Firms' financial statements and competitiveness: an analysis for European non-financial corporations using micro-based data

Nicola Benatti, Annalisa Ferrando and Pierre Lamarche

Abstract

In the empirical analysis the use of micro-data often encounters confidentiality issues, especially when data are derived from different countries. One way to tackle this type of problems is known as "distributed micro-data analysis", which proposes an aggregation of data with sufficient information on the distribution of the underlying micro-based data. In this paper we propose an inverse analysis, i.e. a methodology to mimic the anonymised firms' micro-data starting from a distributed micro-dataset and using standard equations and assumptions about the distribution of the residuals that are most likely to reproduce the original micro dataset. As a result this paper offers an easy tool to analyse the firm-level financial ratios, such as firms' leverage, profitability, and productivity performance of firms across country, sector and firm size even when firm-level data are not readily available.

Keywords: firm-level data, distributed micro-data analysis, simulation-based inference

JEL classification: D22, D24, C53

Introduction

In order to maintain data confidentiality, the information and summary statistics of institutional micro datasets are often published in aggregated form, i.e. providing totals and averages. However, given the increasing need for studying agents' heterogeneity, a number of additional statistics has been included in some publications in the few past years. In general, second and third moments of the distributions and percentiles of the studied variables could provide useful information on their underlying distributions (Bartelsman et al. 2004). The heterogeneity in firms' balance-sheet variables is addressed, for example, in the Bank for the Account of Companies Harmonized (BACH) dataset of the European Committee for Central Balance Sheet Data Offices (ECCBSO), which provides weighted averages, medians, standard deviations and the 25th and 75th percentiles for several balance-sheet items in nine euro-area countries. In a similar way, the Eurosystem Competitiveness Research Network (CompNet) develops a much more detailed dataset providing the values of each decile of the distributions of both balance-sheet items and competitiveness indicators (Lopez Garcia et al. 2014). Last but not least the DynEmp OECD dataset on employment dynamics provides distributed micro-data analysis of business and employment dynamics and firm demographics (Criscuolo et al. 2014).

In this paper, we propose then a methodology to mimic the anonymised firms' micro-data using standard equations and assumptions about the distribution of the residuals that are most likely to reproduce the original dataset. The methodology we propose derives to some extent from other methods often used when facing incomplete information. Indeed the data we are using are only some kind of summary of the complete information we would like to reproduce.

In particular, in surveys, item non-response is often treated so to replace the missing values with values drawn from the appropriate distribution conditional on the information available in the data. For instance, Little and Rubin (2002) have developed Bayesian techniques to deal with incomplete data, addressing in particular the issue of the non-response mechanism. More specifically, household surveys often provide information that can be also considered as incomplete in the sense that the households give answers about e.g. income or assets owned only in brackets. This kind of data collection is used mainly to ease the answer of the households, who are less reluctant to provide fuzzy information that ensures the de facto anonymisation of their answers. Our situation is quite similar in the sense that in the CompNet kind of data we only have information in brackets. Various techniques have been developed to address this issue, such as simulated residuals (see Gourieroux et al. (1987) or Lollivier and Verger (1987)).

We try different approaches to determine the most efficient way of mimicking the data by taking into account the correlations between variables and the bulk of information about the distributions we have at our disposal. We expect the chosen method to be easily implemented and to reproduce as best as possible the expected economic results carried out from the actual micro-data. To do so, we first implement a very simple method but very demanding in terms of hypotheses, assuming a log-normal distribution for each variable. We then try to relax the assumptions by modelling the joint distribution of two given variables in order to avoid using variance-covariance matrices. Finally, for the third model we adopt a Bayesian approach, following for example Gautier (2008). This method is quite computationally demanding, but in theory is able to reproduce any distribution. The different Bayesian techniques that we experimented are described for example by Arnold (1993) or Robert and Casella (2004).

To test the proposed methodology, we first construct an aggregated dataset starting from firm-level observations from Bureau van Dijk's AMADEUS dataset using the CompNet methodology. Our dataset contains information on the distribution of a set of indicators of interest for three euro area countries (France, Italy and Spain) in the period 2006-2011.

Overall, our methodology could be applied to different datasets provided they contained enough information on the distribution of the variables of interest.

The paper is structured as follows: the second section describes the construction of the dataset and its main characteristics while the third section introduces the empirical strategy. In the fourth section we present the results. Then we apply our data to estimate a leverage function and finally we conclude.

Construction of the dataset and data description

As the aim of our analysis is to develop a methodology to retrieve firm-level information from an aggregated database that provides sufficient information on distributions, we decided to use the approach followed by the Eurosystem Competitiveness Research Network (CompNet). The CompNet was established by the European System of Central Banks (ESCB) in 2012 to study competitiveness-related dynamics by collecting indicators and statistics from several European countries. As firm-level balance-sheet data is treated as confidential in most of European countries and cannot be exchanged across different national entities, the network decided to use a common analysis tool (i.e. a Stata script, thereafter CompNet toolkit) to aggregate firm-level data with information on the distribution of the indicators of interest. The script is then sent to National Central Banks (NCBs), where the confidentiality issue does not apply and therefore, the data is available at firm-level.

The characteristic of the database resulting from the data collection is that even though the indicators are computed as aggregates at a country-sector-year level (meso-level), other statistics beyond averages are collected in it, in particular statistics regarding standard-deviations, deciles, skewness and joint moments of the distributions of each indicator.

The main advantage of this procedure is clearly that by applying a common analysis tool to different countries allows using exactly the same treatment for outliers across countries and the same techniques to compute the indicators to produce comparable results.[1]

This type of information, collected via a remote-based analysis, allows studying firms' behaviour in terms of productivity, export, employment, and mark-ups, as well

[1] The network is aware of the many caveats associated to the sample comparability and harmonization of variables' definitions. In this respect a detailed preliminary work has been done to avoid as much as possible discrepancies in the definition (Lopez Garcia et al., 2014).

as studying the impacts on the distribution of firms for the values of each indicator computed. Most recently a "financial module" has been added to the main exercise in order to enlarge the analysis on the impact of the financial situation of firms on their real decisions. For this scope, financial ratios have been collected and indicators on financial constraints have been computed. This paper focuses in particular on this new – not yet published – part of the analysis.

Given the structure of the overall exercise, the available information related to the distributions comes from specific moments (deciles, averages, skewness and standard deviations), making it impossible to further investigate firm heterogeneity within deciles (e.g. to investigate how the distribution of firms between the 1st and 2nd deciles of labour productivity actually looks like). More simply, the limitation of the database derived through the CompNet analysis toolkit is the fact that it cannot be used for a fully micro-based economic analysis.

As the final CompNet dataset is not yet publicly available, we decided to replicate the dataset by applying the CompNet analysis toolkit on a firm-level sample derived from the Bureau van Djik Amadeus database. The aim is twofold. On one side, we apply the CompNet toolkit and we create the aggregated database with information on the distribution of the indicators we have chosen for the analysis. This allows us to apply our methodology to retrieve in a direct way anonymised firm level data. On the other side, the fact that we do have at disposal the original dataset implies that we will be able to prove the goodness of the algorithm we propose in the paper.

