# How well do data from the Global Entrepreneurship Monitor and Amadeus reflect the global firm size distribution?

Markus Poschke*

McGill University,
CIREQ and IZA

May 2015

**Abstract**

The Global Entrepreneurship Monitor (GEM) and Amadeus are appealing sources for information on the firm size distribution across a broad set of countries due to their harmonized data collection protocols. This note compares information on the firm size distribution to that from other, official data sources with categorical data and more limited country coverage. The main conclusion is that the GEM provides a reliable picture of the size distribution for firms with less than 250 employees, whereas Amadeus provides a reliable picture for firms with 50 or more employees.

See Section 2.1 in Poschke (2015) for a description of the GEM and Amadeus.

The differential focus of the GEM and Amadeus raises a concern of representativeness. Since the GEM focusses on new firms, which tend to be small, and Amadeus focusses on larger firms, it is possible that each provides a good picture of part of the firm size distribution, while missing the middle of the distribution. To address this concern, this note documents how the distribution of firm size in these two data sets (measured as employment) compares to those documented in the only official sources containing somewhat comparable data for several countries, namely data from Eurostat. In addition, we compare data from the GEM on firms in the U.S. to data from the U.S. Census Statistics of U.S. Businesses, and data from Amadeus on French firms to French official data. (Amadeus only contains data on European countries.)

## 1 The GEM data versus other sources

### 1.1 Eurostat

Eurostat provides data on the firm size distribution for a very limited number of size classes in its Structural Business Statistics.[1] The data is collected by national statistical agencies via

*Contact: McGill University, Economics Department, 855 Sherbrooke St West, Montreal QC H3A 2T7, Canada. e-mail: markus.poschke@mcgill.ca

[1] See `http://ec.europa.eu/eurostat/cache/metadata/en/sbs_esms.htm` for a detailed description.

surveys or by drawing on business registers or other administrative sources. It covers all sectors except agriculture and personal services. In principle, each national statistical agency treats its data to comply with data quality requirements.

The data used in the analysis here are for the year 2010, which is the one with the largest country coverage at the moment. The broadest sectoral aggregate available for that year is the total business economy, except financial and insurance activities, agriculture and personal services (code sbs_sc_sca_r2). Eurostat publishes counts of firms for five broad size categories: 0-9, 10-19, 20-49, 50-249 and 250 and more employees.

We can compare GEM and Eurostat data for countries for which data is available in both data sets. This restriction limits us to twelve European countries.[2]

Given small sample sizes for some countries in the GEM, the discussion below focusses on unweighted cross-country averages. For individual countries, differences between Eurostat data and the GEM are only to be expected, as the GEM is a survey. Still, we also provide information on individual countries in a set of figures.

Table 1: Firm size distribution, GEM versus Eurostat, average across 12 countries

| share (%) of firms with... | Eurostat | GEM |
|---|---|---|
| $<$ 10 employees | 93.3 | 86.2 |
| 10-249 employees | 6.5 | 12.8 |
| *among these:* | | |
| 10-19 employees | 54.2 | 50.0 |
| 20-49 employees | 30.9 | 34.0 |
| 50-249 employees | 14.9 | 15.9 |
| 250 and more employees | 0.2 | 1.0 |

Note: Figures are arithmetic averages of the data for Austria, Belgium, Spain, France, Croatia, Hungary, the Netherlands, Norway, Poland, Portugal, Sweden and Slovenia.

In this spirit, Table 1 shows a comparison of the firm size distribution in Eurostat to that computed from the GEM, averaging across the twelve countries. The table shows, somewhat surprisingly, that relative to Eurostat data, the GEM data slightly understates the prevalence of firms with less than 10 employees. (We discuss below why this might be the case.) Given that this size group accounts for the bulk of firms everywhere, this implies that the fraction of firms with 10 to 249 or with more than 250 employees in the GEM is substantially larger than that in the Eurostat data. This is particularly pronounced for the largest size group. Information from the GEM on the share of large firms (250 or more employees) thus is not reliable. Information on medium-sized firms (10 to 249 employees), in contrast, is good: the more detailed breakdown shown in the table reveals that for more detailed size classes within the group of firms with 10 to 249 employees, GEM and Eurostat data are very close. The GEM thus slightly overstates

[2]The countries are Austria (AT), Belgium (BE), Spain (ES), France (FR), Croatia (HR), Hungary (HU), the Netherlands (NL), Norway (NO), Poland (PL), Portugal (PT), Sweden (SE) and Slovenia (SI).

the share of firms with fewer than 10 employees, and captures the distribution of firms with 10 to 249 employees well.

