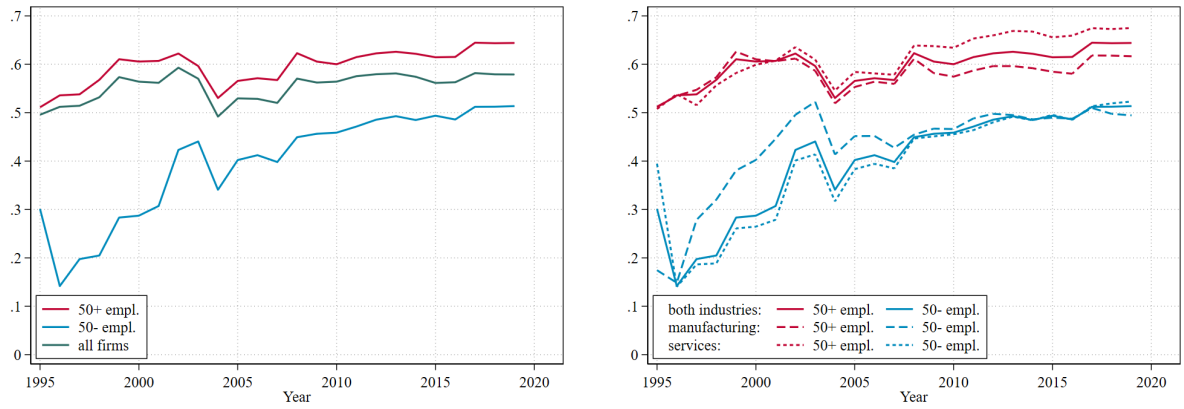
In the second part of Table 1 we present coverage of our data. Our sample contains notable parts of Polish firms with substantial representation of firms with 50+ employees. We show the number of firms as a percentage of firms included in data collected by Statistics Poland in their surveys. Statistics 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, they collect data on roughly 10 percent of firms with up to 9 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 available years, our sample possesses between 7 to 37 percent of

what Statistics Poland collects. For 50+ firms, the percentage of firms included in Orbis relative to Statistics Poland oscillates from 30 percent up to over 50 percent in the most recent years. For firms with less than 50 employees, our sample has between 4 percent and 35 percent of the number of firms included in official surveys.

IV. The evolution of labor share

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

Figure 1: Labor share



(a) by size **(b)** by industry

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

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

ison with companies with 50+ employees. Second, from the early years in the sample it rises consistently, excluding the temporary decline around 2004.

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

In general, the evolution of labor 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 levels of labor share than their counterparts. Industry comparisons show that overall labor share in services in the last years of the sample is at its highest levels, while in manufacturing labor share still makes up after the decline in the early 2000s.

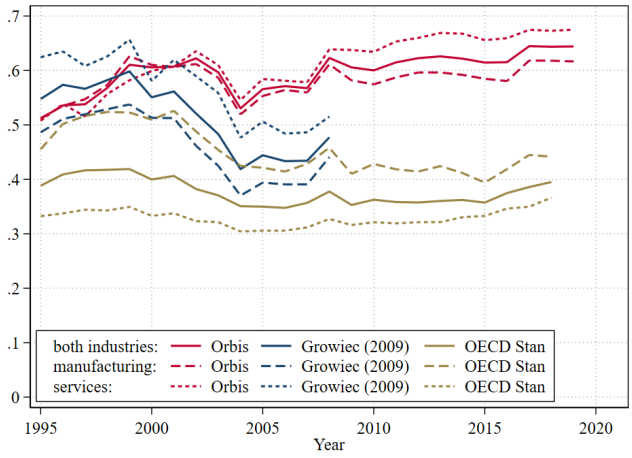
Comparing labor share in Orbis to other data sources. The next step we take is to compare our labor share estimates to other available data. The only research which presents estimates of labor share from firm-level data is Growiec (2009). Since estimates presented by Growiec (2009) break off in 2008,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

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

[4]We compare our estimates to OECD STAN, however there are other available sources of industry-level data, like EU KLEMS or Eurostat. These sources give almost identical estimates of labor share as OECD STAN. This is documented on Figure A.2 in Appendix.

recent observed years, labor share form OECD STAN shows a slight but stable increase, which is also observed in Orbis data.

