OECD Science, Technology and Industry Working Papers
2020/06

Coverage and representativeness of Orbis data

Matej Bajgar,
Giuseppe Berlingieri,
Sara Calligaris,
Chiara Criscuolo,
Jonathan Timmis

https://dx.doi.org/10.1787/c7bdaa03-en

OECD

OECD Science, Technology and Industry Working Papers

OECD Working Papers should not be reported as representing the official views of the OECD or of its member countries. The opinions expressed and arguments employed are those of the authors. Working Papers describe preliminary results or research in progress by the author(s) and are published to stimulate discussion on a broad range of issues on which the OECD works. Comments on Working Papers are welcomed, and may be sent to Directorate for Science, Technology and Innovation, OECD, 2 rue André-Pascal, 75775 Paris Cedex 16, France.

Note to Delegations:

This document is also available on O.N.E under the reference code:

DSTI/CIIE/WPIA(2017)5/FINAL

This document, as well as any data and any map included herein, are without prejudice to the status of or sovereignty over any territory, to the delimitation of international frontiers and boundaries and to the name of any territory, city or area.

© OECD 2020

You can copy, download or print OECD content for your own use, and you can include excerpts from OECD publications, databases and multimedia products in your own documents, presentations, blogs, websites and teaching materials, provided that suitable acknowledgment of OECD as source and copyright owner is given. All requests for commercial use and translation rights should be submitted to rights@oecd.org.

COVERAGE AND REPRESENTATIVENESS OF ORBIS DATA

Matej Bajgar, Giuseppe Berlingieri, Sara Calligaris,
Chiara Criscuolo and Jonathan Timmis (OECD)

Firm-level data covering many countries have the potential to provide key insights for understanding global economic trends and the role of policies across firms within and across countries. This is particularly true in an increasingly globalised economy where multinational corporations operating in several countries account for a large share of economic activity. Commercial databases based on company financials represent a rare source of such cross-country firm-level data, but are likely to suffer from incomplete coverage. While their limitations are well-known in theory, there is little information on how important the limitations are in practice. This paper describes coverage and representativeness of one of the most popular of such datasets – Orbis. As a benchmark, it uses industry-level data from the OECD STAN dataset as well as micro-aggregated data from the OECD MultiProd and DynEmp projects, which draw on official microdata representative of the entire firm population. It documents that firms in Orbis are disproportionately larger, older and more productive, even within each size class. They also display a reduced productivity dispersion in the lower half of the productivity distribution. Consequently, Orbis struggles to replicate variation over time, across industries and across countries observed in official data. However, steps such as focusing on a small number (fewer than 10) best-covered European countries and imputing value added substantially improve the representativeness of the population of firms with at least 10 employees. Overall, Orbis seems more suitable for studies that: i) take a global perspective rather than making comparisons across countries; ii) analyse top performers and multinationals rather than underperforming firms; iii) and focus on mean performance or within-firm changes rather than on the entire firm distribution or entry and exit.

Keywords: Firm-level data; cross-country analysis; distributed microdata analysis

JEL Classifications: D22, O47, Y1

Foreword

We are grateful to Luiz De Mello, Peter Gal, Giuseppe Nicoletti, Cyrille Schwellnus, Dirk Pilat and Andrew Wyckoff for helpful comments on earlier versions of this paper.

Table of contents

Foreword **4**

Chapter 1. Introduction **7**

Chapter 2. Data and methodology **11**

2.1. Orbis 11
2.2. OECD STAN 12
2.3. OECD MultiProd 12
2.4. OECD DynEmp 14
2.5. Definitions of basic variables 14
2.6. Methodology 15

Chapter 3. Baseline representativeness **17**

3.1. Matching industry aggregates 17
3.2. Firm coverage 20
3.3. Matching firm distribution characteristics 23

Chapter 4. Choices on data construction **32**

4.1. Imputation 32
4.2. Coverage thresholds 35
4.3. Industries 38
4.4. Size thresholds 40
4.5. Weighting 41

Chapter 5. Representativeness in the "preferred sample" **44**

Chapter 6. Preparing Orbis data: selected issues **49**

6.1. Rounded values 49
6.2. Account types and consolidation 50
6.3. Ownership module 52

Chapter 7. Summary of results **54**

References **59**

Annex A. Appendix **62**

Figures

Figure 3.1. Orbis data capture around 60% of aggregate employment and output and around 40% of aggregate value added 18
Figure 3.2. Share of total employment, output and value added captured by Orbis is somewhat higher in manufacturing 19
Figure 3.3. Industry-level aggregates based on Orbis show positive but moderate correlations with STAN 20
Figure 3.4. Orbis covers only a minority of firms in most countries 21
Figure 3.5. Firm coverage in Orbis changes sharply over time 22
Figure 3.6. Larger firms are better covered in Orbis 22

Figure 3.7. Orbis tends to have higher coverage in manufacturing than in services 23
Figure 3.8. Orbis sample is skewed towards large firms 24
Figure 3.9. Firms in Orbis are disproportionately large, old, productive and high-wage..... 25
Figure 3.10. Firms in Orbis are disproportionately productive even *conditional* on their size 26
Figure 3.11. Orbis underestimates dispersion in the lower half of the labour productivity distribution 27
Figure 3.12. Average firm size in an industry in Orbis shows high correlations with MultiProd across industries but low across time 28
Figure 3.13. Average productivity, average wage and productivity dispersion show lower correlations over time than across industries..... 29
Figure 3.14. Within firm growth rates show somewhat higher correlations over time and lower correlations across industries and countries 29
Figure 3.15. For many countries, Orbis better captures variation in average productivity than in productivity dispersion 30
Figure 3.16. Orbis does poorly measuring entry and exit 31
Figure 4.1. Internal imputation of value added substantially increases coverage for several countries 33
Figure 4.2. Imputation of value added makes Orbis firm distribution more similar to the population but external imputation reduces productivity dispersion..... 34
Figure 4.3. Imputation of value added increases correlations for average labour productivity but not for average value added or productivity dispersion 35
Figure 4.4. Country-years with 5000+ firms are similar to the hand-picked sample 36
Figure 4.5. Orbis sample in better covered country-years is less skewed toward large firms but keeps underestimating productivity dispersion 37
Figure 4.6. Better-covered country-years show somewhat tighter correlation with MultiProd 38
Figure 4.7. The sample skewness towards large firms tends to be weaker in manufacturing..... 39
Figure 4.8. Correlations in labour productivity over time tend to be higher in manufacturing..... 39
Figure 4.9. Orbis firms are substantially more representative of the population of firms with at least 10 employees than of the full firm population 40
Figure 4.10. Applying size thresholds does not increase the correlation between Orbis and MultiProd over time..... 41
Figure 4.11. Applying external weights makes average firm size in Orbis more similar to the population but not average productivity or productivity dispersion..... 42
Figure 4.12. External weighting increases correlations with MultiProd in average firm size but not in productivity levels, growth or dispersion 43
Figure 5.1. The “preferred sample” keeps country-periods with relatively high and stable coverage .. 45
Figure 5.2. The “preferred sample” is only slightly skewed toward large firms 46
Figure 5.3. Orbis firms in the “preferred sample” are on average more similar to the population 46
Figure 5.4. Productivity dispersion of Orbis firms in the “preferred sample” is also similar to the population 47
Figure 5.5. The “preferred sample” generally shows higher, but not very high, correlations between Orbis and MultiProd..... 48
Figure 5.6. Orbis does poorly at measuring entry and exit even with the “preferred sample”..... 48
Figure 6.1. Rounding is a major issue for several Orbis countries..... 50
1Figure 6.2. A vast majority of accounts in Orbis are unconsolidated..... 51
Figure 6.3. Firms with multiple accounts available most often belong to consolidation codes U2 and C2 52
Figure A.1. Share of total output and input captured by Orbis by country over time 62
Figure A.2. Firm coverage without conditioning on employment information availability..... 63

Chapter 1. Introduction

The last two decades have seen an explosion of empirical studies exploring firm-level or plant-level data. Unlike aggregate or sector-level data, such microeconomic data allow examining differential impact of economic shocks and policies on firms that differ, for instance, by their size, age, ownership or export status. The rich variation of firm characteristics and the possibility to control for firm-specific factors presents additional avenues to pin down causal relationships compared to more aggregate data. Furthermore, representative firm-level data enable analysing microeconomic drivers of aggregate trends in business dynamics (Decker et al., 2014), microfoundations of aggregate fluctuations (Gabaix, 2011), sources of aggregate productivity growth (Griliches and Regev, 1995; Foster et al., 2008), the role of firms in increased income inequality (Card et al., 2013; Song et al., 2015), resource misallocation (Hsieh and Klenow, 2009) and other key issues.

An important limitation of studies based on representative firm-level and plant-level data is that they are typically limited to a single country. This makes it difficult to explore the role of differences in policy and economic conditions in explaining cross-country variation and to reliably compare firm-level trends across countries. In addition, even accessing firm-level data for one country is usually far from straightforward because of the legal frameworks and the infrastructure in place to protect confidentiality; it may require special clearance, payment of an access fee, collaborations with employees in the agency holding the data or it may not be possible at all. Even if access is provided, researchers often can only work with the data at specialised computers located within the premises of the relevant agency. The access restrictions and logistical complications, together with cross-country differences in data structure, coverage and quality, mean that a vast majority of firm-level studies only analyse an individual country.

Orbis is a commercially available dataset that covers over 200 million firms across the globe. It provides company financials (e.g. operating revenue, employment, fixed capital) together with detailed information on firm ownership structure, 4-digit industry and other firm characteristics. Given these advantages and the fact that Orbis is the largest database of its kind, it is not surprising that Orbis and its European version, Amadeus, have been widely used to analyse a variety of issues including: firm-level productivity divergence (Andrews, Criscuolo, and Gal 2016), the impact of product market regulations on allocative efficiency (Arnold, Nicoletti, and Scarpetta, 2008), business dynamics and entry regulations (Klapper, Laeven, and Rajan, 2006), misallocation (Gopinath et al., 2017), persistence of unproductive "zombie" firms (Adalet McGowan, Andrews, and Millot, 2017), spillovers from foreign direct investment (Javorcik and Spatareanu, 2008, 2009, 2011), the importance of industries' position in the global value chains (Criscuolo and Timmis, 2018) and many others.

However, Orbis's advantages of flexibility and cross-country breadth come with some limitations. Orbis covers only a sample of all firms and its coverage varies from over 50% of firms in the best-covered countries to very few observations in others. The sample also changes over time within countries, which may make it difficult to analyse changes in the distribution of firms. Firms appearing in the data and disappearing from it may correspond to genuine entry and exit but may simply be a result of changes in data coverage. Other challenges of the data stem from its inclusion of consolidated accounts that may combine the financials of subsidiaries across different countries and industries, and the absence of clear rules for which firms are covered.

While the limitations of Orbis are often well-known amongst practitioners, there is little information on how important they are in practice. Are they largely innocuous or not? Does this depend on country, period, industry or firm size? What can a researcher do to make Orbis as representative as possible? This paper aims to address these questions.

It makes three main contributions.

First, it examines how representative Orbis (vintage from February 2017) is for different countries and periods and for firms of different characteristics.

Second, it examines the effects of several important choices that a researcher needs to make when preparing the Orbis data for analysis:

1. Should value added be imputed to correct for the large number of missing values?
2. Should the sample be restricted to the better covered countries, industries and years?
3. Which industries should be included in the sample?
4. Should the sample be restricted to firms over a certain size?
5. Should weights be used to make the data more representative?

Third, it discusses three aspects of Orbis data that have so far not received enough attention in the literature; (i) use of rounded values for key variables; (ii) the choice between unconsolidated and consolidated accounts; and (iii) cleaning of the ownership information in Orbis.

Analysing the representativeness of the results obtained from Orbis requires benchmark data against which Orbis can be compared. The challenge here is that cross-country firm-level data are hard to find – indeed, that is exactly the reason why Orbis is so widely used. This paper overcomes this challenge by relying on three datasets constructed at the Organisation for Economic Co-operation and Development (OECD) with the invaluable contribution of national delegates and experts from Statistical Offices, Government Departments and Research Institutions: MultiProd, DynEmp and STAN.[1]

MultiProd and DynEmp are constructed through a "distributed microdata" approach, which involves running a harmonised statistical code in a decentralised manner by national experts who have access to representative national microdata. The key advantage of such micro-aggregated datasets is that for most countries they cover the entire firm population or apply weights, which make their results more representative and comparable across countries. The MultiProd dataset is based on administrative data or the combination of production surveys and business registers; it contains a rich set of statistics describing firm growth, productivity, concentration and wages. MultiProd serves as the benchmark for most of the paper. It is complemented with information from the DynEmp dataset, which is derived from national business registers and provides information on firm entry and exit.

The MultiProd and DynEmp cross-country datasets contain observations at the country, year, and STAN A38 industry level, possibly refined according to other dimensions, such as firm size or productivity quantiles. The statistics are calculated from the official firm-level but the datasets do not contain information at the firm level, therefore they cannot be compared to Orbis at the level of individual firms. Instead, this paper applies the MultiProd statistical routines on the Orbis data, and compares the firm outcomes for each data cell. This approach has the important advantage of ensuring that differences in results are driven by differences in the data rather than in the methodology.

In addition to MultiProd and DynEmp, the STAN database, which contains harmonised national statistics describing industrial performance at the level of 2-digit industries, is used to test the extent to which Orbis can account for sectoral measures of economic activity and for sectoral trends.

The paper focuses on applications of Orbis to the study of productivity and business dynamism. Specifically, the coverage and distribution statistics are calculated over firms with non-missing employment variable. In some countries, many firms have missing employment but non-missing output or capital information, and sometimes most firms listed in Orbis have no financial information at all. Depending on the variables required, the counts of firms in Orbis reported in other sources may, therefore, be different from those reported here.

The paper finds that the suitability of Orbis data depends upon the type of analysis for which the data are used, and on the particular countries and periods considered. Firms in Orbis are disproportionally large, old and productive, even *within* their size class. Thus, weighting or undertaking analysis at the country-industry-year-size level does not solve the representativeness issues. Orbis has good coverage of larger or top performing firms, but the productivity dispersion in the bottom half of the productivity distribution appears to be reduced in Orbis, as underperforming firms are likely to be missing. Therefore, Orbis also struggles to replicate patterns observed in official aggregate and firm-level data.

Three steps are found to make Orbis significantly more representative: restricting the sample to periods of stable and high coverage within the best-covered European countries; imputing value added, using firm wage bill and earnings information; and focusing on firms with more than 10 employees. Jointly applying these three steps substantially improves the representativeness (of the population of firms with at least 10 employees), although its ability to capture evolution of the entire firm distribution remains limited, and it continues to perform poorly at measuring entry and exit.

In contrast, re-weighting, frequently suggested as a remedy for Orbis representativeness issues, is found lacking, since firms that appear in Orbis are non-randomly selected even conditional on their size.

Overall, Orbis seems more suitable for studies which focus on top performers; take a global perspective; examine mean performance; and analyse within-firm reaction to shocks. It is less suitable for studying underperforming firms; making comparisons across countries; analysing the whole distribution; and studying entry and exit.

Importantly, these findings do not imply that other cross-country firm-level datasets should be preferred over Orbis. Indeed, there is a reason why researchers who wish to work directly at the firm level and to cover multiple countries and private as well as listed companies tend to turn to Orbis. It is quite likely that alternative commercial datasets would underperform Orbis should they be subjected to similar testing. The results can be interpreted as an argument for relying on representative official microdata whenever possible, but such data normally cannot be pooled at the firm-level across multiple countries, so they only support single-country analysis or analysis at a somewhat more aggregated level (as in MultiProd and DynEmp). Where pooling firm-level data across countries is required, Orbis may well be the best option at hand, although great caution is still needed when selecting the Orbis sample and interpreting the results.

This paper builds on and develops two earlier papers concerned with the coverage and quality of Orbis data. First, the study by Kalemli-Özcan et al. (2019) provides an overview of the database and key steps needed to prepare it for analysis, and examines its coverage

using Eurostat data and the CompNet database as benchmarks. Second, Gal (2013) focuses on calculating productivity using the Orbis data. He describes availability of key variables and outlines procedures for imputation of value added, weighting and deflation of financial variables. These studies focus on data preparation and on the coverage in terms of the share of firms or the share of economic aggregates for each country and each industry or size category. The present paper also describes the coverage of Orbis (using different benchmark datasets than the earlier studies), but its main focus lies in examining moments of the firm-level data that go beyond counts and totals of basic variables and are often the outcomes of interest in firm-level studies (e.g. mean productivity, wages and age, productivity growth rates, productivity dispersion, entry and exit). Its contribution is, thus, to evaluate Orbis's performance in the task for which it is mainly used – microdata analysis. The aim is then not to give a definite black and white statement as to when Orbis data should be used, but rather put approximate numbers on its limitations, provide guidance on its use and point out analysis for which it is more (or less) appropriate.

The rest of the paper is organised as follows. Section Chapter 2. introduces Orbis and the comparison datasets, summarises the paper's methodological approach, describes variables and statistics used for the analysis and describes the baseline data construction choices. Section Chapter 3. describes coverage and representativeness of Orbis in the baseline sample. Section Chapter 4. then asks, one choice at a time, how alternative methodological choices affect Orbis's coverage and performance. Based on these findings, Section Chapter 5. defines a "preferred sample" that might be used in an applied analysis and examines its performance. Section 6 discusses three important issues related to the preparation of Orbis data: the presence of rounded values, the existence of multiple account types and the Orbis ownership module. Section 7 summarises the main findings from the analysis.

Chapter 2. Data and methodology

This section provides an introduction to Orbis data and explains how the final dataset used in this paper is constructed. It then introduces the OECD MultiProd, DynEmp and STAN datasets against which Orbis is compared in the subsequent sections. Finally, it briefly explains the methodology adopted to compare results based on the different datasets.

2.1. Orbis

Orbis is the largest cross-country firm-level database that is available and accessible for economic and financial research. It is a commercial database provided to the OECD by the electronic publishing firm, Bureau Van Dijk. The industry coverage reflects the non-farm business sector, i.e. all industries excluding agriculture and public services. However, as elaborated in more detail later, the coverage varies by country, industry, over time and across variables within the data. Furthermore, different vintages of Orbis can rely on different data providers for each country, meaning the coverage can also vary across vintages for the same country. This paper focuses on the OECD Orbis vintage from February 2017, the most recent available at the time of analysis.

The financial information primarily derives from company accounts, so some cleaning is required before using the data. We undertake a number of cleaning steps, closely following the suggestions by Kalemli-Özcan, et al. (2019) and described in more detail in previous OECD analyses (Gal, 2013; Andrews et al., 2016). In particular, we keep accounts that refer to entire calendar years, dropping observations with missing information on key variables as well as outliers identified as implausible changes or ratios and firms that are inactive. We also (i) set operating revenues and the number of employees to missing when a large number of firms in a country and year have the same value of a given variable, suggesting rounded or imputed values, and (ii) we drop duplicate accounts of the same firm in the same year. These two steps are important and not sufficiently acknowledged in the literature, so we discuss them in more detail in Section Chapter 6. .

Value added is often missing within Orbis, but it is possible to extend the coverage by imputing missing values as suggested by Gal (2013). Value-added can be imputed internally using the factor incomes provided within Orbis, i.e. the cost of employees and earnings before interest, taxes, depreciation and amortisation. However, the cost of employees is not always available within Orbis. An alternative is to impute the cost of employees combining external information on industry-average wages (for instance, from STAN) and firm-level employment from Orbis. Note that external imputation will tend to understate (overstate) the value-added of larger (smaller) firms, since larger (smaller) firms tend to pay above (below) average wages.[2] The impact of imputation on the sample size is discussed in Section 4.1 below.

Capital is measured in terms of book values, and is divided into tangible and intangible assets. However, intangible assets in Orbis reflect the accounting basis of the company reports (such as “goodwill” payments above the accounting book value upon acquisition of other firms), rather than necessarily an economic measure of intangible assets.[3]

In terms of labour input, Orbis reflects the number of employees, but does not contain information on the hours worked or the types of workers employed.

Financial information is available at different levels of aggregation within Orbis. Information may be available at a group-level, as "consolidated accounts", or at the firm-level, as "unconsolidated accounts", or both. The choice of accounts to include is an important one and it is discussed in Section 6.2 below.

In principle, Orbis contains data for more than 100 countries. In practice, however, its coverage is very low for many of them (a dozen observations per year in some cases), so only data for a minority of countries tend to be used for analysis. To reflect Orbis as it is typically used for productivity analysis spanning multiple countries, this paper focuses on countries which appear in at least two out of a list of seven studies involving such analysis.[4] It further excludes the Czech Republic, Poland and Slovakia, which in theory offer good coverage but lose most observations once rounded values are dropped (see Section 6.1). Based on these criteria, the paper analyses Orbis data for the following 20 countries: Austria, Belgium, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Italy, Japan, Korea, Netherlands, Norway, Portugal, Slovenia, Spain, Sweden, the United Kingdom and the United States. The time period it examines is 2002-2015.[5]

2.2. OECD STAN

The OECD STructural ANalysis Database (STAN) describes industrial performance for a large set of countries. Depending on country and year, the information is available at the level of 2-digit industries or at a somewhat more aggregate level. It is primarily based on national accounts, occasionally complemented by other sources such as business surveys or censuses.

STAN appears to be the most suitable aggregate point of comparison for Orbis data as analysed in this paper. Its important advantage in this context over other industry-level datasets (EU KLEMS; the World Input Output Database (Timmer et al., 2015)) is that the MultiProd code used here is based on industry definitions and industry-specific price deflators from STAN. Differences between Orbis and STAN are, therefore, not due to differences in industry classifications or deflators used (although a firm can be assigned to a different industry in each dataset). STAN is also more readily available and internationally comparable than National Accounts.

The STAN variables used in this paper include the total gross output, value added and employment. In addition, average labour productivity and wages have been calculated by dividing total value added and total wage bill by total employment, respectively.

Information from STAN is available for all 20 countries in the Orbis sample specified above, and for all years 2002-2015. Analysis comparing Orbis against STAN focuses on manufacturing; electricity, gas, water, and waste; construction; and non-financial market services.

2.3. OECD MultiProd

MultiProd is an OECD project based on a "distributed microdata analysis" and is the primary source of data against which Orbis is compared. It applies a harmonised statistical routine to confidential microdata in more than 20 countries to produce a database of statistical moments describing detailed data cells defined at the level of country, year, STAN A38 industry and, in some parts of the output, also further refined by firm characteristics, such as size or age classes, demographics (e.g. entrants, exiting firms) and productivity or sales quantiles.