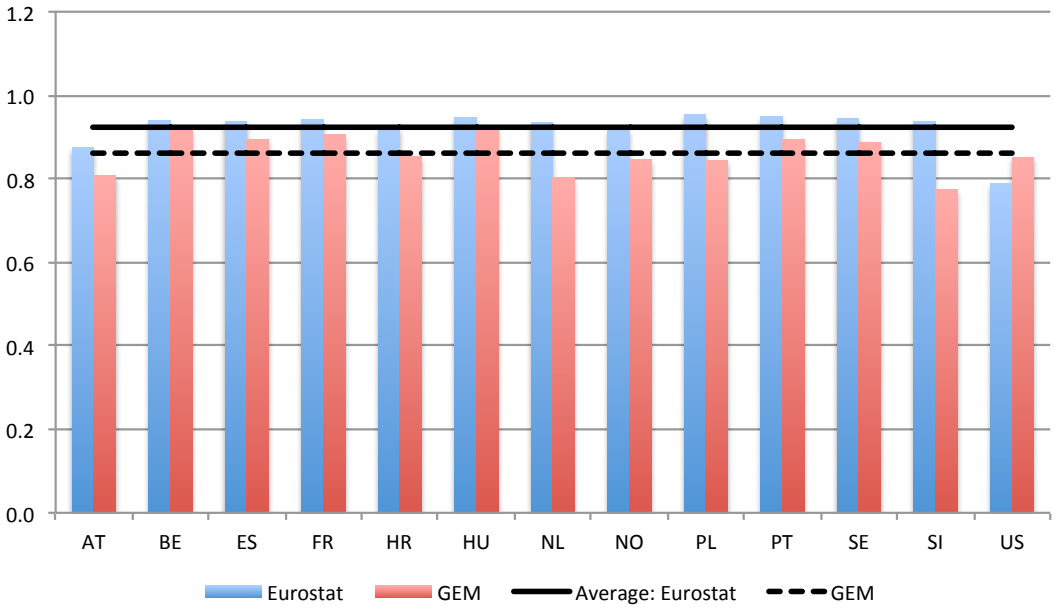


Figure 1: Shares of firms with less than 10 employees, GEM and Eurostat

Figures 1 to 3 show data for individual countries. For comparison purposes, we also include data for the U.S. in the figures (details on the source below). Figures 1 and 2 depict the shares of firms with less than 10 and with 250 or more employees, respectively. Figure 3 depicts ratios of GEM to Eurostat data for the shares of firms with 10 to 19, 20 to 49 and 50 to 249 employees, respectively. That is, a value of one indicates that the share of firms in a group is identical in the two data sets. In each graph, the horizontal lines show cross-country averages. (In Figure 3, the horizontal lines for the larger two groups lie on top of each other, so that only one is visible.)

It is clear from the Figures, as from Table 1, that on average, the share of firms with less than 10 employees in the two data sets is close. At the same time, the share of firms with 250 or more employees is larger in the GEM for many countries (but not for the U.S.). The size distribution of firms with 10 to 249 employees is very close on average (all horizontal lines lie between 0.9 and 1.1). For individual countries, there are deviations from this pattern which, however, do not appear to be systematic.

The bottom line is that on average, the GEM reports a smaller share of small firms (0 to 9 employees) than Eurostat data. Because these firms are so numerous (around 90% of firms on average), this implies that the share of larger firms differs more. However, among firms with 10 to 249 employees, the distributions in the GEM and in Eurostat data are extremely close. The share of firms with 250 or more employees in the GEM is several times larger than