Figure 2: Labor share: Orbis vs Growiec (2009) and OECD STAN



Note: We compare estimates of labor share from Orbis with Growiec (2009) and with indicators from OECD STAN sector level data. In order to make the comparison, indices presented from 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.

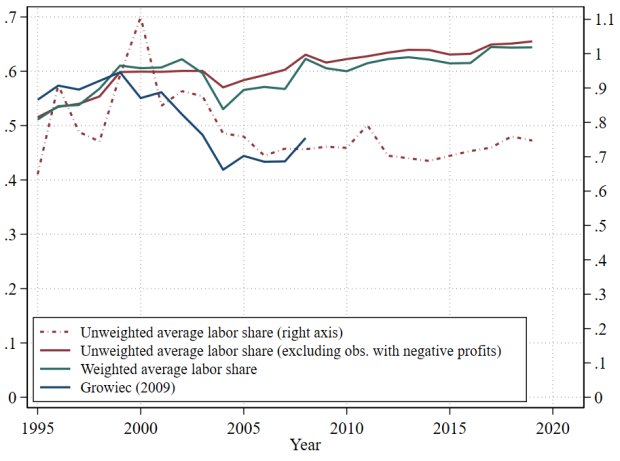
The differences between labor share from Orbis, Growiec (2009) and OECD STAN occur perhaps due to Orbis sample properties. As pointed out in Bajgar, Berlingieri, Calligaris, Criscuolo, and Timmis (2020), Orbis, as compared with nationally representative micro-data, only partially covers firm populations and the distribution of firms in Orbis is skewed towards particular types of firms. Because of partial coverage, Orbis has limited ability to reproduce indices computed from official aggregate statistics. In comparison with Growiec (2009), who worked with the census of 50+ employees firms, our sample under-represents the population of firms, which should explain the observed differences.

The role of aggregation: weighted vs unweighted. So far we studied aggregate labor share: a measure which presents a ratio between aggregate labor cost and aggregate value added. This measure gives higher weight to labor share in larger firms (both in terms of employment and in terms of value added). This measure is not sensitive to several important features occurring at firm level. First, firms which exhibit loss in a given year may mechanically display labor share in excess of 1, which is clearly not micro-founded. This is relevant if firms engaged in

carry-forward optimization of profits over years. Second, the standard aggregate measure is not susceptible to structural and cyclical fluctuations of 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 labor share. This measure is juxtaposed to 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, mean labor share significantly exceeds 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, with smaller firms becoming more prevalent in the sample in the late 1990s. Similar phenomenon is observed when we study the manufacturing and service sectors separately (see Figure A.3 in appendix).

Figure 3: Weighted vs unweighted average labor share.

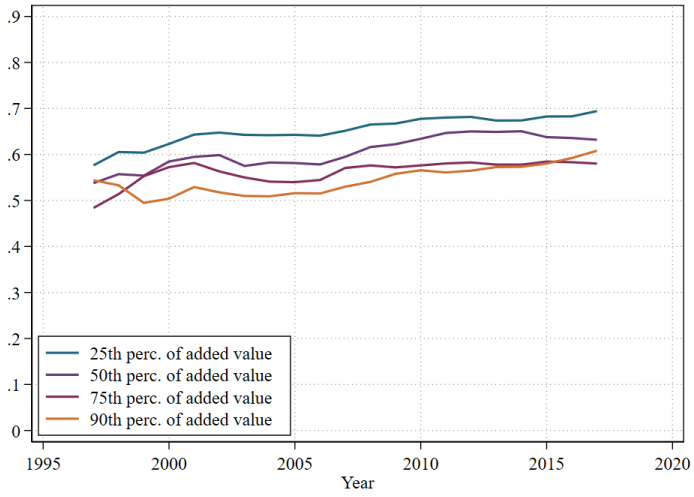


Note: This figure presents unweighted average labor share, unweighted average labor share excluding observations with negative profits, average labor share weighted by share of value added. These indices from Orbis were computed on a sample of large companies (with 50+ employees). We also add estimates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

The difference between weighted (aggregate) and unweighted mean labor share occurred due to changes in labor share among the largest firms. This supposition is supported by Fig-

ure 4, which shows the labor share across time and some percentiles of value added distribution. First, there is a striking difference between the labor share in the 25th percentile and in the 90th percentile. Labor share across high value added companies is lower than in low value added companies. Second, observed difference between the 25th and 90th percentile was stable until the beginning of the 2000s and then expanded in the mid 2000s. This suggests that the difference between the weighted and unweighted average is explained by the fact that labor share among the largest firms declined in comparison with smaller companies. In the 2010s, the difference between labor share in the 25th and 90th percentile of value added remained steady or diminished in comparison with the 2000s, and in both percentile groups labor share increased symmetrically. This resulted in a smaller difference between weighted and unweighted mean labor share in 2010s.