It relies on two types of data. The first type consists of administrative data or production surveys, which provide the firm-level variables needed for analysing productivity and wages: output, value added, employment, fixed capital, investment into fixed capital and wage bill. For most countries the dataset is built drawing on administrative data covering the entire population of employing businesses, but in some countries (e.g. Italy and the Netherlands) such data sources are not available and surveys based on a sample of firms have to be used. For this reason, MultiProd is complemented, where necessary, with information from business registers, which contain a more limited set of variables but cover the entire firm population. The population structure captured by the business register is then used to re-weight the production surveys to obtain representative statistics at the level of each single variable.

At the time of writing, the MultiProd output covers 24 countries, among which the following 13 overlap with the Orbis country sample described in the previous subsection: Austria, Belgium, Denmark, Finland, France, Germany, Hungary, Italy, Japan, Netherlands, Norway, Portugal and Sweden. The time period covered varies across countries, but most often it starts in early 2000s and ends between 2012 and 2015.

Statistics in the MultiProd output cover a broad range of topics which largely coincide with the types of analysis for which Orbis is typically used. These include, among others, distribution of firm size and firm growth as measured by the number of employees, stock of fixed capital, sales or value added; distribution of labour productivity (LP) and multi-factor productivity (MFP) levels and growth; aggregate productivity; aggregate productivity decompositions à la Melitz and Polanec (2015) and Petrin and Levinsohn (2012); and characteristics of firms at the productivity frontier. The data, methodology and output of the MultiProd project are described in detail in Berlingieri et al. (2017). Initial research based on MultiProd output focuses on the growing dispersion of productivity and wages across firms (Berlingieri, Blanchenay, and Criscuolo, 2016; Berlingieri, Calligaris, and Criscuolo, 2017).

This paper heavily relies on the MultiProd data as a benchmark for comparison, taking them as representative of the entire firm population. The data generally are representative of the entire population of firms with at least one employment unit (i.e. employee in most cases). Exceptions are the following: i) the data for Austria and the Netherlands cover only firms with at least 10 employees, and the data for Germany cover only firms with at least 20 employees; ii) the data for Finland exclude firms with less than one full-time equivalent in terms of persons engaged; iii) the data for Japan include only manufacturing.[6] For these countries, we apply the corresponding thresholds also to the Orbis data.

For many of the statistics calculated, the MultiProd data constitute the most representative information source covering a large number of countries. The underlying microdata typically account for 80-100% of the aggregate output, value added and employment in the sectors considered. The underlying microdata for Italy and the Netherlands are a sample rather than a population and account for only about 50% of the aggregate output, value added and employment; however, they are made representative and comparable through re-weighting based on business registers information.[7] To date, studies based on the MultiProd data have focused on manufacturing and non-financial market services, and the parts of the analysis presented here that are based on MultiProd data follow this practice. They also exclude “coke and refined petroleum”, “real estate” and “scientific R&D,” which appear to have somewhat less reliable information. These caveats should be kept in mind when interpreting the overall findings of this paper and the reported comparisons across countries and industries. For more information, including MultiProd coverage for

individual countries and industries, see the comprehensive overview of the MultiProd dataset (Berlingieri et al., 2017) and a paper comparing MultiProd to aggregate figures reported in the OECD STAN database (Bajgar, Berlingieri, Calligaris, and Criscuolo, 2019).

2.4. OECD DynEmp

Similar to MultiProd, the OECD DynEmp project is based on a distributed data collection that creates a harmonised cross-country micro-aggregated database. DynEmp differs from MultiProd in focus and the underlying microdata. It analyses employment dynamics and its primary sources of firm and establishment data are national business registers and, in some cases, social security records. The most recent wave of data collection, named DynEmp version 3, includes a disaggregated analysis of the growth patterns of incumbents and start-ups, following cohorts of entrants for three, five, seven and ten years after their entry in the market. It allows separately identifying different channels of employment, distinguishing between gross job creation and job destruction, and between the role of firm entry and exit and post-entry growth. The role of firm age and size can also be examined. The DynEmp output is now available at the level of the same cells defined by country, A38 industry and year as MultiProd's output. For example, the data has been used for examining overall dynamics of employment growth (Criscuolo, Gal, and Menon; 2014), start-up dynamics (Calvino, Criscuolo, and Menon; 2015), the effect of policies on start-up dynamics and employment growth (Calvino, Criscuolo, and Menon; 2016) and a range of country-specific analyses (OECD; 2017). The Dynemp code is described in detail by Criscuolo, Gal, and Menon (2015).

An important advantage of DynEmp is that its use of business registers, which contain the universe of firms, allows examining entry and exit more reliably than MultiProd. For this reason, this paper uses DynEmp rather than MultiProd as a benchmark for the rates of entry and exit observed in Orbis.

2.5. Definitions of basic variables

There are some differences in the definition of variables across the data sources used.

- **Employment.** Orbis contains information on the "number of employees", but the exact definition of the employment information in Orbis depends on each country and data provider. STAN contains both the "number of employees" and "persons engaged"; to better match Orbis, "number of employees" is used here. In MultiProd, employment is also primarily measured as the "number of employees", but it is measured as "persons engaged" for some countries.[8]
- **Gross output.** Both STAN and MultiProd contain information on gross output. In Orbis, it is proxied by operating revenue, which may differ from gross output, for instance, due to changes in inventory, capitalised production and purchases of goods that are resold.
- **Value added.** STAN, MultiProd and Orbis all contain a variable for value added.[9]
- **Intermediate inputs.** National microdata underlying MultiProd contain a variable for intermediate inputs. In Orbis, intermediate inputs are calculated as the difference between operating revenue and value added.[10]

- **Capital stock.** In Orbis, a measure of capital stock is based on the book value of tangible fixed assets. A measure of intangible assets is intentionally excluded, since this largely reflects accounting measures of "goodwill" resulting from mergers and acquisitions. The MultiProd code constructs capital stock using a perpetual inventory method, see Berlingieri et al. (2017) for more detail. The two measures might substantially differ, as the Orbis measure does not capture intangible assets whereas some intangible assets can be captured in the microdata underlying MultiProd, and as the use of accelerated depreciation rates for tax purposes might reduce the capital stock observed in Orbis.
- **Wages.** Orbis, STAN and MultiProd all report the wage bill of each firm. In all sources, the wage bill tends to be more broadly defined (e.g. including social security contributions), with what exactly is included depending on each country.

2.6. Methodology

The idea of this paper is to uncover the representativeness of Orbis data by comparing statistics calculated using Orbis to the same statistics calculated with data that is known to be representative (STAN, MultiProd, DynEmp).

The comparison with STAN aims to evaluate (1) the share of total number of employees, output and value added captured by Orbis and (2) the ability of Orbis data to capture variation in industry-level economic outcomes over time, across industries and across countries. The statistics used in this part of the analysis, calculated for each combination of a country, A38 industry and year, are the following:

- **Total number of employees, output and value added.**
- **Labour productivity** calculated as total value added divided by total number of employees.
- **Average wage** calculated as total wage bill divided by total number of employees.

The comparison with MultiProd and DynEmp then tries to understand (1) the firm coverage of Orbis data, (2) the representativeness of Orbis data in terms of the distribution of firm characteristics and (3) the ability of Orbis data to capture variation in industry-level moments of firm-level data over time, across industries and across countries. It applies exactly the same methodology to Orbis and to representative national microdata (through MultiProd and DynEmp) and then compares various moments of the conditional distributions. The methodology and output statistics are provided by the MultiProd and DynEmp codes. Of the many measures produced by the MultiProd and DynEmp codes, the analysis focuses on the following subset.[11]

- **Distribution of employment, gross output, value added, capital stock and age.**
- **Distribution of labour and multi-factor productivity and average wage.** Labour productivity is defined as value added divided by employment. The measure of MFP used in this paper is based on Solow residuals using external cross-country industry-specific labour, capital and intermediate input shares.[12] Average wage is calculated as total wage bill of each firm divided by the number employees.
- **Dispersion of productivity across firms** is defined as the difference between natural logarithms of different percentiles of the productivity distribution. The

measures employed here compare 90th to 10th percentile, 90th to 50th percentile and 50th to 10th percentile.

- **Distribution of within-firm growth rates in firm size, capital stock and labour productivity.** The growth rates are calculated as annual changes in natural logarithms of the variables.
- **Entry and exit rates.** The entry and exit rates represent the number of entrants and exitors over the total number of firms.[13] The definition of entry applied here takes advantage of information on firm age rather than simply considering all firms newly appearing in the data as entrants; firms which newly appear in the data but are older than 1 year are not considered as entrants.

Several considerations help ensure that any differences between statistics calculated using Orbis and those calculated using the other data sources are due to differences in the underlying data rather than methodology used to generate the statistics. First, all statistics for Orbis data are based on the output obtained by running the MultiProd code on Orbis data. That means that Orbis-based and MultiProd-based statistics are calculated using exactly the same statistical and data-cleaning routines. Second, all data are deflated using the same industry-specific deflators from the STAN database. Third, any comparisons of coverage and distributions are performed only using those country-industry-year observations that appear in both datasets under consideration. Finally, in the cases where the benchmark data is representative only of firms above a certain size threshold (MultiProd data for Austria, the Netherlands and Germany), the same threshold is applied to the Orbis data for comparisons with MultiProd (but not with STAN).

The MultiProd code, applied here to Orbis, drops observations with missing employment information. As a result, employment is the best covered variable by construction. In the raw Orbis data, many countries offer better coverage for output, and occasionally also for other variables, than for employment. This means that results throughout this paper should be interpreted as describing Orbis firms with non-missing employment information. This seems reasonable for any analysis that looks at productivity. Nevertheless, studies that only require information on gross output will find better coverage for some countries than described here. There are many decisions that an Orbis user has to make, as noted earlier. For example, should value added be imputed to correct for the large number of missing values? Should the sample be restricted to the better covered firm size classes, countries, industries and years? Should weighting be applied? As separately testing each combination of these choices is not possible due to the large number of such combinations, this paper takes a two-step approach. In Section Chapter 3. , it evaluates Orbis in a baseline specification, with no imputation of value added, no weighting and no firm size threshold. In Section Chapter 4. , it then tests what happens when the baseline specification is altered in one dimension at a time.

Chapter 3. Baseline representativeness

This section analyses the representativeness of Orbis in the baseline specification. First, it examines how closely Orbis can match industry-level aggregates as observed in the STAN data. It then evaluates the coverage of Orbis, using MultiProd data as a benchmark. Finally, it describes how well Orbis can match firm distribution characteristics observed in the official microdata, again using MultiProd, together with DynEmp.

3.1. Matching industry aggregates

This subsection compares Orbis data, aggregated up to the level of country-industry cells, to OECD STAN. It first examines the share of the total employment, output and value added captured by Orbis, and then it evaluates the extent to which Orbis data can capture industry-level variation over time, across industries and across countries.

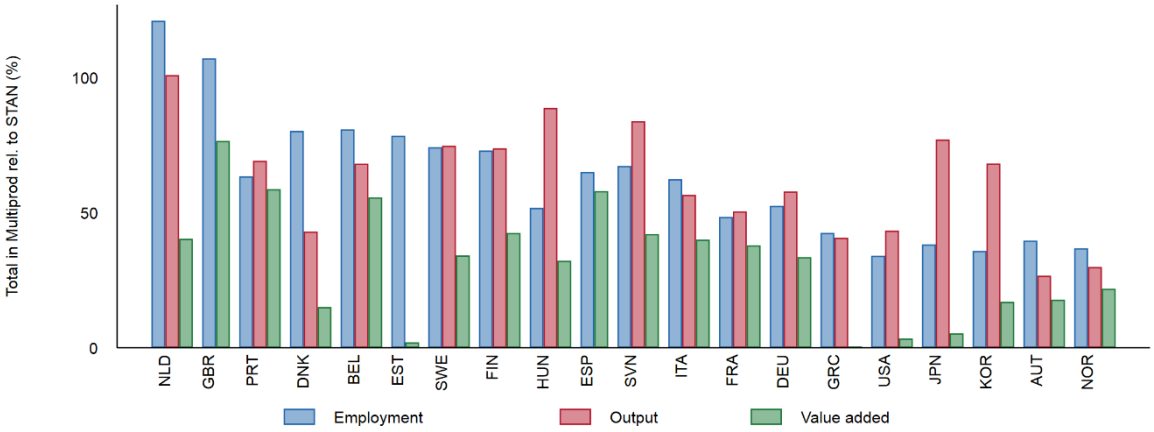
The sample for the analysis in this section spans manufacturing, utilities, construction and non-financial market services for the period 2002-2015. The countries covered are Austria, Belgium, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Italy, Japan, Korea, Netherlands, Norway, Portugal, Slovenia, Spain, Sweden, the United Kingdom and the United States.

Orbis data in the sample (which essentially consists of the best-covered Orbis countries) typically capture around 60% of aggregate employment and output and around 40% of aggregate value added (Figure 3.1).[14] There is, however, a large variation across countries and variables. The share of aggregate output captured by Orbis ranges from over 100% in the Netherlands and the United Kingdom (results likely driven by consolidated accounts of multinationals deriving a large share of their income from abroad) to just 30-40% in Austria, Norway and all three non-European countries in the sample: Japan, Korea and the United States. The share of employment captured is somewhat more evenly distributed, but is rather low for Austria and Norway. Value added tends to be the least well captured by Orbis, with just 10% of value added or less captured for Japan and the United States.

Importantly, note that Orbis offers better coverage for large firms (see Section Chapter 3. below), so it accounts for a larger share of the total employment, output and value added than of the number of firms.

Figure 3.1. Orbis data capture around 60% of aggregate employment and output and around 40% of aggregate value added

Total employment, output and value added relative to STAN, by country (mean over years, 2002-2015)



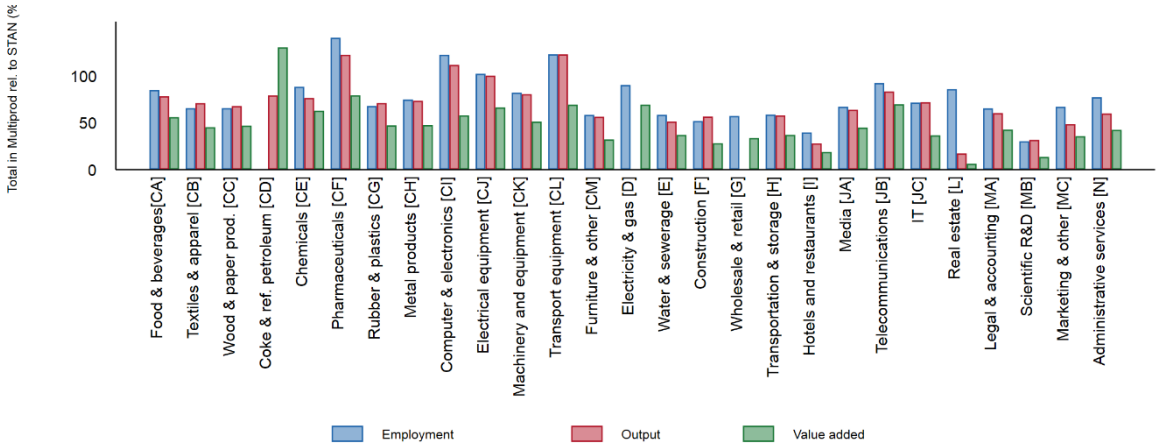
Note: The graphs shows the total employment, output and value added in Orbis relative to STAN. Only Orbis firms with non-missing employment included. Manufacturing, utilities, construction and non-financial services. Figures for gross output exclude "Wholesale and retail".
Source: Orbis and OECD STAN.

For many countries, the share covered by Orbis varies dramatically over time (see Figure A.1). For example, for all variables in Portugal, it increases in 2006 from less than 5% to more than 80%.

Looking at variation across industries, the share of employment, output and value added captured by Orbis is somewhat larger for manufacturing than for the other macro-sectors considered. The variation seems to reflect the size composition of each sector, where "pharmaceuticals" and "telecommunications" show relatively good coverage and are also industries with a strong presence of large firms.[15]

Figure 3.2. Share of total employment, output and value added captured by Orbis is somewhat higher in manufacturing

Total employment, output and value added relative to STAN, by industry (mean over years, 2002-2015)



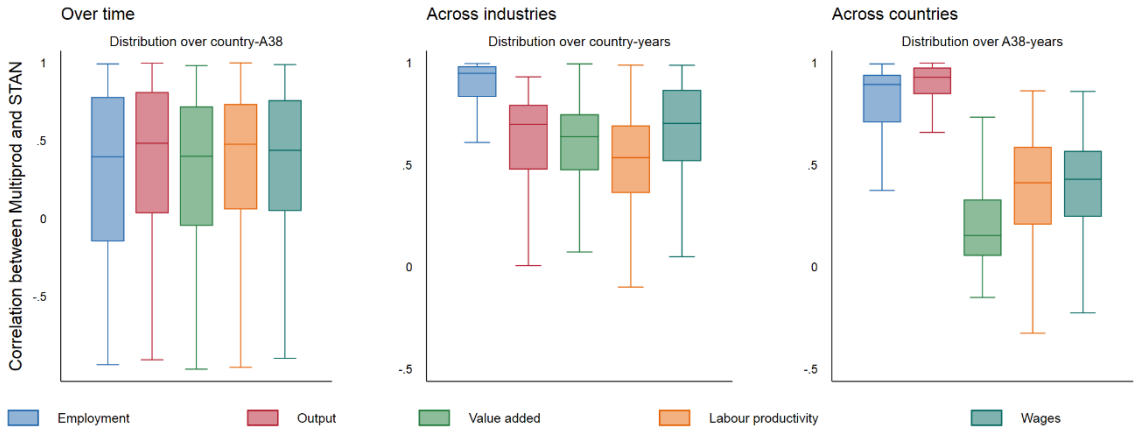
Note: The graphs shows the total employment, output and value added in Orbis relative to STAN. Only Orbis firms with non-missing employment included. Manufacturing, utilities, construction and non-financial services. Figures for gross output exclude "Wholesale and retail". Countries: AUT, BEL, DEU, DNK, ESP, EST, FIN, FRA, GBR, GRC, HUN, ITA, JPN, KOR, NLD, NOR, PRT, SVN, SWE, USA.
Source: Orbis and OECD STAN.

Correlations between industry-level values observed in Orbis and STAN are mostly positive and moderately high, with the median correlation around 0.5 for most variables and types of variation. Figure 3.3 displays three types of correlations between Orbis and STAN in terms of total employment, total output, total value added, aggregate labour productivity and average wage. The left panel shows the distribution of correlations which have been calculated over time within each country-A38 industry pair. The middle panel focuses on correlations calculated across industries within each country-year pair. The right panel describes correlations calculated across countries for each A38 industry-year pair. Correlations over time have median values around 0.4-0.5 for all variables. Correlations across industries are very high for employment and also fairly high (medians 0.5-0.7) for the other variables. Correlations across countries are very high for both employment and output, but rather low for the other three variables (medians below 0.5).

It would be unreasonable to expect a correlation of 1 between firm-level data and industry-level values from national accounts. Indeed, there are other reasons besides incomplete coverage why aggregate values calculated directly from microdata can differ from official statistics. For example, when firms change industries, official statistics will tend to count them in a different industry in each year, whereas Orbis assigns firms to the latest available industry for all years and MultiProd uses the most frequent (mode) industry of each firm. However, Bajgar et al. (2019) show that representative national microdata can approximate variation in industry aggregates reasonably well, as they find corresponding correlations between MultiProd and STAN data to be greater than 0.75 for all variables and types of variation and close to 1 in several cases (see Bajgar et al., 2019).

Figure 3.3. Industry-level aggregates based on Orbis show positive but moderate correlations with STAN

Correlations between Orbis and STAN by variable and type of variation (2002-2015)



Note: Left panel: correlations across years, calculated separately for each country-A38 pair. Middle panel: correlations across industries, calculated separately for each country-year pair. Right panel: correlations across countries, calculated separately for each A38-year pair. The graph plots the dispersion over country-A38s (left), country-years (middle) or A38-years (right) of these correlations. Manufacturing, utilities, construction and non-financial services. Countries: AUT, BEL, DEU, DNK, ESP, EST, FIN, FRA, GBR, GRC, HUN, ITA, JPN, KOR, NLD, NOR, PRT, SVN, SWE, USA.
Source: Orbis and OECD STAN.

3.2. Firm coverage

This subsection compares the firm coverage of Orbis to that observed in official microdata as reflected in MultiProd output. The sample used for this and the following subsections covers Austria, Belgium, Denmark, Finland, France, Germany, Hungary, Italy, Japan, Netherlands, Norway, Portugal and Sweden for a period starting in early 2000s and ending between 2012 and 2015, depending on the country.

Unless otherwise stated, all coverage and distribution statistics reported here are calculated *conditional on non-missing employment*. This is important to keep in mind because firms with non-missing employment often represent only a fraction of all firms listed by Orbis, and, for some countries, other variables (e.g. output) are available for more firms than employment.

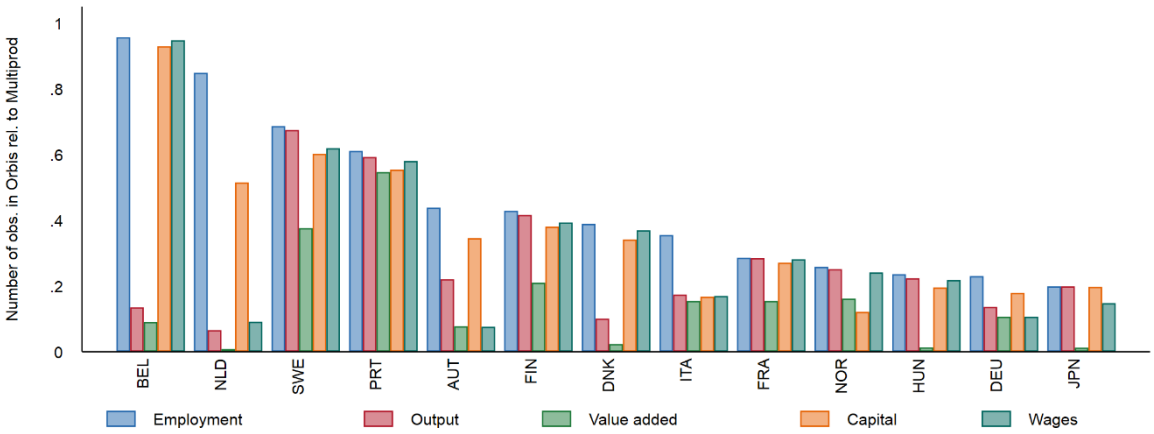
Conditional on non-missing employment, the firm coverage varies from over 80% for some countries, years and variables to close to zero for others. Figure 3.4 shows the number of firms in Orbis relative to the number of firms in the microdata underlying MultiProd. It separately shows relative coverage in terms of observations which have each of the following variables available: employment, output, value added, capital and wages. For each country, the graph shows average coverage over the years.[16] The coverage is not far from 100% for some variables in Belgium and the Netherlands, but it is quite low for both output and value added in these countries. Sweden and Portugal are the countries with the highest coverage consistently across all variables, at 40-60%. On the other hand, about half the countries in the sample have coverage in the 10-30% range, and sometimes even lower for value added (Denmark, Hungary, Japan).