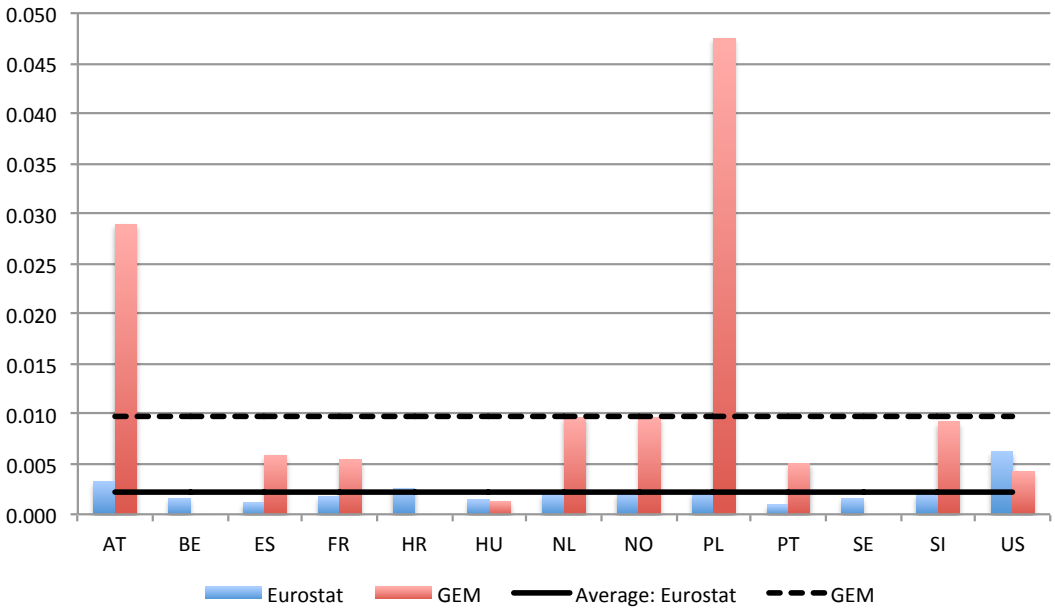


Figure 2: Shares of firms with 250 or more employees, GEM and Eurostat

the corresponding share in Eurostat data. As a consequence, the GEM appears to provide a reliable picture of the firm size distribution for firms with less than 250 employees, while slightly understating the share of very small firms (0 to 9 employees).

**Understanding potential sources of differences.** It is surprising that the GEM, a data set that is geared towards the measurement of emerging entrepreneurial activity, should count a smaller fraction of small firms than Eurostat data. This difference then causes ripples in the reported size distributions, particularly affecting the reported share of large firms (250+ employees). What is the source of this difference?

There are three potential sources for this difference. First, individuals interviewed for the GEM might not report self-employment or firm ownership even when their firm is captured in official data. Whether individuals correctly report self-employment or firm owner status depends on how well the survey question is phrased. The GEM survey questions are very clear (see footnote for details). Therefore, this probably is not an important source of differences.[3]

Second, sample selection criteria may disproportionately remove small firms from the GEM

[3]To be potentially counted as an entrepreneur in the author's treatment of GEM data, respondents need to answer "Yes" to at least one of the following three questions (multiple affirmative answers are possible): 1. "You are, alone or with others, currently the owner of the company you help manage, self-employed, or selling any goods or services?" (true for 93% of self-employed in the 2005 data) 2. "You are, alone or with others, currently trying to start a new business, including any type of self-employment or selling any goods or services?" (true for 17% of self-employed in the 2005 data) 3. "You are, alone or with others, trying to start a new business or a new venture with your employer – an effort that is part of your normal work?" (true for 7% of self-employed in the 2005 data)