Our starting point is the Amadeus database, which is a commercial product from Bureau Van Dijk that collects business registers' balance-sheets information about non-financial corporations in Europe. For our exercise we selected a sub-sample of the database which includes all the available firms in France, Italy and Spain for the period 2006 to 2011. These three countries are those with the highest number of companies in the original dataset and their representativeness across firm size and sectors is relatively high compared with other countries.

We run the CompNet toolkit[2] on this initial sample. The program removes erroneous observations and runs an outlier treatment, which consists of removing observations beyond the 1st or 99th percentile. In addition it is verified that certain ratios (such as collateral) cannot exceed unity. Table 1 summarises some characteristics of the resulting sample. The number of observations is around 4 million and the number of firms 1.3 million. The observations are uniformly distributed across countries with Spanish firms covering 36% of the sample, French companies 33% and Italian firms the remaining 31%. Most of the sample consists of micro and very small firms.[3]

[2] In particular, in order to obtain a set of micro-based aggregated indicators and the statistics of their distributions, we re-adapted the CompNet analysis programme (".do file") to be able to run it on a dataset which contains also a country dimension and we focus only on a set of indicators we are interested in implementing our algorithms.

[3] Firms are divided in five categories according to the number of employees: micro firms are firms with less than 10 employees, very small firms between 10 and 19, small firms between 20 and 49, medium between 50 and 249 and large firms have more than 250 employees.

Number of observations and firms, broken down by country

Table 1

Country	Number of observations	Number of firms	%				
			Micro	Very small	Small	Medium	Large
Spain	1,474,803	446,201	70.2	15.1	9.7	4.2	0.9
France	1,380,743	454,811	74.7	11.1	8.6	4.5	1.1
Italy	1,290,529	419,171	63.7	18.7	10.5	6.0	1.1
Total	4,186,069	1,320,183	69.7	14.9	9.6	4.9	1.0

The statistics refer to the number of observations and firms when the leverage ratio is not missing.

After having cleaned the data, the CompNet toolkit creates the set of indicators and their distributions. In this paper we choose to focus on the following indicators: (a) financial leverage, which is defined as the sum of short- and long-term debt, divided by total assets; (b) cash holding defined as cash and cash equivalents divided by total assets; (c) cash flow to total assets; (d) return on assets (ROA) as net income over total assets and (e) labour productivity as real value added over number of employees. Table 2 presents some of their characteristics across the three countries. Spanish firms are more indebted than French and Italian firms, they are less profitable and their productivity is also lower. French firms are holding more cash and cash equivalents and they produce more internal funds than Italian and Spanish firms.

We focus on eight macro-sectors based on NACE rev. 2 codes. Firms whose code is not available are excluded from the dataset. Furthermore, firms operating in agriculture, fishing, mining, financial activities, public sector, education, health, entertainment, and other services (sections A, B, K, O, P, Q, R, and S) are excluded. The detailed sectorial classification used in the analysis is as follows (in parentheses we report the percentage of observations in our sample): 1) Manufacturing (20%); 2) Construction (18%); 3) Wholesale trade and retail trade (28%); 4) Transportation and storage (5%); 5) Accommodation and food (8%); 6) real estate (5%); 7) Professional, scientific and technical activities (8%) and 8) Administrative and support service activities (8%). Most of the companies in our sample are in the trade, manufacturing and construction sectors, covering altogether two-thirds of the whole sample. The details of the breakdown of the financial indicators by macro-sector are reported in Table A1 in the annex.

As explained above, the final result of the CompNet tool is a set of indicators computed as aggregates at a country-sector-year level (meso-level) with information on their averages, standard-deviations, deciles, and skewness.

Figure 1 on the left displays in detail the heterogeneity of a specific variable, labour productivity, as derived from the CompNet toolkit. We observe that the mean of labour productivity within different countries is always statistically different from the median and it always lays closer to the 70th percentile rather than the 50th.

Summary statistics of financial indicators and productivity, broken down by country

Table 2

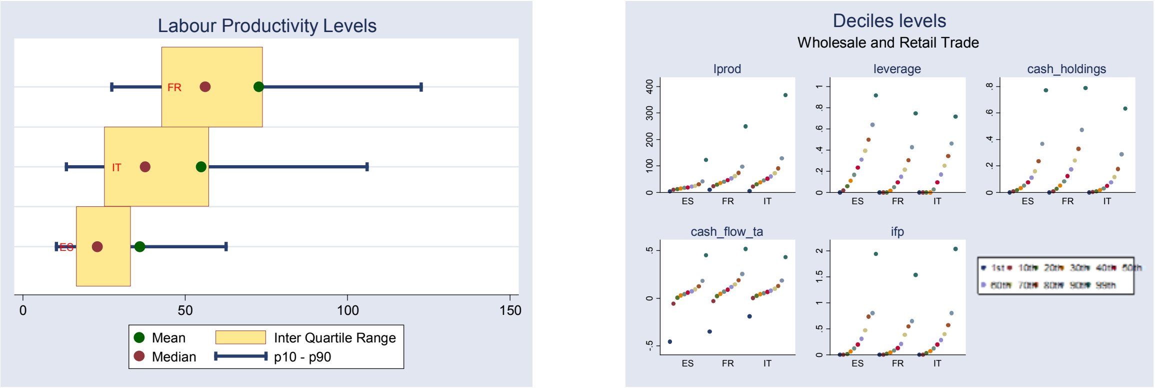
Country		Leverage	Return on Assets	Cash holdings	Cash flow	Labour productivity
Spain	Mean	0.3	0	0.14	0.07	36.02
	Median	0.25	0.03	0.07	0.06	22.88
	Std.	0.25	0.33	0.18	0.16	56.41
France	Mean	0.15	0.07	0.22	0.12	72.63
	Median	0.08	0.07	0.16	0.11	56.07
	Std.	0.19	0.22	0.21	0.17	71.58
Italy	Mean	0.17	0.05	0.1	0.09	54.95
	Median	0.09	0.04	0.04	0.08	37.66
	Std.	0.2	0.14	0.14	0.12	65.54

Notes: All indicators are unitless, except labour productivity in thousands of euro per employee.

Furthermore, by plotting the deciles for the variables of interest, we can easily observe how skewed the distributions actually are (Figure 1 on the right). This is often a characteristic of firm's balance-sheet data.

Labour productivity levels and decile levels[4]

Figure 1



[4] Given the nature of the collected data (i.e. deciles), the Inter Quartile Range of the boxplots is computed using the difference between the 7th and the 3rd deciles.

Empirical strategy

The main aim of this paper is to simulate micro-data that would in the end reflect as much as possible the results given by the CompNet database. We have then at our disposal a set of moments (some percentiles, standard deviation, totals and inter-quartile ranges and joint moments) for a few given variables (namely cash, net income, financial debt cash flow, total assets, employees, real value added, depreciation, labour cost, real turnover, interest rates and capital). We have also for a set of financial ratios derived from these variables (namely financial leverage, cash holding, cash flow over total assets, return on assets, labour productivity, unit labour cost, and labour cost per employee).