Out of the key variables, capital and wages are available for most firms with non-missing employment. A substantial share of firms in several countries lack information on output (Belgium, the Netherlands, Austria, Denmark, Italy and Germany). Value added is the least well covered variable, although this improves to some extent when it is imputed based on wage bill and earnings information (see Subsection 4.1 below).

For several countries, the coverage increases when we do not condition on non-missing employment variable (Figure A.2). In particular the availability of the output variable is significantly higher than the availability of the employment variable for Finland, France, Hungary, the earlier years in Germany and Italy and the later years in Norway.

Figure 3.4. Orbis covers only a minority of firms in most countries

Firm coverage by country, (mean over years, 2002-2015)



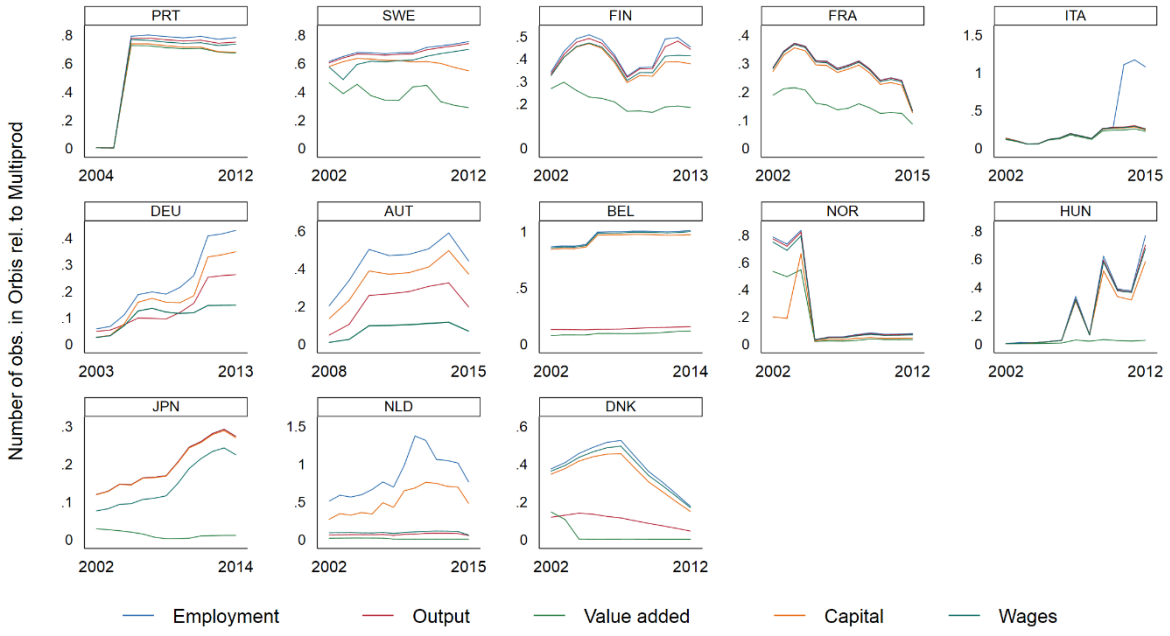
Note: The graph shows the number of observations in Orbis with employment and given variable available relative to the number of observations in MultiProd with employment available. Manufacturing and non-financial services (excluding "Coke and refined petroleum", "Real estate" and "Scientific R&D").
Source: Orbis and OECD MultiProd.

Coverage in Orbis changes sharply over time in many countries (see Figure 3.5). While coverage in some countries shows a steady increase (Germany, Japan), in others it displays high volatility (Hungary, Finland), a large one-off increase (Portugal) or a deep drop (Norway). The large changes in coverage over time are important because they are likely to entail changes in the composition of the sample which may confound analysis of changes in firm distribution over time.

The Orbis sample is heavily skewed towards larger firms. Figure 3.6 contains a box plot[17] showing the distribution of firm coverage across cells defined by country, industry and year, separately for each firm size class. The median coverage in terms of employment, output and value added is, respectively, only 43%, 29% and 7% for micro-firms with less than 10 employees but 83%, 79% and 50% for large firms with at least 250 employees.

Figure 3.5. Firm coverage in Orbis changes sharply over time

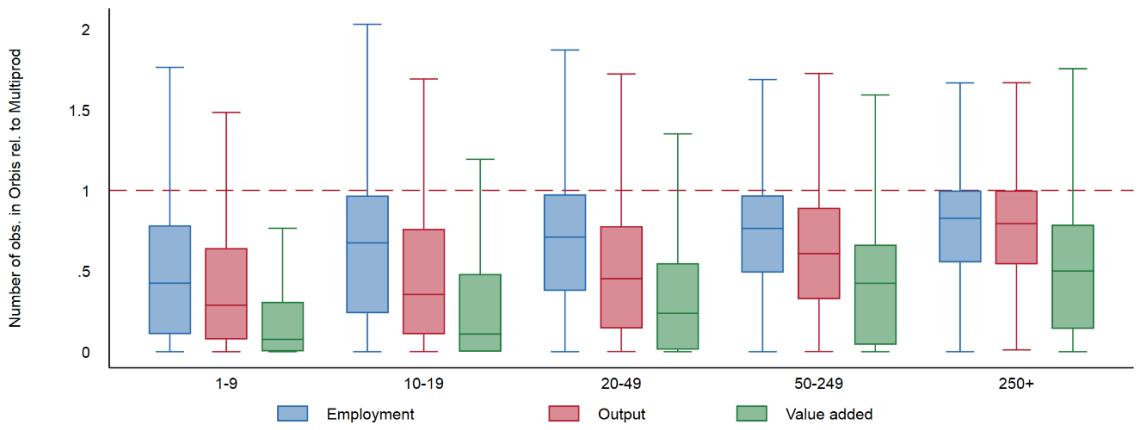
Firm coverage by country over time (2002-2015)



Note: The graph shows the number of observations in Orbis with employment and given variable available relative to the number of observations in MultiProd with employment available. Manufacturing and non-financial services (excluding "Coke and refined petroleum", "Real estate" and "Scientific R&D").
Source: Orbis and OECD MultiProd.

Figure 3.6. Larger firms are better covered in Orbis

Firm coverage by size class, distribution over country-industry-years (2002-2015)

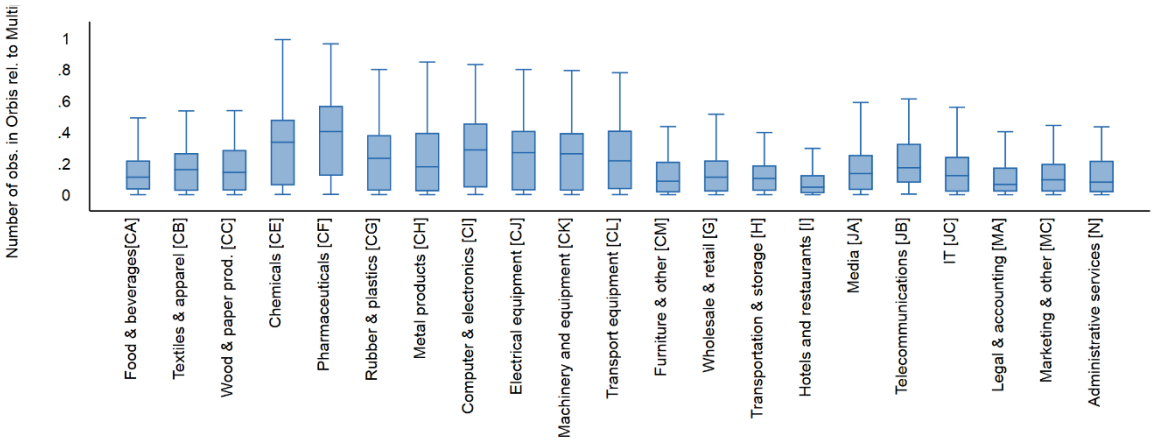


Note: The graph shows the number of observations in Orbis with employment and given variable available relative to the number of observations in MultiProd with employment available. For each firm size class and variable, it shows the distribution of the coverage over country-A38-year combinations. Manufacturing and non-financial services (excluding "Coke and refined petroleum", "Real estate" and "Scientific R&D"). Countries: AUT, BEL, DNK, FIN, FRA, DEU, HUN, ITA, JPN, NLD, NOR, PRT, SWE.
Source: Orbis and OECD MultiProd.

Orbis is not only more representative for larger firms but also for industries which consist of larger firms to a greater extent. Figure 3.7 shows how within-cell coverage varies across individual A38 industries. It tends to be greater for manufacturing than for services, but there is also a substantial variation within each of these broad sectors. The differences are likely to be at least partly driven by size composition of each industry. For example, the pharmaceutical sector, which contains relatively few small firms, is relatively well covered in Orbis, whereas wholesale and retail and hotels and restaurants sectors, which contain a large number of small firms (alongside some giant chains), have only a small share of firms covered in Orbis.

Figure 3.7. Orbis tends to have higher coverage in manufacturing than in services

Firm coverage by A38 industry, distribution over country-years (2002-2015)



Note: The graph shows the number of observations in Orbis with employment and value added available relative to the number of observations in MultiProd with employment available. For each industry, it shows the distribution of the coverage over country- year combinations. Countries: AUT, BEL, DNK, FIN, FRA, DEU, HUN, ITA, JPN, NLD, NOR, PRT, SWE.
Source: Orbis and OECD MultiProd.

3.3. Matching firm distribution characteristics

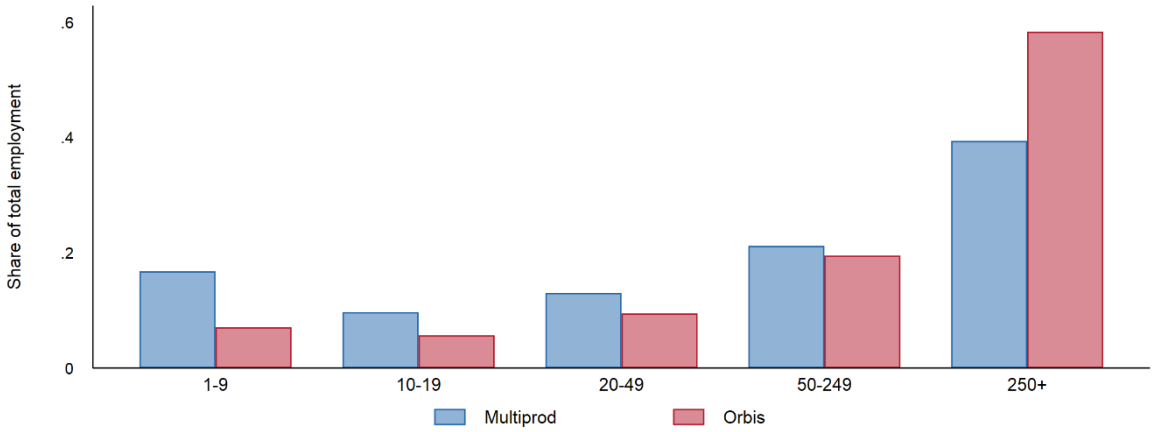
After examining firm coverage, the analysis now turns to the ability of Orbis to match properties of the firm distribution which are observed in official microdata. It proceeds in four steps. First, it compares firms observed in Orbis to those in MultiProd in terms of within-cell means of basic variables such as employment, physical capital and age, as well as constructed measures of firm performance, including productivity and average wage. Second, it examines the amount of firm heterogeneity captured by Orbis by comparing the productivity dispersion between top-performing and lagging firms in the two datasets. Third, it examines correlations between the two datasets along various dimensions. Fourth, it examines entry, exit and their correlation between Orbis and DynEmp.

Firms in Orbis are systematically larger. Figure 3.8 describes the firm employment size distribution in MultiProd and in Orbis. It shows the share of the total employment observed in each dataset accounted for by firms in each of five size categories. The displayed values represent the means over all countries and years. The graph clearly shows that the firm size distribution in Orbis is skewed towards large firms. Large firms with 250 or more employees account, on average, for about 40% all employment according to the MultiProd

data but for about 60% in Orbis. On the contrary, small firms with less than 10 employees account, on average, for 17% of aggregate employment in MultiProd but just 7% in Orbis.

Figure 3.8. Orbis sample is skewed towards large firms

Distribution of employment over firm size categories (mean over country-years, 2002-2015)



Note: Each bar represents the share of total employment observed in a given dataset accounted for by firms in a given firm size category. Firm size categories are defined by employment. Manufacturing and non-financial services (excluding "Coke and refined petroleum", "Real estate" and "Scientific R&D"). Countries: BEL, DNK, FIN, FRA, HUN, ITA, JPN, NOR, PRT, SWE.
Source: Orbis and OECD MultiProd.

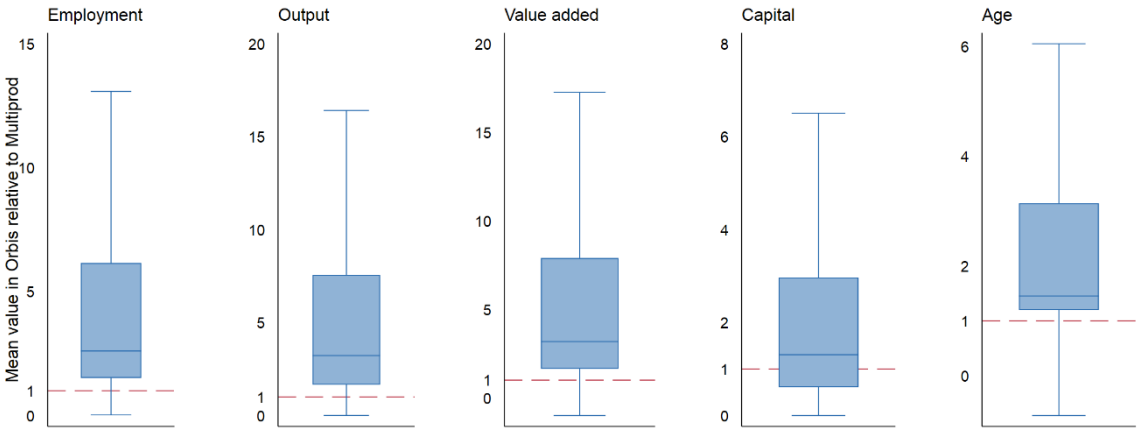
Firms in Orbis are larger when size is measured by employment, output, value added or physical capital, and they are also older. Panel A, in Figure 3.9, shows the means of these variables as observed in Orbis relative to their means in the MultiProd output. It describes the distribution of the relative values over country-A38-year combinations. For a median cell, the average firm in Orbis is two-and-half times larger in terms of employment compared to the average firm in MultiProd. It also has more than three times larger gross output and value added, 30% larger capital[18] and it is 50% older. In many cells, the differences are even larger. For a quarter of country-A38-year combinations, the average firm size is more than six times larger in Orbis than in MultiProd. This means that Orbis captures a much larger share in the total employment or output than in the total number of firms, but also that the Orbis sample is far from representative for most countries, industries and years.

Firms in Orbis are also more productive and pay higher wages. Panel B, in Figure 3.9, shows that, compared to MultiProd, an average firm in Orbis has 35% greater labour productivity, 16% greater total factor productivity and pays 20% higher wages.[19]

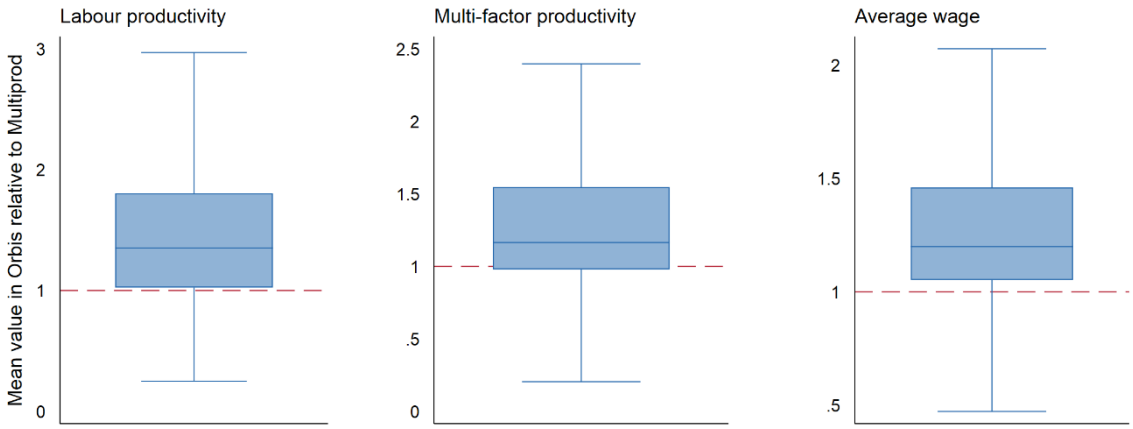
Figure 3.9. Firms in Orbis are disproportionately large, old, productive and high-wage

Industry averages in Orbis relative to MultiProd, distribution over country-industry-years (2002-2015)

Panel A: Firm size and age



Panel B: Firm productivity and average wage

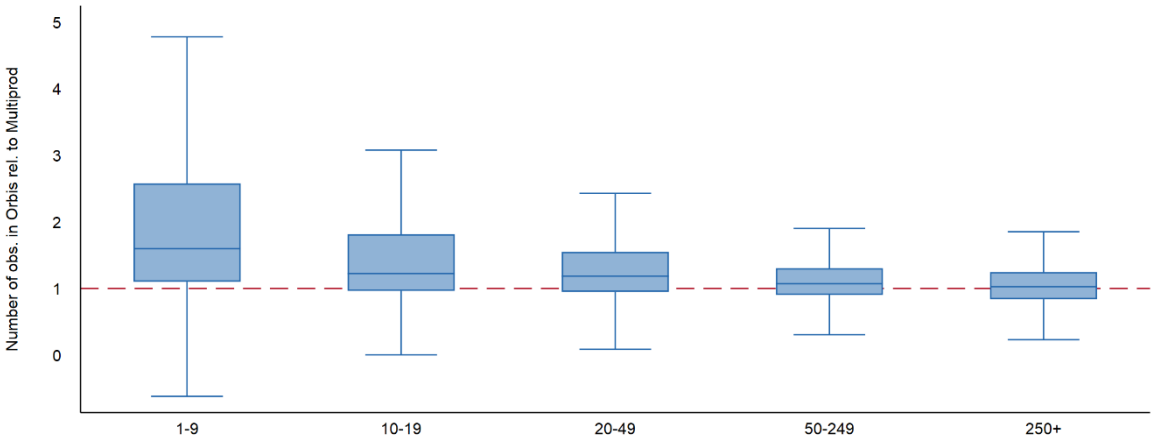


Note: The graph describes ratios of average employment, output, value added, capital and age, labour productivity, multi-factor productivity and average wage in Orbis and in MultiProd for each industry. It shows a distribution of the ratios over country-A38-year combinations. Manufacturing and non-financial services (excluding "Coke and refined petroleum", "Real estate" and "Scientific R&D"). Countries: AUT, BEL, DNK, FIN, FRA, DEU, HUN, ITA, JPN, NLD, NOR, PRT, SWE.
Source: Orbis and OECD MultiProd.

These differences can be partially attributed to the larger average size of firms in Orbis. However, firms in Orbis are disproportionately productive even *conditional on firm size*. Figure 3.10 shows the mean productivity in Orbis relative to MultiProd separately for individual firm size classes. It shows that Orbis firms are on average more productive within each size class. This is particularly true of the smaller firms, which are underrepresented in Orbis and presumably only appear in the data if they are particularly productive relative to peers in the same size class. For the median country-industry-year, the labour productivity difference between an average firm in Orbis and in MultiProd is just 3% for large firms above 250 employees but 23% for small firms with 10-19 employees and 60% for micro-firms with less than 10 employees.

Figure 3.10. Firms in Orbis are disproportionately productive even *conditional* on their size

Industry average labour productivity in Orbis relative to MultiProd by firm size class, distribution over country-industry-years (2002-2015)

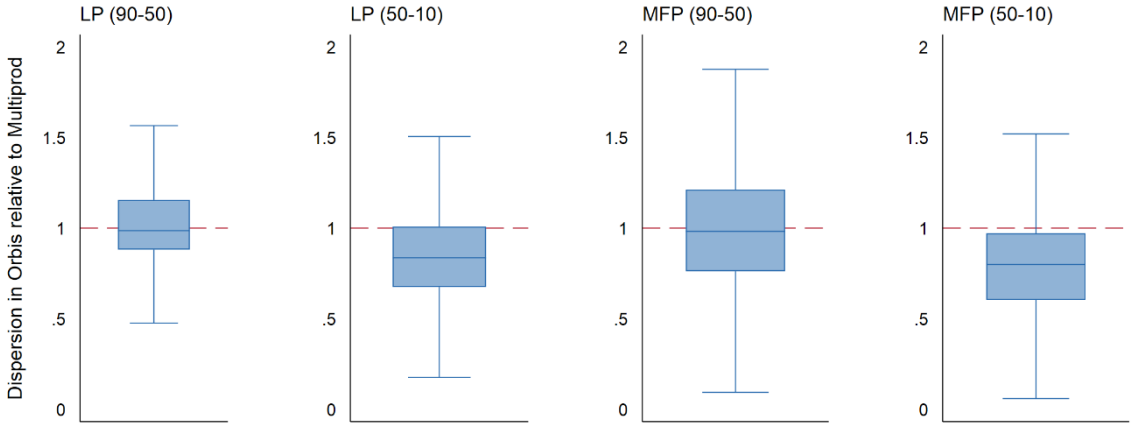


Note: The graph describes the ratio of average labour productivity in Orbis and in MultiProd. For each firm size class, it shows a distribution of the ratios over country-A38-year combinations. Manufacturing and non-financial services (excluding "Coke and refined petroleum", "Real estate" and "Scientific R&D"). Countries: AUT, BEL, DNK, FIN, FRA, DEU, HUN, ITA, JPN, NLD, NOR, PRT, SWE.
Source: Orbis and OECD MultiProd.