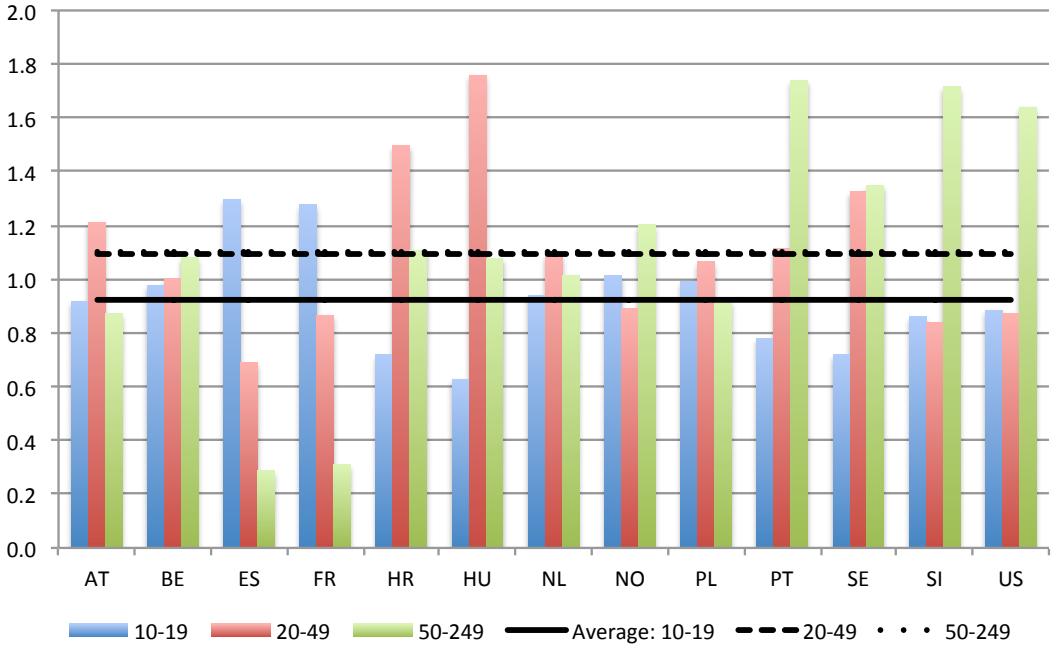


Figure 3: Shares of firms with 10 to 249 employees, subcategories, share in GEM relative to share in Eurostat

data set. While the GEM survey design aims to also capture "nascent" firms that do not yet operate, these are not relevant for a comparison to the size distribution of active firms. Therefore, our sample selection criterion for the GEM firm size data is to only include start-ups (firms started within the last three and a half years) that answer "Yes" to the following question: "Has the new business paid any salaries, wages, or payments in kind, including your own, for more than three months?" Startups in the GEM account for 25% of all firms. Among these, only about 40% report already having made payments for long enough to be included in the sample as an active firm. The criterion thus is rather strict. It is in principle possible that Eurostat data includes firms that do not fulfill this criterion, for example because they are making losses. Of course, Eurostat also attempts to only count economically active firms. Inclusion criteria are country-dependent. In the Netherlands, for example, only firms with at least 15 weekly hours of labor input are counted. It is unclear which criterion is stricter. The larger share of small firms in Eurostat might thus be explained by differences in the stringency of the selection criterion.

Third, the GEM data do not provide information on how many firms an individual owns or operates. In particular for small firms, this may affect counts. This is particularly relevant since Eurostat defines an enterprise as "the smallest combination of legal units that is an organisational unit producing goods or services, which benefits from a certain degree of autonomy in decision-making, especially for the allocation of its current resources. An enterprise carries out one or more activities at one or more locations. An enterprise may be a sole legal unit." This type of definition tends to result in more, smaller firms than one that consolidates firms with a common

owner. In particular the last two factors may raise the share of small firms measured by Eurostat relative to that to be found in GEM data.

An important question is how the firm size distribution differs in countries with different levels of per capita income. Differences across data sources that are common across countries are not a concern for that question. However, the country-level pattern of differences may be. We next consider this question.

**Are differences systematic?** Due to sampling, it is normal that there could be differences between the GEM data and Eurostat data, in particular for countries with small samples. It would be a concern if these deviations were correlated with per capita income of the country. Figure 4 shows that this is not the case. The figure plots the ratio of the share of firms with zero to nine employees from the GEM relative to that from Eurostat and the U.S. Census against 2005 per capita income from the Penn World Tables. The ratio is on average slightly below one, in line with the differences just discussed. However, it is obvious that there is no relationship to per capita income.