The difficulty is to generate micro-data matching all the information listed above. For instance, finding a law whose parameters could be adjusted so to respect percentiles as well as mean, standard deviation and inter-quantile range. We have to impose at least 14 constraints[4] on each variable, without any regards to the ratios that impose also indirect constraints. Finding such a law turns to be challenging. At this point, we have at our disposal two options: the first one is to determine the optimal parameters of a joint distribution thanks to some kind of Newton-Raphson algorithm,[5] the second one is to use a Bayesian approach like in Gautier (2008). We choose the second approach, given the complexity of the likelihood, as described hereafter.

As a starting point, we chose not to impose any constraint on each expected percentile. We rather simulate each variable as the realisation of a random variable following a log-normal law, whose parameters are given by the data. We are able to reproduce the correlation between variables thanks to the variance-covariance matrix, but we have no insurance so far that we will reproduce as expected the distribution of ratios. We generate the variables for each country, sector and year to take into account as much as possible the heterogeneity between firms.

Since the covariance between different variables is not entirely available through the CompNet database, we investigate other methods that would be likely to take into account the link between variables through the distribution of ratios. In a second method, we divide the population according to the percentiles for two given variables (say real value added and labour) and look for the most likely repartition of the population between these joint percentiles with respect to the expected distribution of the ratio (in our example, labour productivity). Doing so, we do not make any assumption about the shape of the joint distribution. However, we assume a uniform distribution and the independence of the variables within each cell of the joint distribution. This method can be seen as a linear interpolation between each given percentile when computing the distribution of the variables. Moreover, the number of parameters to be estimated is higher, making the computation process more demanding.

Finally we relax the assumption for the distribution for each variable per decile. More exactly, we have at our disposal the 1st, 10th, 20th, …, 90th and 99th percentiles. Within each of the 12 strata defined by these percentiles, we simulate the expected variables following a bounded law. Thanks to this method we mechanically generate data that respect at least the given percentile for each variable. However, this method does not ensure at all that the other constraints will be respected. In particular, the data generation shall also be conditioned by the distribution of the financial ratios that have been computed with the variables.

As a first assumption, we consider a Beta distribution for each variable within each stratum of the population. This distribution is conditioned on two parameters, α and β. These parameters have to be set so as to respect as much as possible the

[4] We have at our disposal the mean, the standard deviation, the inter-quantile range, the 1st, the 10th, the 20th,…, the 90th and the 99th percentiles.

[5] See for instance Atkinson (1989).

distribution of the different ratios. We therefore use Bayesian algorithms and in particular Monte Carlo Markov Chain algorithms (MCMC) so to find the proper parameters α and β giving the expected distribution not only for the variables but also for the ratios.

We derive here our algorithm from the idea proposed by Metropolis (1953) and improved by Hastings (1970), under the so-called Hastings-Metropolis algorithm. We set a prior distribution denoted $\pi(\theta)$ for $\theta = (\alpha_1, \dots, \alpha_n, \beta_1, \dots, \beta_n)$ the parameters of our Beta distributions. As a matter of fact we have $n = v * s$, where v denotes the number of variables we want to simulate, and s the number of strata for the total population (in this first approach, we take s=12). For $k \in [1\colon s]$, we consider the different indicators $\gamma_k^{i,j}$ that we want to reproduce in the data. These indicators can be estimations for percentiles, inter-quantile ranges, means or standard deviations. We here make the assumption that each of them follows a normal law. Thanks to this assumption, we can write the likelihood associated to a set of v random variables $(x_1^1, \dots, x_M^1, \dots, x_M^v)$ following a Beta law:

$$L_k(x|\theta) = \prod_{j=1}^{J} \varphi_{i,j}\left(\widehat{\gamma_k^{i,j}}\right)$$

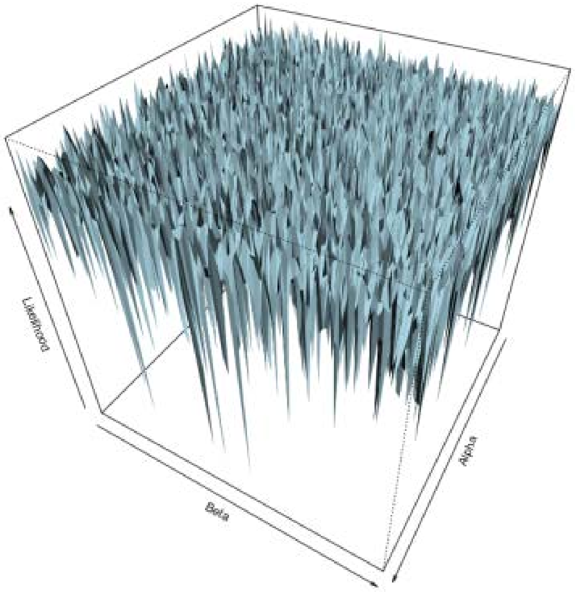
where J denotes the number of constraints in the generation of the model. Then in a Bayesian perspective we use the prior distribution of θ to improve the likelihood associated to the observations $(x_1^1, \dots, x_M^1, \dots, x_M^v)$. We then obtain the posterior distribution $\pi(\theta|x)$ that enables to generate observations that will fit as much as possible the constraints defined a priori. This method, also known as importance sampling, enables to maximize a likelihood which is difficult to compute analytically. Indeed the shape of the likelihood (Figure 2) makes it very difficult to derive analytically. Hence a Bayesian approach is completely justified, given the high number of constraints applied to the model.

Following the Metropolis-Hastings algorithm, we generate θ according to a prior symmetric distribution iteratively and retain a new set of θ only if the likelihood associated to the new set of observations $(\mathrm{x}_1^1, \dots, \mathrm{x}_\mathrm{M}^1, \dots, \mathrm{x}_\mathrm{M}^\mathrm{v})$ is above the previous one. We determine the number of iterations according to an expected acceptance rate and, also, since this method is computationally demanding, taking into account the time of computation. We test different ways of simulating the data to impose the least constraints as possible, in particular in terms of correlations between the v variables. The only source of information we use so far on potential links between say, number of employees and real value added is the information on the distribution of the ratio of labour productivity. Then generating micro-data should make explicit the implicit link between these two variables through information on labour productivity.

The first way of doing it is to take into account potential correlations between the variables thanks to an expectation-maximization algorithm. Indeed, once the v variables generated according to the Beta law, we merge the different columns according to an order that has to be defined. In the initialisation phase, we randomly sort the variables so there is almost no correlation between the different variables. During each iteration, we regress each variables on the v-1 other ones and use the fitted values to sort the variables. However this method does not prove to be conclusive and we rather introduce correlations as part of the vector θ. This method shows significantly better results, in particular in terms of acceptance rate. This new component of θ is generated thanks to a truncated normal law.