Orbis seems to be more representative for frontier firms than for firms in the lower half of the productivity distribution. Going beyond averages, recent research has devoted a lot of attention to differences across firms and in particular productivity dispersion.[20] Figure 3.11 describes productivity dispersion in Orbis relative to MultiProd. For both labour productivity and multi-factor productivity, Orbis captures fairly well the productivity dispersion between the top performers (90th percentile of the productivity distribution) and the typical (median) firm, but it underestimates the dispersion between the typical firm and underperforming firms (10th percentile). This is in line with the idea that Orbis is more representative of larger and more productive firms but is largely missing firms which are towards the lower end of the performance spectrum.

Figure 3.11. Orbis underestimates dispersion in the lower half of the labour productivity distribution

Labour productivity dispersion in Orbis relative to MultiProd, distribution over country-industry-years (2002-2015)



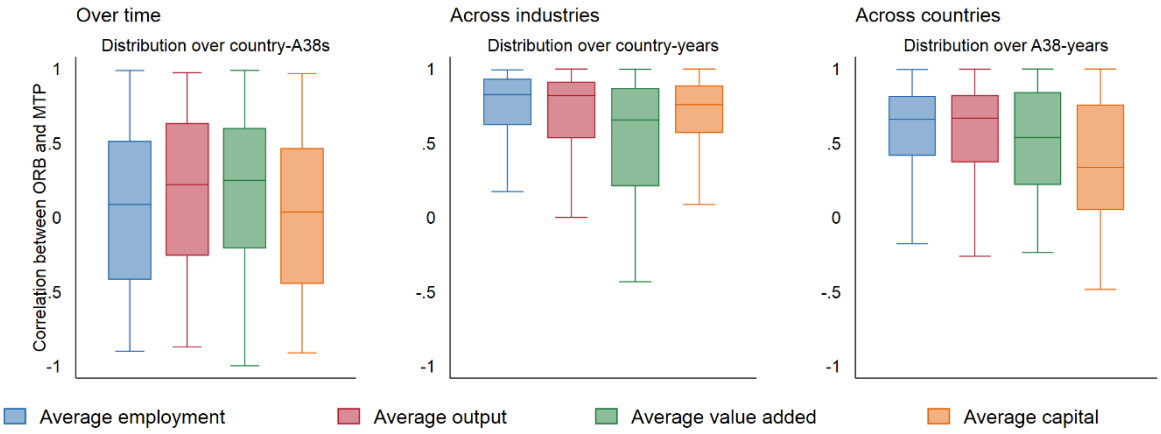
Note: The graph describes ratios of labour productivity and multi-factor productivity dispersions between 90th and 50th percentile and between 50th and 10th percentile of firm productivity distribution. It shows a distribution of the ratios over country-A38-year combinations. Manufacturing and non-financial services (excluding "Coke and refined petroleum", "Real estate" and "Scientific R&D"). Countries: AUT, BEL, DNK, FIN, FRA, DEU, HUN, ITA, JPN, NLD, NOR, PRT, SWE.
Source: Orbis and OECD MultiProd.

The analysis so far has focused on static comparisons of firm distributions. It is also important to understand to what extent Orbis can capture patterns across countries, across industries and, probably most important, over time.

Orbis can capture differences in average firm size across industries and, to a smaller extent, across countries, but not over time. Figure 3.12 displays three types of correlations between Orbis and MultiProd in terms of average employment, average output, average value added and average capital. The left panel shows the distribution of correlations which have been calculated over time within each country-A38 pair. The middle panel focuses on correlations calculated across industries within each country–year pair. The right panel describes correlations calculated across countries within each A38 industry-year pair. Correlations over time (left panel) are close to zero for a median country-industry, suggesting a poor ability of Orbis to capture evolution of average firm size over time. In contrast, correlations across industries are relatively high, with median between 0.7 and 0.8. Correlations across countries tend to be positive but moderate (around 0.5). The results are in line with the idea that differences in coverage are generally smaller across industries than over time and across countries.

Figure 3.12. Average firm size in an industry in Orbis shows high correlations with MultiProd across industries but low across time

Correlations between Orbis and MultiProd by variable and type of variation (2002-2015)



Note: Left panel: correlations across years, calculated separately for each country-A38 pair. Middle panel: correlations across industries, calculated separately for each country-year pair. Right panel: correlations across countries, calculated separately for each A38-year pair. The graph plots the dispersion over country-A38s (left), country-years (middle) or A38-years (right) of these correlations. Manufacturing and non-financial services (excluding "Coke and refined petroleum", "Real estate" and "Scientific R&D"). Countries: AUT, BEL, DNK, FIN, FRA, DEU, HUN, ITA, JPN, NLD, NOR, PRT, SWE.
Source: Orbis and OECD MultiProd.

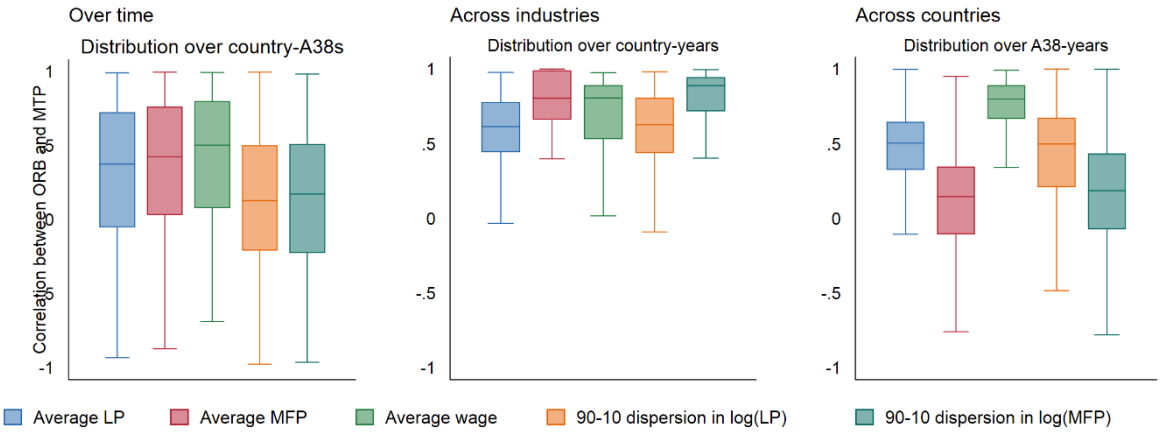
Orbis has a limited ability to capture the evolution of average productivity and wages, and it struggles to capture the evolution of productivity dispersion (Figure 3.13). Also in the case of average productivity, average wage and productivity dispersion, correlations across industries are highest and correlations over time lowest. Correlations over time are somewhat higher for average productivity and wages than for average size, but they are very low for productivity dispersion, highlighting poor suitability of Orbis (at least in the sample analysed here) for analysing changes in the shape of the distribution of firm productivity. Correlations across industries are very low for average multi-factor productivity and multi-factor productivity dispersion.

Orbis captures variation over time in average within-firm growth rates of firm size and productivity better than variation over time in the average levels of these variables (Figure 3.14). At the same time, correlations across industries and countries are somewhat lower for growth rates than for levels.

Countries differ in the ability of Orbis data to explain changes in the productivity distribution over time (Figure 3.15). However, even the countries with good and relatively stable coverage where Orbis is able capture variation in average productivity quite well (e.g. Sweden, Finland, Belgium) show low correlations in terms of productivity dispersion.

Figure 3.13. Average productivity, average wage and productivity dispersion show lower correlations over time than across industries

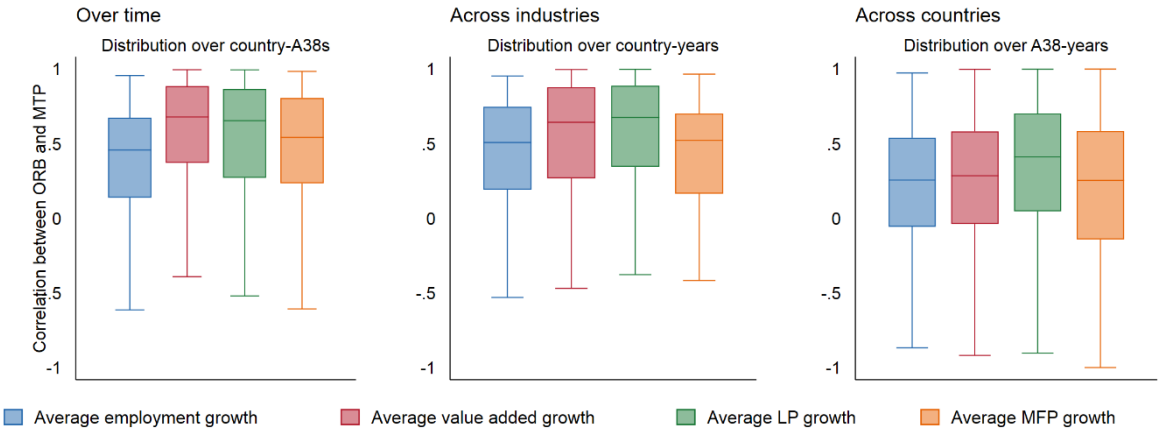
Correlations between Orbis and MultiProd by variable and type of variation (2002-2015)



Note: Left panel: correlations across years, calculated separately for each country-A38 pair. Middle panel: correlations across industries, calculated separately for each country-year pair. Right panel: correlations across countries, calculated separately for each A38-year pair. The graph plots the dispersion over country-A38s (left), country-years (middle) or A38-years (right) of these correlations. Manufacturing and non-financial services (excluding "Coke and refined petroleum", "Real estate" and "Scientific R&D"). Countries: AUT, BEL, DNK, FIN, FRA, DEU, HUN, ITA, JPN, NLD, NOR, PRT, SWE.
Source: Orbis and OECD MultiProd.

Figure 3.14. Within firm growth rates show somewhat higher correlations over time and lower correlations across industries and countries

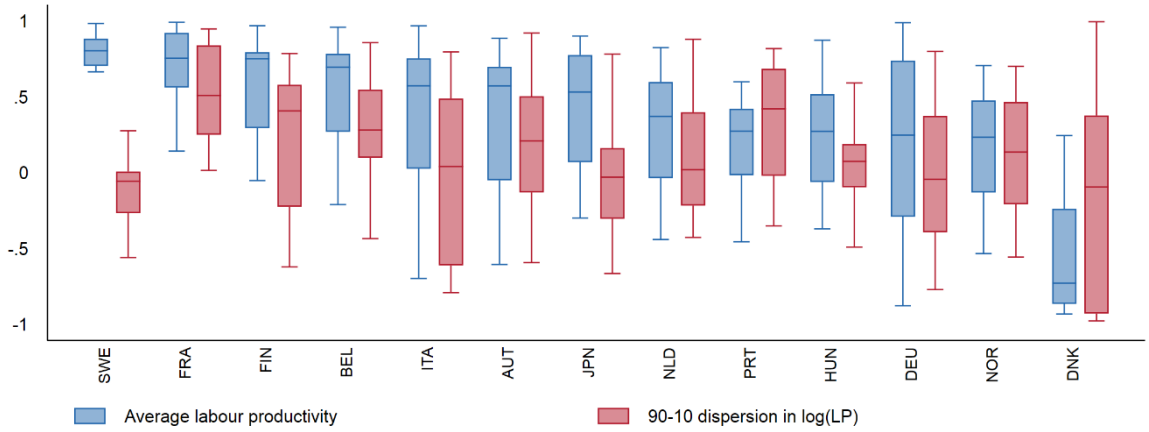
Correlations between Orbis and MultiProd by variable and type of variation (2002-2015)



Note: Left panel: correlations across years, calculated separately for each country-A38 pair. Middle panel: correlations across industries, calculated separately for each country-year pair. Right panel: correlations across countries, calculated separately for each A38-year pair. The graph plots the dispersion over country-A38s (left), country-years (middle) or A38-years (right) of these correlations. Manufacturing and non-financial services (excluding "Coke and refined petroleum", "Real estate" and "Scientific R&D"). Countries: AUT, BEL, DNK, FIN, FRA, DEU, HUN, ITA, JPN, NLD, NOR, PRT, SWE.
Source: Orbis and OECD MultiProd.

Figure 3.15. For many countries, Orbis better captures variation in average productivity than in productivity dispersion

Correlations between Orbis and MultiProd over time, by variable (2002-2015)

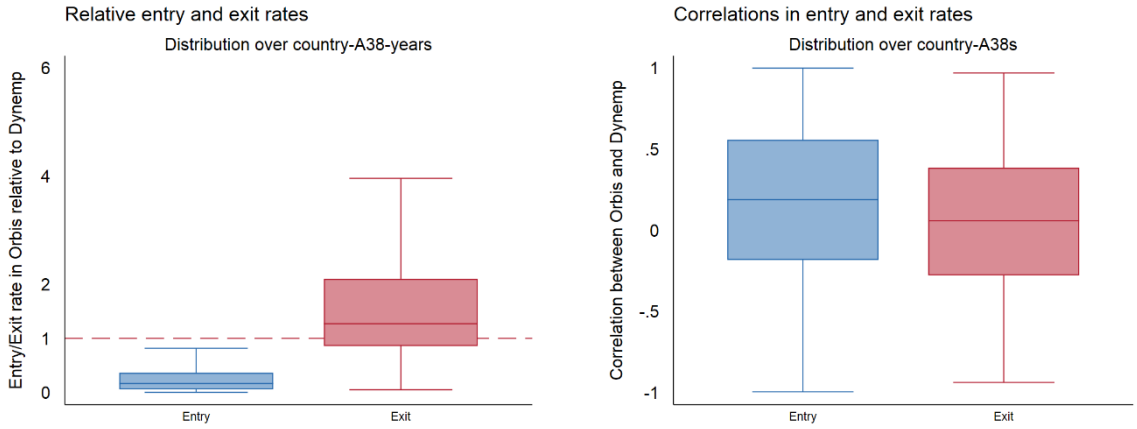


Note: Correlations across years, calculated separately for each country-industry pair. The graph plots the distribution of correlation coefficients over country-A38 pairs. Manufacturing and non-financial services (excluding "Coke and refined petroleum", "Real estate" and "Scientific R&D").
Source: Orbis and OECD MultiProd.

Entry and exit are not well-measured using Orbis. For exit, this is because it is hard to distinguish between exit from the market and merely leaving the dataset, i.e. attrition. True entry can be identified using the date of incorporation, but new firms are underrepresented in Orbis data. The first two panels of Figure 3.16 compare entry and exit rates in Orbis to those in the DynEmp dataset, which is based on business registers and, as such, reflects actual entry and exit into/from the market. The left panel compares the entry and exit rates themselves and the right panel shows correlations in terms of those rates over time. Entry rates in Orbis are only a small fraction of those observed in DynEmp. This is because entry rates considered here use information on year of birth to identify entry and young firms are underrepresented in Orbis. Entry rates calculated purely based on appearing in the dataset for the first time would likely be much higher in Orbis than in DynEmp, due to firm coverage increasing over time. The exit rates in Orbis are more similar than entry rates to their DynEmp counterparts, but they tend to be higher. This might be surprising given that Orbis firms are large and productive, so their true exit rates are likely to be lower than those in the population. The explanation is that, unlike the entry rates based on year of birth, the exit rates confound true exit with mere dropping out of the sample. Although coverage increases over time, and as a result spurious exit is less common than spurious entry, the spurious exit is still large enough to push exit rates in Orbis above those observed in business registries. As for variation over time, median correlation between Orbis and DynEmp is about 0.2 for entry rates and close to zero for exit rates. This suggests that Orbis is not suitable for studying entry and exit, at least using its full sample.

Figure 3.16. Orbis does poorly measuring entry and exit

Entry and exit in Orbis relative to DynEmp and MultiProd (2002-2015)



Note: Left panel: ratios of entry and exit in Orbis and MultiProd. Right panel: correlations in entry and exit rates across years, calculated separately for each country-A38 pair. The graph shows the distribution of the statistics over country-A38-year (left panel) or country-A38 (right panel) combinations. Manufacturing and non-financial services (excluding "Coke and refined petroleum", "Real estate" and "Scientific R&D"). Countries: AUT, BEL, DNK, FIN, FRA, DEU, HUN, ITA, JPN, NLD, NOR, PRT, SWE.
Source: Orbis, OECD MultiProd and OECD DynEmp.

Chapter 4. Choices on data construction

The analysis in Section Chapter 3. documented that firms observed in Orbis typically represent only a small share of all firms; they are systematically different from the overall firm population; and Orbis has only a limited ability to capture firm distribution characteristics and dynamics such as productivity dispersion, trends over time and the amount of entry and exit. It also suggested that there are large differences in the coverage and performance of Orbis across countries, firm sizes and other dimensions.

This section explores in more detail to what extent various choices in the preparation of the data and selection of the sample can improve representativeness using Orbis data. In particular, the following data choices are presented and discussed: (i) imputation of value added; (ii) restricting the country sample; (iii) only including selected industries; (iv) only including firms above a certain size threshold; and (v) weighting.

4.1. Imputation

As documented above, value added is missing in Orbis for many firms, even for firms where other key variables are available. For this reason, some researchers choose to impute value added in Orbis. Gal (2013) describes two types of imputation: internal, which uses only firm-specific variables available in Orbis, and external, based also on external aggregate data.

The internal imputation calculates value added as a sum of profits and remuneration of employees. Profits are measured by the earnings before interest, taxes, depreciation and amortisation (EBITDA) variable and employee remuneration as the total costs of employees. Importantly, to make the variable definition consistent across firms, the imputed value added is used for all firms, including those for which value added is directly observed in Orbis.

A drawback of the internal imputation is that the wage information is also missing for some firms. The external imputation aims to address this problem. It obtains employee costs by multiplying firm-specific employment by average wage for a given country, year and 2-digit industry, as reported in the OECD STAN data. External imputation is also applied to all observations.

Imputation substantially increases coverage in value added for about half of the countries (Figure 4.1). Internal imputation substantially increases coverage for about half the countries in the sample, and it reduces it for Norway.[21] External imputation has little effect on coverage beyond internal imputation. In proportional terms, it significantly raises coverage for Japan, but the coverage (in terms of value added) for Japan remains low in absolute terms even with the imputation.

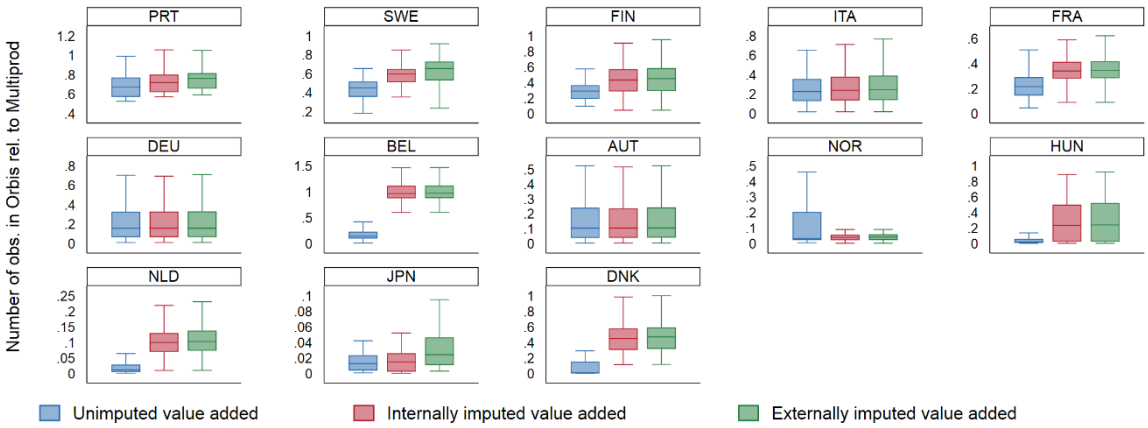
The internal imputation makes the firm distribution in Orbis more representative in terms of mean firm characteristics, and it moves the level of heterogeneity in the bottom half of the productivity distribution closer to that observed in the population (Figure 4.2). Using internal imputation moves average firm size and labour productivity closer to that observed in official microdata. It also increases the productivity dispersion in the lower half of the distribution, reducing the differences between the dispersion observed in Orbis and that observed in MultiProd by about a half.

The external imputation has similar effects on mean size and productivity as internal imputation, but it leads to a much more pronounced reduction in productivity dispersion.[22] This is not surprising given that the wage component of the externally imputed value added is based on aggregate data and is the same for all firms in each country, industry and year, so that all variation in value added per worker is driven by profits.

Both the internal and external imputation methods slightly improve correlations of Orbis with MultiProd over time for average labour productivity, but otherwise they have a limited effect on the correlations between Orbis and MultiProd (Figure 4.3).

Figure 4.1. Internal imputation of value added substantially increases coverage for several countries

Firm coverage by country, distribution over industry-years (2002-2015)

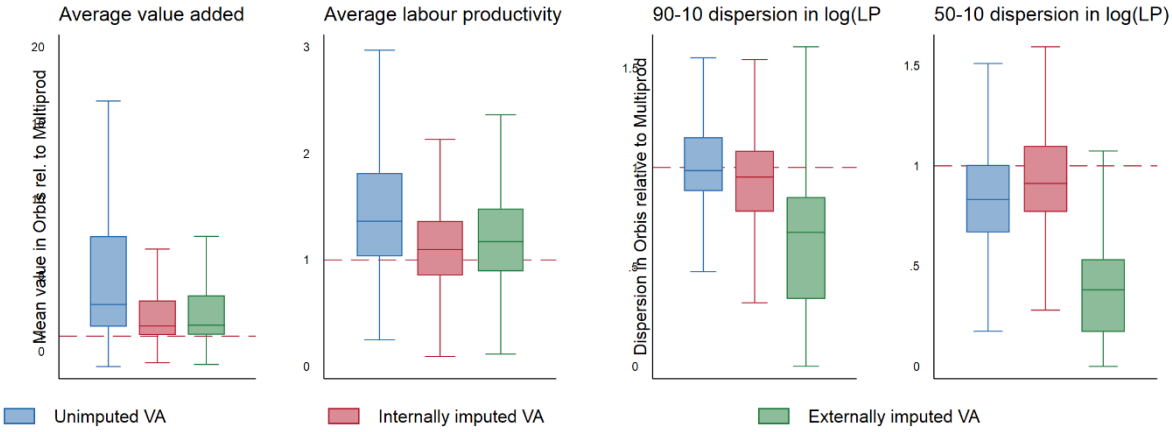


Note: The graph shows the number of observations in Orbis with employment and given variable available relative to the number of observations in MultiProd with employment available. For each country and variable, it shows the distribution of the coverage over A38-year combinations. Manufacturing and non-financial services (excluding "Coke and refined petroleum", "Real estate" and "Scientific R&D"). Internal imputation calculates value added as a sum of profits and remuneration of employees. External imputation additionally obtains employee costs by multiplying firm-specific employment by average wage for a given country, year and 2-digit industry, as reported in the OECD STAN data. In each case, the same method for calculating value added is applied to all observations.
Source: Orbis and OECD MultiProd.