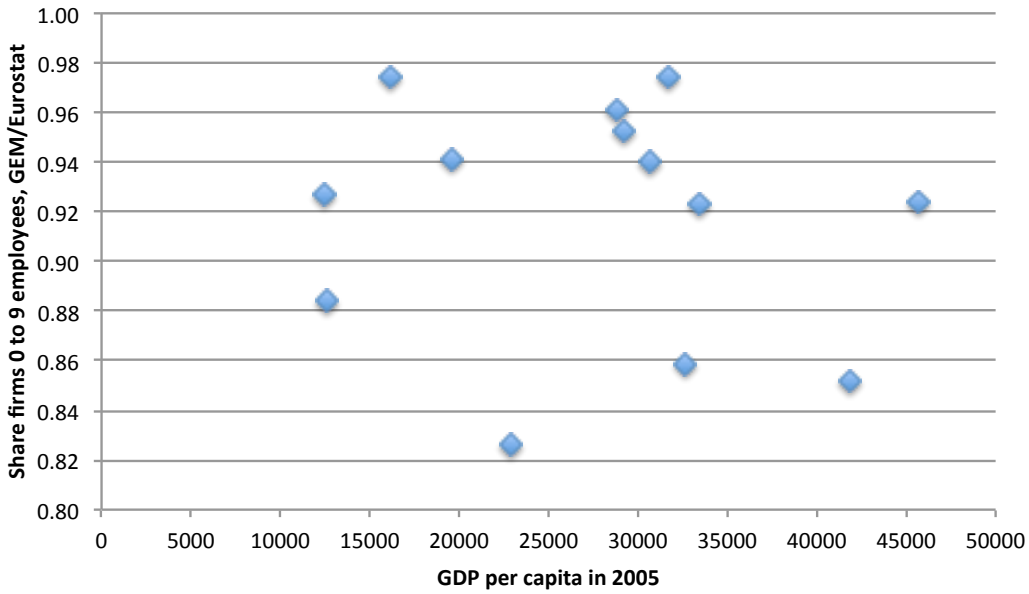


Figure 4: Share of firms with 0 to 9 employees, GEM/Eurostat, vs income per capita in 2005 (Penn World Tables)

## 1.2 The GEM data and U.S. Census data

While the small sample size implies that comparing GEM data on the firm size distribution to data for a single, specific country may be hazardous, it is still instructive. We therefore provide a comparison to firm size data from the U.S. Census Statistics of U.S. Businesses. For the

years in the sample, our GEM data contain 2151 observations on U.S. firms with information on employment. Of course, most of these are on small firms, making observations on detailed size classes of large firms unreliable. The comparison below takes this into account by comparing cumulative distribution functions and shares in broader size classes.

Figure 5 shows the cumulative share of firms in different size classes. In contrast to the European countries, here, the GEM slightly overstates the fraction of small firms. However, the difference shrinks to 2 percentage points once one considers all firms with size between zero and nineteen. For firms larger than this, the two lines quickly become indistinguishable.


Figure 5: The U.S. firm size distribution, U.S. Census Statistics of U.S. Businesses and the GEM

Of course, due to the small size of bins of firms with large size, a *cdf* may not be the best way of making the comparison. Therefore, Table 2 shows the fraction of firms in different broad size classes.[4] Again, it is clear that even for larger size classes, the distributions in the Census data and in the GEM are similar.

[4]Due to sample size, this is done for a popular aggregation of size classes, and not for the smallest size classes reported by the Census. Note that the Census does not report data for a size class 200-249 or 250-299, but only 200-299. For the figure in the previous subsection, the firm count in the 200-249 size class was imputed assuming a Pareto distribution with tail index 1.165 for firms with 200 or more employees. This is the tail index implied by average employment of firms with 200 or more employees.

Table 2: The firm size distribution, broad groups, U.S. Census and GEM

| share (%) of firms with ... employees | Census | GEM |
|---|---|---|
| 0-9 | 78.8 | 85.2 |
| 10-19 | 10.7 | 6.6 |
| 20-49 | 6.7 | 4.1 |
| 50-99 | 2.1 | 2.2 |
| 100-299 | 1.3 | 1.6 |
| 300+ | 0.5 | 0.4 |

# 2 Amadeus versus other data sources

## 2.1 Eurostat

Since the Amadeus dataset covers 44 European countries, it is only appropriate to compare its depiction of the firm size distribution with that to be found in Eurostat data. Since in Europe, Amadeus covers more countries than the GEM, we can do this comparison for 19 countries where data is available from both sources.[5,6] Our comparison uses Amadeus data for the year 2007.

Amadeus focusses on larger firms. To compare it to Eurostat data, it is therefore not helpful to compare shares of firms in different size classes, as Amadeus by construction does not include a large number of very small firms. We will therefore conduct a more demanding exercise and compare the *number* of large firms in Amadeus to that in Eurostat data. This comparison clearly is not sensitive to the treatment of small firms.

Figure 6 shows the number of firms with 250 or more employees in Amadeus relative to that in Eurostat for the 19 countries. That is, a value of one indicates that the number of large firms is identical in the two datasets. The horizontal line shows the average of the ratio across countries. Since the Eurostat data draw on national registers in which large firms should be fully represented, this ratio should in principle be bounded above by one. Because Amadeus also draws on these registers, it should be close to one. The Figure shows that on average, the ratio is indeed close to one, implying that Amadeus does an excellent job at capturing the universe of large firms.