Figure 4: Labor share by percentiles of added value.



Note: This figure presents labor shares in the 25th, 50th, 75th and 90th percentile of value added. All four indices are smoothed with 5-year moving average.

Overall, exploring firm-level measures in addition to aggregate labor share measures reveals that aggregation is not necessarily innocuous. On the one hand, aggregate measures are automatically weighted, hence they mask the importance of firms-level tax optimization (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 explain the of behavior of aggregate labor share.

V. 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 including observations where employment level is missing.[5] To this end we deploy a battery of imputation methods, which we describe in detail in Appendix 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 employment level for those firms-years for which employment data is missing. Thus, we compare the estimates of labor share obtained in the sample of 540 thousand observation to the full sample of 720 thousand observations.

V.A. Missingness mechanism

Missing values are ubiquitous in financial data across different datasets and imputing them is one of the solutions extensively studied in Bryzgalova, Lerner, Lettau, and Pelger (2022) for the COMPUSTAT, as well as in White, Reiter, and Petrin (2018) for 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 the 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 - 2016 (Figure A.1). For firms available in the sample before that period or after it – the 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 a simple unconditional mean imputation could produce biased estimators of population parameters under missing at random (MAR) or missing not at random (MNAR) mechanisms (e.g., Van Buuren, 2018). Both technical terms refer to the systematic missingness. The former describes unconfounded missingness, i.e., such that it can be modeled

[5] Note that we do not need employment data to obtain labor share measures, merely to classify firms as small or large.

with the observed data. The latter missingness depends on unobserved values, potentially on unobserved variables or the values themselves being missing. We direct our readers to Appendix B for more details on the missingness mechanisms. Bajgar et al. (2020) shows that smaller firms are under-reported in Orbis in comparison to the population of firms, which in the light of our question plays a crucial role and hints at non-uniformly random missings.

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

Small firms have lower values for certain covariates that are observable. A strong proxy for the value of employment could be labor cost. The Spearman correlation for the two is 0.92 in the observed part. A 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 probability to miss by observed characteristics. The regression takes into account sector and year indicators as well as value added, payroll, turnover, fixed assets, other current assets and value of stocks. Thus, we assume MAR mechanism is present and next we proceed with imputation.

V.B. Evaluation of imputation methods performance

In order to approximate imputation error, as well as 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 reduces to a one dimensional imputation problem. Fortunately, the panel setting is in this case quite helpful for modeling, since some variables do not change much between sectors or within firms across years. We have considered 8 models for imputation. We compare naive imputation resorting to economic identities such as capital-labor ratio and 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 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 results of RMSE averaged over 100 simulations for systematic MAR setting. The sample is further divided into inside and outside samples to enable the comparison of linear interpolation with other methods for the variables that lie inside two observable years. Bolded values are the lowest RMSE in a given category of our interest.

To test the quality of imputation we have simulated MAR missingness mechanisms. In the MAR setting, for every observation we have drawn 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 subsection subsection V.A.. This way we mimic the missingness mechanism observed in the data as closely as possible. We train the methods described in Appendix B and predict the amputed part. On each set we have chosen hyperparameters fitted to training data. We run the simulation scheme 100 times and average the results. Due to the fact that linear interpolation can only work for observations between two observable years, we further divide the sample into values missing inside two observable years that we call "inside" sample and the rest, that we call "outside" sample. For the criterion of quality we calculate 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 test both MCAR and MAR was undertaken by Bryzgalova et al. (2022). In the MCAR design we have randomly selected four firms from every sector in all of the years to be amputed and

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

The simulation presents a few insights. First, methods preserve their rank in terms of 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, next random forest, followed by XGB, then linear regression, sector wage, CART, K-L ratio and Cobb-Douglas. Third, linear interpolation performs better than the alternatives on the data to which it can be applied, so inside two given observed years for a given firm. Furthermore, this confirms stability of employment in firms. Fourth, average sector wage performs well in comparison with other methods, suggesting that firms are similar in employment in a given industry in a given year, however that is not true in the case of capital-labor ratio. Finally, production function estimates employment poorly, showing 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, as those have proven to be the best among the methods considered, in terms of sample RMSE for missingness relative to the data, in the simulation study.