Overall, using internal imputation seems to be a sensible choice, as it significantly improves coverage for many countries and brings average firm characteristics and differences across firms closer to those observed in representative data. On the contrary, external imputation dramatically reduces variation without offering substantial additional benefits. This suggests that external imputation should be used only for analyses specifically focusing on countries where neither unimputed nor internally imputed value added is available.

Figure 4.2. Imputation of value added makes Orbis firm distribution more similar to the population but external imputation reduces productivity dispersion

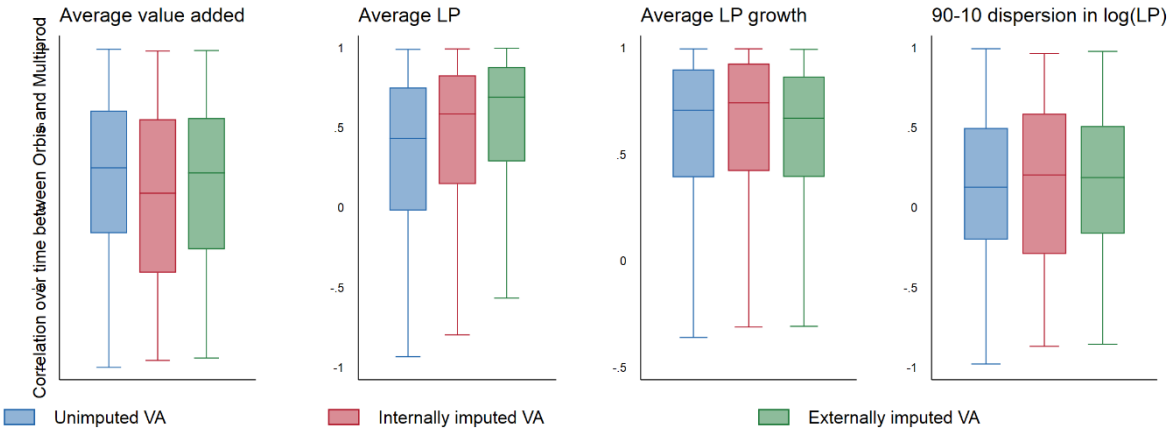
Industry averages and productivity dispersion in Orbis relative to MultiProd by value added imputation method, distribution over country-industry-years (2002-2015)



Note: The graph describes ratios of average employment, average labour productivity and labour productivity dispersion between 90th and 50th percentile and between 50th and 10th percentile of firm productivity distribution in Orbis and in MultiProd. It shows a distribution of the ratios over country-A38-year combinations. Internal imputation calculates value added as a sum of profits and remuneration of employees. External imputation additionally obtains employee costs by multiplying firm-specific employment by average wage for a given country, year and 2-digit industry, as reported in the OECD STAN data. In each case, the same method for calculating value added is applied to all observations. Manufacturing and non-financial services (excluding "Coke and refined petroleum", "Real estate" and "Scientific R&D"). Countries: AUT, BEL, DNK, FIN, FRA, DEU, HUN, ITA, JPN, NLD, NOR, PRT, SWE.
Source: Orbis and OECD MultiProd.

Figure 4.3. Imputation of value added increases correlations for average labour productivity but not for average value added or productivity dispersion

Correlations over time between Orbis and MultiProd by variable and type of variation (2002-2015)



Note: Correlations across years, calculated separately for each country-A38 pair. The graph plots the distribution of correlation coefficients over country-A38 pairs. Internal imputation calculates value added as a sum of profits and remuneration of employees. External imputation additionally obtains employee costs by multiplying firm-specific employment by average wage for a given country, year and 2-digit industry, as reported in the OECD STAN data. In each case, the same method for calculating value added is applied to all observations. Manufacturing and non-financial services (excluding "Coke and refined petroleum", "Real estate" and "Scientific R&D"). Countries: AUT, BEL, DNK, FIN, FRA, DEU, HUN, ITA, JPN, NLD, NOR, PRT, SWE.
Source: Orbis and OECD MultiProd.

4.2. Coverage thresholds

Results presented so far were based on an Orbis sample consisting of countries that have previously appeared in studies on productivity based on Orbis data. It is, however, possible to construct a more restrictive Orbis sample, and some studies have done so (e.g. Gopinath et al., 2017; Fons-Rosen et al., 2017). This section investigates whether restricting the sample to the best-covered data cells improves the representativeness of Orbis.

To this end, it defines two more restricted samples. It defines them at the level of country-years, rather than another level such as country-industry-years. This is for two reasons. First, the most dramatic differences in Orbis coverage are those between countries and between years, rather than those between industries. Second, defining the sample in terms of country-years seems to be what studies using Orbis typically do.

The first restricted sample, "5000+", includes only country-years with at least 5000 observations for which value added is available. The second restricted sample, the "hand-picked" sample, is manually selected and represents, broadly speaking, continuous country-periods with coverage that is north of 20% and reasonably stable over time. The country-years included in each sample are summarised in Table 4.1.[23] The "hand-picked" sample is a strict subset of the "5000+" sample, differing in three ways: (i) it excludes early years for Finland, Germany and Sweden when coverage was high but growing; (ii) it excludes Denmark, Japan and Norway which have reasonable coverage for only a few early years; and (iii) it excludes Hungary, where coverage fluctuates and the better-covered years are not consecutive. Note that the table lists country-years which are also covered by the

MultiProd data. Some other countries in Orbis, such as Spain, could also fit in the restricted samples but are not shown here.

Table 4.1. Restricted Orbis samples

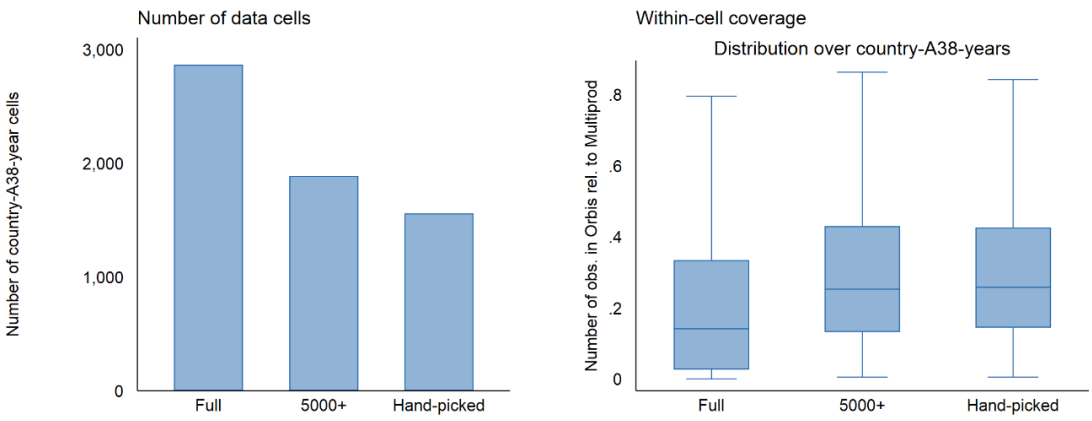
	5000+ sample	Hand-picked sample
Belgium	2002-2014	2002-2014
Denmark	2002-2003	
Finland	2002-2013	2004-2013
France	2002-2015	2002-2015
Germany	2005-2013	2006-2013
Hungary	2007, 2009, 2010, 2012	
Italy	2002-2015	2002-2015
Japan	2002-2004	
Norway	2002-2004	
Portugal	2006-2012	2006-2012
Sweden	2002-2012	2004-2012

Note: Only country-years also covered by MultiProd are listed.

Applying the 5000 value-added observations threshold reduces the number of country-industry-year data cells by about a third (left panel of Figure 4.4). Using the "hand-picked" sample further reduces it by about a fifth.

Limiting the sample to better-covered countries and years increases firm coverage. While firm coverage for a median country-industry-year in the full sample is 14%, in the "5000+" and the "hand-picked" samples it is 25% and 26%, respectively (right panel of Figure 4.4). The effect is particularly marked at the lower end of the coverage distribution: the 25th percentile of coverage increases from less than 3% in the full sample to 13% and 14% in the restricted samples.

Figure 4.4. Country-years with 5000+ firms are similar to the hand-picked sample



Left panel: number of country-A38-year data cells in the sample. Right panel: number of observations in Orbis with employment and value added available relative to the number of observations in MultiProd with employment available (distribution over country-A38-years). The "5000+" sample includes country-years where at least 5000 firms have non-missing value added. See Table 4.1**Error! Reference source not found.** for country-years included in each sample. Manufacturing and non-financial services (excluding "Coke and refined petroleum", "Real estate" and "Scientific R&D").
Source: Orbis and OECD MultiProd.

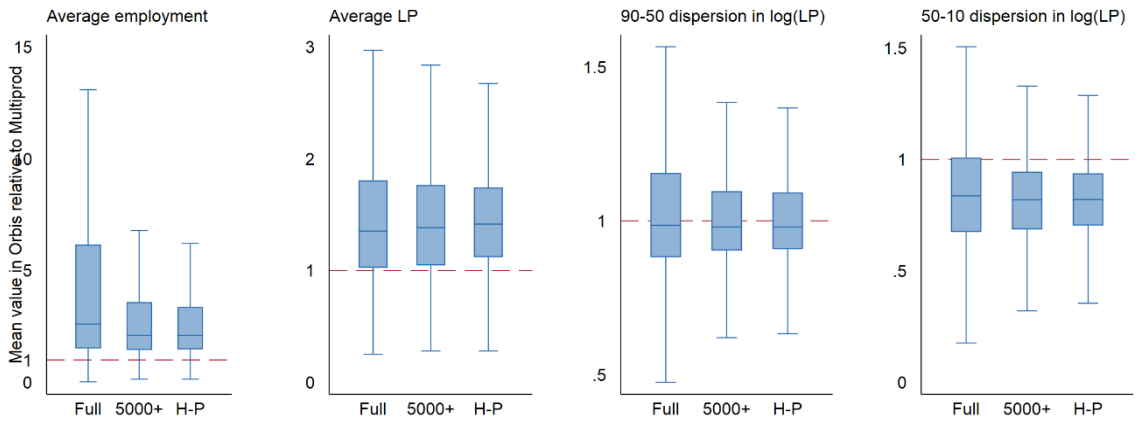
Restricting the sample also makes average characteristics of firms more similar to those in official microdata (Figure 4.5). It reduces the ratios of average firm sizes in Orbis to those in MultiProd, in particular reducing the number of cells with a very high difference in average firm size between Orbis and MultiProd. It also reduces mean labour productivity, but only slightly and mainly through reducing the number of cell outliers at the top. It has little effect on productivity dispersion for a median cell but it reduces the number of data cells with a very large difference in productivity dispersion between Orbis and MultiProd.

Most importantly, restricting the sample significantly increases correlations with MultiProd over time in terms of average firm size, average labour productivity and average labour productivity growth (Figure 4.6). It also increases the correlations over time for labour productivity dispersion, but these remain quite low even in the restricted samples.

Overall, the results indicate that reducing the sample to better-covered countries and years may be desirable, as it leads to substantially more representative results. The performance of the "5000+" sample is actually quite similar to that of the "hand-picked" sample. In some cases, the manual selection might still be preferable, as blindly applying a simple rule such as the 5000 threshold may keep in countries for which the included period is very short or has gaps (see Table 4.1).

Figure 4.5. Orbis sample in better covered country-years is less skewed toward large firms but keeps underestimating productivity dispersion

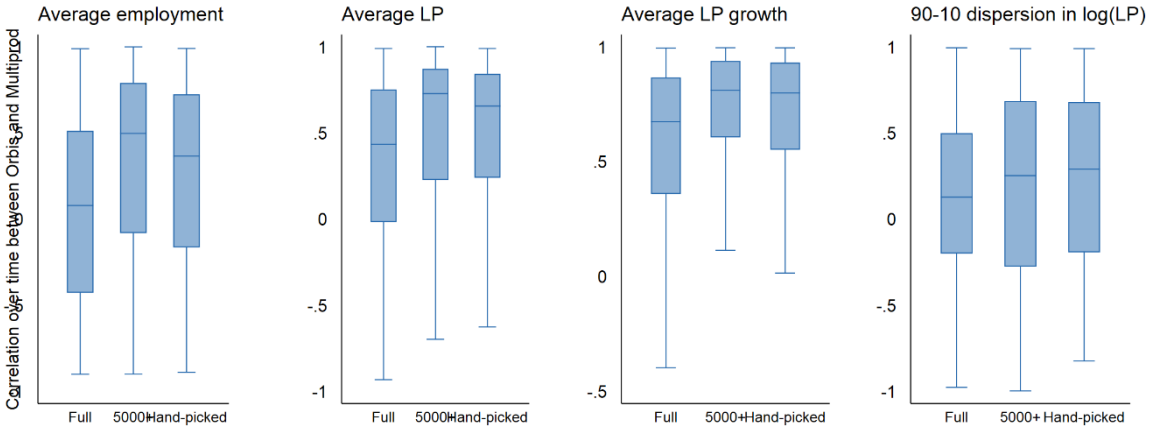
Industry averages and productivity dispersion in Orbis relative to MultiProd by sample, distribution over country-industry-years (2002-2015)



Note: The graph describes ratios of average employment, average labour productivity and labour productivity dispersion between 90th and 50th percentile and between 50th and 10th percentile of firm productivity distribution in Orbis and in MultiProd. It shows a distribution of the ratios over country-A38-year combinations. The "5000+" sample includes country-years where at least 5000 firms have non-missing value added. "H-P" marks "hand-picked" sample. See Table 4.1 for country-years included in each sample. Manufacturing and non-financial services (excluding "Coke and refined petroleum", "Real estate" and "Scientific R&D").
Source: Orbis and OECD MultiProd.

Figure 4.6. Better-covered country-years show somewhat tighter correlation with MultiProd

Correlations over time between Orbis and MultiProd by variable and type of variation (2002-2015)



Note: Correlations across years, calculated separately for each country-A38 pair. The graph plots the distribution of correlation coefficients over country-A38 pairs. The "5000+" sample includes country-years where at least 5000 firms have non-missing value added. "H-P" marks "hand-picked" sample. See Table 4.1 for country-years included in each sample. Manufacturing and non-financial services (excluding "Coke and refined petroleum", "Real estate" and "Scientific R&D").
Source: Orbis and OECD MultiProd.

4.3. Industries

If some industries are systematically better represented than others, focusing on these industries would be another way of improving the representativeness of Orbis. This subsection examines representativeness of individual industries.

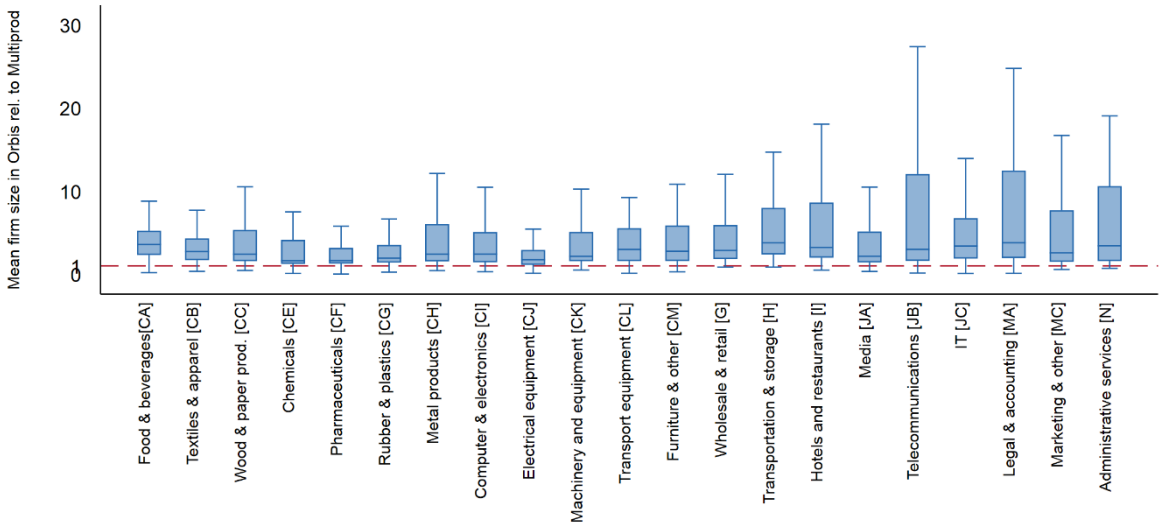
Firm characteristics in Orbis tend to be more representative in manufacturing than in services. Figure 4.7 compares the average firm size in Orbis and MultiProd for each A38 industry. The relative firm size is closely related to the relative coverage in each industry as shown earlier in Figure 3.7, which is also higher for manufacturing. For example, the average firm size in pharmaceuticals, an industry generally dominated by few large players, is comparatively similar in Orbis and in MultiProd for most country-year cells. On the contrary, a large number of cells in the two datasets display a very different average firm size in hotels and restaurants as well as legal and accounting.

The correlations in average labour productivity in Orbis and MultiProd over time are generally greater for manufacturing than for service sectors (Figure 4.8). Within manufacturing, industries are not very different in terms of the correlation for the median cell but they differ significantly in terms of correlations at the 25th percentile.

Overall, experimentation with other measures of representativeness suggests that it is not possible to single out industries that appear consistently less representative across many measures. Focusing on manufacturing will lead to a somewhat greater representativeness but may be too restrictive for many types of analysis.

Figure 4.7. The sample skewness towards large firms tends to be weaker in manufacturing

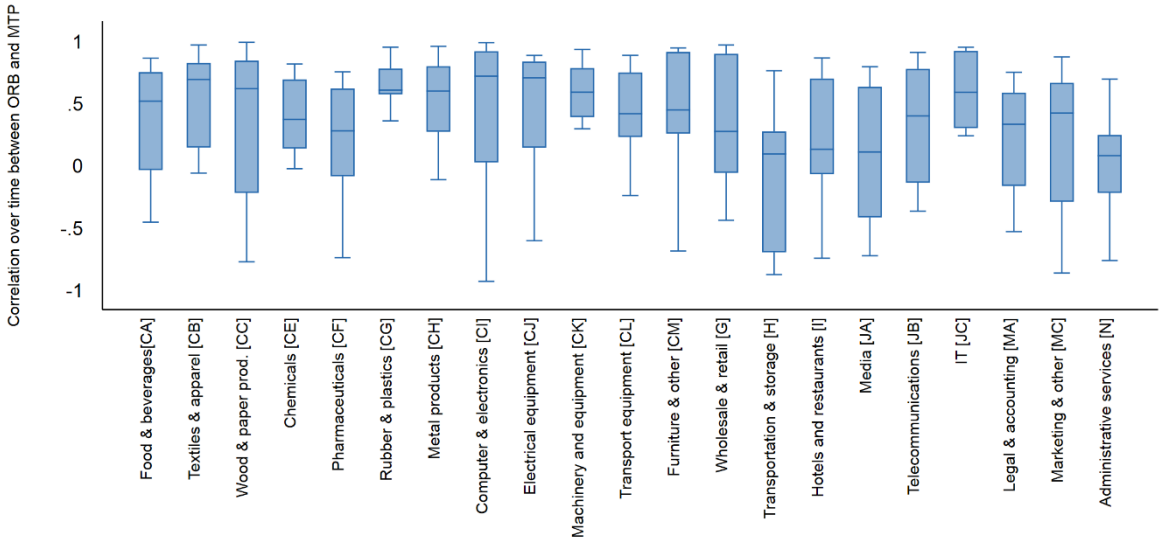
Industry averages employment in Orbis relative to MultiProd by industry, distribution over country-years (2002-2015)



Note: The graph describes the ratio of average employment in Orbis and in MultiProd. For each A38 industry, it shows the distribution of the ratio over country-year pairs. Countries: AUT, BEL, DNK, FIN, FRA, DEU, HUN, ITA, JPN, NLD, NOR, PRT, SWE.
Source: Orbis and OECD MultiProd.

Figure 4.8. Correlations in labour productivity over time tend to be higher in manufacturing

Correlations in labour productivity over time between Orbis and MultiProd by A38 industry (distribution over countries, 2002-2015)



Note: Correlations in labour productivity across years, calculated separately for each country-industry pair. For each A38 industry, it shows the distribution of the correlation over countries. Countries: AUT, BEL, DNK, FIN, FRA, DEU, HUN, ITA, JPN, NLD, NOR, PRT, SWE.
Source: Orbis and OECD MultiProd.

4.4. Size thresholds

As Orbis coverage is particularly poor for the smallest firms, researchers sometimes restrict the sample to firms above a certain size threshold. This will make the sample less representative of the overall firm population, but better defined and more stable over time.[24]

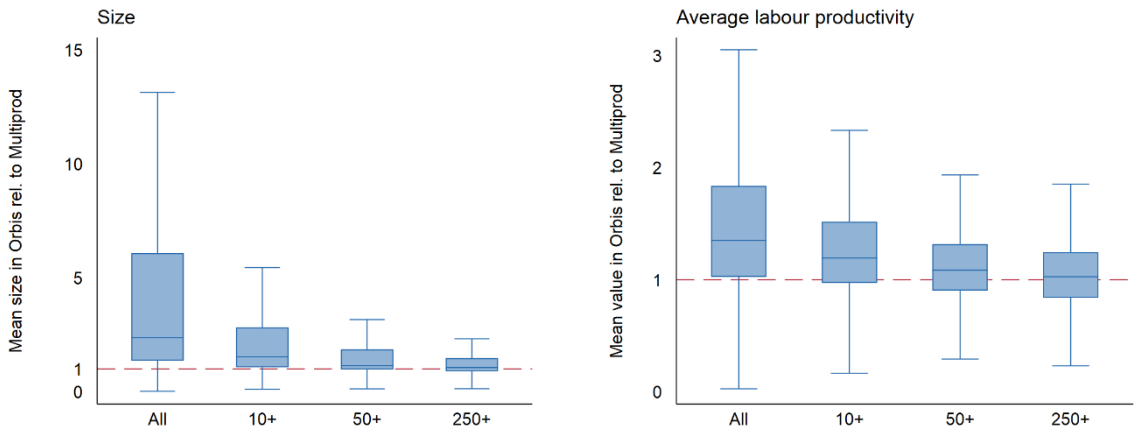
Applying size thresholds make average firm size and average labour productivity in Orbis more similar to those in MultiProd (Figure 4.9). The thresholds imply higher firm coverage, as Orbis has a better coverage for larger firms (see Figure 3.6). The higher coverage, in turn, means that the average firm size in Orbis and MultiProd become much more similar. Average labour productivity also becomes more similar with increasing size thresholds, as introducing thresholds both reduces productivity differences due to the different size composition of Orbis and official microdata and also removes the smallest firms, which show largest within-size-class productivity differences between Orbis and MultiProd (see Figure 3.10).