What about smaller firms? Figure 7 shows the ratio of the number of firms with 50 to 249 employees to that with 250 or more employees, for both Amadeus and Eurostat. This measure captures in a very simple way concentration of the size distribution. Again, the horizontal lines show cross-country averages. It is clear that on average, the ratios are very close, with about five

[5]The countries are Austria (AT), Belgium (BE), the Czech Republic (CZ), Estonia (EE), Spain (ES), France (FR), the United Kingdom (GB), Croatia (HR), Hungary (HU), Lithuania (LT), Luxembourg (LU), the Netherlands (NL), Norway (NO), Poland (PL), Portugal (PT), Romania (RO), Sweden (SE), Slovenia (SI) and Slovakia (SK). In the main set of results in the paper, I omit countries where Amadeus only covers a low fraction of employment. However, results in the paper change little as that cutoff is changed from 10 to 33 to 50%. Of course, statistical significance is harmed by reducing the sample.

[6]Note that the U.S. is not represented in Amadeus.

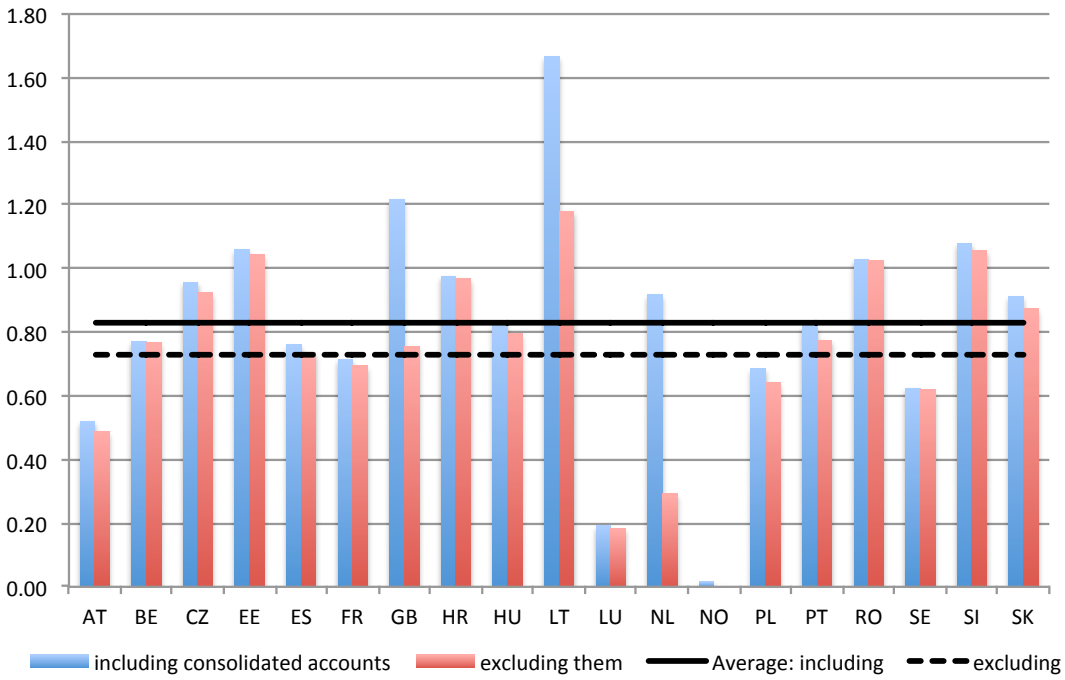


Figure 6: Number of firms with more than 250 employees, Amadeus/Eurostat

medium-sized firms for each large firm in the 19 countries under consideration. More than this, it is clear that for many countries, the ratios computed from the two data sets are extremely close. This implies that the shares of large and medium sized firms among firms with 50 or more employees are very close in Amadeus and Eurostat data. In addition, there is no systematic pattern; even in countries with large deviations the deviations go one way in some cases, and the other way in others. Amadeus thus provides an excellent picture of the size distribution of firms with 50 or more employees across the 19 countries under consideration.