Representation of the likelihood

Figure 2



Furthermore, to improve the convergence process and to shrink the time of computation, we add an annealing part to the algorithm. The intuition behind this classical method in Bayesian inference is to reduce the area in which the optimal parameters θ are more likely to be found so to avoid a huge number of iterations. To implement statistical annealing in our algorithm, we decrease the variance associated to the prior law by for instance a half after *n* iterations for which none of the generated parameters have succeeded in improving the likelihood.

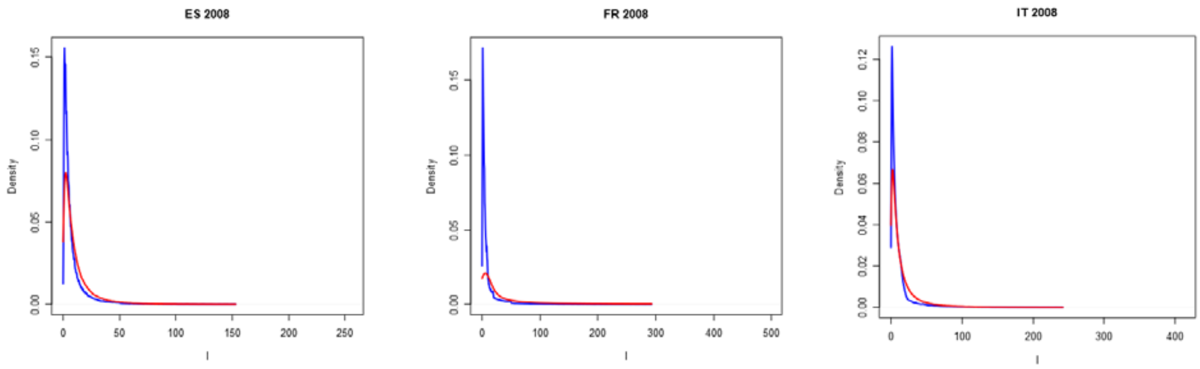
Finally, we have also to consider the law we need to use for simulating the last percentile of each variable. Indeed, for the rest of the distribution, we can use as lower and upper bounds the different percentiles that we have at our disposal. Then the maximisation process is just about finding the proper parameters that will give the expected shape for the distribution of the variable within the deciles. However, for the last percentile we do not have any upper bound at our disposal. We could set of course an upper bound related to the treatment of outliers that is made on the original data. We rather set a different law that do not use any upper bound. Given the high concentration observed for the variables we want to simulate, we choose a Pareto law, whose parameter k also follows a prior distribution.

Results

For the first method we generate data for Spain, France and Italy. Since we impose strong assumptions on the joint distributions of the variables, the obtained marginal distribution do not completely match with the expected ones (figure 3). Indeed, the link between the variables is generated thanks to the variance-covariance matrix, which imposes strong assumptions on the shape of the joint distribution and deteriorates the modelling of each variable.

Distributions of labour for 2008, Spain, France and Italy

Figure 3

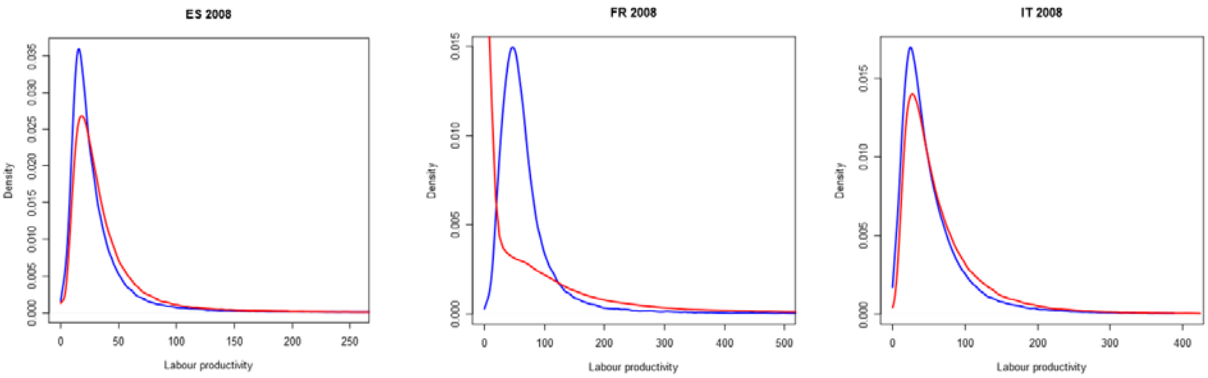


The blue line stands for the expected distribution, the red one shows the simulated distribution.

As a result, the ratios may not have the expected distribution, which weakens the simulation. We take here the example of the labour productivity which is defined as the ratio of real value added over productivity. As shown on Figure 5, the obtained distribution does not properly fit the expected one, in particular for France, where the labour is poorly reproduced.

Distributions of labour productivity for 2008, Spain, France and Italy

Figure 4



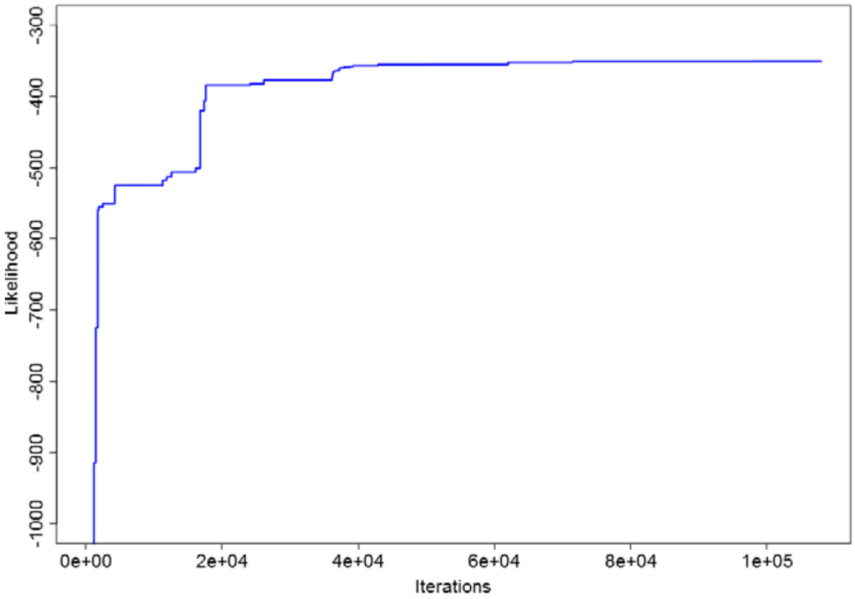
The blue line shows the expected distribution, the red one the simulated distribution.

We also implemented the second method, with no conclusive results. However, the estimation of the parameters can be done thanks to Bayesian procedures or through a Newton-Raphson algorithm. The latter one offers more guarantees in terms of convergence, but has not been yet implemented.

We run the maximisation program over 252 parameters for each country, year and sector. Such an algorithm is very demanding in terms of computations, and the computations for one given country, year and sector may take several hours. Regarding such duration for computing, the annealing part of the algorithm enables to gain both time and accuracy. We set the number of iterations here to 10,000 before starting the annealing process. This means, the algorithm starts the annealing only after 10,000 iterations without any change. As shown on figure 5, the annealing has enabled to increase the likelihood in a more efficient way.