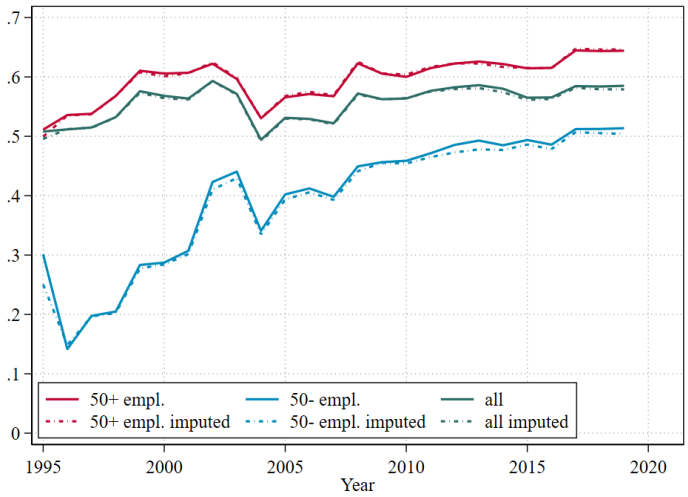
V.C. Results of the imputation

In accordance with the ranking presented in Table 2 we continue with linear interpolation for gaps inside and random forest for gaps outside. For the final imputed values we fit our method of choice to the whole data this time and tune hyperparameters on 5 fold cross-validation.

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

Figure 5 shows the labor share before and after imputation, for all firms and for firms with both more and less than 50 employees. The levels of labor share for firms with 50+ employees before and after imputation are almost identical. In the case of firms with less than 50 employees, labor share estimates are slightly lower, especially in in years 2010-2015. Still, this difference does not change the fact that labor share among firms with less than 50 employees increases throughout the considered period. Thus, lack of any serious differences in labor share

Figure 5: Labor share: baseline sample vs imputed sample



Note: We are comparing estimates of labor share using our baseline sample described in subsection III.B. and full sample with size determined by our preferred imputation method; linear interpolation inside two observable years for a given firm and random forest outside.

before and after imputation of missing employment suggests that the method of classifying firms based on historically observed employment values gives similar results to more sophisticated imputation methods. In general, our findings described in section IV. remain robust to sample enlargement achieved by application of imputation procedure.

VI. Conclusion

A large body of literature attempted to investigate declining labor share using available aggregate or firm level micro-data and find notable decline of labor share in many economies and sectors all around the world. In this paper we look into the case of Poland. We construct a new firm level dataset including 720 thousands 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 labor share in Poland. Before, the only available estimates of labor share from firm level data for Poland were provided by Growiec (2009).

In general, we show that there was no systematic decline of labor share in Poland between 1995-2019. On the contrary, we show evidence suggesting that over the timeframe of 20 years, labor share in Poland was fairly stable. First, we document the evolution of the labor share between 1995-2019. In line with findings from Growiec (2009), we also observed a labor share

decline during the mid 2000's. For later years, we find that labor share has recovered since the late 2010s. Second, utilizing available information on employment, we can distinguish between firms with more than 50 employees and firms with less than 50 employees in our data. According to our estimates, labor share from firms with less than 50 employees feature stable growth, but its level is lower than labor share from firms with 50+ employees. We then contrast aggregate labor share (weighted average) with unweighted average labor share and analyze labor share by distribution of added value. This unveils firm heterogeneity in labor share in different parts of the added value distribution. Firms with lower added value have higher labor share, and the majority of firms have an individual labor 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 labor share from Orbis with Growiec (2009) (when available) and for the remaining years with OECD STAN data. In general, time patterns in labor share estimates from Orbis are similar to 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. Labor share estimates from all 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.

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)

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.

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* (pp. 785–794).

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., ... others (2015). Xgboost: extreme gradient boosting. *R package version 0.4-2*, *1*(4), 1–4.