On the other hand, the size thresholds do not increase correlations between Orbis and MultiProd over time (Figure 4.10).

Overall, as micro-firm with fewer than 10 employees are particularly poorly covered in Orbis, excluding this group is a sensible choice. An analysis performed on data that are representative of the population of firm with at least 10 employees may be preferable to trying to make claims about the entire firm population using Orbis.

Figure 4.9. Orbis firms are substantially more representative of the population of firms with at least 10 employees than of the full firm population

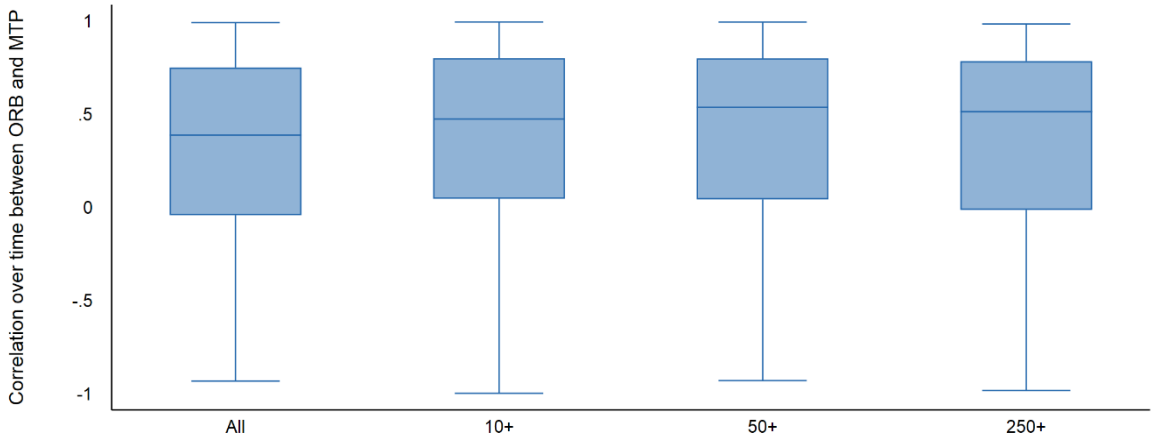
Industry averages firm size and labour productivity in Orbis relative to MultiProd by firm size threshold, distribution over country-industry-years (2002-2015)



Note: The graph describes ratios of average employment and average labour productivity in Orbis and in MultiProd. For each firm size threshold, it shows a distribution of the ratios over country-A38-year combinations. Manufacturing and non-financial services (excluding "Coke and refined petroleum", "Real estate" and "Scientific R&D"). Countries: AUT, BEL, DNK, FIN, FRA, DEU, HUN, ITA, JPN, NLD, NOR, PRT, SWE.
Source: Orbis and OECD MultiProd.

Figure 4.10. Applying size thresholds does not increase the correlation between Orbis and MultiProd over time

Correlations between Orbis and MultiProd over time, by variable (2002-2015)



Note: Correlations across years, calculated separately for each country-industry pair. For each firm size threshold, the graph plots the distribution of correlation coefficients over country-A38 pairs. Manufacturing and non-financial services (excluding "Coke and refined petroleum", "Real estate" and "Scientific R&D"). Countries: AUT, BEL, DNK, FIN, FRA, DEU, HUN, ITA, JPN, NLD, NOR, PRT, SWE.
Source: Orbis and OECD MultiProd.

4.5. Weighting

Weighting is a standard way to ensure the aggregate representativeness of a sample of firms. For example, inverse probability weighting is used for stratified random samples. This subsection examines whether reweighting can improve representativeness in the case of Orbis. Statistics produced by the MultiProd code applied to the Orbis data are re-calculated to match the population number of firms in each cell defined by country, year, industry and firm size category.

The analysis here explores two types of weights. Reweighting with "internal weights" assumes that firms in Orbis represent the population of firms and it only corrects for observations which have some variables (most often value added) missing. Reweighting with "external weights" uses information on the population of firms put together within the MultiProd project based on business registers or other administrative. This data, at least in principle, covers all firms.[25]

Surprisingly, weighting does *not* improve representativeness of Orbis data beyond the mechanic effect on the firm size distribution. The effect of reweighting is captured in Figure 4.11 and Figure 4.12. Reweighting with internal weights has mostly negligible effects, as the coverage issues of Orbis are, for the most part, due to entire observations being missing rather than values of particular variables. Reweighting with external weights, by construction, makes average firm size similar to that in MultiProd. However, as for average productivity and productivity dispersion, it does not make them more similar to MultiProd for a median country-industry-year cell and it increases their variance across the data cells (Figure 4.11). Along similar lines, reweighting with external weights leads to a substantially higher correlation with MultiProd over time for average firm size but has little systematic effect on corresponding correlations for average labour productivity, average labour productivity growth and labour productivity dispersion (Figure 4.12).

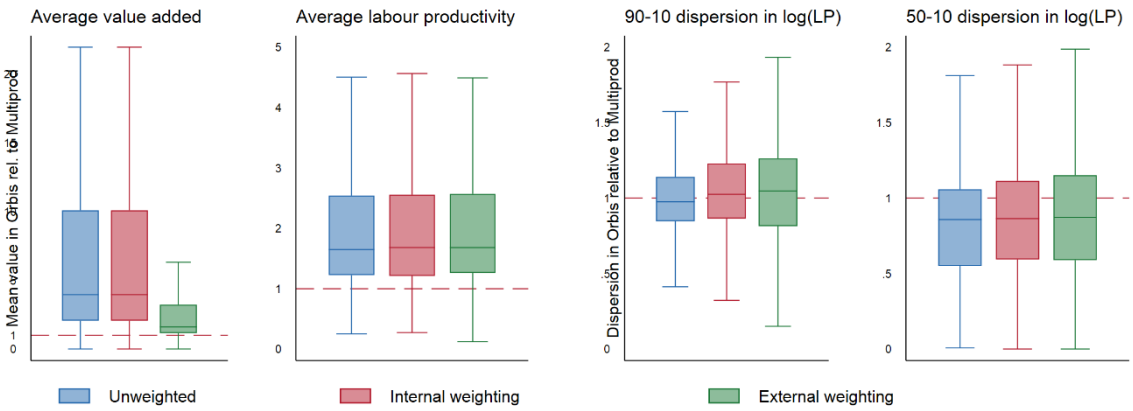
To understand the underwhelming effects of reweighting, it is good to consider that reweighting corrects for the fact that some cells may be better covered than others within Orbis, but, crucially, assumes that data *within* each cell are representative (i.e. firms within that cell should be randomly selected). That is to say, for the reweighting to solve the representativeness issue, small firms of a particular country-industry-year are allowed to be poorly covered in Orbis, but the average small firm in Orbis has to be similar to the average small firm in the true population (for that country-industry-year).

As documented above in Figure 3.10, the assumption is not satisfied for Orbis, where firms are disproportionately more productive even within each size class. The higher productivity of firms in Orbis is largely due to their greater productivity given their firm size rather than due to the size composition of the sample. Reweighting does not solve this source of unrepresentativeness and it could even exacerbate it by putting more weight on small firms. Indeed, selection issues are particularly severe for small firms, which are especially underrepresented in Orbis and may only appear in the data if they are particularly productive relative to peers in the same size class.

Overall, reweighting based on firm size does not solve the problem at hand.

Figure 4.11. Applying external weights makes average firm size in Orbis more similar to the population but not average productivity or productivity dispersion

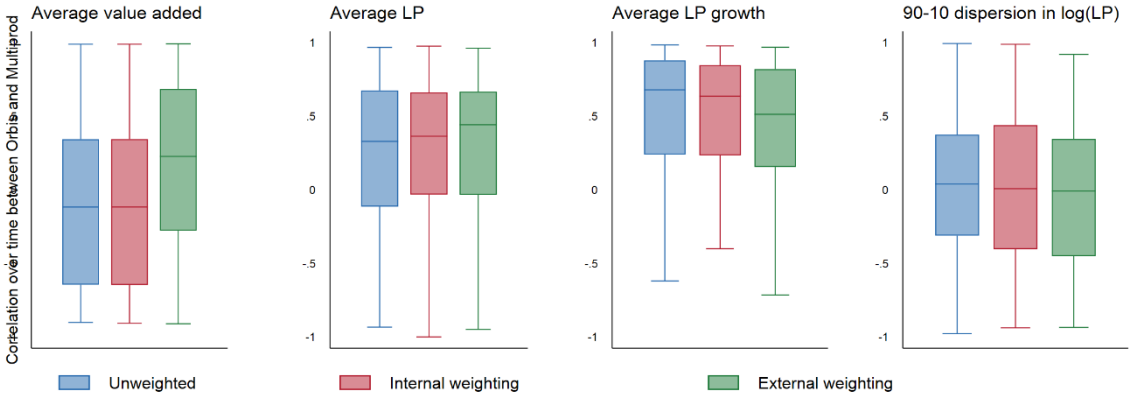
Industry averages and productivity dispersion in Orbis relative to MultiProd by weighting method, distribution over country-industry-years (2002-2015)



Note: The graph describes ratios of average employment, average labour productivity and labour productivity dispersion between 90th and 50th percentile and between 50th and 10th percentile of firm productivity distribution in Orbis and in MultiProd. It shows a distribution of the ratios over country-A38-year combinations. Weighting is variable-specific. Internal weighting takes firms with non-missing employment in Orbis as the population and re-weights for firms with a given variable missing. External weighting reweights based on business registry information as used in the MultiProd project. Manufacturing and non-financial services (excluding "Coke and refined petroleum", "Real estate" and "Scientific R&D"). Countries: AUT, BEL, DNK, FIN, FRA, DEU, HUN, ITA, JPN, NLD, NOR, PRT, SWE.
Source: Orbis and OECD MultiProd.

Figure 4.12. External weighting increases correlations with MultiProd in average firm size but not in productivity levels, growth or dispersion

Correlations over time between Orbis and MultiProd by variable and type of variation (2002-2015)



Note: Correlations across years, calculated separately for each country-A38 pair. The graph plots the distribution of correlation coefficients over country-A38 pairs. Weighting is variable-specific. Internal weighting takes firms with non-missing employment in Orbis as the population and re-weights for firms with a given variable missing. External weighting reweights based on business registry information as used in the MultiProd project. Manufacturing and non-financial services (excluding "Coke and refined petroleum", "Real estate" and "Scientific R&D"). Countries: AUT, BEL, DNK, FIN, FRA, DEU, HUN, ITA, JPN, NLD, NOR, PRT, SWE.
Source: Orbis and OECD MultiProd.

Chapter 5. Representativeness in the "preferred sample"

The previous section took the baseline sample as a starting point and documented the impact of various data construction choices, examining one choice at a time. It identified three steps that significantly improve representativeness of Orbis: internally imputing value added, restricting the sample to better-covered country-years and focusing on firms above a certain size threshold.

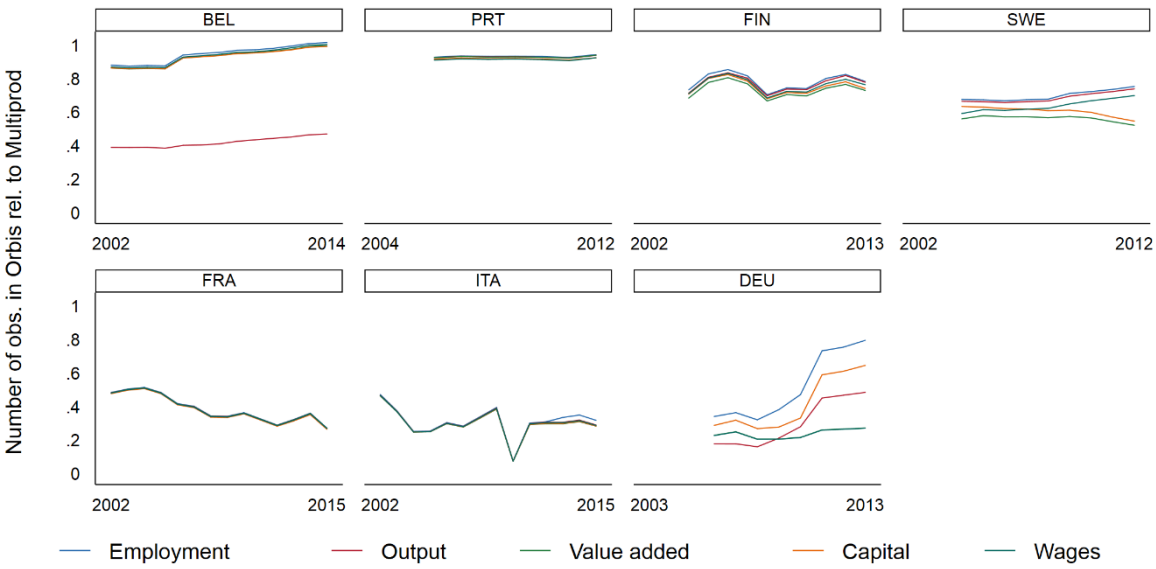
This section examines performance of Orbis when these three steps are implemented at the same time. This represents the "preferred sample" that a researcher might want to use in an applied analysis. In particular, it relies on the internal imputation of value added, on the "hand-picked" country sample and on firms with at least 10 employees. The sample consists of seven countries: Belgium, Finland, France, Germany, Italy, Portugal and Sweden.

Note that we write "preferred sample" with quotation marks to acknowledge the fact that the sample that is actually to be preferred inevitably depends on the question at hand. For example, very different samples will be appropriate for studying a distribution of sales in a cross-section and for studying evolution of MFP over time.

There is a substantial variation in coverage even within this reduced sample (Figure 5.1). Across countries, the coverage ranges from over 70% for Belgium (all variables except output), Portugal and Finland to about 40% for France, Italy and Germany. Over time, the coverage is relatively stable for most countries, but there is some downward trend for France, an upward trend for Germany and a dip for Italy in 2010. The coverage is also similar across all variables for most countries, with the exception of some variation across variables for Germany and the case of output in Belgium, which is available for just half as many observations as the other variables.

Figure 5.1. The "preferred sample" keeps country-periods with relatively high and stable coverage

Firm coverage by country over time (2002-2015)



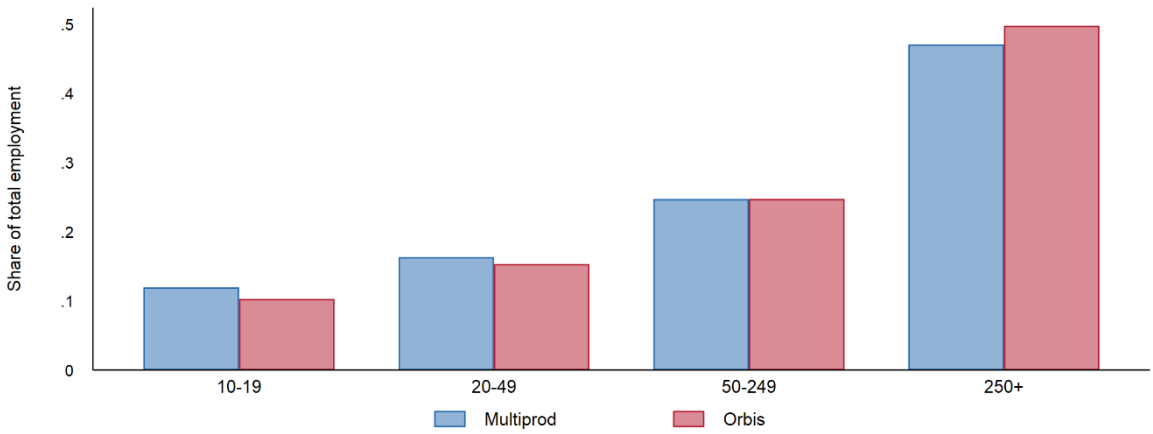
Note: The graph shows the number of observations in Orbis with employment and given variable available relative to the number of observations in MultiProd with employment available. Manufacturing and non-financial services (excluding "Coke and refined petroleum", "Real estate" and "Scientific R&D"). Firms with at least 10 employees. Value added is calculated as a sum of profits and remuneration of employees.
Source: Orbis and OECD MultiProd.

The sample is still slightly skewed towards larger, more productive and older firms, but much less so than in the baseline specification (Figure 5.2 and Figure 5.3). Firms with more than 250 employees account for 50% of the total employment captured by Orbis, as compared to 47% in MultiProd. For a median country-industry-year, firms in Orbis have, on average, 21% greater employment, 15% higher age, 5% greater labour productivity, 13% greater multi-factor productivity and 5% higher wages than firms in MultiProd.

Importantly, Orbis in the "preferred sample" displays a similar dispersion of productivity as MultiProd (at least for the median country-year-industry combination), both in terms of LP and MFP and both in the upper and lower half of the productivity distribution (Figure 5.4).

Figure 5.2. The "preferred sample" is only slightly skewed toward large firms

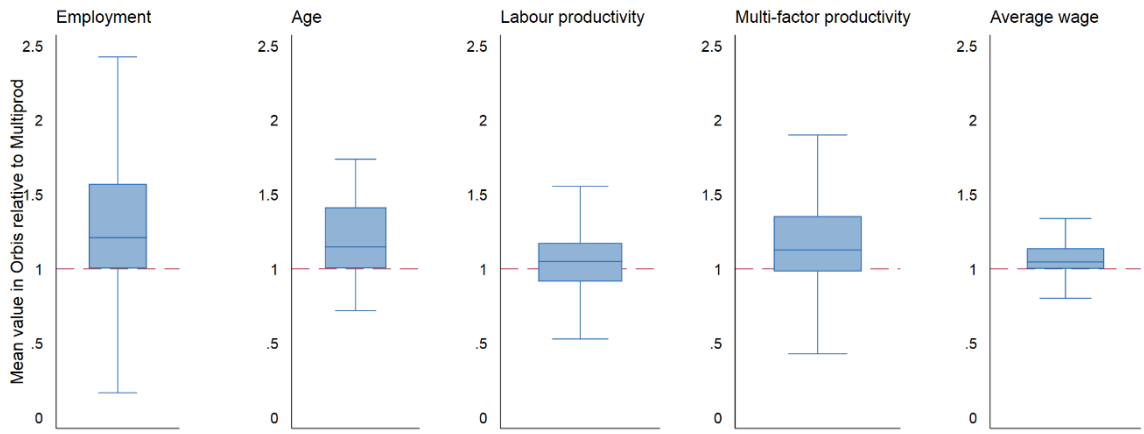
Distribution of employment over firm size categories (mean over country-years, 2002-2015)



Note: Each bar represents the share of total employment observed in a given dataset accounted for by firms in a given firm size category. Firm size categories are defined by employment. Manufacturing and non-financial services (excluding "Coke and refined petroleum", "Real estate" and "Scientific R&D"). Countries: BEL, FIN, FRA, ITA, PRT, SWE. Firms with at least 10 employees. Value added is calculated as a sum of profits and remuneration of employees.
Source: Orbis and OECD MultiProd.

Figure 5.3. Orbis firms in the "preferred sample" are on average more similar to the population

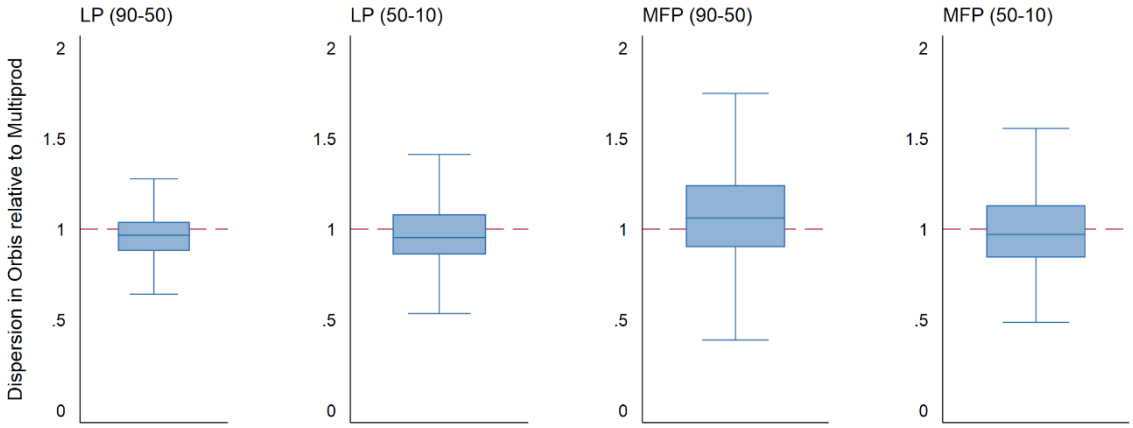
Industry averages in Orbis relative to MultiProd, distribution over country-industry-years (2002-2015)



Note: The graph describes ratios of average employment, age, labour productivity, multi-factor productivity and average wage in Orbis and in MultiProd for each industry. It shows a distribution of the ratios over country-A38-year combinations. Manufacturing and non-financial services (excluding "Coke and refined petroleum", "Real estate" and "Scientific R&D"). Countries: BEL, DEU, FIN, FRA, ITA, PRT, SWE. Firms with at least 10 employees. Value added is calculated as a sum of profits and remuneration of employees.
Source: Orbis and OECD MultiProd.

Figure 5.4. Productivity dispersion of Orbis firms in the "preferred sample" is also similar to the population

Labour productivity dispersion in Orbis relative to MultiProd, distribution over country-industry-years (2002-2015)



Note: The graph describes ratios of labour productivity and multi-factor productivity dispersion between 90th and 50th percentile and between 50th and 10th percentile of firm productivity distribution. It shows a distribution of the ratios over country-A38-year combinations. Manufacturing and non-financial services (excluding "Coke and refined petroleum", "Real estate" and "Scientific R&D"). Countries: BEL, DEU, FIN, FRA, ITA, PRT, SWE. Firms with at least 10 employees. Value added is calculated as a sum of profits and remuneration of employees.
Source: Orbis and OECD MultiProd.