Evolution of the log-likelihood for the third algorithm

Figure 5



Evolution of the log-likelihood over iterations.

We choose for the annealing part to reduce the variance associated to the prior law of the parameters by 90%, and we iterate the decrease 4 times. The whole algorithm implies about 100,000 iterations, or more, depending on the number of constraints we apply to the model.

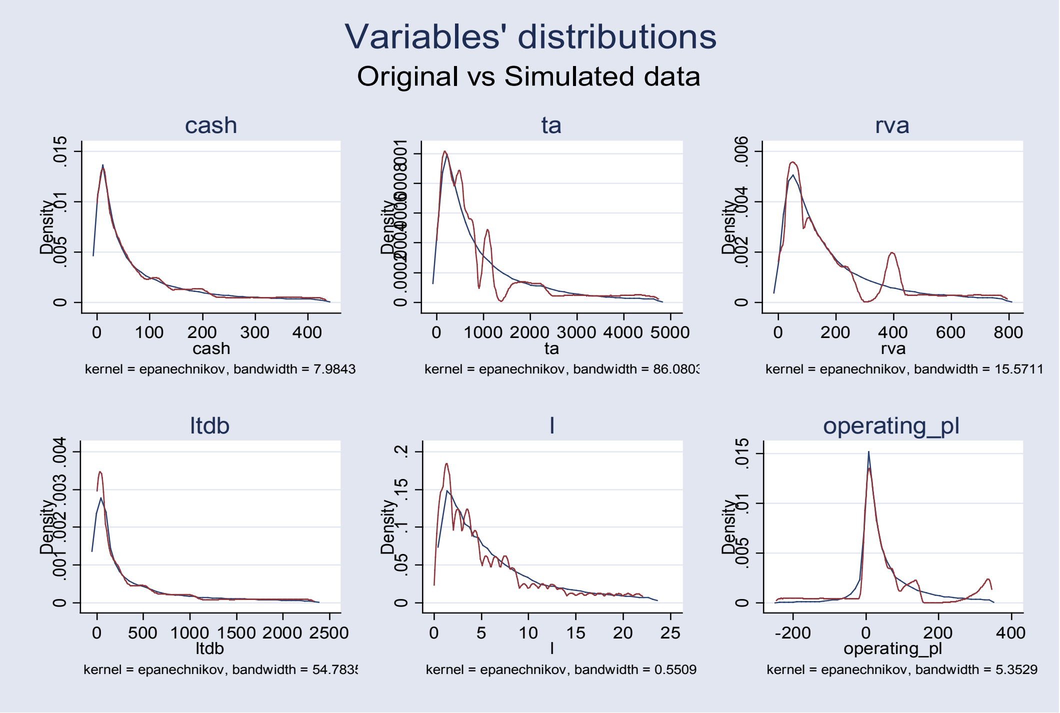
As the algorithm is highly time-consuming, we only present results for a very restricted sample of combinations of countries, years and sectors for which we performed the computation.

When looking at the results of the simulations and comparing them to the original data, we observe that the absolute percentage differences between the two vary both across variables and percentiles. In particular, results improve significantly when adding constraints related to joint moments of the distributions, as shown in Table A1 and Table A2 for the Wholesale trade sector in Spain in the year 2007.

Differences are especially strong in terms of mean. From this point of view the algorithm has difficulties to properly reproduce the high concentration of cash, total assets, long term debt, loans and operating profits and losses. The specification of the Pareto law for the very top of the distribution of each of these variables was intended to precisely reproduce the high concentration; however convergence of the simulation toward properly concentrated variables is not achieved. With this regard, another approach should perhaps be chosen for the simulation of the top the distribution. Indeed, the simulation in the case of the Pareto law is highly volatile. Then the parameter could be set thanks to another approach (e.g. maximum likelihood) and the simulated observations could be chosen with respect to given constraints. However, this method implies to get more information about the top of the distribution, which can be difficult with regards to the anonymisation rules that may be applied.

Distributions of the variables for the third algorithm compared to the expected ones

(data for Spain, NACE 46, year 2007) Figure 6



The blue line stands for the expected distribution, the red one shows the simulated distribution.

However, although the resulting simulated data seem to properly fit their original distributions (see Figure 6), this does not happen when looking at the correlations among those. The ordering of the observations, in fact, does not allow so far letting us find the expected correlations. Given the higher number of constraints we will be able to add, especially on joint moments, we expect this not to be a problem in the future replications of the exercise. Indeed information carried out by only the distribution of ratios such as labour productivity seem not to be sufficient to drive the simulation toward the expected link between labour and real value added for example. From this point of view, adding more constraints in terms of joint distribution should solve at least partially this issue.

An empirical application: leverage model

As an assessment of the goodness of the data produced by the algorithms of the previous sections, we estimate a simple static leverage function using first the firm-level data derived from Amadeus and second the simulated firm-level data as derived from the aggregated dataset. The purpose of the exercise is to compare the estimated coefficients derived from the two datasets.

Following the literature,[6] we estimate the following leverage model for each country c in our sample:

$$\text{Leverage}_{ict} = \sum\nolimits_{k=1}^{K} \beta_k X_{kict-1} + \eta_i + \eta_{ts} + \nu_{ict}$$

where Leverage_{ict} is the leverage of company i in country c at time t. Among the control variables we include: Sales growth, Size, Collateral and Cash flow. We also include Cash holdings to control for other factors that may allow firms to attain a degree of financial flexibility. All variables are lagged once to avoid endogeneity. Finally, we include firms fixed effects (ηi) that account for the potential correlation between firm-specific characteristics and regressors; and time-sectoral effects (ηts) that account for macro-economic factors (such as market shocks) related also to economic sectors.

Leverage results from firm-level dataset (Amadeus)

Table 3

Variables	(1)	(2)	(3)
	Italy	France	Spain
Cash flow/total assets	–0.03809***	–0.03978***	–0.06590***
	(0.003)	(0.002)	(0.002)
Collateral	0.05384***	0.15714***	0.00174
	(0.004)	(0.004)	(0.002)
Cash holdings	–0.02477***	0.01738***	–0.02054***
	(0.003)	(0.002)	(0.002)
Sales growth	0.00121*	0.00972***	0.00205***
	(0.001)	(0.001)	(0.000)
Size dum2	0.00029	0.00409***	–0.00707***
	(0.001)	(0.001)	(0.001)
Size dum3	0.00329*	0.00591***	–0.01588***
	(0.002)	(0.002)	(0.001)
Size dum4	0.00704***	0.01075***	–0.01744***
	(0.002)	(0.003)	(0.002)
Constant	0.17963***	0.09714***	0.29416***
	(0.001)	(0.001)	(0.001)
Observations	434,465	453,253	749,699
Number of firms	170,209	198,769	293,851
Firm fixed effects	yes	yes	yes
Year-sector fixed effects	yes	yes	yes

Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1

[6] See among others, Wanzeried (2006) and Ferrando, Marchica and Mura (2014).