Dao, M. C., Das, M., & Koczan, Z. (2020, 07). 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*(1049).

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

Kaldor, N. (1961). Capital accumulation and economic growth. In D. C. Hague (Ed.), *The theory of capital: Proceedings of a conference held by the international economic association* (pp. 177–222). London: Palgrave Macmillan UK.

Kalemli-Ozcan, S., Sorensen, B., Villegas-Sanchez, C., Volosovych, V., & Yesiltas, S. (2022). How to construct nationally representative 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*.

Mućk, 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.

Smith, M., Yagan, D., Zidar, O., & Zwick, E. (2022, September). The rise of pass-throughs and the decline of the labor share. *American Economic Review: Insights*, *4*(3), 323-40.

Statistics Poland. (2020a). Activity of enterprises with up to 9 persons employed in 2019.

Statistics Poland. (2020b). Activity of non-financial enterprises in 2019.

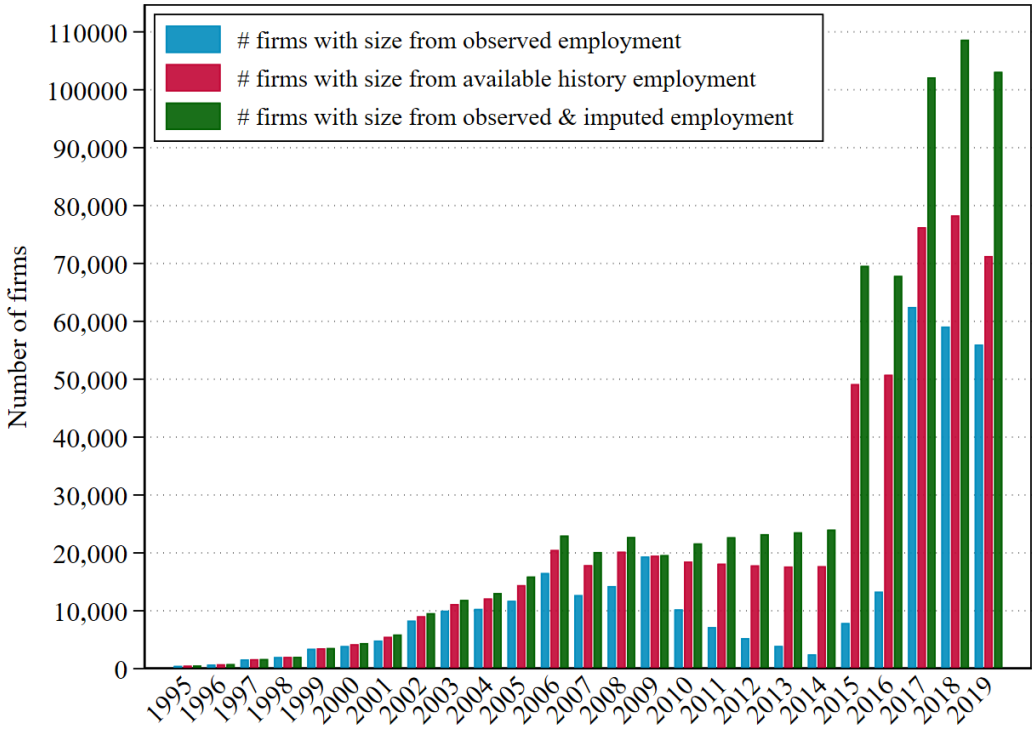
Van Buuren, S. (2018). *Flexible imputation of missing data*. CRC press.