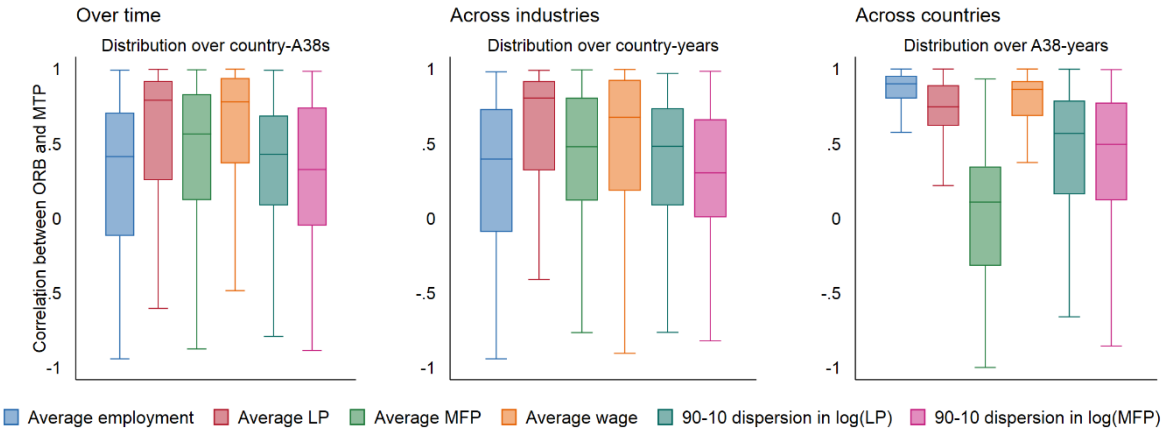
The correlations between Orbis and MultiProd are generally higher in the "preferred sample" than in the baseline specification, although they are still only moderately high for the productivity dispersion (Figure 5.5). Correlations in firm size are high across countries but only moderate across industries and over time. Correlations in terms of average labour productivity and average wages are quite high for all types of variation. Correlations in multi-factor productivity are lower than for labour productivity and very low for correlations across countries. Correlations in productivity dispersion, for all types of variation, are now predominantly positive but still only moderate, at around 0.5

Finally, Orbis continues to perform poorly when it comes to measuring entry and exit, even in the "preferred sample" (Figure 5.6). As in the baseline sample, entry rates are too low and only weakly correlated with DynEmp figures; exit rates are of the right magnitude on average but almost uncorrelated with DynEmp.

In summary, Orbis in the "preferred sample" is not quite representative of the population of firms with at least 10 employees, but it is substantially improved. Its firms are only slightly larger, more productive and older than those in official microdata, and it is able to capture variation in firm performance and heterogeneity across countries, across industries and over time, albeit imperfectly. On the other hand, even the "preferred sample" is not well suited for analysing entry and exit dynamics.

Figure 5.5. The "preferred sample" generally shows higher, but not very high, correlations between Orbis and MultiProd

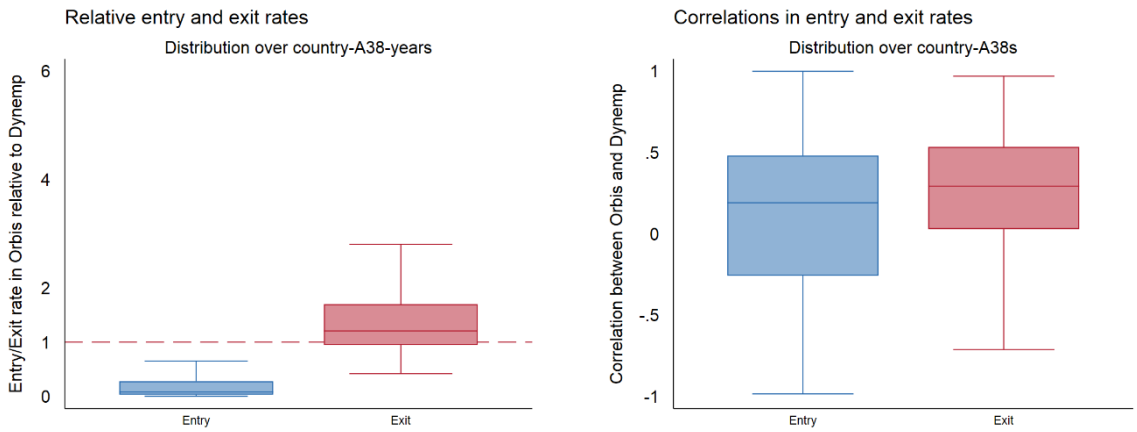
Correlations between Orbis and MultiProd by variable and type of variation (2002-2015)



Note: Left panel: correlations across years, calculated separately for each country-A38 pair. Middle panel: correlations across industries, calculated separately for each country-year pair. Right panel: correlations across countries, calculated separately for each A38-year pair. The graph plots the dispersion over country-A38s (left), country-years (middle) or A38-years (right) of these correlations. Manufacturing and non-financial services (excluding "Coke and refined petroleum", "Real estate" and "Scientific R&D"). Countries: BEL, DEU, FIN, FRA, ITA, PRT, SWE. Firms with at least 10 employees. Value added is calculated as a sum of profits and remuneration of employees.
Source: Orbis and OECD MultiProd.

Figure 5.6. Orbis does poorly at measuring entry and exit even with the "preferred sample"

Entry and exit in Orbis relative to DynEmp and MultiProd (2002-2015)



Note: Left panel: ratios of entry and exit in Orbis and MultiProd. Right panel: correlations in entry and exit rates across years, calculated separately for each country-A38 pair. The graph shows the distribution of the statistics over country-A38-year (left panel) or country-A38 (right panel) combinations. Manufacturing and non-financial services (excluding "Coke and refined petroleum", "Real estate" and "Scientific R&D"). Countries: BEL, DEU, FIN, FRA, ITA, PRT, SWE. Firms with at least 10 employees. Value added is calculated as a sum of profits and remuneration of employees.
Source: Orbis, OECD MultiProd and OECD DynEmp.

Chapter 6. Preparing Orbis data: selected issues

This section first highlights two important data issues in Orbis which have so far not received enough attention in the literature: the use of rounded values for key variables, and the choice between unconsolidated and consolidated accounts. It then discusses ownership information within Orbis.

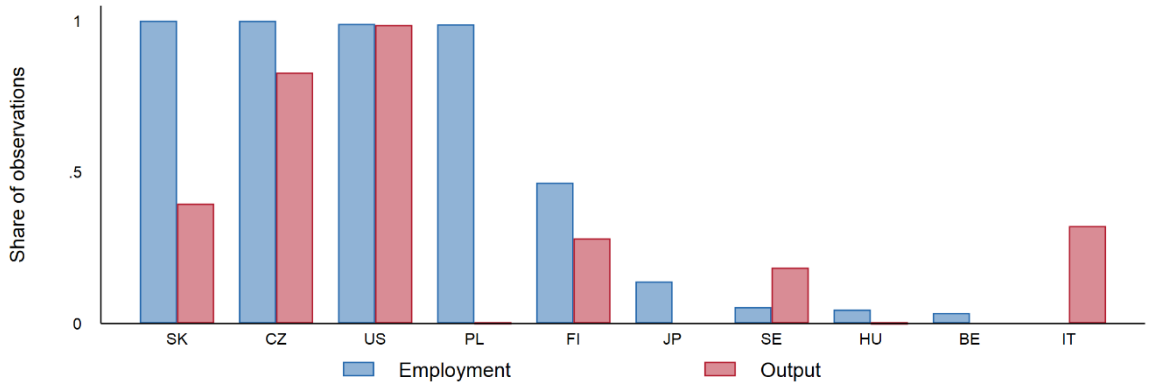
6.1. Rounded values

The Orbis user guide notes that for some countries and consolidation codes (e.g. those with limited financials) the number of employees and operating revenues are reported as the middle of a range rather than an exact value. To see how important this issue is in practice, we inspect the distribution of employment and operating revenues for each country and consolidation code to identify a mass of firms at specific values of employment or operating revenues. We consequently remove those cases as they are most likely the result of such rounding/imputation.

Rounding in Orbis is a major issue for the Czech Republic, Poland, Slovakia and the United States. Figure 6.1 summarises the share of firm-year observations in each country for which employment and output were set to missing because of rounding. It features all countries included in the Orbis-STAN comparison in Subsection 3.1, and additionally also the Czech Republic, Poland and Slovakia. The latter three countries appear in several multi-country productivity studies based on Orbis but were not included in the analysis in the previous sections exactly because correcting for the rounding of employment leaves them with very few observations. In addition to these three countries, rounding is also a major issue for the United States, where both employment and output are rounded for the vast majority of unconsolidated accounts, and, as a result, only consolidated results are left after the cleaning (see the following subsection for a discussion of unconsolidated and consolidated accounts). Rounding also plays a smaller, but non-negligible, role for Finland, Japan, Sweden and Italy.

Figure 6.1. Rounding is a major issue for several Orbis countries

Share of firm-year observations with employment/output set to missing due to rounding (all years in Orbis)



Note: Separately for employment and output, the graph shows the share of firm-year observations in Orbis for which the variable was set to missing because of rounding issues. Countries: AUT, BEL, CZE, DEU, DNK, ESP, EST, FIN, FRA, GBR, GRC, HUN, ITA, JPN, KOR, NLD, NOR, POL, PRT, SVK, SVN, SWE, USA. Countries with no rounding issues are omitted from the graph.
Source: Orbis.

6.2. Account types and consolidation

Although Orbis is often considered to be a firm-level dataset, it should rather be thought of as an account-level dataset. This is because sometimes multiple financial accounts are available for the same firm in a given year. In addition, the accounts may correspond to an individual firm (unconsolidated accounts) or to an entire business group (consolidated accounts). In order to prevent double-counting and use account types appropriately to the task at hand, it is important to drop duplicate accounts of the same firm and understand the different account types.

Accounts reported in Orbis data come in five main types:

- **U1.** Unconsolidated accounts of companies for which consolidated accounts are not available.
- **U2.** Unconsolidated accounts of companies for which consolidated accounts are also available.
- **C1.** Consolidated accounts of companies for which unconsolidated accounts are not available.
- **C2.** Consolidated accounts of companies for which unconsolidated accounts are also available.
- **LF.** Accounts with limited financial information.

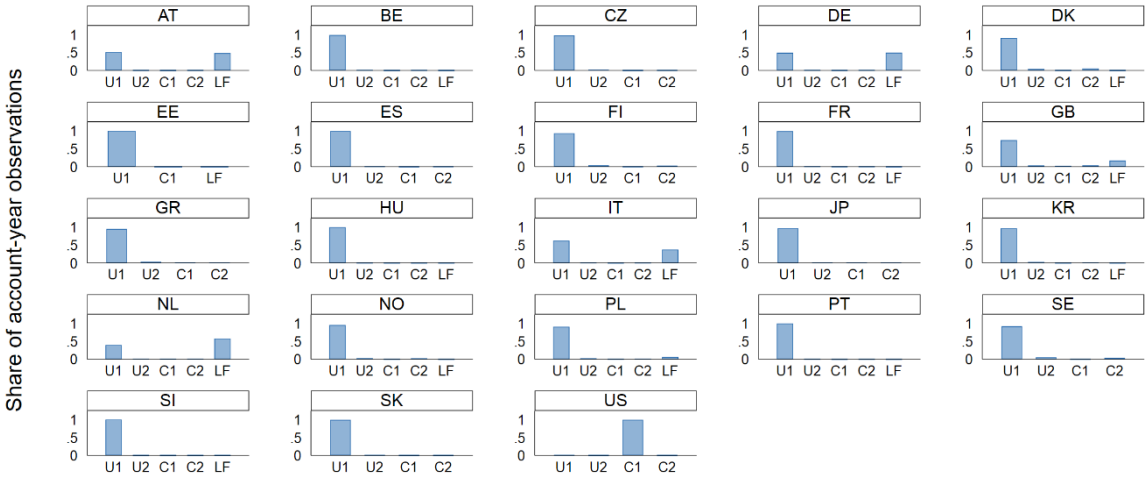
A dominant majority of accounts belong to type U1. Figure 6.2 shows the distribution of account types across accounts reported in Orbis, separately for each country. In most countries, a vast majority of accounts are of the U1 type. In the case of the United States, virtually all unconsolidated accounts are dropped due to rounding of employment and output (see the previous subsection), so most accounts shown in Figure 6.2 are consolidated.

Companies filing consolidated information are comparatively few, but they represent a large part of the economy because they tend to be the largest companies, with output that is, on average, twenty or even one-hundred times greater than that of companies with unconsolidated accounts only. Both consolidated accounts (C1 and C2) and unconsolidated accounts (U2) of companies that file consolidated accounts report on average much larger output than accounts of companies which only file unconsolidated accounts (U1).

Some countries (particularly Austria, Germany and the Netherlands) also have a large number of LF accounts. LF accounts sometimes have information on employment and output, but they almost never have information on value added, capital or wages. They are most likely unconsolidated, although the Orbis user guide does not explicitly say so.

Figure 6.2. A vast majority of accounts in Orbis are unconsolidated

Distribution of observations over consolidation codes, by country (2002-2015)



Note: The graph shows the share of account-year observations with each account type, by country. U1 = unconsolidated accounts of companies for which consolidated accounts are not available. U2 = unconsolidated accounts of companies for which consolidated accounts are also available. C1 = consolidated accounts of companies for which unconsolidated accounts are not available. = C2 = consolidated accounts of companies for which unconsolidated accounts are also available. LF = accounts with limited financial information.
Source: Orbis.

Multiple accounts of different types are available for some firms. By far the most common combination of account types existing for the same firm is U2+C2, which can be expected given the definitions of these consolidation codes (Figure 6.3). There is also a small fraction of firms with less expected combinations, U2+C1+C2, C1+C2, U2+C1 and U1+C1, and a few firms with other combinations of consolidation codes.

To avoid duplicating accounts for the same firm, the analysis in this paper relies primarily on unconsolidated accounts. It drops consolidated accounts whenever unconsolidated accounts for the same company are indicated to exist in Orbis (consolidation code C2) and consolidated accounts of companies whenever unconsolidated accounts are actually observed in the data for the same business ID (this happens for some of the accounts with consolidation code C1). If duplicate accounts for the same ID still exist and one of them has the consolidation code U2, the other accounts are dropped. Finally, if duplicate account

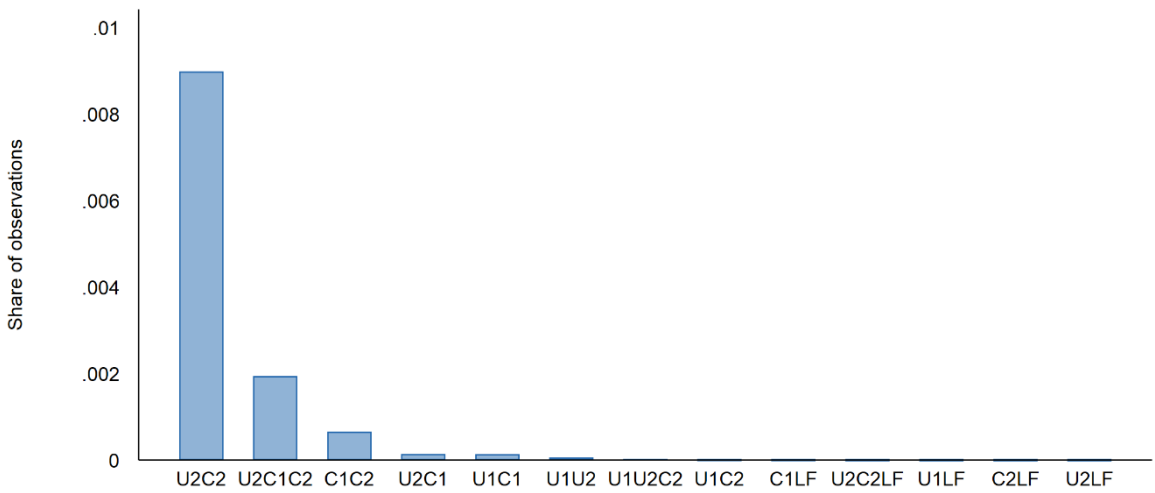
types still remain, the account types with limited financial information (consolidation code LF) are dropped.

An alternative approach would be to rely primarily on consolidated accounts. This would involve dropping unconsolidated accounts whenever a consolidated account for the same firm should exist in Orbis (code U2) and, when dropping duplicates, giving preference to code C1.

Each approach has advantages and disadvantages, and the choice between them depends on the type of analysis. Consolidated accounts cover business groups that may span many industries and countries. As a result, it can happen that a firm is classified in a particular industry and registered in a particular country, but most of the output and employment covered in its consolidated accounts belong to other industries and other countries. For this reason, unconsolidated accounts appear more suitable for analyses of dynamics within particular countries and industries. On the contrary, consolidated accounts may be more appropriate, for example, for studying activities of multinationals in the global context. Giving preference to consolidated accounts could also allow capturing a larger part of the total economy, but at a risk of double-counting activity of subsidiaries.[26]

Figure 6.3. Firms with multiple accounts available most often belong to consolidation codes U2 and C2

Distribution of observations corresponding to duplicate firm identifiers, by consolidation-code combinations (2002-2015)



Note: The figure shows the share of duplicate firm identifiers among all account-year observations for each combination of account types. For example, about 0.9% of all account-year observations correspond to firms with both account type U2 and account type C2 in the data in a given year. U1 = unconsolidated accounts of companies for which consolidated accounts are not available. U2 = unconsolidated accounts of companies for which consolidated accounts are also available. C1 = consolidated accounts of companies for which unconsolidated accounts are not available. C2 = consolidated accounts of companies for which unconsolidated accounts are also available. LF = accounts with limited financial information.
Source: Orbis and OECD MultiProd.

6.3. Ownership module

Orbis contains comprehensive information on ownership linkages between firms, which has been extensively used in the literature on multinationals (e.g. Cravino and Levchenko,

2017) or measures of vertical integration (e.g. Alfaro et al, 2018). Orbis contains information on both ownership linkages between shareholders and subsidiaries and the ultimate owners of subsidiaries calculated by Bureau van Dijk at the end of each calendar year. The ultimate owners are calculated by following the ownership pyramid beyond the immediate direct owners, to their owners and so on.

There is a breadth of ownership information available. The Orbis data contain both direct and total ownership linkages between shareholders and subsidiaries. Total linkages encompass both direct and indirect ownership – where the latter reflects indirect ownership through subsidiaries. For example, firm A may own 100% of firm B, which in turn controls 50.01% of C, therefore firm A indirectly owns 50.01% of firm C. There are different definitions of global ultimate owner, depending upon the minimum ownership percentage (25.01% or 50.01%) or the type of ultimate owner (reflecting either firms or individuals).[27] The 50.01% ownership criteria is a commonly used threshold for the definition of control of another firm and hence whether the subsidiary financials are consolidated into the parent accounts.[28]

The partial and improving coverage over time can generate severe challenges for researchers. There are ownership linkages since the early 1990s, but coverage is better from 2007 onwards. Furthermore, the global ultimate owner data only begins in 2007. Whilst other researchers have noted the improving coverage over time, these issues are further complicated by churning in the data, with many living firms leaving the ownership data each year. We find that around 5-10% of firms enter the Orbis ownership data each year and up to 4% leave each year.

Common approaches in the literature are either to assume that firms without an Orbis ultimate owner are independent or to take data from a recent year - assuming ownership has not changed over time. Both of these approaches are problematic. With increasing coverage in Orbis ownership over time, the former approach will falsely equate missing data with independence and lead to an overstatement of ownership changes over time. The latter approach will clearly lead to an understatement of ownership changes over time and will typically overstate the number of markets and countries in which a firm operates.

A better solution is to use Zephyr M&A data to improve coverage. Firstly, one can use M&A data to identify changes in immediate ownership not available in Orbis. Secondly, one can roll-forward and backward known ultimate owners at specific points in time. In particular, if we know firm A is an ultimate owner of firm B in 2010, and from Zephyr M&A data that firm A was acquired in 2008, we can roll backwards the ultimate ownership until 2008.

The coverage issues above may imply that some very large groups change from having few subsidiaries to a large number of subsidiaries from one year to the next, or that some extremely large firms never have subsidiaries. In such cases, using additional M&A data is unlikely to be informative, and manual checks may be needed if these large firms are important for the analysis (such as analysis of industry concentration in Bajgar et al, 2019). Secondary issues that may require correction include: firms that are majority owned but are missing an ultimate owner; ultimate owners that are in fact majority owned by another firm (so by definition cannot be an ultimate owner), temporary (one or two year) ownership changes that reverse themselves.

Chapter 7. Summary of results

Firm-level data covering a large number of countries are essential for understanding key economic trends and investigating the role of policies across countries. This paper explores a widely used cross-country firm-level dataset – Orbis – and compares it to statistics calculated within the OECD MultiProd and DynEmp projects, which are based on official microdata that are representative and often cover the entire firm population. It examines how Orbis compares, overall, to the official microdata and how the coverage and representativeness of Orbis depend on several choices that researchers need to make when preparing the data for analysis. Several lessons stand out from the analysis.

Firm distribution. Firms included in Orbis represent only a fraction of the entire firm population. In addition, they do not form a representative sample of the firm population: they are, on average, larger, older and more productive. Importantly, they are more productive even within each size class. Productivity dispersion appears underestimated in Orbis, primarily because firms at the bottom of the labour productivity distribution are often missing from Orbis. The firm coverage of Orbis is not only partial and skewed towards certain types of firms, it also varies across countries and years. At the same time, the large average size of firms in Orbis means that they represent a much larger share of total employment, output and value added compared to their share in firm population.

Correlations with official data. A consequence of the partial coverage is that Orbis has only a limited ability to replicate variation over time and across industries and countries observed in official aggregate statistics, with median correlations mostly around 0.5. Representativeness issues also mean that Orbis has only a partial success in replicating variation in characteristics of firm distributions that can be observed in official microdata, with median correlations over time often between 0 and 0.5.

Representativeness. Restricting the sample to periods with consistent and relatively high coverage for a small number of best covered countries, and imputing value added based on profits and wages within Orbis, substantially improves the representativeness of the data. Orbis is also more representative of the population of firms with at least 10 employees (after excluding firms below the threshold) than of the entire firm population. However, it is important to keep in mind that the same absolute size threshold will trim the sample at rather different percentiles of the full firm distributions in different countries. Finally, Orbis tends to be more representative for manufacturing than for other sectors, but the advantage is unlikely to be large enough to justify, on its own, excluding a larger part of the economy from the analysis.

Weighting. Application of weights does not seem to significantly improve performance of Orbis beyond its mechanical effects on the firm size distribution. Firms in Orbis are more productive than official microdata even within each size class, thus applying weights to each size class does not recover the official microdata statistics.