Results on the leverage model derived from the first dataset are reported in Table 3 where each column shows the results at country level. Results are in line with previous findings in the capital structure literature (e.g., Rajan and Zingales (1995); Flannery and Rangan (2006); Wanzenried (2006)). Firms whose sales are growing faster need more leverage and this result is robust across countries. The results show also a clear size effect as larger firms are less opaque and may raise external finance more easily, and at more favorable rates. This appears to be true for Italian and French companies while in the case of Spanish firms the coefficient is negative. Collateral is also important when it comes to enabling firms to obtain external finance, hence the positive sign. The coefficient of Cash Flow is negative. According to the pecking theory, firms should prefer internal to external finance. Hence the more profitable the firm, the lower the need for external finance. Finally, leverage is negatively affected by Cash Holdings signaling that the availability of liquid assets may reduce the need of external debt.

We then use the same analysis approach on the simulated data using the first simulation method (Table 4). Although the difficulties described above when ordering the observations in order to reproduce the proper correlations across variables, the results deriving from an economic analysis are overall convincing. The coefficients are always statistically significant but the signs might differ and we can conclude that although the simulated data properly reproduce the distributions of the raw variables, the joint distributions of those variables are not completely reliable and regressions on these data should be run carefully.

Leverage results from simulated dataset (1st Method) Table 4

Variables	(1) Italy	(2) France	(3) Spain
Cash flow/ total assets	–0.02115***	0.06219***	0.04271***
	(0.002)	(0.003)	(0.003)
Collateral	0.06838***	0.14125***	0.10723***
	(0.001)	(0.003)	(0.002)
Cash Holdings	–0.12482***	–0.05284***	–0.11942***
	(0.001)	(0.002)	(0.002)
Sales Growth	–0.00964***	–0.01672***	–0.00495***
	(0.001)	(0.001)	(0.001)
Size dum2	–0.01221***	–0.02644***	–0.04110***
	(0.001)	(0.001)	(0.001)
Size dum3	–0.01501***	–0.03060***	–0.06114***
	(0.001)	(0.001)	(0.001)
Size dum4	–0.01988***	–0.00029	–0.06718***
	(0.001)	(0.001)	(0.002)
Constant	0.15548***	0.10155***	0.32955***
	(0.001)	(0.002)	(0.004)
Observations	635,808	181,513	418,662
Firm fixed effects	no	no	no
Year-sector fixed effects	yes	yes	yes

Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Conclusions

In this paper we proposed a methodology to mimic anonymised firms' micro-data using standard equations and assumptions about the distribution of the residuals that are the most likely to reproduce the original dataset

This exercise turned to be highly demanding in terms of computations, which makes its assessment quite difficult. Indeed, simulating the entire database for Italy, Spain and France appeared to be highly time consuming. Moreover, the results obtained for all given years, countries and sectors are not completely convincing yet, since the algorithm seems not to be completely able to reproduce expected correlations between variables. The idea of sorting variables so to reproduce the correlations relies on the assumption that conditionally to the fact that distributions are properly reproduce, there is an order for each variable that should ensure to get closer to the expected correlation matrix. This order can be of course found in the original data; for the time being seems not completely convincing to achieve reproducing this order.

Annexes

Absolute percentage differences between percentiles of original and simulated data before using joint moments as constraints

Table A.1

Statistics	Cash	Tot. assets	Real value added	Long term debt	Loans	Op. profits & loss
mean	0.29	0.37	0.08	0.57	0.38	0.46
p1	–	0.00	0.00	–	0.00	0.00
p10	0.00	0.00	0.00	0.00	0.00	0.00
p20	0.10	0.00	0.00	0.00	0.50	1.00
p30	0.50	0.35	0.31	0.56	0.00	0.64
p40	0.39	0.31	0.27	0.01	0.00	0.08
p50	0.35	0.00	0.26	0.31	0.17	0.38
p60	0.04	0.30	0.25	0.43	0.29	0.37
p70	0.37	0.06	0.17	0.42	0.19	0.05
p80	0.40	0.27	0.00	0.00	0.00	0.43
p90	0.00	0.51	0.48	0.54	0.01	0.56
p99	0.05	0.87	0.00	0.00	0.00	0.89
sd	0.36	0.47	0.25	0.35	0.74	0.35

Absolute percentage differences between percentiles of original and simulated data after using one joint moment as constraint

Table A.2

Statistics	Cash	Tot. assets	Real value added	Long term debt	Loans	Op. profits & loss
mean	0.11	0.01	0.43	0.27	0.42	0.11
p1	–	0.00	0.01	–	0.00	0.00
p10	0.00	0.00	0.00	0.00	0.00	0.00
p20	0.00	0.00	0.00	0.00	0.08	0.00
p30	0.00	0.00	0.16	0.00	0.00	0.00
p40	0.00	0.00	0.00	0.00	0.00	0.00
p50	0.00	0.03	0.00	0.00	0.00	0.00
p60	0.00	0.11	0.00	0.00	0.00	0.00
p70	0.00	0.06	0.00	0.00	0.00	0.00
p80	0.00	0.00	0.00	0.00	0.00	0.00
p90	0.00	0.00	0.00	0.00	0.00	0.00
p99	0.00	0.00	0.00	0.00	0.00	0.00
sd	0.11	0.73	0.60	0.03	0.30	0.51

Summary statistics of financial indicators and productivity, broken down by country and macrosectors