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

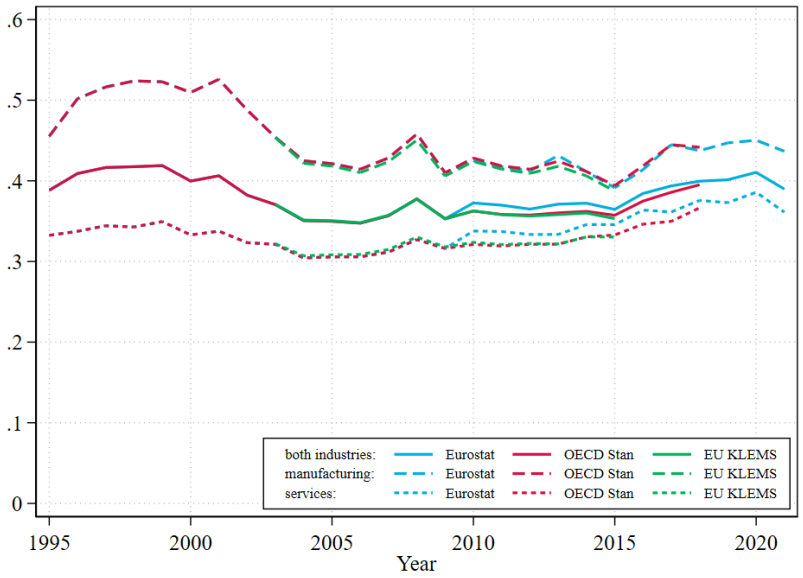
A Data appendix

Figure A.1: Sample size by year



Note: Blue bars show number of observations (firms) with non-missing value added, employment and total labor cost. Green bars show number of observations with non-missing value added and total labor cost. There is substantial difference between green bars and blue bars, especially between 2010 and 2016 which occurs due to missing employment data. Red bars show number of observations after distinguishing firms with less than 50 employees as discussed in subsection III.B..

Figure A.2: Comparison of labor share measures from industry-level databases.

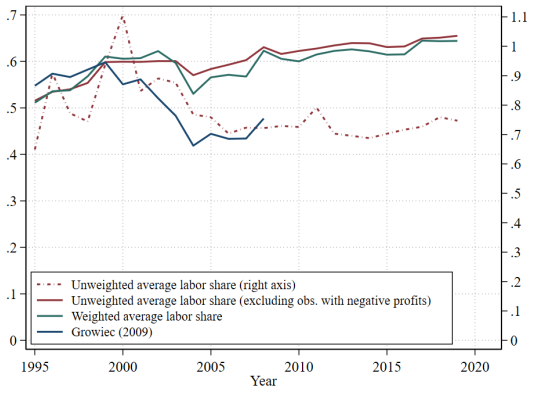


Source: Eurostat, EU KLEMS and OECD STAN.

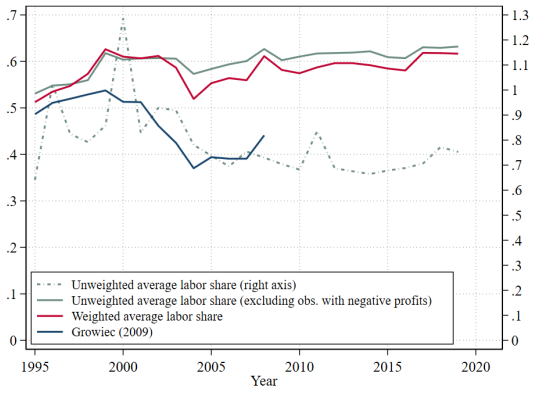
Note: On this figure we compare measures of labor share obtained from three industry-level macroeconomic databases: OECD STAN, Eurostat and EU KLEMS. Labor share is computed as ratio of total compensation spending over gross value added. In the first half of the considered period, there is no difference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

Figure A.3: Mean labor share

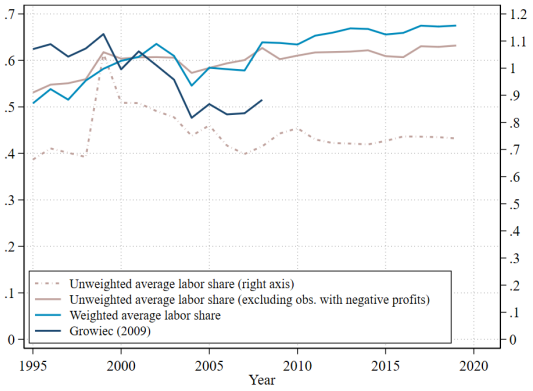
(a) All



(b) Manufacturing



(c) Services



Source: Orbis, Growiec (2009).

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

B Imputation appendix

II.A. Missingness mechanisms

We briefly clarify the meaning of different types of missing 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 missing mechanism as an assignment mechanism, where the treatment is the missing mechanism, we would refer to the regime as a randomized 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 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 variable to be missing and/or on unobserved variables. In causal inference terminology this is close to a confounded assignment mechanism.