**Understanding sources of differences.** Why is the number of large firms in Amadeus smaller than that in Eurostat data? The reason again is the definition of the firm. When there are ownership links among firms, the design of Amadeus data aims to attribute information on a firm to the ultimate owner. Eurostat data, in contrast, defines a firm as "the smallest combination of legal units ..." (full definition above). As a result, a group of firms with ownership links may appear as a single record in Amadeus, but as multiple firms in Eurostat data. Consider two examples. First, consider a firm with 500 employees that owns another firm of the same size. These would appear as one large firm in Amadeus, but as two in Eurostat, resulting in a smaller count of large firms in Amadeus. Secondly, consider a firm with 200 employees that owns a firm with 100 employees. Eurostat will put both in the 50 to 249 category, whereas in Amadeus, there will be an entry for a single consolidated firm with 300 employees. As a result, it

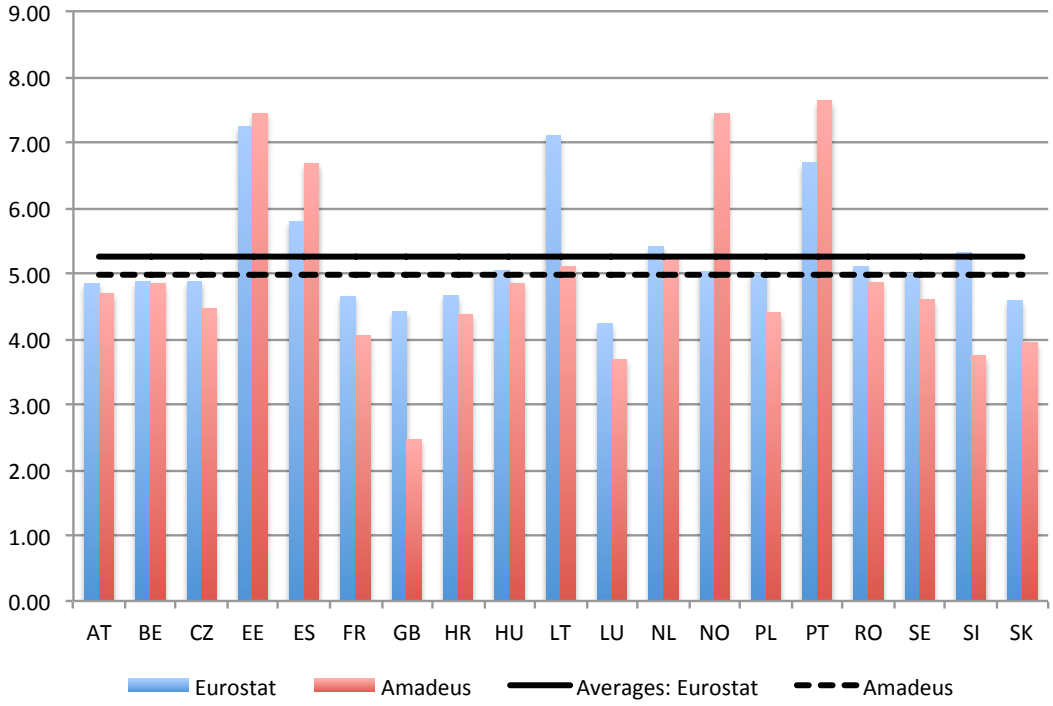


Figure 7: Number of medium (50-249 employees) relative to large (>250) firms, Amadeus and Eurostat

is in principle also possible for the number of large firms in Amadeus to exceed that in Eurostat (as is the case for the United Kingdom).

In Figure 6, the blue bars are based on Amadeus data including data for "consolidated accounts". (Amadeus consolidation code C1 included, C2 excluded.) The red bars exclude these. (Only consolidation code U1 included.) By definition, the red bars therefore understate the true number of firms, since they exclude a group of firms. But even for the blue bars, the interpretation needs to bear in mind that differences in the definition of the firm tend to reduce the measured number of large firms in Amadeus relative to Eurostat.

## 2.2 A detailed look at French data

Finally, we compare the firm size distribution in Amadeus to that in official data in more detail for a single country. We can do this using the extremely detailed data on firm sizes for France reported in Gourio and Roys (2014).[7] This allows assessing how well Amadeus represents the French size distribution not just for large firms, but also for smaller ones.

[7]Thanks to Nicolas Roys for providing the detailed firm size counts underlying Figure 1 in Gourio and Roys (2014). The data is complied by INSEE, the French statistical institute, based on a combination of administrative data and surveys. It includes essentially all firms with 30 or more employees. The data Gourio and Roys use is for the years 1994-2000.