Performance of Orbis in a "preferred sample". Jointly restricting the sample to a small number of best-covered European countries (fewer than 10), imputing value added and focusing on firms with at least 10 employees, substantially improves the representativeness (of the population of firms with at least 10 employees). However, its ability to capture the evolution of the entire firm distribution remains limited and it continues to perform poorly at measuring entry and exit.

Rounded values and account consolidation. Researchers using Orbis should pay attention to the heavy presence of rounded values for some key variables in certain countries, and to the presence of multiple account types, often for the same firms.

Ownership information. The ownership data is a particular novelty of the Orbis database, presenting the opportunity to examine multinationals and business groups. However, the coverage of the data is partial and improves substantially over time, which raises particular challenges for measuring changes in ownership without additional cleaning or inclusion of additional sources of data.

To use or not to use? The answer depends upon the type of analysis at hand. Orbis is better suited to an analysis which

- examines performance of large and high-performing firms (e.g. firms at the productivity frontier, multinationals, patenting firms...);
- focuses on statistics at a global level;
- focuses on the best covered country-years;
- examines mean performance; and
- examines within-firm responses to policies and other shocks.

At the same time, caution is required when using Orbis for an analysis which

- examines firms in the lower half of the performance distribution;
- makes comparisons across countries;
- examines properties of the entire firm distribution (e.g. productivity dispersion);
- studies entry and exit.

End notes

[1] Further information on MultiProd, DynEmp and STAN is available at http://www.oecd.org/sti/ind/MultiProd.htm, http://www.oecd.org/sti/ind/dynemp.htm and http://www.oecd.org/sti/ind/stanstructuralanalysisdatabase.htm respectively.

[2] Gal (2013) proposes a regression-based adjustment that takes into account the size-wage premium during the external imputation of value added.

[3] See Haskel and Westlake (2017) for a thorough review of this issue.

[4] Alfaro and Chen (2018); Andrews et al. (2016); Fons-Rosen et al. (2017); McGowan et al. (2015, 2017a,b); and Gopinath et al. (2017).

[5] All studies listed in the previous footnote except those by McGowan et al. (2015, 2017a,b) start their sample in 2002 or earlier.

[6] For a detailed description of the thresholds see Desnoyers-James et al. (2019).

[7] For the Netherlands, this is possible only for firms above 10 employees because the production survey is not representative for smaller firms.

[8] See Berlingieri et al. (2017) for more detail. This paper refers to the measure simply as "employment".

[9] In STAN and, for most countries, in MultiProd, value added is measured in basic prices. In the case of Orbis, the use of basic vs. producer prices is not clear and is likely to vary by country and data source.

[10] Orbis contains a variable for material costs. However, material costs represent only a part of intermediate inputs, so use of intermediate inputs defined as the difference between operating revenue and value added is preferred.

[11] Entry and exit rates are based on DynEmp; all other measures rely on MultiProd. See Berlingieri et al. (2017) and Criscuolo, Gal, and Menon (2015) for more detail.

[12] Studies based on MultiProd data typically rely on productivity based on factor-shares estimated with the method proposed by Wooldridge (2009). This paper instead uses external factor shares to ensure that productivity in Orbis and MultiProd differs due to different firm samples and different output and input information for each firm rather than due to different estimates of factor elasticities. Using the Wooldridge productivity measure (not shown here) leads to larger MFP differences between the two datasets.

[13] The exact definitions differ somewhat between the DynEmp v2 output and the output of the MultiProd code applied to Orbis data. In DynEmp v2, entrants are defined as firms born at time t, exitors are defined as firms last appearing in the data in time t-1 and the denominator consists of all firms present in time t plus the exitors. In MultiProd code, as applied to Orbis here, entrants are defined as above, but exitors are defined as firms last appearing in the data in time t and the denominator consists only of all firms present in time t.

[14] Figures for gross output exclude "wholesale and retail", because this sector has output adjusted downwards in STAN, which would lead to an overestimation of the coverage in Orbis.

[15] The share of value added covered is very high in "coke & petroleum" and very low in "real estate" and "scientific R&D". Measurement of value added in these sectors is generally problematic (e.g. due to imputed rents in the case of "real estate"), and they are dropped in the subsequent analysis benchmarking Orbis against MultiProd.

[16] Coverage data at the country-industry-year level is available from the authors upon request.

[17] In the box plots shown in this paper, the box extends from the 25th to 75th percentile (with a line at the median) and the whiskers mark values that are 1.5 times inter-quartile range below the 25th percentile and above the 75th percentile.

[18] Examining the best-covered data cells, where average firm size measured by employment is similar in Orbis and MultiProd, suggests that capital in Orbis is on average about 35% smaller than in MultiProd. This may appear surprising given that firms in Orbis are likely to be *more* capital intensive than firms in the population. The difference is likely due to a faster depreciation of capital in the books of firms covered by Orbis compared to the depreciation rates applied when constructing capital stock in MultiProd. Rincon-Aznar et al. (2017) similarly find that, for the United Kingdom, asset lives implied by the perpetual inventory method tend to be longer than those implied by company accounts.

[19] The difference in wages may be partly due to a difference in definitions of wages between Orbis and MultiProd.

[20] See, for example, Berlingieri et al. (2017).

[21] When imputation is used, the imputed value added is used for all firms, including the ones where unimputed value added is available. This is done to obtain a measure that is consistent across firms. One implication is that if variables used for constructing the imputed value added are available for fewer firms than the "raw" value added, the coverage with imputed value added can be lower than without imputation.

[22] White, Reiter and Petrin (2017) document that mean-imputation reduces total factor productivity dispersion in the US Census of Manufactures.

[23] The table lists country-years which are also covered by the MultiProd data. Some other countries in Orbis, such as Spain, could also fit in the restricted samples but are not shown here.

[24] Note, however, that cutting the bottom part of the size distribution according to a fixed threshold means comparing different parts of the distribution across different countries. For example, by excluding firms with less than 20 employees, one drops about 97% (55% of employment) of employing firms in Italy service sector and 89% of firms (35% of employment) in the Norwegian service sector.

[25] For more information, see Berlingieri et al. (2017).

[26] A potential solution to this issue is to rely on the Orbis ownership module and drop subsidiaries of mother companies that have consolidated accounts. The challenge of this approach is that ownership information in Orbis is incomplete and restricted to the more recent years (see Subsection **Error! Reference source not found.**).

[27] The firm-type ultimate owners include industrial, financial and insurance companies and banks. This definition excludes individuals, as well as mutual and pension funds, foundation and research institutes, public / state owners, employees/managers/directors, self-ownership, private equity firms, unnamed shareholders, venture capital or hedge funds.

[28] For example, a majority of voting rights is often a sufficient condition for control under international accounting practices. Whilst we do not observe voting rights, we assume these are reflected in shareholder ownership percentages, such that a majority (50.01%) of shares reflects a

majority of voting rights. This is clearly a first-order approximation. For example, some shares can carry more voting rights than others and some types of shares have no voting rights at all. In addition, definitions of control can vary across accounting practices and having a minority of the voting rights can still imply de facto control if the remaining shares are spread across a large number of parties.

References

Adalet McGowan, Muge and Dan Andrews. 2015. 'Labour Market Mismatch and Labour Productivity: Evidence from PIAAC Data'. OECD Economics Department Working Paper 1209. Paris: Organisation for Economic Co-operation and Development. http://dx.doi.org/10.1787/5js1pzx1r2kb-en.

Adalet McGowan, Muge, Dan Andrews, and Valentine Millot. 2017. 'Insolvency Regimes, Zombie Firms and Capital Reallocation'. OECD Economics Department Working Paper 1399. Paris: Organisation for Economic Co-operation and Development. https://doi.org/10.1787/18151973.

Adalet McGowan, Müge, Dan Andrews, and Valentine Millot. 2018. "The Walking Dead? Zombie Firms and Productivity Performance in OECD Countries." *Economic Policy* 33 (96): 685–736. https://doi.org/10.1093/epolic/eiy012.

Alfaro, Laura, and Maggie X. Chen. 2018. "Selection and Market Reallocation: Productivity Gains from Multinational Production." *American Economic Journal: Economic Policy* 10 (2): 1–38. https://doi.org/10.1257/pol.20150437.

Andrews, Dan, Chiara Criscuolo, and Peter N. Gal. 2016. 'The Best versus the Rest: The Global Productivity Slowdown, Divergence across Firms and the Role of Public Policy'. OECD Productivity Working Papers 5. Paris: OECD Publishing. https://doi.org/10.1787/63629cc9-en.

Arnold, Jens, Giuseppe Nicoletti, and Stefano Scarpetta. 2008. 'Regulation, Allocative Efficiency and Productivity in OECD Countries: Industry and Firm-Level Evidence'. 616. OECD Economics Department Working Papers. OECD Publishing. https://ideas.repec.org/p/oec/ecoaaa/616-en.html.

Bajgar, Matej, Giuseppe Berlingieri, Sara Calligaris, and Chiara Criscuolo. 2019. 'Can Business Micro Data Match Macro Trends? Comparing Multiprod Data with STAN'. OECD STI Working Paper.

Berlingieri, Giuseppe, Patrick Blanchenay, Sara Calligaris, and Chiara Criscuolo. 2017. 'The MultiProd Project: A Comprehensive Overview'. OECD Science, Technology and Industry Working Papers 2017/04. Paris: OECD Publishing.

Berlingieri, Giuseppe, Patrick Blanchenay, and Chiara Criscuolo. 2016. 'The Great Divergence'. OECD Science, Technology and Industry Policy Papers. Paris: OECD Publishing.

Berlingieri, Giuseppe, Sara Calligaris, and Chiara Criscuolo. 2017. 'Firm-Level Productivity Differences: Insight from the OECD's MultiProd Project'. International Productivity Monitor 32: 97–115.

Calvino, Flavio, Chiara Criscuolo, and Carlo Menon. 2015. 'Cross-Country Evidence on Start-up Dynamics'. OECD Science, Technology and Industry Working Papers. Paris: OECD Publishing.

———. 2016. 'No Country for Young Firms? Start-up Dynamics and National Policies'. 29. OECD Science, Technology and Industry Policy Papers. Paris: OECD Publishing.

Card, David, J/"org Heining, and Patrick Kline. 2013. 'Workplace Heterogeneity and the Rise of West German Wage Inequality'. The Quarterly Journal of Economics 128 (3): 967–1015. https://doi.org/10.1093/qje/qjt006.

Criscuolo, Chiara, Peter N. Gal, and Carlo Menon. 2014. 'The Dynamics of Employment Growth: New Evidence from 18 Countries'. 14. OECD Science, Technology and Industry Policy Papers. OECD Publishing. https://ideas.repec.org/p/oec/stiaac/14-en.html.

———. 2015. 'DynEmp: A Routine for Distributed Microdata Analysis of Business Dynamics'. Stata Journal 15 (1): 247–274.

Criscuolo, Chiara, and Jonathan Timmis. 2018. 'GVCs and Centrality: Mapping Key Hubs, Spokes and the Perphery'. OECD Productivity Working Papers 12.

Decker, Ryan, John Haltiwanger, Ron Jarmin, and Javier Miranda. 2014. 'The Role of Entrepreneurship in US Job Creation and Economic Dynamism'. *Journal of Economic Perspectives* 28 (3): 3–24. https://doi.org/10.1257/jep.28.3.3.

Desnoyers-James, I., S. Calligaris, and F. Calvino (2019), "DynEmp and MultiProd: Metadata", *OECD Science, Technology and Industry Working Papers*, forthcoming.

Fons-Rosen, Christian, Şebnem Kalemli-Özcan, Bent E. Sorensen, Carolina Villegas-Sanchez, and Vadym Volosovych. 2017. "Foreign Investment and Domestic Productivity: Identifying Knowledge Spillovers and Competition Effects", NBER Working Paper No. 23643. https://www.nber.org/papers/w23643

Foster, Lucia, John C. Haltiwanger, and Cornell John Krizan. 2001. "Aggregate Productivity Growth. Lessons from Microeconomic Evidence." In *New Developments in Productivity Analysis*, 303–372. University of Chicago Press. http://www.nber.org/chapters/c10129.pdf.

Gal, Peter N. 2013. 'Measuring Total Factor Productivity at the Firm Level Using OECD-ORBIS'. OECD Publishing. http://ideas.repec.org/p/oec/ecoaaa/1049-en.html.

Gabaix, Xavier. 2011. "The Granular Origins of Aggregate Fluctuations." *Econometrica* 79(3): 733-772. https://doi.org/10.3982/ECTA8769

Gopinath, Gita, Şebnem Kalemli-Özcan, Loukas Karabarbounis, and Carolina Villegas-Sanchez. 2017. "Capital Allocation and Productivity in South Europe." *The Quarterly Journal of Economics* 132 (4): 1915–67. https://doi.org/10.1093/qje/qjx024.

Griliches, Zvi, and Haim Regev. 1995. "Firm Productivity in Israeli Industry 1979–1988." *Journal of Econometrics* 65 (1): 175–203. https://doi.org/10.1016/0304-4076(94)01601-U.

Haskel, Jonathan, and Stian Westlake. 2017. Capitalism without Capital: The Rise of the Intangible Economy. Princeton University Press.

Hsieh, Chang-Tai, and Peter J. Klenow. 2009. 'Misallocation and Manufacturing TFP in China and India'. The Quarterly Journal of Economics 124 (4): 1403–48.

Javorcik, Beata S., and Mariana Spatareanu. 2008. 'To Share or Not to Share: Does Local Participation Matter for Spillovers from Foreign Direct Investment?' Journal of Development Economics 85 (1–2): 194–217.

———. 2009. 'Tough Love: Do Czech Suppliers Learn from Their Relationships with Multinationals?' Scandinavian Journal of Economics 111 (4): 811–33.

———. 2011. 'Does It Matter Where You Come From? Vertical Spillovers from Foreign Direct Investment and the Origin of Investors'. Journal of Development Economics 96 (1): 126–38.

Kalemli-Özcan, Şebnem, Bent Sørensen, Carolina Villegas-Sanchez, Vadym Volosovych, and Sevcan Yeşiltaş. 2019. 'How to Construct Nationally Representative Firm Level Data from the ORBIS Global Database'. Working Paper 21558. National Bureau of Economic Research. https://doi.org/10.3386/w21558.

Klapper, Leora, Luc Laeven, and Raghuram Rajan. 2006. 'Entry Regulation as a Barrier to Entrepreneurship'. Journal of Financial Economics 82 (3): 591–629. https://doi.org/10.1016/j.jfineco.2005.09.006.

Melitz, Marc J., and Sašo Polanec. 2015. 'Dynamic Olley-Pakes Productivity Decomposition with Entry and Exit'. The RAND Journal of Economics 46 (2): 362–75. https://doi.org/10.1111/1756-2171.12088.

OECD. 2017. Business Dynamics and Productivity. Paris: OECD Publishing. /content/book/9789264269231-en.

Petrin, Amil, and James Levinsohn. 2012. 'Measuring Aggregate Productivity Growth Using Plant-Level Data'. RAND Journal of Economics 43 (4): 705–725.

Rincon-Aznar, Ana, Rebecca Riley, and Garry Young. 2017. "Academic Review of Asset Lives in the UK." NIESR Discussion Paper No. 474. https://www.niesr.ac.uk/sites/default/files/publications/DP474.pdf

Song, Jae, David J. Price, Fatih Guvenen, Nicholas Bloom, and Till von Wachter. 2019. "Firming Up Inequality." *The Quarterly Journal of Economics* 134 (1): 1–50. https://doi.org/10.1093/qje/qjy025.

Timmer, M. P., E. Dietzenbacher, B. Los, R. Stehrer and G.T. De Vries (2015), "An Illustrated User Guide to the World Input-Output Database: the Case of Global Automotive Production", *Review of International Economics*, Vol. 23/3, pp. 575-605, http://dx.doi.org/10.1111/roie.12178.

Wooldridge, J. M. (2009), "On Estimating Firm-Level Production Functions Using Proxy Variables to Control for Unobservables", *Economics Letters*, Vol. 104, No. 3, pp. 112–114.

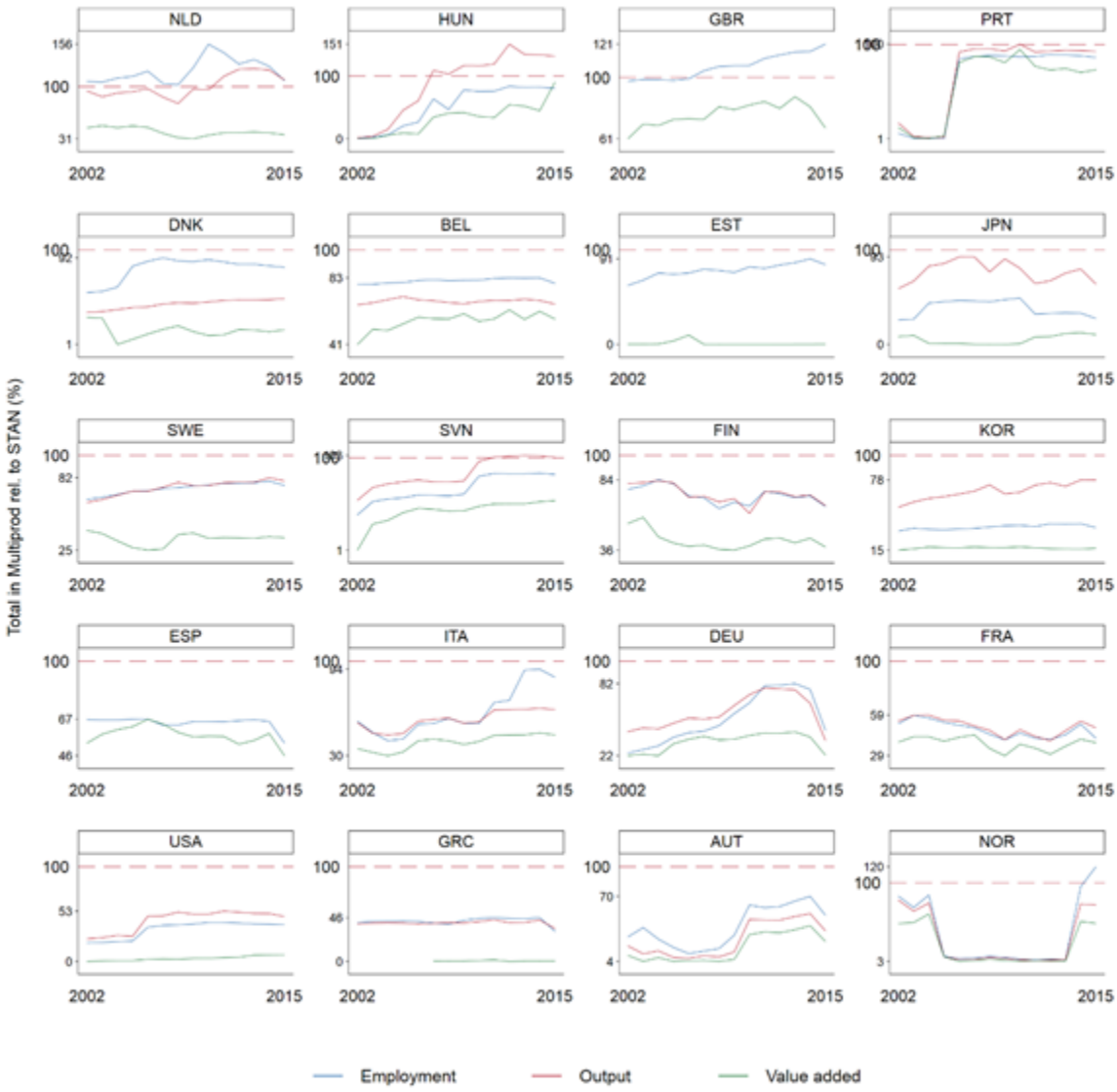
White, T. Kirk, Jerome P. Reiter, and Amil Petrin. 2017. 'Imputation in U.S. Manufacturing Data and Its Implications for Productivity Dispersion'. The Review of Economics and Statistics, June. https://doi.org/10.1162/REST_a_00678.

Annex A. Appendix

Additional results

Figure A.1. Share of total output and input captured by Orbis by country over time

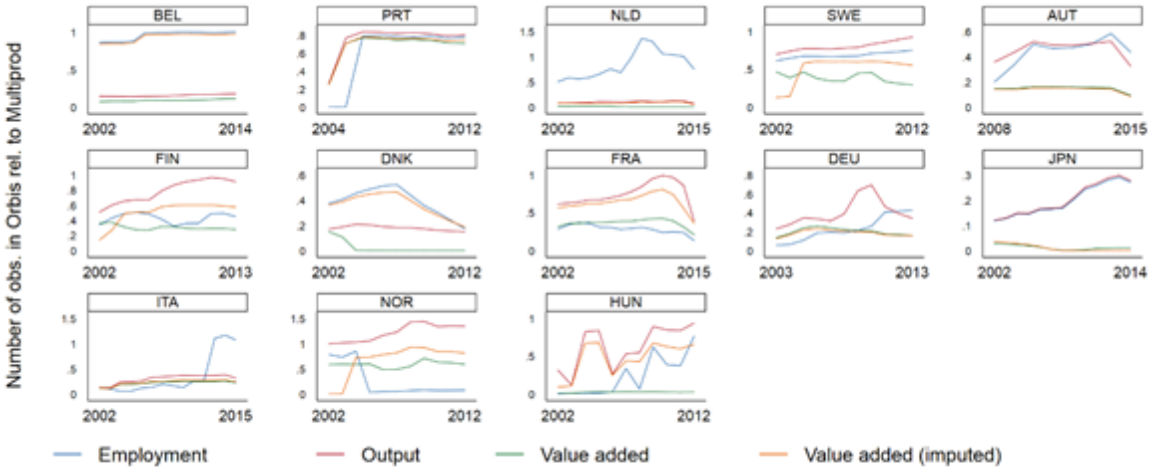
Total employment, output and value added relative to STAN, by country over time (2002-2015)



Note: Manufacturing, utilities, construction and non-financial services. Figures for gross output exclude "Wholesale and retail".
Source: Orbis and OECD STAN.

Figure A.2. Firm coverage without conditioning on employment information availability

Firm coverage by country over time (2002-2015).



Note: The graph shows the number of non-missing observations in Orbis for each variable relative to number of observations in MultiProd with non-missing employment. Manufacturing and non-financial services (excluding "Coke and refined petroleum", "Real estate" and "Scientific R&D").
Source: Orbis and OECD MultiProd.