Table A1

Country	Sector/Variable	Cash Holdings			Leverage			Labour Productivity			Returns on Assets		
		Mean	**Median**	**Sd**	**Mean**	**Median**	**Sd**	**Mean**	**Median**	**Sd**	**Mean**	**Median**	**Sd**
ES	**Manufacturing**	0.12	0.06	0.15	0.30	0.25	0.78	35	27	72	0.01	0.03	0.42
	Construction	0.14	0.07	0.18	0.34	0.26	0.78	47	21	299	0.00	0.03	0.71
	Wholesale and retail trade	0.14	0.08	0.17	0.31	0.24	0.53	25	18	50	0.01	0.03	0.36
	Transportation and storage	0.13	0.07	0.16	0.34	0.29	0.68	63	40	407	0.01	0.03	0.27
	Accommodation and food service activities	0.15	0.07	0.19	0.47	0.41	0.87	42	31	141	-0.02	0.02	0.55
	Real estate activities	0.10	0.03	0.16	0.36	0.26	1.62	108	35	722	0.00	0.02	0.56
	Professional, scientific and technical activities	0.20	0.12	0.22	0.34	0.25	1.26	40	19	311	0.03	0.04	0.40
	Administrative and support service activities	0.18	0.10	0.21	0.36	0.26	1.13	74	35	688	0.00	0.03	0.93
FR	**Manufacturing**	0.18	0.12	0.19	0.16	0.09	0.28	71	53	406	0.07	0.06	0.15
	Construction	0.27	0.23	0.22	0.11	0.06	0.17	82	62	304	0.10	0.08	0.19
	Wholesale and retail trade	0.19	0.12	0.19	0.16	0.09	0.47	63	46	682	0.07	0.06	0.16
	Transportation and storage	0.21	0.16	0.19	0.15	0.08	0.23	88	65	315	0.06	0.05	0.18
	Accommodation and food service activities	0.19	0.12	0.19	0.27	0.21	0.28	69	49	829	0.08	0.07	0.19
	Real estate activities	0.39	0.36	0.30	0.18	0.06	0.58	204	60	2255	0.06	0.04	0.27
	Professional, scientific and technical activities	0.28	0.22	0.24	0.11	0.02	0.22	203	71	3610	0.10	0.08	0.23
	Administrative and support service activities	0.26	0.20	0.23	0.11	0.02	0.22	193	85	1605	0.08	0.06	0.24
IT	**Manufacturing**	0.08	0.03	0.12	0.20	0.15	0.20	43	31	239	0.05	0.04	0.27
	Construction	0.09	0.03	0.13	0.18	0.10	0.21	44	28	161	0.06	0.05	0.10
	Wholesale and retail trade	0.10	0.05	0.14	0.17	0.10	0.20	268	54	47631	-2.26	0.04	427.65
	Transportation and storage	0.10	0.04	0.14	0.15	0.06	0.19	100	76	169	0.04	0.04	0.13
	Accommodation and food service activities	0.11	0.05	0.16	0.22	0.12	0.27	61	46	112	0.03	0.03	0.39
	Real estate activities	0.08	0.02	0.14	0.22	0.12	0.25	94	28	632	0.03	0.03	1.35
	Professional, scientific and technical activities	0.13	0.06	0.16	0.13	0.03	0.20	67	26	998	0.08	0.05	0.15
	Administrative and support service activities	0.13	0.07	0.17	0.13	0.03	0.19	92	59	352	0.07	0.05	0.15

	Number of Firms for which the Leverage indicator is available																	
	ES						***FR***						**IT**					
sector	***2006***	***2007***	***2008***	***2009***	***2010***	***2011***	***2006***	***2007***	***2008***	**2009**	**2010**	**2011**	**2006**	**2007**	**2008**	**2009**	**2010**	**2011**
10	*5934*	*1533*	*6716*	*7067*	*7117*	*7046*	*6314*	*6825*	*6944*	7792	8685	8199	3225	3581	4860	4434	3879	7077
11	*1296*	*376*	*1449*	*1600*	*1618*	*1602*	*468*	*452*	*420*	404	420	446	572	651	779	704	636	958
12	*15*	*8*	*15*	*13*	*13*	*12*	*706*	*712*	*620*	602	679	594	8	9	8	7	11	12
13	*1540*	*258*	*1700*	*1714*	*1722*	*1694*							1861	2078	2622	2232	1890	3499
14	*1155*	*187*	*1310*	*1320*	*1267*	*1204*	*635*	*627*	*579*	606	606	534	1820	2087	2936	2463	2078	4349
15	*886*	*124*	*1009*	*1007*	*1006*	*1006*	*269*	*265*	*281*	264	257	263	1379	1633	2074	1919	1606	3394
16	*2653*	*346*	*2948*	*2961*	*2840*	*2653*	*1472*	*1455*	*1376*	1425	1525	1421	1257	1516	2164	1827	1523	3190
17	*763*	*265*	*828*	*852*	*824*	*823*	*465*	*471*	*414*	399	431	373	914	1000	1261	1128	948	1657
18	*3586*	*523*	*4023*	*4116*	*4028*	*3875*	*1901*	*1881*	*1768*	1873	1914	1780	1415	1624	2345	1982	1600	3237
20	*1627*	*560*	*1748*	*1810*	*1822*	*1845*	*856*	*816*	*732*	723	760	713	1515	1707	2072	1845	1657	2558
21	*202*	*119*	*216*	*217*	*234*	*228*	*170*	*162*	*157*	139	140	139	287	309	342	330	313	404
22	*1956*	*543*	*2179*	*2203*	*2230*	*2155*	*1378*	*1385*	*1237*	1220	1307	1150	2375	2595	3304	2807	2424	4339
23	*3213*	*873*	*3585*	*3572*	*3400*	*3159*	*1370*	*1399*	*1301*	1357	1382	1256	2587	2953	3952	3344	2810	5245
24	*973*	*320*	*1103*	*1111*	*1114*	*1074*	*356*	*344*	*296*	296	342	303	1043	1122	1255	1159	1070	1575
25	*9279*	*1534*	*10614*	*10585*	*10155*	*9479*	*5064*	*5058*	*4778*	4688	5018	4720	8672	10179	13771	11831	9957	20030
26	*584*	*173*	*692*	*726*	*757*	*777*	*846*	*821*	*773*	773	830	705	1587	1823	2364	2077	1760	3261
27	*899*	*277*	*963*	*986*	*1023*	*981*	*639*	*659*	*584*	562	600	552	2001	2236	2875	2552	2205	3928
28	*2649*	*572*	*3060*	*3101*	*3084*	*3051*	*1754*	*1747*	*1556*	1525	1654	1477	5676	6467	8141	6980	6092	10855
29	*776*	*336*	*860*	*874*	*869*	*857*	*586*	*562*	*535*	494	572	530	680	747	891	789	746	1222
30	*221*	*93*	*265*	*271*	*275*	*263*	*194*	*208*	*199*	200	205	190	512	660	882	786	672	1295
31	*3149*	*386*	*3443*	*3318*	*3148*	*2924*	*937*	*936*	*886*	922	953	862	1774	2045	2825	2377	1934	3926
32	*1124*	*193*	*1290*	*1367*	*1353*	*1312*	*1646*	*1698*	*1635*	1713	1820	1630	1437	1710	2473	2062	1730	3263
33	*2211*	*247*	*2695*	*2803*	*2805*	*2763*	*3342*	*3284*	*3308*	3471	3757	3648	1069	1307	1935	1741	1406	3418
41	*29503*	*4664*	*34292*	*31958*	*28654*	*25398*	*3282*	*3409*	*3362*	3708	4012	3773	9715	12473	21136	17813	13791	31684
42	*1136*	*332*	*1390*	*1394*	*1391*	*1280*	*1258*	*1242*	*1150*	1065	1157	1077	1081	1201	1718	1544	1301	2592
43	*22592*	*2606*	*27790*	*27227*	*25733*	*23391*	*33994*	*36398*	*37391*	42248	45685	42918	6973	8880	15052	12990	10258	24707
45	*10415*	*1947*	*12463*	*12929*	*12788*	*12522*	*11963*	*12305*	*11906*	13215	14485	13621	4660	5572	8570	7341	6015	12296
46	*33381*	*7179*	*39171*	*41002*	*41096*	*40381*	*21657*	*21437*	*20222*	20908	22536	20936	19206	22719	34124	29330	23760	47527
47	*26930*	*3207*	*32575*	*34245*	*34114*	*33270*	*30857*	*32765*	*32413*	37005	40504	37731	10351	12789	21321	18236	14476	34437
49	*9801*	*1534*	*11783*	*12088*	*11997*	*11737*	*6491*	*6670*	*6578*	7068	7787	7323	3624	4079	6368	5565	4744	11403
50	*410*	*77*	*476*	*492*	*504*	*542*	*151*	*155*	*167*	199	207	216	186	199	237	227	197	339
51	*92*	*28*	*109*	*107*	*111*	*107*	*53*	*46*	*54*	48	59	52	46	52	54	53	49	75
52	*2608*	*662*	*3168*	*3351*	*3435*	*3404*	*1571*	*1614*	*1520*	1570	1713	1568	2748	2916	4184	3681	3197	6433