We present a more formal description in table 4 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, ψ – vector of parameters of the model of missing data mechanism,

Table B.1: Missing mechanisms

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	$P(R_{ij} = 1\|X_{obs}, X_{mis}, \psi)$

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

II.B. Panel missingness

Analyzing 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 summarized in Table B.2.

Table B.2: Quantiles of missingness index distribution.

Quantile	Missingness index
25%	0.25
50%	0.52
75%	0.8

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

Source: Author's own calculation using Orbis data.

II.C. Simulation

For completeness we have conducted both MCAR and MAR scenario simulations. In the MCAR setting we have randomly selected four firms from every sector in all of the years to be amputed and form the test sample. In the MAR setting for every observation we have drawn Bernoulli random variable with probability to ampute equal to propensity to miss scores. The miss score was the predicted outcome taken from a logistic regression that included added value, labor costs, NACE two 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 rest (60,125) of observations were filled with a logistic regression of added value labor 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 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 hyperparameters 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 optimized over small space of hyperparameters, so the two inferences remain connected. The results of the simulations are present in Table B.3.

Table B.3: Raw Mean Square Error of imputation methods

	Cobb-Douglas	K-L ratio	Sector wage	Linear Regression	Linear interp.	Random Forest	CART	XGB
				Random missings (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 missings (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 results of RMSE averaged over 100 simulations for MCAR and MAR settings respectively. The sample is further divided into inside and outside samples to allow to compare linear interpolation with other methods for the variables that lie inside two observable years. Bolded values are the lowest RMSE in a given category of interest.

II.D. Methods

Below we present 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. What is more, we simplify our notation, such that when we write an industry subscript for a variable X, we mean that

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

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

Capital-labor ratio imputation From firm-level data we compute capital-labor ratio:

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

Then we create average capital-labor ratio over over two-digit NACE codes and year, denoted as $klratio_{st}$. We can back out employment by dividing firm-level payroll and averaged capital-labor 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 wage for each firm:

$$wage_{i,t} = \frac{payroll_{i,t}}{employment_{i,t}}.$$

Then we compute average firm-level wage over two-digit NACE industry codes and year. Next we use those industry-level means to impute for missing observations on firm level within particular 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 Cobb-Douglas production function, widely used in economic literature, of following form:

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

We can calculate employment with observing total assets and value added, which are our proxies for capital and production. However have to calculate α and TFP. We obtain those in following way. Firstly, assume constant returns to scale and we observe that with Cobb-Douglas production function α is indeed factor share. Hence we can back out α as 1 diminished by 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 TFP term, so we derive it from production function and calculate it with industry level variables and factor share α, obtained 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 obtained $\widehat{\alpha}_{st}$ and $\widehat{A}_{st}$ and observed total assets and added value we are able to back out firm-level employment:

$$\widehat{employment}_{i,t} = \left(\frac{added\,value_{i,t}}{\widehat{A}_{s,t}\,total\,assets_{i,t}^{\widehat{\alpha}_{s,t}}} \right)^{\frac{1}{1-\widehat{\alpha}_{s,t}}},$$

given $i \in s$.

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

$$ln(employment_{i,t}) = \beta_0 + \beta_1 ln(added\,value_{i,t}) + \beta_2 ln(total\,assets_{i,t}) + \beta_3 ln(payroll_{i,t}) + \delta_{industry} + \delta_{year} + \varepsilon_{i,t}$$

where $\delta_{industry}$ and δ_{year} are fixed effects for two-digit NACE industries and years respectively. With estimated equation, we predict for employment, $\widehat{employment}_{i,t}$ and use it to impute missing employment.

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

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

CART The decision tree is a widely used method Breiman, Friedman, Olshen, and Stone (2017). It was used in the case of productivity estimation under 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, non-linear, with naturally existing interactions methods, thus widely recommended. The tuning of the hyperparametrs was done on the parameter of maximum depth of a tree, the rest were set default. We included the same variables as in 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 the CART. The hyperparameter tuned was the "mtry" from the R package, so 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 linear regression above when creating the trees.

XGBoost The XGBoost algorithm has consistently outperformed even deep neural networks for tabular data. It is a combination of two techniques of ensambling algorithms, bagging and boosting 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 optimized over maximum depth of a tree and number of rounds setting the rest of the parameters to default. We included the same set of variables as in linear regression above when creating the trees.