Figure 8 compares the size distribution for firms with twenty to one hundred employees. Because Gourio and Roys aggregate data across years, the figure plots the ratio of the number of firms at each integer level of employment from 20 to 100 to the number of firms with employment of exactly 100. It does so both for the official French data and for the Amadeus data. It is clear from the figure not only that the Amadeus data shows essentially exactly the same size distribution for firms with employment of more than 50 employees, but that it even reveals the kink in the size distribution at employment of 50 that is the focus of Gourio and Roys (2014). Below the threshold of 50 employees, the accuracy of the size distribution obtained from Amadeus gradually deteriorates.

The fit is similarly good for larger firms. In the INSEE data, the ratio of the number of firms with more than 200 (1000) employees to that with between 50 and 100 employees is 0.57 (0.074). In Amadeus, this figure is 0.52 (0.081).

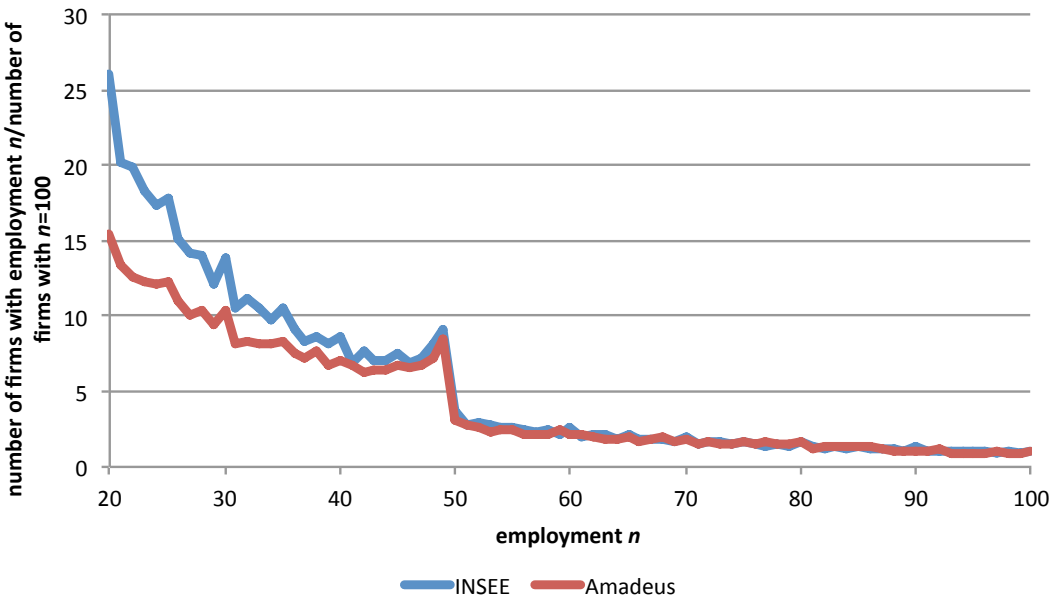


Figure 8: Number of firms of size $n$ relative to number of firms of size 100, data for France from INSEE and from Amadeus.

Source for the INSEE data: Gourio and Roys (2014).

## 3 Conclusion

For countries with the necessary data, the comparison of firm size distributions from different sources has led to the following conclusions:

1. GEM data slightly understate the fraction of firms with 0 to 9 employees. They provide

an excellent picture of the size distributions of firms with 10 to 249 employees.

2. Amadeus data capture a very large fraction of firms with 250 or more employees, and provide an excellent picture of the size distribution of firms with 50 or more employees.

These statements are based on cross-country averages, and on comparisons of broad size classes. Sampling error may lead to differences for individual countries, or for narrower size classes. However, more detailed results for the U.S. for the GEM and for France for Amadeus suggest that even here, the GEM and Amadeus provide a good picture.

Each of the two data sets thus provides a reliable image of part of the size distribution, with a substantial area of overlap in which both do a good job.

## References

Gourio, F. and Roys, N. (2014), 'Size dependent regulations, firm size distribution and reallocation', *Quantitative Economics* **5**(2), 377–416.

Poschke, M. (2015), 'The firm size distribution across countries and skill-biased change in entrepreneurial technology', *mimeo, McGill University* .