53	*297*	*48*	*391*	*434*	*450*	*451*	*78*	*80*	*77*	102	113	104	52	68	101	115	89	245
55	*4873*	*1001*	*5803*	*6363*	*6503*	*6531*	*6606*	*6865*	*6603*	7371	7803	7110	2535	3168	4861	4128	3221	7848
56	*10867*	*1249*	*14504*	*15918*	*16072*	*16119*	*15935*	*18021*	*19042*	23136	25525	24291	3779	5196	9717	8269	6381	18744
58	*934*	*249*	*1189*	*1250*	*1219*	*1166*	*1762*	*1712*	*1677*	1693	1769	1657	815	906	1275	1099	941	1661
59	*844*	*176*	*1073*	*1147*	*1157*	*1140*	*1495*	*1496*	*1466*	1694	1779	1558	450	518	767	692	539	1211
60	*296*	*65*	*338*	*384*	*398*	*388*	*121*	*112*	*113*	114	113	117	352	415	657	522	423	787
61	*570*	*162*	*693*	*737*	*752*	*777*	*262*	*262*	*274*	314	379	353	200	252	402	387	341	668
62	*2521*	*551*	*3401*	*3741*	*3909*	*3953*	*2983*	*3194*	*3272*	3492	3812	3622	2828	3468	5073	4571	3700	7687
63	*338*	*55*	*477*	*528*	*534*	*550*	*606*	*602*	*589*	634	665	618	2597	3272	5625	4746	3676	7957
68	*19141*	*2710*	*20642*	*21361*	*20448*	*20072*	*8588*	*8765*	*8335*	8713	9191	8327	5616	6967	10439	8769	6434	13506
69	*6151*	*899*	*8292*	*9499*	*9555*	*9588*	*2923*	*2967*	*2951*	3023	3362	3118	1088	1304	1992	1684	1358	2622
70	*2490*	*457*	*3269*	*3667*	*3719*	*3754*	*5360*	*5539*	*5450*	5868	6364	5871	2547	3050	4560	4148	3473	6966
71	*4539*	*738*	*6489*	*6827*	*6619*	*6370*	*7115*	*7475*	*7473*	8143	8914	8508	2017	2507	3993	3546	2850	6098
72	*264*	*56*	*409*	*469*	*513*	*554*	*323*	*332*	*340*	328	343	326	345	400	578	528	507	931
73	*2294*	*377*	*3034*	*3238*	*3188*	*3002*	*2217*	*2283*	*2232*	2333	2560	2323	1266	1538	2457	2119	1667	3615
74	*3470*	*552*	*4683*	*5075*	*5089*	*4916*	*1340*	*1355*	*1328*	1538	1675	1631	1254	1571	2520	2275	1790	4478
75	*327*	*36*	*499*	*550*	*563*	*600*	*151*	*154*	*196*	256	307	327	11	22	42	32	23	69
77	*2300*	*370*	*2757*	*2835*	*2771*	*2664*	*1430*	*1473*	*1446*	1562	1699	1590	676	852	1387	1179	955	2272
78	*282*	*91*	*359*	*387*	*392*	*365*	*631*	*686*	*690*	700	736	795	136	162	203	187	164	280
79	*1135*	*205*	*1503*	*1598*	*1582*	*1513*	*959*	*944*	*936*	975	1027	937	1067	1300	2067	1658	1290	2731
80	*427*	*101*	*554*	*619*	*629*	*640*	*602*	*649*	*656*	749	852	755	228	285	365	287	292	714
81	*2401*	*406*	*3152*	*3402*	*3329*	*3349*	*3746*	*3998*	*4202*	5070	5483	5169	1649	1663	2599	2550	2201	5790
82	*2204*	*388*	*2782*	*3086*	*3081*	*3098*	*2232*	*2291*	*2210*	2386	2637	2387	1925	2321	3829	3464	2789	6195

References

Arnold S. F. (1993), “Gibbs sampling”, *Handbook of Statistics*, 9,pp. 599–625.

Bartelsman, E, Haltiwanger, J. and S. Scarpetta (2004): "Microeconomic evidence of creative destruction in industrial and developing countries." The World Bank, Policy Research Working Paper No. 3464, December.

Criscuolo, C., P. N. Gal and C. Menon (2014), "DynEmp: A Stata® Routine for Distributed Micro-data Analysis of Business Dynamics", *OECD Science, Technology and Industry Working Papers*, No. 2014/02, OECD Publishing.

Ferrando, A, M.T: Marchica and R. Mura (2014), “Financial Flexibility across the Euro Area and the UK”, ECB WP n 1630.

Flannery, M. and K. Rangan, 2006, “Partial adjustment toward target capital structures,” *Journal of Financial Economics* 79,469–506.

Gautier, E. (2008), “Bayesian estimation of inequalities with non-rectangular censored survey data”.

Gourieroux, C., Monfort, A., Renault, E., & Trognon, A. (1987). Generalised residuals. *Journal of Econometrics*, *34*(1), 5–32.

Little, R. J., & Rubin, D. B. (2002). Statistical analysis with missing data.

Lollivier S. and Verger D., (1987), “D’une variables discrete à une variable continue: la technique des résidus simulés“.

Paloma Lopez-Garcia, Filippo di Mauro, Nicola Benatti, Chiara Angeloni, Carlo Altomonte, Matteo Bugamelli, Leandro D’Aurizio, Giorgio Barba Navaretti, Emanuele Forlani, Stefania Rossetti, Davide Zurlo, Antoine Berthou, Charlotte Sandoz-Dit-Bragard, Emmanuel Dhyne, João Amador, Luca David Opromolla, Ana Cristina Soares, Bogdan Chiriacescu, Ana-Maria Cazacu, Tibor Lalinsky, Elena Biewen, Sven Blank, Philipp Meinen, Jan Hagemejer, Patry Tello, Antonio Rodríguez-Caloca, Urška Čede, Kamil Galušč□ák, Jaanika Merikyll and Péter Harasztosi, 2014, “Micro-based evidence of EU competitiveness: the CompNet database” ECB WP n 1634.

Rajan, R. G. and L. Zingales, 1995, “What Do We Know about Capital Structure? Some Evidence from International Data,” *Journal of Finance* 50, 1421–1460.

Robert, C. P., & Casella, G. (2004). *Monte Carlo statistical methods* (Vol. 319). New York: Springer.

Wanzenried,, G., 2006, “Capital Structure Dynamics in the UK and Continental Europe,” *European Journal of Finance*, 12, 693–716.